UNIVERSIDAD DE CÓRDOBA

Programa de doctorado:
Computación avanzada, energía y plasmas

TÍTULO:

MODELOS GENÉRICOS PARA LA PREDICCIÓN DE LAS NOTAS FINALES EN CURSOS A PARTIR DE LA INFORMACIÓN DE INTERACCIÓN DE LOS ESTUDIANTES CON EL SISTEMA MOODLE

Tesis presentada por:

Javier López Zambrano

Directores:

Dr. D. Cristóbal Romero Morales

Dr. D. Juan A. Lara Torralbo

Córdoba Junio, 2021

TITULO: *MODELOS GENÉRICOS PARA LA PREDICCIÓN DE LAS NOTAS FINALES EN CURSOS A PARTIR DE LA INFORMACIÓN DE INTERACCIÓN DE LOS ESTUDIANTES CON EL SISTEMA MOODLE*

AUTOR: *Javier López Zambrano*

© Edita: UCOPress. 2021
Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

https://www.uco.es/ucopress/index.php/es/
ucopress@uco.es

UNIVERSITY OF CÓRDOBA

Doctoral Programme:

Advanced computing, energy and plasmas

TITLE:

GENERIC MODELS FOR PREDICTING FINAL MARKS STARTRING FROM THE STUDENTS' INTERACTION INFORMATION IN MOODLE

A Thesis presented by:

Javier López Zambrano

Advisors:

Dr. D. Cristóbal Romero Morales

Dr. D. Juan A. Lara Torralbo

Córdoba June, 2021

TÍTULO DE LA TESIS: MODELOS GENÉRICOS PARA LA PREDICCIÓN DE LAS NOTAS FINALES EN CURSOS A PARTIR DE LA INFORMACIÓN DE INTERACCIÓN DE LOS ESTUDIANTES CON EL SISTEMA MOODLE

DOCTORANDO: Javier López Zambrano

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

El doctorando (Javier López Zambrano) ha progresado enormemente como investigador desde que comenzara la tesis doctoral en el año 2017 en la Universidad de Córdoba. Durante estos 4 años el doctorando ha realizado todas las actividades obligatorias y opcionales, trabajado duro seguido siempre las pautas de trabajo que le hemos marcado los directores y el plan de investigación que se estableció. Como principales frutos del trabajo realizado se han derivado los tres siguientes artículos publicados o aceptados en revistas incluidas en los tres primeros cuartiles de la relación de revistas del ámbito de la especialidad y referenciadas en la última relación publicada por el Journal Citation Reports (SCI y/o SSCI):

1.- López-Zambrano, J., Lara, J. A., & Romero, C. (2020). Towards portability of models for predicting students' final performance in university courses starting from Moodle Logs. Applied Sciences, 10(1), 354-377. https://doi.org/10.3390/app10010354. Impact Factor: 2.474 (Q2).

2.- López-Zambrano, J., Lara, J. A., & Romero, C. (2021). Improving the portability of predicting students' performance models by using ontologies. Journal of Computing in Higher Education,1-19. Published on line: 24 March 2021. https://doi.org/10.1007/s12528-021-09273-3. Impact factor: 2.271 (Q1).

3.- López-Zambrano, J., Lara, J. A., & Romero, C. (2021). Early Prediction of Student Learning Performance through Data Mining: A systematic review. Psicothema. 33(3) 1-10. https://doi.org/10.7334/psicothema2021.62Impact Factor: 2.632 (Q1).

Por todo ello, se autoriza la presentación de la tesis doctoral por compendio de artículos.

Córdoba, Junio de 2021

Firma del/de los director/es

ROMERO MORALES CRISTOBAL - 30800430A

Firmado digitalmente por ROMERO MORALES CRISTOBAL - 30800430A
Fecha: 2021.06.04 00:46:55 +02'00'

Fdo.: Cristóbal Romero Morales

Firmado digitalmente por LARA TORRALBO JUAN ALFONSO - 30970803S el día 04/06/2021 con un certificado emitido por AC FNMT Usuarios

Fdo.: Juan A. Lara Torralbo

La tesis titulada "MODELOS GENÉRICOS PARA LA PREDICCIÓN DE LAS NOTAS FINALES EN CURSOS A PARTIR DE LA INFORMACIÓN DE INTERACCIÓN DE LOS ESTUDIANTES CON EL SISTEMA MOODLE", que presenta Javier López Zambrano para optar al grado de doctor, ha sido realizada dentro del programa de doctorado Computación Avanzada, Energía y Plasmas, en la línea de investigación Aprendizaje Automático, Modelado de Sistemas y Minería de Datos, del Departamento de Informática y Análisis Numérico de la Universidad de Córdoba, bajo la dirección de los doctores Cristóbal Romero Morales y Juan A. Lara Torralbo cumpliendo, en su opinión, los requisitos exigidos a este tipo de trabajos.

Córdoba, junio de 2021

El Doctorando

JAVIER HERNAN LOPEZ ZAMBRANO

Firmado digitalmente por JAVIER HERNAN LOPEZ ZAMBRANO Fecha: 2021.06.04 10:37:42-05'00'

Fdo: Javier López Zambrano

El Director

ROMERO MORALES CRISTOBAL - 30800430A

Firmado digitalmente por ROMERO MORALES CRISTOBAL - 30800430A Fecha: 2021.06.04 00:47:33 +02'00'

Fdo: Dr. Cristóbal Romero Morales

El Director

Firmado digitalmente por LARA TORRALBO JUAN ALFONSO - 30970803S el día 04/06/2021 con un certificado emitido por AC FNMT Usuarios

Fdo: Dr. Juan A. Lara Torralbo

Esta tesis ha sido parcialmente subvencionada con los proyectos del Ministerio de Ciencia, Innovación y Universidades: TIN2017-83445-P y por el proyecto Titulado “Improving Data Science User's Experience with Computational Intelligence (INTENSE):

AGRADECIMIENTOS

Desde que retorné a la docencia en el 2015, tenía la firme convicción de realizar un estudio doctoral, aspiración que se tornó realidad hace cuatro años cuando la UCO firmó un convenio con la Universidad donde trabajo, ¿casualidad o conspiración del Universo?, sin lugar a duda iniciaba un nuevo objetivo para mí, objetivo que parecía difícil de alcanzar, sin embargo otra conspiración del Universo fue el de contar con un tutor muy reconocido en el campo de investigación como lo es el Doctor Cristóbal Romero (docente de la UCO), quien con su vasta experiencia y experticia en el área hizo que todo pareciera fácil, guiándome y apoyándome incondicionalmente; y para atribuirle otra más al Universo, luego pude contar también con la guía y apoyo incondicional del Doctor Juan A. Lara (docente de la UDIMA) como cotutor.

A ambos les ofrezco mi admiración, consideración y amistad, expresándoles mi enorme gratitud por compartir conmigo sus experiencias y conocimientos, convirtiéndolos en coautores de este, mi logro alcanzado.

Javier López Zambrano

"Cuando quieres realmente una cosa, todo el Universo conspira para ayudarte a conseguirla"

Paulo Coelho

TABLA DE CONTENIDOS

RESUMEN

Esta tesis propone el desarrollo de modelos genéricos de predicción del rendimiento académico de los estudiantes en Sistemas de Aprendizaje en Línea que puedan ser portados o transferidos a otros cursos o asignaturas diferentes de los cursos origen a partir de los cuales se obtuvieron dichos modelos, pero de contexto similar y así no tener una pérdida de precisión excesiva en las predicciones y que pueda ser aceptable. El problema de los modelos actuales de predicción es que son específicos a cada curso o asignatura y no pueden ser replicados en otros contextos con atributos diferentes. Se pretende buscar un balance entre el aumento de la generalidad de los modelos y disminuir la pérdida de la precisión de la predicción. Para resolver este problema primero se realizó una revisión sistemática del estado del arte, y conocer las anteriores y principales investigaciones en este campo desde el punto de vista de la precisión de los modelos de predicción, las técnicas y algoritmos utilizados, atributos utilizados en la generación de los modelos, y determinar si estos modelos podían ser replicados, transferidos, generalizados manteniendo una precisión aceptable en cursos diferentes a los que habían sido originalmente generados.

A continuación, se planteó una primera propuesta o aproximación para el desarrollo de modelos genéricos en base a los eventos de bajo nivel directamente proporcionados por los logs o registros de Moodle. Para ello se construyeron dos tipos de conjuntos de datos (numéricos y discretizados) a partir de 24 asignaturas diferentes de cursos Universitarios y se realizaron dos experimentos de portabilidad de los modelos generados en cada curso, el primero consistió en agrupar los cursos por su titulación, y el segundo por niveles similares de uso de recursos/actividades en Moodle. En ambos experimentos se consiguieron resultados relativamente similares, y los mejores valores con menor pérdida de AUC (área bajo la curva ROC) se obtuvieron con conjuntos de datos discretizados, lo que nos indicó que la discretización de los atributos mejoraba la portabilidad de los modelos genéricos.

Finalmente, se propuso la utilización de ontologías de alto nivel definidas a partir de los eventos de bajo nivel para comprobar si podían mejorar la portabilidad de dichos modelos de predicción genéricos. Para este segundo experimento se utilizaron 16 asignaturas diferentes de cursos Universitarios, se mantuvo la lógica de agrupación de asignaturas por niveles similares de uso de recursos/actividades de Moodle, nuevamente se generaron dos conjuntos de datos (numéricos y discretizados) para cada asignatura. Los resultados obtenidos mostraron que la utilización de atributos discretos de más alto nivel aplicando una ontología mejora significativamente la portabilidad de los modelos de predicción.

ABSTRACT

This thesis proposes the development of generic models for predicting students' final marks in Learning Management Systems (LMSs). These models must be portable or transferable to different courses from the source/initial course which the model was obtained and so, the context of the courses should be similar in order to obtain a lower loss of accuracy. The problem is that current prediction models are specifics for each course or subject and so, they can not be replicated in different contexts due to the attributes are different and the loss of accuracy. In this thesis, we try to do a balance between the increase of generality of the prediction models and the loss of accuracy in the transferability.

In order to resolve this problem, we started by doing a systematic review of the state of art about the prediction models in order to know all the research in this topic, the most used techniques and algorithms, the most used attributes for generating the models, and to know if some other researchers have tried previously to replicate or transfer a generic prediction model to different courses/subjects from the original generated.

Then, we proposed a first approximation for developing generic models based on the low level event provided directly from the Moodle's logs files. We developed two different datasets (numeric and discretized) starting from 24 subjects of University courses. We carried out two experiments to test the transferability of the generated models for each course by grouping the data in two ways. The first by grouping by the same subject or grade and the second by grouping by the same level of usage of resources and activities in Moodle. In both cases, we obtained similar results, and the best results with lower loss AUC (Area Under the ROC Curve) was obtained with the discretized data, and so, the transferability of generic prediction models was improved by using discretized data.

Finally, we proposed to use an ontology by defining high level attributes based on the previous low level Moodle's events in order to test if we can improve the portability/transferability of the generated prediction models. In this second experiment we used 16 subjects of University courses, and we also grouped the data by the same subject or grade and by the same level of usage of resources and activities in Moodle. The results obtained shown us that the discretized high level attributes using the ontology improved significatively the portability/transferability of the prediction models.

Parte I. Tesis Doctoral

1

INTRODUCCIÓN

Uno de los principales retos tecnológicos de los sistemas educativos en línea o basados en la Web es el desarrollo de sistemas automáticos para la predicción del rendimiento académico de los alumnos con el objetivo de poder ayudar a los alumnos con algún tipo de problema a tiempo para que no lleguen a abandonar o suspender las asignaturas y cursos. Existen multitud de investigaciones (Castro et al., 2007) relacionadas con el análisis de los datos que genera cada estudiante mientras interactúa con dichos sistemas. Según Koedinger et al., (2010) los estudiantes en línea pueden generar grandes repositorios de datos, los cuales reflejan el proceso de aprendizaje de los mismos en la denominada educación basada en la web (e-learning).

Para analizar estos grandes volúmenes de datos, se ha propuesto la utilización de técnicas de Minería de Datos o DM por sus siglas en inglés (Data Mining), y de análisis de datos o LA por sus siglas en inglés (Learning Analytics) con el objetivo de la extracción de información interesante, interpretable, útil y novedosa (Fayyad et al., 1996). Esta aplicación concreta de técnicas de minería de datos a la información generada en los entornos educativos se le conoce como Minería de Datos Educativa o EDM por sus siglas en inglés (Educational Data Mining) (Romero & Ventura, 2020). EDM utiliza las mismas técnicas que DM pero con ciertas adaptaciones de acuerdo con los problemas específicos que se intenten resolver (Romero & Ventura, 2010).

Una de las técnicas más populares en EDM, es la clasificación que se utilizar para predecir el rendimiento o nota final de los estudiantes en los cursos (Romero & Ventura, 2019). La Clasificación es la técnica de minería de datos que empareja o asocia datos a grupos predefinidos (aprendizaje supervisado), encuentra modelos (funciones) que describen y distinguen clases o conceptos para futuras predicciones y es probablemente la tarea más familiar y más popular de la minería de datos (Chen et al., 2000). Esta técnica se basa en el uso de etiquetado para realizar un mapeo desde un espacio de características (discreto o continuo) a un conjunto discreto de etiquetas (Duda et al., 2000) y en nuestro caso se utiliza para la predicción de desempeño o rendimiento de los estudiantes y su calificación final. Existen multitud de algoritmos de clasificación que se pueden agrupar en: Estadísticos (Regresión simple, regresión múltiple, bayes, ...), Distancia (k vecinos más cercanos,...), Árboles de decisión (ID3, C4.5, CART,...), Redes neuronales (Retropropagación, ...), Reglas (Class Association Rules), etc.

Por otro lado, de los innumerables sistemas actuales de educación basados en la web o e-learning o LMSs de las siglas en inglés (Learning Management Systems) , el más utilizado a nivel mundial es Moodle que es una “plataforma de aprendizaje diseñada para proporcionar a educadores, administradores y estudiantes un sistema integrado único, robusto y seguro para crear ambientes de aprendizaje personalizados” (Sánchez, 2009). En Moodle, un docente puede utilizar varios recursos y actividades para incluir en sus cursos. Los recursos (Archivo, Carpeta, Etiqueta, Libro, Página, entre otros) son objetos con los que se asiste al proceso de enseñanza – aprendizaje, mientras que una actividad (Tareas, Chat, Foros, Lecciones, Wikis, entre otros) es un trabajo que el alumno realizará, de forma individual o grupal, bien sea interactuando con otros compañeros y/o docente.

Moodle registra todas las actividades de los estudiantes en una base de datos relacional que almacena toda la información, tales como: datos personales del usuario (perfil), resultados académicos (grados) y datos de interacción del usuario (leer, escribir, realizar pruebas y tareas en entornos virtuales, comentar los eventos con sus compañeros, etc.), toda esta información está distribuidas en una gran cantidad de atributos/variables, por lo tanto, puede ser importante seleccionar solo un grupo representativo de atributos para reducir la dimensionalidad de los datos y crear una nuevo conjunto de datos (resumen) que proporcione toda la información importante relacionada con los estudiantes enrolados en el curso. Por ejemplo, la Tabla 1 muestra una lista de las características o atributos que han utilizados con éxito para predecir la nota final de estudiante en un curso de Moodle, (Cristóbal Romero et al., 2008).

Tabla 1. Ejemplo de lista de atributos seleccionado por estudiante en cursos de Moodle

Nombre	Descripción
id_student	Número de identificación del estudiante
id_course	Número de identificación del curso
num_sessions	Número de sesiones
num_assigment	Número de tareas realizadas
num_quiz	Número de pruebas tomadas
a_scr_quiz	Puntaje promedio en pruebas
num_posts	Número de mensajes enviados a foros
num_read	Número de mensajes leídos en foros
t_time	Tiempo total utilizado en Moodle.
t_assignment	Tiempo total utilizado en las tareas
t_quiz	Tiempo total utilizado en pruebas
t_forum	Tiempo total utilizado en foros
f_scr_course	Resultado final del estudiante obtenido en el curso

Pero estos atributos son sólo una propuesta específica de posibles variables que se pueden utilizar para capturar, recoger o modelar la información de interacción de los usuarios cuando utilizan Moodle con el objetivo de predecir su nota final en el curso. De hecho, existen una gran cantidad de trabajos sobre este mismo problema que utilizan otros atributos diferentes. Por ejemplo, en varios artículos se han enumerado una lista de otros posibles atributos, variables o características (ver Tabla 2) que se pueden utilizar para predecir la nota de los alumnos en Moodle (Conijn et al. ,2016), (Macfadyen & Dawson, 2010). Aunque existen muchas herramientas específicas de minería de datos que utilizan datos de Moodle (Luna et al., 2017), no se ha encontrado ninguna diseñada para preprocesar los logs de Moodle directamente y exportarlos a un formato estándar tipo CSV (Comma Separated Values) para poder ser utilizados por las herramientas más típica de minería de datos como Weka, rapidMiner, etc. Por ello, durante la primera etapa de la tesis se ha desarrollado una herramienta que nos permitiera preprocesar y preparar los datos de todos los cursos de Moodle que se utilizaron posteriormente en las pruebas experimentales.

Tabla 2. Otras variables utilizadas para la predicción de notas en Moodle.

Descripción
Número total de clicks
Número de páginas de curso vistas
Número de páginas de contenidos vistas
Número de recursos consultados
Número de enlaces vistos
Número de archivos vistos
Número total de debates(foros) posteados
Número de nuevos mensajes de foros posteados
Número de respuestas a mensajes de foros
Número de visitas al área de chat del curso
Número de pruebas vistas
Número de pruebas aprobadas
Número de intentos por pruebas
Número de ediciones wiki
Número de vistas wiki
Número de mensajes de mail leídos
Número de mensajes de mail enviados
Número de uso de la herramienta “Compile”
Número de uso de la función “Search”
Número de visitas a la herramienta “MyGrades”
Número de visitas a la herramienta “MyProgress”
Número de usos del visor “Who is online”
Irregularidad del tiempo de estudio
Irregularidad del intervalo de estudio
Periodo más largo de inactividad
Tiempo hasta la primera actividad
Tiempo promedio por sesión

Existen muchos trabajos de investigación en esta misma línea, donde cada autor utiliza sus propios conjuntos diferentes de atributos heterogéneos recogidos de Moodle para predecir la nota final de sus alumnos (Conijn et al., 2016) (Macfadyen & Dawson, 2010b). Este supone un grave problema a la hora de querer transferir o portar los modelos de predicción generados con los datos de un curso a otros cursos diferentes, debido a la especificidad de los mismos respecto a las características de los atributos. Esto plantea la necesidad de generalizar los atributos a un determinado nivel, de forma tal que sean comunes a todos los cursos de Moodle y así poder crear modelos lo más genéricos posible.

En esta tesis se propone una solución al problema de la especificad de los datos y modelos de predicción de los cursos de Moodle. Para ello se propone aplicar algoritmos de clasificación basada en datos discretizados primeramente y después en ontologías con niveles de granularidad más alta.

Para ello se propone comprobar si un modelo de predicción de las notas finales de los estudiantes obtenido a partir de los datos de interacción con un curso o asignatura dentro del sistema Moodle basado en atributos genéricos (atributos de alto nivel y ontologías) puede ser transferido (con una precisión aceptable) en lugar de un típico modelo basado en atributos específicos (atributos de bajo nivel relacionadas con frecuencias de eventos) a otros cursos diferentes. Hasta ahora los modelos de predicción de la nota final de los estudiantes a partir de los datos de utilización de Moodle descubiertos en diferentes trabajos/papers (Cristóbal Romero et al., 2008) (Romero et al., 2013) (Cerezo et al., 2016) (Won, 2016) utilizan atributos específicos basados en frecuencias para cada curso concreto y por tanto los modelos de clasificación descubiertos con los datos de estudiantes de un curso no se pueden utilizar para predecir a estudiantes de un curso distinto. Esto es debido principalmente a que los atributos utilizados en cada modelo son distintos dependiendo del investigador, herramienta, etc. Normalmente estos atributos suelen ser de muy bajo nivel (demasiado concretos), basados en frecuencias, de tipo numérico (con diferentes rangos) y específicamente seleccionados por el usuario con un nombre determinado. Todo esto impide que los modelos sean generalizables y transferibles, sino totalmente específicos. Para solucionar este problema, se propone en esta Tesis generar modelos de predicción genéricos que estén basados en un conjunto común de atributos de alto nivel o granularidad, que puedan estar relacionados con comportamientos o actividades (del tipo: nivel de utilización, nivel de interacción, nivel de comunicación y nivel de evaluación) y además con valores discretizados en una misma escala (muy básica e intuitiva tipo Alto, Bajo y Medio). De esta forma se podría reutilizar los modelos generados con un curso para predecir la nota de otros cursos diferentes del curso original donde se ha obtenido. Esto permitiría transferir el mismo modelo de predicción a cursos donde o bien no existan todavía datos al ser un curso nuevo, o bien tenga muy pocos alumnos anteriores o simplemente no se tenga acceso a los datos de años anteriores por cualquier motivo. También se podrían utilizar para hacer comparaciones directamente de los modelos de predicción/clasificación obtenidos por diferentes trabajos sobre los mismos datos, cosa que actualmente no se suele hacer.

Finalmente, es importante comprobar la validez, desde el punto de vista de la exactitud o perdida de precisión que se produce cuando transferimos los modelos genéricos con respecto a los específicos, y si esta puede variar mucho o no cuando se aplican en diferentes tipos de cursos, estudios y titulaciones, etc. Para ello se propone utilizar tanto datasets con atributos específicos obtenidos directamente de diferentes logs de cursos de Moodle, como datasets con atributos más genéricos obtenidos del preprocesado y aplicación de ontologías y así poder comparar las precisiones obtenidas con algoritmos de clasificación al utilizar ambos modelos cuando se aplica sobre cursos de distintas topologías o niveles de uso: cursos básicos donde los profesores sólo tienen contenidos teóricos o recursos de visualización (ficheros Word o PDF, Power Point, etc.), cursos más avanzados que también usan foros, y que además usan tareas (assigments), hasta cursos más completos que incluso usan otros recursos/actividades no tan habituales como: wikis, diarios, cuestionarios/test y otras herramientas de evaluación.

1.1 Objetivos

El objetivo principal de esta tesis es la generación de modelos genéricos de predicción de las notas finales de los estudiantes, en cursos a partir de la información de interacción con el sistema MOODLE, que puedan ser transferibles o portables a otros cursos diferentes de los originales. Para alcanzar este objetivo principal se definieron los siguientes 3 subjetivos:

- **O_1**: Realizar una revisión sistemática del estado del arte sobre la Predicción del rendimiento del aprendizaje de los estudiantes mediante técnicas de minería de datos. A partir de ella, se podrá conocer cuáles son tanto los algoritmos como los datos y/o atributos más utilizados que deberemos utilizar en la tesis.
- **O_2**: Evaluar la portabilidad de los modelos clásicos de predicción del rendimiento académico en cursos universitarios a partir de los eventos que proporcionan los ficheros logs sobre la interacción de los alumnos con la plataforma Moodle. Esto nos permitirá calcular cual es la perdida de exactitud que se produce al transferir los modelos de predicción a otros cursos.
- **O_3**: Mejorar la portabilidad de los modelos de predicción del rendimiento de los estudiantes mediante el uso de ontologías. Se propondrá una ontología en base a atributos de alto nivel que permita generalizar los modelos y se comprobará la mejora en la perdida de exactitud al transferir dichos modelos.

1.2 Hipótesis

Cada uno de los anteriores Objetivos, tiene asociado un conjunto de hipótesis. Concretamente las siguientes hipótesis $H_{1.1}$, $H_{1.2}$ y $H_{1.3}$ están relacionadas con el objetivo O_1, y estas se abordan en el artículo titulado: "*Early Prediction of Student Learning Performance through Data Mining: A systematic review*" (López-Zambrano et al. 2021b) donde se hace una revisión del estado de arte sobre los principales modelos de predicción temprana del rendimiento académico, sirviendo como base para el desarrollo del objetivo principal de la tesis:

- **$H_{1.1}$:** La mayoría de los trabajos de investigación sobre predicción temprana de rendimiento académico se han realizado en los sistemas de educación en línea y de nivel Universitario.

- **$H_{1.2}$:** La técnica y algoritmos mayormente utilizados en investigaciones relacionadas con la predicción temprana de rendimiento académico son clasificadores de tipo caja blanca.
- **$H_{1.3}$:** Los trabajos de investigación sobre predicción temprana de rendimiento académico realizados hasta la actualidad no definen modelos genéricos que sean portables a otros cursos, sino que son específicos a cada curso y utilizan variables o atributos de bajo nivel muy concretos.

Las hipótesis $H_{2.1}$ y $H_{2.2}$ están relacionadas con el objetivo O_2, y estas se abordan en el artículo titulado: "*Towards portability of models for predicting students' final performance in university courses starting from Moodle Logs*" (López-Zambrano at al., 2020):

- **$H_{2.1}$**: La portabilidad de los modelos de predicción clásicos (basados en atributos numéricos y discretos de bajo nivel) tiene una menor pérdida de exactitud cuando se realiza entre cursos de una misma titulación.
- **$H_{2.2}$**: La portabilidad de dichos modelos predictivos tiene una menor pérdida de exactitud cuando se realiza entre cursos similares desde el punto de vista del número de recursos y actividades de Moodle que utilizan.

Las hipótesis $H_{3.1}$ y $H_{3.2}$ están relacionadas con el objetivo O_3, y esta se aborda en el artículo titulado: "*Improving the portability of predicting students' performance models by using ontologies*" (López-Zambrano et al. 2021a):

- **$H_{3.1}$**: La utilización de una ontología con atributos de alto nivel de granularidad mejora la portabilidad de los modelos predictivos con respecto a los modelos clásicos que utilizan atributos de bajo nivel basados en eventos.
- **$H_{3.2}$**: La portabilidad de los modelos predictivos mejora cuando se transfieren los modelos entre cursos de la misma titulación o área y que además son similares desde el punto de vista del número de actividades/recursos de Moodle que utilizan.

2

CONCLUSIONES

Tras el desarrollo de la presente tesis se han obtenido varias conclusiones que podemos agrupar en 3 grandes grupos y que han sido abordados en los 3 artículos con índice de impacto que se han publicado.

Primeramente, tras realizar una revisión sistemática del estado del arte en la predicción del rendimiento académico de los estudiantes utilizando técnicas de minería de datos (López-Zambrano et al. 2021b) se obtuvieron las siguientes conclusiones:

- Respecto a la hipótesis $H_{1.1}$ se puede concluir que de los artículos revisados, hay un 57.3% que han utilizado datos principalmente del aprendizaje en línea, y un 86.6% de artículos describieron estudios realizados con estudiantes de educación terciaria lo que indica que la mayor parte del esfuerzo hasta la fecha ha sido en la predicción temprana con estudiantes universitarios, lo que también está de acuerdo con la accesibilidad de los datos, debido a que en estos entornos de aprendizaje es más fácil recopilar, gestionar y analizar datos.

- Respecto a la hipótesis $H_{1.2}$, este estudio revela que la clasificación es la técnica más utilizada, seguida de regresión (ambas son técnicas supervisadas), considerándose como las dos técnicas principales de DM que se han aplicado tradicionalmente a la predicción temprana del rendimiento académico de los estudiantes; sin embargo, cabe señalar que la aplicación de asociación y agrupamiento en conjunto con las dos anteriores puede implicar una cierta tendencia. También se evidencia que los algoritmos más utilizados fueron Naive Bayes, Decision Tree, Support Vector Machine y Logistic Regression, lo cual es concordante con las técnicas más utilizadas ya que los tres primeros corresponden a algoritmos de clasificación y el último es de regresión.

- Respecto a la hipótesis $H_{1.3}$, el estudio realizado, evidencia que las variables y los atributos de los estudiantes utilizados para la predicción varían según el entorno educativo, e incluso dentro del mismo entorno, las variables varían entre estudios. Los investigadores, han utilizado diferentes grupos de variables en cada artículo, lo que dificulta la tabulación de las variables por frecuencia de uso. En general, estas variables provienen de las mismas fuentes de datos, como la demografía de los estudiantes, las actividades de los estudiantes y las interacciones de los estudiantes, lo cual nos lleva a inferir que es muy difícil poder transferir los modelos de predicción generados en un curso a otros cursos diferentes, debido a que utilizan atributos diferentes y específicos, y además que se obtendrá una pérdida importante de exactitud.

A continuación, tras aplicar nuestro primera propuesta o enfoque experimental (López-Zambrano et al. 2020) aplicando el algoritmo de clasificación J48 (versión Java del clásico algoritmo C4.5) sobre los logs de Moodle de 24 asignaturas de la Universidad de Córdoba, se obtuvieron los valores de AUC (área bajo la curva ROC) y la pérdida de AUC de los modelos de predicción del rendimiento académico al aplicarlos a diferentes cursos del mismo grupo usando conjuntos de datos numéricos y discretizados. Como conclusiones respecto al planteamiento de nuestras hipótesis se obtuvo que:

- Respecto a la hipótesis $H_{2.1}$, se pudo observar en los experimentos evaluamos, habiendo considerado cuatro grupos diferentes (Informática, Educación, Ingeniería y Física), que la portabilidad de modelos de predicción entre cursos pertenecientes a la misma titulación según los valores promedio de AUC no son muy altos (tanto en conjuntos de datos numéricos como discretizados). También se pudo observar que la pérdida de AUC es mejor en los conjuntos de datos discretizados que en los numéricos, consiguiendo pérdidas de 0.003 (la más baja) y 0.126 (la más alta), cuyo valor más bajo está muy cerca de la portabilidad perfecta.

- Respecto a la hipótesis $H_{2.2}$, se pudo observar que la portabilidad de modelos de predicción entre cursos con un nivel similar de uso de actividades de Moodle considerado tres grupos diferentes (Alto, Medio y Bajo), los mejores valores de AUC se obtienen nuevamente con los conjuntos de datos discretos. Y, también se observó que la pérdida de AUC es mejor en los conjuntos de datos discretizados que en los numéricos, consiguiendo pérdidas de 0.009 (la más baja) y 0.061 (la más alta). Además, en ambos experimentos se consiguen valores de perdida de portabilidad muy buenos con algunos modelos predictivos, en concordancia con Baker, (2019) quien indica que los modelos de predicción son portables siempre que sus valores de pérdida de portabilidad se mantengan alrededor de 0,1 (y el AUC se mantenga por encima de la aleatoriedad).

Finalmente, se propuso un segundo enfoque (López-Zambrano et al. 2021a) donde se utilizaron atributos de alto nivel con un significado semántico más alto mediante el uso de una ontología que utiliza una taxonomía de acciones que resume las interacciones de los estudiantes con el sistema de gestión del aprendizaje Moodle. Tras realizar una comparación con los resultados del anterior enfoque inicial (hipótesis $H_{2.2}$) que utilizaba atributos de bajo nivel con respecto al nuevo enfoque propuesto que utiliza atributos de alto nivel basados en ontología, obtuvimos las siguientes conclusiones sobre las hipótesis:

- Respecto a la hipótesis $H_{3.1}$, los resultados obtenidos muestran que el uso de la ontología con atributos de alto nivel y discretizados mejora significativamente la portabilidad de los modelos en cuanto a su exactitud predictiva y que se pueden aplicar a otros cursos diferentes con niveles de uso de actividades y recursos de Moodle similares sin perder mucha exactitud en la predicción (pérdida de AUC).

- Respecto a la hipótesis $H_{3.2}$, los resultados obtenidos muestran además que la portabilidad de los modelos predictivos mejora cuando se transfieren los modelos obtenidos entre cursos de la misma titulación o área y que además utilizan similares actividades y recursos en Moodle.

2.1 Futuras mejoras y líneas

El tema tratado en esta tesis de la generación de modelos genéricos y su portabilidad o transferibilidad a asignaturas diferentes de las cuales han sido obtenidos es de gran interés y futuro. Creemos que esta es una línea novedosa y muy prometedora, donde se puede avanzar mucho y donde nosotros vemos las siguientes líneas potenciales o futuras de investigación donde llevar a cabo nuevos experimentos:

1. Utilizar una cantidad mucho mayor con respecto al número de asignaturas (y no sólo 24 asignaturas), de muchas más titulaciones diferentes (no sólo 5 titulaciones) de otros campos como ciencia, biología, medicina, filosofía y literatura, incluso de otras Universidades (y no sólo de una) para así poder comprobar de una forma mucho más fiable como de buenos son nuestros resultados obtenidos cuando se aplican a un conjunto de datos mucho mayor y más genérico.

2. Obtener modelos predictivos lo antes posible que puedan ser portables en las primeras etapas o semanas de los cursos. Esto significa que no tendríamos que esperar hasta el final del curso para tener disponibles todos los datos de uso de Moodle, y los modelos obtenidos podrían usarse como modelos generales de predicción de alerta temprana para diferentes cursos similares (Cristóbal Romero & Ventura, 2019). Para ello se deben obtener datasets de las asignaturas en etapas incrementales de tiempo: la primera semana, segunda, etc. o el primer mes, segundo mes, etc. y así poder comparar la transferibilidad de dichos modelos de predicción temprana con respecto a los modelos completos (utilizando los datos de todo el curso completo).

3. Aplicar nuestra propuesta de modelos predictivos con ontologías no solo a otros Sistemas de Gestión del Aprendizaje diferentes a Moodle como pueden ser CANVAS, Ilias, atutor, Claroline, etc. sino también a otros dominios o tipos de sistemas educativos basados en computador y Web diferentes como los Sistemas de Tutoría Inteligente (ITSs), Cursos Abiertos Masivos Online (MOOCs), entornos educativos presenciales tradicionales, entornos de aprendizaje mixto (Blended Learning) y aprendizaje multimodal, etc.

4. Utilizar otros criterios para agrupar todas las asignaturas (además de la titulación y el nivel de uso de recursos/actividades de Moodle) de diferentes formas o grupos y así analizar qué tan portables son los modelos dentro de esos grupos y qué formas de agrupación son las ideales desde el punto de vista de mejora de la portabilidad y transferibilidad de modelos genéricos de predicción. Por ejemplo, se podría utilizar el número de estudiantes, el número de tareas de evaluación, la metodología utilizada por el instructor, etc.

2.2 Contribuciones científicas

Se indican a continuación toda la producción investigativa generada en la presente tesis, misma que se anexan en la parte II de publicaciones:

- **Comunicación en Congreso Internacional (CORE B):**

C1. López-Zambrano, J., Martinez, J. A., Rojas, J., & Romero, C. (2018). A tool for preprocessing moodle data sets. In Proceedings of the 11th International Conference on Educational Data Mining, Buffalo, NY, USA (pp. 15-18).

- **Ponencia en Congreso Internacional:**

P1. López-Zambrano, Lara, J.A., & Romero, C. (2019). MODELOS GENÉRICOS PARA LA PREDICCIÓN DE LAS NOTAS FINALES EN CURSOS A PARTIR DE LA INFORMACIÓN DE INTERACCIÓN DE LOS ESTUDIANTES CON EL SISTEMA MOODLE. I Congreso Internacional y Multidisciplinario de Investigadores en Formación en Ecuador. Manta, Ecuador (pp. 110-111).

- **Artículo en Revista con factor de impacto (incluida en el JCR):**

A1. López-Zambrano, J., Lara, J. A., & Romero, C. (2020). Towards portability of models for predicting students' final performance in university courses starting from Moodle Logs. Applied Sciences, 10(1), 354-377.

A2. López-Zambrano, J., Lara, J. A., & Romero, C. (2021). Improving the portability of predicting students' performance models by using ontologies. Journal of Computing in Higher Education, 1-19. In Press.

A3. López-Zambrano, J., Lara, J. A., & Romero, C. (2021). Early Prediction of Student Learning Performance through Data Mining: A systematic review. Psicothema. 33(3) 1-10.

REFERENCIAS

Ajlan, A.A. & Zedan, H. (2008). Why moodle. 12th IEEE International Workshop on Future Trends of Distributed Computing Systems. Suzhou, China, 58-64. https://doi.org/10.1109/FTDCS.2008.22

Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. *Studies in Computational Intelligence*. https://doi.org/10.1007/978-3-540-71974-8_8

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers and Education*, *96*, 42–54. https://doi.org/10.1016/j.compedu.2016.02.006

Chen, G.-D., Liu, C.-C., Ou, K.-L., & Liu, B.-J. (2000). Discovering Decision Knowledge from Web Log Portfolio for Managing Classroom Processes by Applying Decision Tree and Data Cube Technology. *Journal of Educational Computing Research*, *23*(3), 305–332. https://doi.org/10.2190/5JNM-B6HP-YC58-PM5Y

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2016). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, *PP*(99), 1. https://doi.org/10.1109/TLT.2016.2616312

Duda, R. O., Stork, D. G., & Hart, P. E. (2000). *Pattern classification: Pt. 1: Pattern classification*. New York: John Wiley \& Sons.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*. https://doi.org/10.1609/aimag.v17i3.1230

Koedinger, K., Cunningham, K., Skogsholm, A., & Leber, B. (2008). An open repository and analysis tools for fine-grained , longitudinal learner data. *Proceedings of the 1st International Conference on Educational Data Mining*. https://doi.org/Research supported by the National Science Foundation award number SBE-0354420.

López-Zambrano, J., Lara, J. A., & Romero, C. (2020). Towards portability of models for predicting students' final performance in university courses starting from Moodle Logs. Applied Sciences, 10(1), 354-377.

López-Zambrano, J., Lara, J. A., & Romero, C. (2021a). Improving the portability of predicting students' performance models by using ontologies. Journal of Computing in Higher Education, 1-19. In Press.

López-Zambrano, J., Lara, J. A., & Romero, C. (2021b). Early Prediction of Student Learning Performance through Data Mining: A systematic review. Psicothema. 33(3) 1-10.

Luna, J. M., Castro, C., & Romero, C. (2017). MDM tool: A data mining framework integrated into Moodle. *Computer Applications in Engineering Education*, *25*(1), 90–102. https://doi.org/10.1002/cae.21782

Macfadyen, L. P., & Dawson, S. (2010a). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Macfadyen, L. P., & Dawson, S. (2010b). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, *21*(1), 135–146. https://doi.org/10.1002/cae.20456

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1355

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2019). Guest Editorial: Special Issue on Early Prediction and Supporting of Learning Performance. *IEEE Transactions on Learning Technologies*, *12*(2), 145–147. https://doi.org/10.1109/TLT.2019.2908106

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers and Education*, *51*(1), 368–384. https://doi.org/10.1016/j.compedu.2007.05.016

Sánchez, I. (2009). Plataforma educativa Moodle: administración y gestión. *Editorial Alfaomega, México-México*.

Won, J. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *Internet and Higher Education*, *29*, 23–30. https://doi.org/10.1016/j.iheduc.2015.11.003

Parte II. Publicaciones

A Tool for Preprocessing Moodle data sets

Javier López-Zambrano[1], José Antonio Martinez[2], Jesús Rojas[2], Cristóbal Romero[2]
[1]ESPAM MFL, Faculty of Computation, 131106, Calceta, Ecuador
[2]University of Cordoba, Dept. of Computer Science, 14071, Córdoba, Spain
jlopez@espam.edu.ec, i32marej@uco.es, i32roraj@uco.es, cromero@uco.es

ABSTRACT

This paper describes a desktop Java tool for allowing instructors to preprocess Moodle data sets. Our idea is to provide instructors with an easy to use tool for preparing the raw Excel students data files directly downloaded from Moodle's courses interface. Several traditional preprocessing techniques are considered to transform input data into well-formatted data sets that can be later used by most of the popular data mining frameworks.

Keywords

Moodle's students data, data preprocessing, data mining tool.

1. INTRODUCTION

Nowadays, there is a great interest in analyzing and mining any students' usage/interaction information gathered by Learning Management Systems (LMS) such as Moodle [1]. However, to obtain and preprocess these data can be an arduous and tedious task [2]. Generally, it is necessary to know SQL language as well as to be an user with administrator role in order to have access to all the course information. And to our knowledge there isn't any specific Moodle data mining tool for preprocessing [2]. So, in order to resolve these problems, we have developed an easy to use Java GUI application oriented to be used by non-expert users in data mining and SQL, such as instructors. Our idea is to provide the instructor of a Moodle course the possibility of using Excel files directly downloaded from Moodle's interface without a labor and time-intensive preprocessing step. Finally, the obtained files from our desktop tool are well-formatted datasets that can be used by most of the well-known data mining frameworks (Weka, RapidMiner, Knime, R, etc.) for applying data mining algorithms.

2. TOOL DESCRIPTION

Our Moodle data preprocessing desktop tool has been developed in Java language and it includes six main steps and taps (see Figure 1).

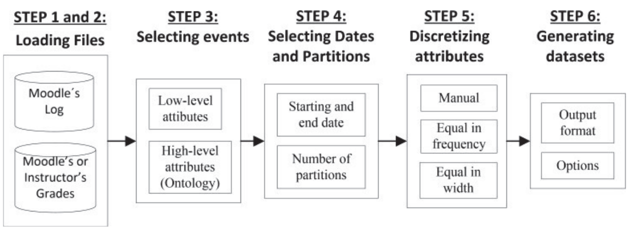

Figure 1: Preprocessing flow

2.1 Log file selection

This tab enables a log file (directly downloaded from Moodle's course interface in spreadsheet Excel format) to be opened/loaded. After that, it shows the content of the file and allows selecting the specific columns where the required information is located (Name of the students, Date and Events). This tab also provides basic information about the loaded file such as the total number of records, and the first and last update for all the records (see Figure 2).

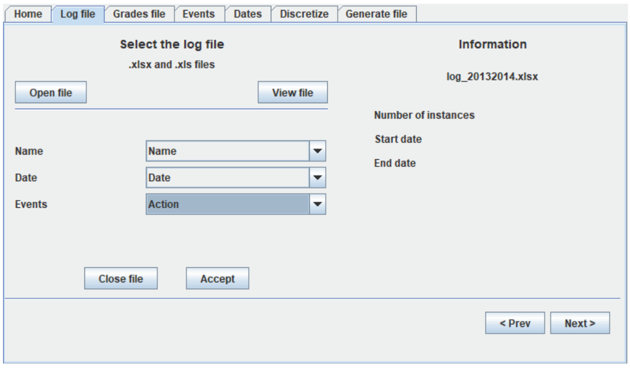

Figure 2: Selecting a log file.

2.2 Grades file selection

This tab is used by instructors to load a file (in spreadsheet Excel format) containing the students' grades (directly downloaded from Moodle or provided by the own instructors). Instructors can also fill in the students' mark manually (see Figure 3). Finally, those students with no final mark in the course can be removed, set as fail or even set as withdraw.

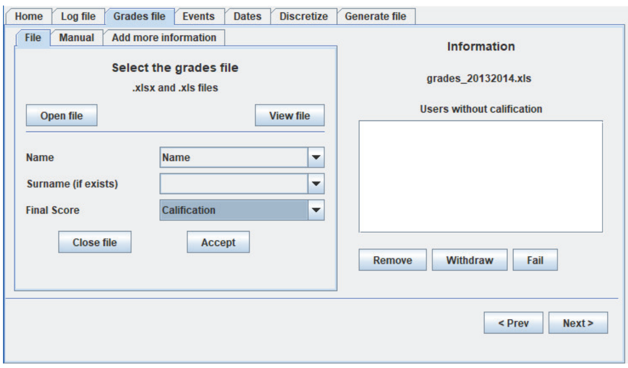

Figure 3: Loading a grades file.

2.3 Events selection

This tab allows the instructor to select what events (all of them or just a few) should be used as attributes in the final dataset. It is also possible to group these raw events in new high level attributes manually or automatically by using an ontology (see Figure 4). This ontology can be created, saved, loaded and viewed.

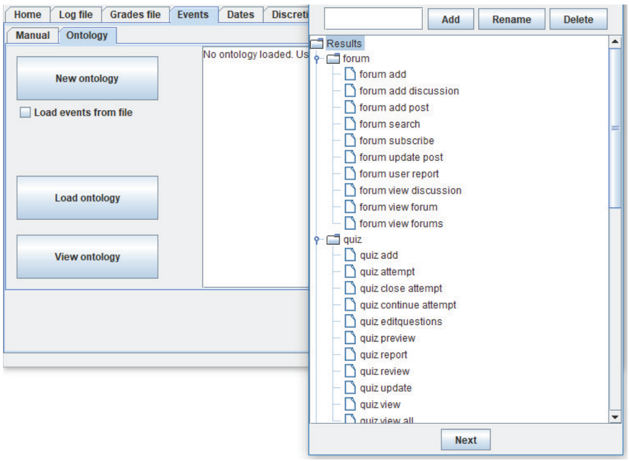

Figure 4: Selecting events using an ontology.

2.4 Date and partitions selection

The specific starting and ending date of the course can be established from this tab in order to use only the events that occurred between these dates (see Figure 5). It is also possible to specify whether the user requires a single summarization file or a number of cumulative data partitions (e.g. one per week/month).

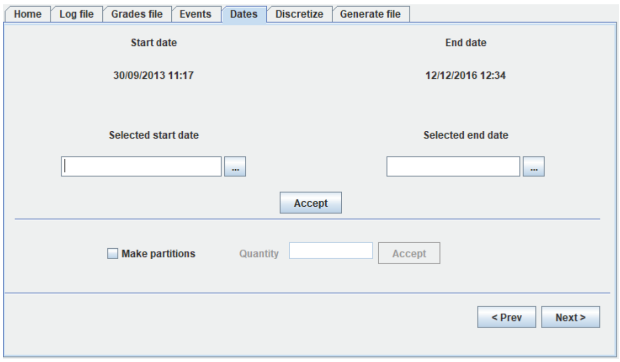

Figure 5: Selecting dates and partitions.

2.5 Discretization

For the sake of transforming those attributes or variables defined in a continuous domain/range into discrete values, this tab provides the option of performing a manual discretization as well as traditional techniques such as equal-width or equal-frequency (see Figure 6).

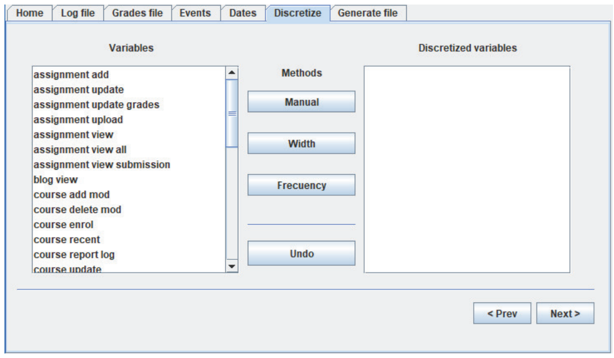

Figure 6: Discretizing variables.

2.6 Dataset generation

Finally, this last tab allows the instructor to generate the preprocessed data file, or several data files in case he/she selected several partitions that can be downloaded in three different file formats: .ARFF (Attribute-Relation File Format), .CSV (Comma-Separated Values) and .XLS (eXceL Spreadsheet). This tab includes additional options such as data anonymization and previous discretization techniques (see Figure 7). It also gives the possibility to generate a student's engagement variable that unifies the time, in minutes and days that each student has been connected in Moodle, as well as the total number of records/instances of each student in the log file.

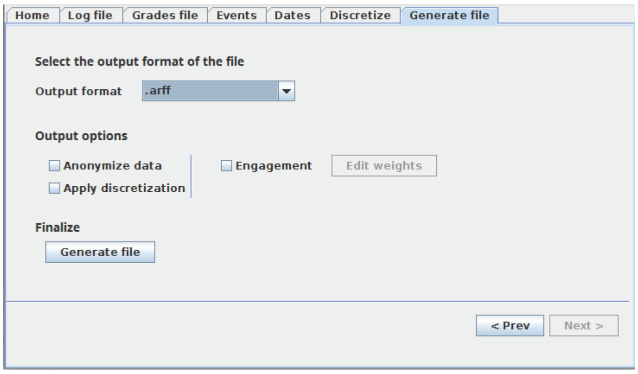

Figure 7: Generating preprocessed datasets.

3. ACKNOWLEDGMENTS

The authors acknowledge the financial subsidy provided by Spanish Ministry of Science and Technology TIN2017-83445-P.

4. REFERENCES

[1] Luna, J. M., Castro, C., Romero, C. 2017. MDM tool: A data mining framework integrated into Moodle. Computer Applications in Engineering Education, 25(1), 90-102

[2] Romero, C., Romero, J.R., Ventura, S. 2014. A Survey on Pre-processing Educational Data. Educational Data Mining: Applications and Trends, Springer Series Studies in Computational Intelligence. 29-64, 2014.

applied sciences

MDPI

Article

Towards Portability of Models for Predicting Students' Final Performance in University Courses Starting from Moodle Logs

Javier López-Zambrano [1,2], Juan A. Lara [3] and Cristóbal Romero [2,*]

1 Escuela Superior Politécnica Agropecuaria de Manabí (ESPAM MFL), Faculty of Computing, SISCOM Group, 131205 Calceta, Ecuador; jlopez.ec@outlook.com or jlopez@espam.edu.ec

2 Department of Computer Science and Numerical Analysis, University of Córdoba (UCO), 14014 Córdoba, Andalucía, Spain

3 Department of Computer Science, Madrid Open University (UDIMA), 28400 Madrid, Spain; juanalfonso.lara@udima.es

* Correspondence: cromero@uco.es

Received: 27 November 2019; Accepted: 31 December 2019; Published: 3 January 2020

check for updates

Abstract: Predicting students' academic performance is one of the older challenges faced by the educational scientific community. However, most of the research carried out in this area has focused on obtaining the best accuracy models for their specific single courses and only a few works have tried to discover under which circumstances a prediction model built on a source course can be used in other different but similar courses. Our motivation in this work is to study the portability of models obtained directly from Moodle logs of 24 university courses. The proposed method intends to check if grouping similar courses by the degree or the similar level of usage of activities provided by the Moodle logs, and if the use of numerical or categorical attributes affect in the portability of the prediction models. We have carried out two experiments by executing the well-known classification algorithm over all the datasets of the courses in order to obtain decision tree models and to test their portability to the other courses by comparing the obtained accuracy and loss of accuracy evaluation measures. The results obtained show that it is only feasible to directly transfer predictive models or apply them to different courses with an acceptable accuracy and without losing portability under some circumstances.

Keywords: Educational Data Mining; predicting student performance; student model portability

1. Introduction

The use of web-based education systems or e-learning systems has grown exponentially in the last years, spurred by the fact that neither students nor teachers are bound to any specific location and that this form of computer-based education is virtually independent of a specific hardware platform. Adopting these e-learning systems in higher educational institution can provide us with enormous quantities of data that describe the behavior of students. In particular, Learning Management Systems (LMSs) are becoming much more common in universities, community colleges, schools, and businesses, and are even used by individual instructors in order to add web technology to their courses and supplement traditional face-to-face courses. One of the most popular LMS is Moodle [1], a free and open-source learning management system that allows the creation of completely virtual courses (electronic learning, e-learning) or courses that are partially virtual (blended learning, b-learning). Moodle accumulate a vast amount of information, which is very valuable for analyzing students' behavior and could create a gold mine of educational data. Moodle keeps detailed logs of all events that students perform and keeps track of what materials students have accessed. However, due to the huge quantities of log data that Moodle can generate daily, it is very difficult to analyze them, thus, it is necessary to use Educational Data Mining (EDM) and Learning Analytics (LA) tools [2]. EDM and LA techniques discover useful, new, valid, and comprehensible knowledge from educational data in order to resolve educational problems [3]. There is a wide range of EDM/LA tasks or applications, but one of the oldest and most important ones is to predict student performance [4]. The objective of prediction is to estimate the unknown value of a variable that describes the student. In education the values normally predicted are performance, knowledge, score, or mark [5]. This value can be numerical/continuous value (regression task) or categorical/discrete value (classification task). In fact, nowadays, there is a great interest in analyzing and mining students' usage/interaction information gathered by Moodle for predicting students' final mark in blended learning [6,7]. Blended learning combines the e-learning and the classical face-to-face learning environments. It has been termed as blended learning, hybrid, or mixed learning [8]. Since either pure e-learning or traditional learning hold some weaknesses and strengths, it is better to mix the strengths of both learning environments into a new method of instruction delivery called blended learning.

Most of the research about predicting students' performance has focused on scenarios that assume that the training and test data are drawn from the same course [9]. As a matter of fact, the obtained/discovered models are mostly built on the samples that researchers have ready at hand, whether it is the current population of students at a university developing a model, the current user base of the adaptive learning system for which the model is being built, or just students who are relatively easy to survey or observe [10]. However, in real educational environments, we historical data are not always available from all the courses. Let us imagine, for example, the case of a new course that is taught for the first time in an institution. Here, we would not have data for training model for predicting student performance. Yet, it is fair that the tutors and students of this new subject have the chance to work with predictive models that notify them of possible unwanted at risk situations such as student drop out. Thus, model portability can be very useful to create and use transferable models of other similar course in which we have a prediction model.

The idea of Portability is that knowledge extracted from a specific environment can be applied directly to another different environment. Within the educational sphere, this idea has great applicability, as it permits to use a model discovered on a previous course (source) to an ongoing course (target) that does not have a model for any reason whatsoever, and to apply these models with certain guarantees to this new course [11]. Most of the previous works related with model portability use a Transfer Learning (TL) approach in which there is a tune-up process, usually based on deep learning approaches, so that the updated model is transferred from one course to another, as shown in [12,13]. Other different works use a Generalization approach that tries to discover one single general model that fit to all the exited courses [14,15]. This is the reason why, in this paper, we have used the term "portability" instead of the related terms "transferability" and "generalization", since we think that it better describes the direct application of a model obtained with one dataset to a different dataset. In this regard, the goal of this research is to study the portability of predictive models between courses taught via blended-learning (b-learning) in formal university education. These predictive models try to predict whether a student will succeed or not in a certain course (pass or fail) starting to the log data generated from the student interactions with Moodle LMS. Specifically, the problem we want to resolve is: if we have available data for different university courses, could we use or apply the performance prediction model obtained in one specific course in other different course (in which we do not have enough data or we do not have a prediction model) without losing much accuracy. However, due to that the number of courses in a University can be large, and thus, the number of combinations will be huge, and it seems logical to think that good model portability only occurs between similar courses. This is why, in this paper, we propose to group courses in two different ways; our main objective in this paper is to answer the following two research questions:

Can the models obtained in one (source) course be used in other different similar (target) courses of the same degree, while maintaining an acceptable predictive quality?

Can the models obtained in one (source) course be used in other different (target) courses that make a similar level of usage of Moodle activities/resources?

The rest of the document is arranged in the following order: Section 2 reviews the literature related to this research. Section 3 describes the data and experiments. The results are shown in Section 4. Section 5 discusses the results obtained. Finally, Section 6 presents the conclusions and suggests future lines of research.

2. Background

Within the EDM and LA scientific community, several works have been published that discuss the difficulty of achieving generalizable and portable models. In [14], the authors suggested that it is imperative for learning analytics research to account for the diverse ways technology is adopted and applied in a course-specific context. The differences in technology use, especially those related to whether and how learners use the learning management system, require consideration before the log-data can be merged to create a generalized model for predicting academic success. In [16], the authors stated that the portability of the prediction models across courses is low. In addition, they show that for the purpose of early intervention or when in-between assessment grades are taken into account, LMS data are of little (additional) value.

Nevertheless, Baker [10] considered that one of the challenges for the future of EDM is what he called the "Generalizability" problem or "The New York City and Marfa" problem. In his words, Learning Analytics models are mostly built on the samples that we have ready at hand, whether it is the current population of students at a university developing a model, the current user base of the adaptive learning system we are building the model for, or just students who are relatively easy to survey or observe. However, what happens when the population changes? He defined this problem in three steps: (1) Build an automated detector for a commonly-seen outcome or measure; (2) Collect a new population distinct from the original population; and (3) Demonstrate that the detector works for the new population with degradation of quality under 0.1 in terms of AUC ROC (Area Under the ROC -Receiver Operating Characteristic- Curve) and remaining better than chance (AUC ROC > 0.5).

In this regard, there are works that have demonstrated the possibility of replicating EDM models. In [17], they presented an open-source software toolkit, the MOOC (Massive Open Online Course) Replication Framework (MORF), and show that it is useful for replication of machine learned models in the domain of the learning sciences, in spite of experimental, methodological, and data barriers. This work demonstrates an approach to end-to-end machine learning replication, which is relevant to any domain with large, complex, or multi-format, privacy-protected data with a consistent schema.

What Baker [10] defined as "Generalizability" is, in reality, closely related to the concept of Transfer Learning (TL). Boyer and Veeramachaneni [11] defined TL as the attempt to transfer information (training data samples or models) from a previous course to establish a predictive model for an ongoing course. According to Hunt et al. [18], TL enables us to transfer the knowledge from a related (source) task that has already been learned, to a new (target) task. This idea breaks with the traditional view of attempting to learn a predictive model from the data from the on-going course itself, known as in-situ learning.

As listed in [11], there are various types of TL, among which are: (a) Naive Transfer Learning, when using samples from a previous course to help predict students' performance in a new course; (b) Inductive Transfer Learning, when certain class labels are available as attributes for the target course; and (c) Transductive Transfer Learning, where no labels are available for the target course data.

Transfer learning has been applied in the field of EDM and LA in different applications. In [18], they proposed an approach for predicting graduation rates in degree programs by leveraging data across multiple degree programs. There are also TL-based works for dropout prediction. In [12], they developed a framework to define classification problems across courses, provide proof that ensemble methods allow for the development of high-performing predictive models, and show that these techniques can be used across platforms, as well as across courses. Nevertheless, this study neither mentions each course topic nor does it analyze the transferability of the models. However, in [13] they proposed two alternative transfer methods based on representation learning with auto-encoders: a passive approach using transductive principal component analysis and an active approach that uses a correlation alignment loss term. With these methods, they investigate the transferability of dropout prediction across similar and dissimilar MOOCs and compare with known methods. Results show improved model transferability and suggest that the methods are capable of automatically learning a feature representation that expresses common predictive characteristics of MOOCs. A detailed description of the most relevant works in TL can be found in the survey presented in [9], and more recently, in the survey described in [19].

Domain Adaptation (DA) has gained ground in TL, being a particular case of TL that leverages labeled data in one or more related source domains, to learn a classifier for unseen or unlabeled data in a target domain [20]. In this regard, [21] propose an algorithm, called DiAd, which adapts a classifier trained on a course with labelled data by selectively choosing instances from a new course (with no label data) that are most dissimilar to the course with labelled data and on which the classifier is very confident of classification. A complete review of DA techniques can be found in [20] and [22].

Contextualizing our work in relation to the rest of the related research, we may affirm that our research is innovative and very interesting because it deals with one of the six challenges on EDM/LA community recently presented by Baker [10] named the "The New York City and Marfa Problem". Our work focuses on traditional university courses that use blended learning, while most of the previous works focus on MOOCs [11–13,21]. Although our research is very related to TL, as it fits the definitions of [11,18], it is not our goal to propose or study a specific tune-technique, similar to the latest research on DA [21], but only to study the direct portability of prediction models. To do so, we will follow the idea demonstrated in [13], but instead of carrying out tests with two subjects to prove the reliability of the method, our goal is to carry out a complete study with a greater number of courses in order to study the degree of model portability between subjects. Given that our study does not focus on any concrete technique, rather it studies the degree of portability of models; we use a direct transfer, also called Naive in [11]. This type of transfer has innumerable benefits such as simplicity and immediacy, which can aid other researchers in easily replicating our study with their own data. Additionally, studies such as [13] have demonstrated that this type of direct approach obtains better results than other approaches such as instance-based learning and even in-situ learning approaches. Taking all of this into account, and based on the extent of the authors' knowledge, this is the first study that measures the degree of model portability in blended learning university courses (not MOOCs), focusing on how portability of model is affected when using course of the same degrees and courses with similar level of usage of Moodle.

3. Materials and Methods

In this section, we describe the data used and the experiments we have conducted in order to answer the initial research questions.

3.1. Data Description and Preprocessing

We have downloaded the Moodle log files generated by 3235 students in 24 courses in different bachelor's degrees of University of Cordoba (UCO) in Spain as shown in Table 1. These courses can be from year 1 to year 4 of the bachelor's degree (most of them from year 1) and they have different numbers of students (#Stds in Table 1) ranging from 50 (minimum) to 302 (maximum). We have categorized each course depending on how many different Moodle's activities are used in each course, having three different usage levels (Low, Medium, and High), denoted LMS Level in Table 1, having found a medium level in most courses. We have defined three levels of usage according to the number of activities used in the course:

- Low level: The course only uses one type of activity or even none of them.
- Medium level: The course uses two different types of activities.

- High level: The course uses three or more different types of activities.

Table 1. Information about the courses.

Course Name	Code	Degree	Year	#Stds	LMS Level
Introduction to programming	IP	Computer	1	289	High
Programming methodology	PM	Computer	1	233	High
Professional computer tools	PCT	Computer	1	124	Medium
DataBases	DB	Computer	2	58	Medium
Human Computer Interfaces	HCI	Computer	2	260	High
Information Systems	InS	Computer	2	188	Medium
Software Engineering	SE	Computer	2	58	Medium
Interactive Systems	IS	Computer	3	84	High
Requirement engineering	RE	Computer	3	86	Medium
Software Design and Construction	SDC	Computer	3	50	Medium
Primary Education in the School System	PESS	Education	1	205	Medium
Knowledge of the Social and Cultural Environment	KSCE	Education	1	302	Low
Primary Education Planning and Innovation	PEPI	Education	2	117	Medium
Psychoeducational Care for the Cultural Diversity of Early Childhood Education	PECE	Education	4	55	Medium
Hermeneutics of the Work of Art	HWA	Education	4	83	Low
Spanish Social and Cultural Media	SSCM	Education	4	58	Medium
Introduction to Psychology	IPs	Education	4	91	High
Introduction to Computer Science	ICS1	Electrical Engineering	1	100	Low
Introduction to Computer Science	ICS2	Electronic Engineering	1	198	High
Introduction to Computer Science	ICS3	Civil Engineering	1	85	Low
Introduction to Computer Science	ICS4	Mining Engineering	1	77	Low
Mathematics Analysis I	MA1	Physics	1	155	Low
Mathematical Analysis II	MA2	Physics	1	160	Low
Mathematical Methods	MM	Physics	1	119	Low

Finally, it is important to notice that the class (final marks) of the students in these courses is not unbalanced, that is, there are not many differences between the number of students who pass the course and the number of students who fail the course. In addition, although all courses have a little imbalance (between 50%–70% for each class), this is not a problem for most machine learning algorithms since standard performance evaluation measures remain effective in those scenarios with such a little imbalance rate [23].

In order to preprocess the Moodle's log files and to add the course final marks, we have developed a specific Java GUI (Graphical User Interface) tool for preprocessing this type of files [24]. This is a visual and easy-to-use tool for preparing both the raw Excel students' data files directly downloaded from Moodle's courses interface and the Excel students mark files provided by instructors.

Firstly, it shows the content of the Excel files and allows selecting the specific columns where the required information is located: Name of the students, Date and Events (Moodle events) in the Log file and Name of student and Marks (final mark in the course that has a value between 0 and 10) in the Grades file. It joins the information about each student (events and mark) and it anonymizes the data by deleting the name of the students. Next, it allows the user to select what events (all of them or just a few) should be used as attributes in the final dataset. In our case, we have only selected 50 attributes (see Table 2) from all the events that appear in our logs files (we have removed all the instructor's and administrator's events). As can be seen from Table 2, we have considered attributes related to the interactions of students with assignments, choices, forums, pages, quizzes, wikis, and others.

Then, the specific starting and ending date of the course can be established in order to count only the number of events that occurred between these dates for each student. Next, it is possible to transform these values defined in a continuous domain/range into discrete or categorical values. This tool provides the option of performing a manual discretization (by specifying the cut points) as well as traditional techniques such as equal-width or equal-frequency. In our case, we are going to generate two different datasets for each course: one continuous dataset (with numerical values in all the attributes less the class attribute) and another discretized datasets (with categorical values in all the attributes). We have discretized all the Moodle's attributes using the equal-width method (it divides the data into k intervals of equal size) with the two labels HIGH and LOW. Moreover, we have manually discretized the students' final grade attribute, that is, the class to predict in our classification problem, to two values or labels: FAIL (if the mark is lower than 5) and PASS (if the mark is greater or equal than 5). Finally, this tool allows us to generate a preprocessed data file in. ARFF (Attribute-Relation File Format) format for doing data mining. It is important to notice that all the data used has been treated according to academic ethics. In fact, firstly we requested the instructors of each course to download the log files of their courses from Moodle together with an excel file with the final marks of the students. Then, we signed a declaration for each course stating that we would use the data only for researching purposes and would anonymize them after integrating the students' events with their corresponding final marks as a previous step to the application of data mining algorithms.

Table 2. List of Moodle logs attributes/events used.

Assignments	**Folders**	**Quizzes**
1. assign submit	17. folder view	34. quiz attempt
2. assign submit for grading	18. folder view all	35. quiz close attempt
3. assign view	**Forums**	36. quiz continue attempt
4. assign view all		37. quiz review
5. assignment upload	19. forum add discussion	38. quiz view
6. assignment view	20. forum add post	39. quiz view all
7. assignment view all	21. forum mark read	40. quiz view summary
Choices	22. forum search	**Resources**
	23. forum subscribe	
8. choice choose	24. forum subscribe all	41. resource view
9. choice choose again	25. forum unsubscribe	42. resource view all
10. choice report	26. forum view discussion	**Urls**
11. choice view	27. forum view forum	43. url view
12. choice view all	28. forum view forums	44. url view all
Courses	**Pages**	**Wikis**
13. course enroll	29. page view	
14. course user report	30. page view all	45. wiki edit
15. course view	**Questionnaires**	46. wiki info
16. course view section		47. wiki links
		48. wiki update
	31. questionnaire submit	49. wiki view
	32. questionnaire view	50. wiki view all
	33. questionnaire view all	

3.2. Experimentation

For each of the mentioned 24 UCO courses, we have considered two datasets: one of them in which we have used continuous values of attributes (called Numerical Dataset); and the other one in which we have used discretized values of those attributes (called Discretized Dataset). This means we had 48 datasets in total. In order to answer the two research questions described in the Introduction section, we conducted two types of experiments that we will describe in detail later (denoted "Experiment 1" and "Experiment 2") in which we categorize the courses into different groups. In those experiments, for each of the 48 datasets, we have measured the portability of each obtained model to the rest of the courses of the same group. We have used WEKA (Waikato Environment for Knowledge Analysis) [25] because it is a well-known open-source machine learning tool that provides a huge number of classification algorithms and evaluation measures. In fact, we have compared the portability of the models obtained by using the J48 classification algorithm, the AUC and the loss of AUC (difference in two AUC values) as evaluation performance measures. An explanation of the key points in which this choice is based can be found in the coming paragraphs.

We have used the well-known J48 classification algorithm, namely, the Weka version of the C4.5 algorithm [26]. J48 is a re-implementation in Java programming language of C4.5 release 8 (hence the name J48). We have selected this algorithm for two main reasons. The first one is that it is a popular white box classifier that provides a decision tree as model output. Decision trees are very interpretable or comprehensible models that explain the predictions in the form of IF-THEN rules in a decision tree [27] and it has been widely used in education for predicting student performance. The second one is that C4.5 became quite popular after ranking #1 in the Top 10 Algorithms in Data Mining pre-eminent paper published by Springer LNCS in 2008 [28].

We have used AUC and AUC loss as evaluation measures of the performance of the classifier because: (a) AUC is one of the evaluation measures most commonly used for assessing students' performance predictive models [29–31]; and, (b) AUC loss is also proposed by Baker in his Learning Analytics Prizes [10] as the evaluation measure for testing whether or not his transfer challenge has been solved. The Area Under the ROC Curve (AUC) is a universal statistical indicator for describing the accuracy of a model regarding predicting a phenomenon [32]. It has been widely used in education research for comparing classification algorithms and models [33,34] instead of other well-known evaluation measures such as Accuracy, F-measure, Sensitivity, Precision, etc. AUC can be defined as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'). It is often used as a measure of the quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1. We have also calculated the AUC loss or difference between the two AUC values obtained when applying the model over the source dataset and when applying over the target dataset.

The general procedure of our experiments has been summarized in Figure 1, where we graphically show the main steps of the experiments by using a flow diagram.

An overall explanation of the main steps (see Figure 2) of our experiments is:

- Firstly, Moodle logs have to be pre-processed (step 1) in order to obtain numerical and discretized datasets according to the format expected in the data mining tool to be used, Weka.
- Then, for each course dataset (numerical and discretized), the algorithm J48 is run in order to obtain a general prediction model (step 2) to be used in portability experiments.
- Next, courses are grouped according to 2 different criteria to conduct two types of experiments (step 3); for the first experiment (named "Experiment 1"), related courses are grouped by the area of knowledge (attribute "Degree" in Table 1); for the second experiment ("Experiment 2"), groups of courses are built according to the Moodle usage ("Moodle Usage" in Table 1).
- In each experiment, each model is selected (step 4) and tested against the rest of the datasets of courses belonging to the same group (step 5), repeating this process for each course.
- Finally, AUC values are obtained and AUC loss values are calculated when using the model against the rest of the courses of the same group (step 6).

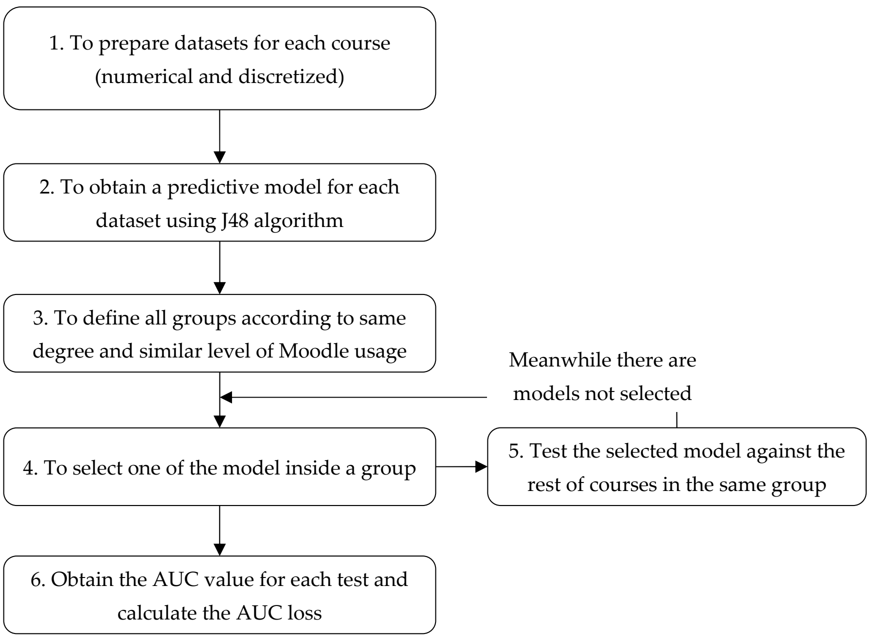

Figure 1. Experiments procedure steps.

Figure 1. Experiments procedure steps.

An overall explanation of the main steps (see Figure 2) of our experiments is:

- Firstly, Moodle logs have to be pre-processed (step 1) in order to obtain numerical and discretized datasets according to the format expected in the data mining tool to be used, Weka.
- Then, for each course dataset (numerical and discretized), the algorithm J48 is run in order to obtain a general prediction model (step 2) to be used in portability experiments.
- Next, courses are grouped according to 2 different criteria to conduct two types of experiments

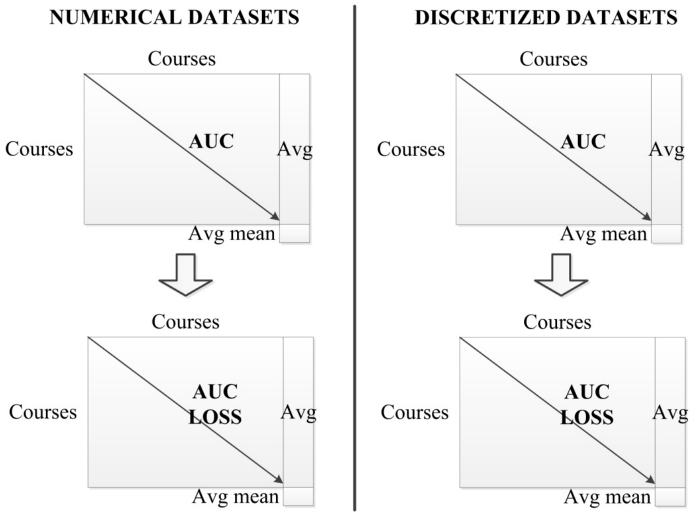

Figure 2. Visual description of the results in four tables.

In the next two subsections, we provide a more detailed explanation of how we have grouped similar courses in the two experiments.

3.2.1. Groups of Experiment 1

In this portability test, four groups of similar courses were used, according to the degree they belong to, as shown in Table 3. Our idea is that all courses of the same degree must be related and can be similar in the subjects.

Table 3. List of groups by degree.

N.	Group	N. of Subjects
1	Computer	10
2	Education	7
3	Engineering	4
4	Physics	3

It is noticeable in Table 3 that there are a higher number of courses in the Computer degree and in the Education degree, than in the Engineering and Physics degrees. In general, both Science and Humanities areas are considered in this study.

3.2.2. Groups of Experiment 2

In this portability experiment, three groups of similar courses were used, according to the respective Moodle usage level, as shown in Table 4. Moodle is an LMS that provides different types of activities (assignments, chat, choice, database, forum, glossary, lesson, quiz, survey, wiki, workshop, etc.). Our idea is that courses that use similar activities will have the same level of usage and these activities are related to the fact of passing or failing the course [2,6].

Table 4. List of groups by Moodle usage.

N.	Group	Number of Subjects
1	High	6
2	Medium	10
3	Low	8

Figure 2. Visual description of the results in four tables.

4.1. Experiment 1

In this experiment we assess the portability of prediction models between courses belonging to the same degree, having considered four different groups (Computer, Education, Engineering, and Physics). Firstly, we have obtained 24 prediction models (one for each course) and then we have tested them with the other courses' datasets of the same group, which in this case is a total of 174 numerical and 174 discretized datasets. Thus, we have carried out a total of 348 executions of J48 algorithm for obtaining each AUC value and then calculated the AUC loss versus the reference model.

For the Computer group, we can observe from Table 5 that the best AUC value (0.896) when transferring a prediction model to a different course corresponds to the PM (refer to Table 1 for course names abbreviations) course model when tested against the DB course numerical dataset. However, we can observe that the overall mean value for AUC measure with discretized datasets is 0.56, which means that the predictive ability of models when used in other subjects of this group is lightly above randomness. Something similar happens with numerical datasets, where the average value is 0.57. We can also observe that the lowest (best) AUC loss in discretized dataset is close to the perfect portability (0.006). This value is obtained when using the PM model against the RE subject. Overall, we can observe that AUC loss is better in the discretized dataset than in the numerical one (0.23 versus 0.33 in average). We can also highlight that the best average values in terms of portability loss are obtained for DB course in numerical dataset and PM course in discretized datasets (0.10 in both cases).

For the Education group, we can observe from Table 6 that the best AUC value (0.708) is obtained when using the prediction model of PESS course against the SSCM course discretized datasets. The overall average AUC for this group's discretized dataset (0.56) is very similar to that for the numerical datasets (0.57). In addition, we noticed that portability loss (AUC loss) is near-perfect (0.003) when testing the PEPI model against HWA course dataset in the discretized dataset. The overall average portability loss for discretized dataset experiments is 0.29, much better than the mean value obtained for numerical dataset experiments (0.39). We can also highlight that the best average values in terms of portability loss correspond to PEPI course (0.30 for numerical datasets and 0.11 for discretized datasets).

It is important to notice that the most popular activities in our 24 courses are assignments, forums, and quizzes. Normally, low level courses only use one of these three kinds of activities, medium level courses use two of them, and high level courses use all three mentioned types of activities or even more.

4. Results

In this section we show the results obtained from the two sets of experiments carried out. We present the AUC and the loss of AUC in four different matrixes (two for numerical datasets and two for discretized datasets) for each group of similar courses (see Figure 2). In the upper part, we show two matrixes containing the AUC metric values that we have obtained when testing each course model (row) against the rest of the courses datasets (columns) using the numerical and the discretized datasets. The matrix main diagonal values correspond to the tests carried out for each course model against its own dataset, which means those AUC values (the highest ones) constitute the reference value (in green color) when compared with the rest of the courses. We have also calculated the average AUC values for each course (column denoted as "*avg*" in the tables) and the overall mean value for the group (cell denoted as "*avg mean*" in the tables). In the lower part, we show two matrixes showing the difference values between the highest AUC (row), which is considered to be the reference value, and the AUC values obtained when applying the corresponding model to each of the rest of the courses in the same group (column) using the numerical and the discretized datasets. Finally, our analysis focused on finding the best or highest AUC values and the best or lowest rates of AUC loss in each group of similar courses. Thus, we highlighted (in bold) the highest AUC values (without considering the reference value) and the lowest AUC loss values, which will represent the lowest portability loss, and thus the best results.

4.1. Experiment 1

In this experiment we assess the portability of prediction models between courses belonging to the same degree, having considered four different groups (Computer, Education, Engineering, and Physics). Firstly, we have obtained 24 prediction models (one for each course) and then we have tested them with the other courses' datasets of the same group, which in this case is a total of 174 numerical and 174 discretized datasets. Thus, we have carried out a total of 348 executions of J48 algorithm for obtaining each AUC value and then calculated the AUC loss versus the reference model.

For the Computer group, we can observe from Table 5 that the best AUC value (0.896) when transferring a prediction model to a different course corresponds to the PM (refer to Table 1 for course names abbreviations) course model when tested against the DB course numerical dataset. However, we can observe that the overall mean value for AUC measure with discretized datasets is 0.56, which means that the predictive ability of models when used in other subjects of this group is lightly above randomness. Something similar happens with numerical datasets, where the average value is 0.57. We can also observe that the lowest (best) AUC loss in discretized dataset is close to the perfect portability (0.006). This value is obtained when using the PM model against the RE subject. Overall, we can observe that AUC loss is better in the discretized dataset than in the numerical one (0.23 versus 0.33 in average). We can also highlight that the best average values in terms of portability loss are obtained for DB course in numerical dataset and PM course in discretized datasets (0.10 in both cases).

For the Education group, we can observe from Table 6 that the best AUC value (0.708) is obtained when using the prediction model of PESS course against the SSCM course discretized datasets. The overall average AUC for this group's discretized dataset (0.56) is very similar to that for the numerical datasets (0.57). In addition, we noticed that portability loss (AUC loss) is near-perfect (0.003) when testing the PEPI model against HWA course dataset in the discretized datasets. The overall average portability loss for discretized dataset experiments is 0.29, much better than the mean value obtained for numerical dataset experiments (0.39). We can also highlight that the best average values in terms of portability loss correspond to PEPI course (0.30 for numerical datasets and 0.11 for discretized datasets).

For the Engineering group, we can observe from Table 7 that the best AUC value (0.636) is obtained for ICS2 course prediction model when tested against ICS1 course discretized dataset. In this experiment, we can observe that the overall average value of AUC is again better in the numerical dataset (0.59) than in the discretized one (0.56), with both values staying above randomness. In addition, we can observe that the best portability loss value of 0.126 is obtained for ICS2 course model when tested against ICS1 course dataset in discretized datasets. Again, we obtain better results in the discretized than in the numerical dataset (0.24 versus 0.30) in terms of portability loss. We can also highlight that the best course average portability loss values are obtained for ICS3 in numerical dataset (0.20) and for ICS1 subject in discretized dataset (0.22).

Finally, for the Physics group, we can see in Table 8 that the highest AUC value (0.641) corresponds to the MM course prediction model when tested against the MA2 course numerical dataset. This value is very close to the overall mean value for the numerical dataset (0.68), which outperforms the overall AUC mean value for the discretized dataset (0.60). If we look at the portability loss values, we notice that the best (the lowest) AUC loss value of 0.009 is obtained when testing the MM course model against MA1 course discretized dataset. In this group, again, the global mean values are better for the discretized dataset than for the numerical one (0.09 versus 0.28), which means that the portability loss rate is particularly lower in this experiment in the discretized dataset compared to the numerical one. We can also highlight that the best course portability loss values are obtained for MM course model in both the numerical (0.21) and discretized (0.04) datasets.

4.2. Experiment 2

In this experiment we assess the portability of prediction models between courses with a similar level of usage of Moodle activities. In fact, we have considered three different groups (High, Medium, and Low). Firstly, we have obtained 24 prediction models (one for each course), and then, we tested them with other courses datasets of the same group, in this case a total of 204 numerical and 204 discretized datasets. Thus, we have carried out a total of 400 executions of J48 algorithm for obtaining each AUC value and then calculating the AUC loss versus the reference model.

For the high level group, as we can see from Table 9, the best value for AUC measure (0.656) is obtained when testing the IS course prediction model against the PM course discretized dataset. In this experiment (and equal than in the previous ones), the overall AUC means values are very similar for numerical (0.58) and discretized datasets (0.57). If we have a look at portability loss values, we can see that the best AUC loss value (0.061) is obtained when testing the ICS2 model against IS discretized dataset. In general, the average mean of AUC loss is better for the discretized datasets than for the numerical datasets (0.24 versus 0.37). Finally, we highlighted the average values of AUC loss for the ICS2 course, which are the lowest both in numerical datasets (0.25) and in discretized datasets (0.10).

For the medium level group, we can observe from Table 10 that the best AUC value of 0.792 corresponds to the prediction model of SDC course when tested against the discretized dataset of the SSCM course. The global average AUC value for this discretized category (0.53) is very similar to the global AUC value for numerical datasets (0.55). Regarding portability loss, we can see that the best value (0.009) belongs to DB prediction model when tested against PEPI discretized dataset. Moreover, again, the portability loss is better in the discretized datasets (0.25) than in the numerical datasets (0.38). Finally, we would also like to highlight the good average AUC loss results obtained by the InS course prediction model with the numerical datasets (0.12) and DB course prediction model in the discretized datasets (0.14).

Finally, for the low level group, we can see from Table 11 that the best AUC measure value (0.758) is obtained when testing the ICS3 prediction model against the HWA numerical dataset. The global average value for the numerical dataset (0.57) is a bit better than the obtained value by the discretized dataset (0.54). We can also notice that the best portability loss value is obtained when testing the MM model against HWA discretized dataset (0.028). The overall mean value for portability loss measure is also better for discretized than for numerical datasets (0.22 versus 0.34). Additionally, the best course prediction model on average values in terms of portability loss correspond to KSCE for numerical dataset (0.16) and MM for the discretized dataset (0.12).

Table 5. AUC Results and Loss in Portability in Computer degree group.

	AUC (Numerical Datasets)											AUC (Discretized Datasets)										
Course	**HCI**	**IS**	**IP**	**PM**	**DB**	**SDC**	**PCT**	**RE**	**SE**	**InS**	**avg**	**HCI**	**IS**	**IP**	**PM**	**DB**	**SDC**	**PCT**	**RE**	**SE**	**InS**	**avg**
HCI	0.943	0.510	0.555	0.524	0.505	0.507	0.576	0.500	0.491	0.460	0.56	0.769	0.621	0.496	0.570	0.590	0.541	0.543	0.510	0.561	0.525	0.57
IS	0.522	0.966	0.526	0.522	0.521	0.604	0.483	0.691	0.491	0.482	0.58	0.557	0.854	0.643	0.551	0.513	0.573	0.545	0.534	0.460	0.687	0.59
IP	0.524	0.616	0.931	0.652	0.608	0.562	0.495	0.493	0.442	0.590	0.59	0.496	0.501	0.827	0.674	0.621	0.478	0.500	0.500	0.434	0.541	0.56
PM	0.514	0.500	0.687	0.915	**0.896**	0.550	0.554	0.518	0.530	0.544	0.62	0.550	0.622	0.598	0.715	0.646	0.682	0.562	**0.710**	0.542	0.597	0.62
DB	0.527	0.249	0.513	0.564	0.601	0.574	0.425	0.544	0.593	0.477	0.51	0.469	0.490	0.450	0.593	0.602	0.508	0.471	0.575	0.444	0.466	0.51
SDC	0.510	0.660	0.492	0.523	0.572	0.844	0.599	0.558	0.467	0.514	0.57	0.475	0.626	0.523	0.484	0.484	0.783	0.544	0.605	0.481	0.579	0.56
PCT	0.510	0.660	0.492	0.523	0.572	0.599	0.844	0.558	0.467	0.514	0.57	0.475	0.626	0.523	0.484	0.484	0.544	0.783	0.605	0.481	0.579	0.56
RE	0.490	0.563	0.491	0.521	0.515	0.525	0.561	0.992	0.491	0.551	0.57	0.499	0.549	0.444	0.516	0.547	0.620	0.514	0.845	0.582	0.508	0.56
SE	0.515	0.408	0.603	0.504	0.508	0.557	0.475	0.623	0.978	0.429	0.56	0.441	0.558	0.462	0.446	0.471	0.477	0.511	0.569	0.729	0.463	0.51
InS	0.479	0.481	0.533	0.551	0.544	0.572	0.501	0.610	0.525	0.673	0.55	0.492	0.634	0.546	0.606	0.570	0.578	0.533	0.565	0.510	0.792	0.58
					avg mean						0.57					avg mean						0.56

	AUC LOSS (Numerical Datasets)											AUC LOSS (Discretized Datasets)										
Course	**HCI**	**IS**	**IP**	**PM**	**DB**	**SDC**	**PCT**	**RE**	**SE**	**InS**	**avg**	**HCI**	**IS**	**IP**	**PM**	**DB**	**SDC**	**PCT**	**RE**	**SE**	**InS**	**avg**
HCI	-	0.432	0.388	0.418	0.437	0.436	0.367	0.443	0.452	0.483	0.43	-	0.148	0.273	0.200	0.179	0.228	0.226	0.260	0.208	0.245	0.22
IS	0.444	-	0.440	0.444	0.445	0.363	0.483	0.276	0.475	0.484	0.43	0.297	-	0.211	0.303	0.341	0.281	0.309	0.321	0.394	0.167	0.29
IP	0.408	0.315	-	0.279	0.323	0.370	0.436	0.439	0.489	0.341	0.38	0.331	0.326	-	0.153	0.206	0.349	0.327	0.327	0.393	0.286	0.30
PM	0.401	0.415	0.228	-	0.019	0.365	0.361	0.397	0.385	0.370	0.33	0.165	0.094	0.118	-	0.069	0.034	0.153	**0.006**	0.173	0.119	**0.10**
DB	0.074	0.352	0.087	0.037	-	0.027	0.176	0.057	**0.008**	0.124	**0.10**	0.134	0.112	0.152	0.009	-	0.094	0.131	0.027	0.158	0.136	0.11
SDC	0.335	0.184	0.352	0.321	0.272	-	0.245	0.287	0.377	0.330	0.30	0.308	0.157	0.260	0.298	0.299	-	0.239	0.178	0.302	0.204	0.25
PCT	0.335	0.184	0.352	0.321	0.272	0.245	-	0.287	0.377	0.330	0.30	0.308	0.157	0.260	0.298	0.299	0.239	-	0.178	0.302	0.204	0.25
RE	0.501	0.429	0.501	0.470	0.476	0.467	0.431	-	0.501	0.441	0.47	0.346	0.296	0.400	0.329	0.297	0.225	0.331	-	0.263	0.337	0.31
SE	0.464	0.570	0.375	0.474	0.470	0.422	0.503	0.356	-	0.549	0.46	0.288	0.171	0.267	0.283	0.258	0.252	0.218	0.160	-	0.266	0.24
InS	0.193	0.192	0.140	0.122	0.128	0.101	0.172	0.063	0.148	-	0.14	0.300	0.158	0.246	0.186	0.222	0.215	0.259	0.227	0.282	-	0.23
					avg mean						0.33					avg mean						0.23

Table 6. AUC Results and Loss in Portability in Education degree group.

	AUC (Numerical Datasets)								AUC (Discretized Datasets)							
Course	**PESS**	**SSCM**	**PEPI**	**PECE**	**HWA**	**KSCE**	**IPs**	**avg**	**PESS**	**SSCM**	**PEPI**	**PECE**	**HWA**	**KSCE**	**IPs**	**avg**
PESS	0.938	0.554	0.553	0.548	**0.667**	0.558	0.535	0.62	0.805	**0.708**	0.526	0.331	0.500	0.525	0.611	0.57
SSCM	0.629	0.843	0.574	0.395	0.667	0.530	0.522	0.59	0.560	0.839	0.466	0.366	0.500	0.500	0.515	0.54
PEPI	0.490	0.587	0.839	0.562	0.556	0.499	0.552	0.58	0.483	0.572	0.670	0.568	0.667	0.460	0.597	0.57
PECE	0.447	0.265	0.463	0.972	0.333	0.467	0.541	0.50	0.308	0.342	0.515	0.749	0.500	0.500	0.465	0.48
HWA	0.493	0.533	0.441	0.574	1.000	0.543	0.534	0.59	0.549	0.569	0.549	0.488	0.778	0.532	0.515	0.57
KSCE	0.531	0.575	0.459	0.516	0.354	0.817	0.500	0.54	0.550	0.679	0.523	0.472	0.608	0.931	0.583	0.62
IPs	0.586	0.322	0.643	0.519	0.528	0.625	0.921	0.59	0.556	0.500	0.505	0.498	0.618	0.542	0.884	0.59
				avg mean				0.57				avg mean				0.56

	AUC LOSS (Numerical Datasets)								AUC LOSS (Discretized Datasets)							
Course	**PESS**	**SSCM**	**PEPI**	**PECE**	**HWA**	**KSCE**	**IPs**	**avg**	**PESS**	**SSCM**	**PEPI**	**PECE**	**HWA**	**KSCE**	**IPs**	**avg**
PESS	-	0.384	0.385	0.390	0.271	0.380	0.403	0.37	-	0.097	0.279	0.474	0.305	0.280	0.195	0.27
SSCM	0.214	-	0.269	0.448	0.176	0.313	0.321	0.29	0.279	-	0.373	0.473	0.339	0.339	0.324	0.35
PEPI	0.349	0.253	-	0.277	0.283	0.340	0.288	**0.30**	0.187	0.099	-	0.102	**0.003**	0.210	0.073	**0.11**
PECE	0.526	0.707	0.509	-	0.639	0.505	0.431	0.55	0.442	0.408	0.234	-	0.249	0.249	0.285	0.31
HWA	0.507	0.468	0.559	0.426	-	0.457	0.466	0.48	0.229	0.210	0.229	0.290	-	0.246	0.263	0.24
KSCE	0.286	**0.243**	0.358	0.301	0.463	-	0.317	0.33	0.381	0.253	0.408	0.459	0.323	-	0.348	0.36
IPs	0.336	0.600	0.278	0.402	0.393	0.296	-	0.38	0.329	0.385	0.379	0.386	0.266	0.342	-	0.35
				avg mean				0.39				avg mean				0.29

Table 7. AUC Results and Loss in Portability in Engineering degree group.

	AUC (Numerical Datasets)					AUC (Discretized Datasets)				
Course	**ICS1**	**ICS2**	**ICS3**	**ICS4**	**avg**	**ICS1**	**ICS2**	**ICS3**	**ICS4**	**avg**
ICS1	0.958	0.477	0.464	0.569	0.62	0.742	0.535	0.554	0.474	0.58
ICS2	**0.576**	0.789	0.504	0.557	0.61	**0.636**	0.761	0.523	0.402	0.58
ICS3	0.544	0.547	0.739	0.525	0.59	0.446	0.506	0.739	0.514	0.55
ICS4	0.410	0.477	0.542	0.790	0.55	0.428	0.455	0.483	0.685	0.51
		avg mean			0.59		avg mean			0.56
	AUC LOSS (Numerical Datasets)					AUC LOSS (Discretized Datasets)				
Course	**ICS1**	**ICS2**	**ICS3**	**ICS4**	**avg**	**ICS1**	**ICS2**	**ICS3**	**ICS4**	**avg**
ICS1	-	0.480	0.494	0.389	0.45	-	0.206	0.187	0.268	**0.22**
ICS2	0.213	-	0.285	0.231	0.24	**0.126**	-	0.238	0.359	0.24
ICS3	0.195	**0.192**	-	0.214	**0.20**	0.293	0.233	-	0.225	0.25
ICS4	0.380	0.314	0.248	-	0.31	0.257	0.230	0.202	-	0.23
		avg mean			0.30		avg mean			0.24

Table 8. AUC Results and Loss in Portability in Physics degree group.

	AUC (Numerical Datasets)				AUC (Discretized Datasets)			
Course	**MM**	**MA1**	**MA2**	**avg**	**MM**	**MA1**	**MA2**	**avg**
MM	0.807	0.559	**0.641**	0.67	0.639	**0.630**	0.563	0.61
MA1	0.542	0.880	0.591	0.67	0.578	0.697	0.603	0.63
MA2	0.574	0.592	0.905	0.69	0.546	0.525	0.642	0.57
		avg mean		0.68		avg mean		0.60
	AUC LOSS (Numerical Datasets)				AUC LOSS (Discretized Datasets)			
Course	**MM**	**MA1**	**MA2**	**avg**	**MM**	**MA1**	**MA2**	**avg**
MM	-	0.249	**0.166**	**0.21**	-	**0.009**	0.076	**0.04**
MA1	0.337	-	0.288	0.31	0.119	-	0.094	0.11
MA2	0.331	0.313	-	0.32	0.096	0.117	-	0.11
		avg mean		0.28		avg mean		0.09

Table 9. AUC Results and Loss in Portability in high level of usage of Moodle group.

	AUC (Numerical Datasets)							AUC (Discretized Datasets)						
Course	**HCI**	**IS**	**ICS2**	**IP**	**PM**	**IPs**	**avg**	**HCI**	**IS**	**ICS2**	**IP**	**PM**	**IPs**	**avg**
HCI	0.943	0.510	0.522	0.538	0.524	0.457	0.58	0.769	0.621	0.569	0.417	0.570	0.550	0.58
IS	0.485	0.927	0.494	0.470	0.606	0.520	0.58	0.479	0.816	0.577	0.555	**0.656**	0.596	0.61
ICS2	0.514	0.590	0.783	0.500	0.569	0.513	0.58	0.503	0.558	0.619	0.485	0.516	0.552	0.54
IP	0.484	0.420	0.472	0.862	0.490	**0.627**	0.56	0.519	0.576	0.535	0.761	0.491	0.409	0.55
PM	0.514	0.489	0.530	0.618	0.899	0.610	0.61	0.574	0.488	0.522	0.592	0.793	0.480	0.57
IPs	0.516	0.529	0.514	0.427	0.597	0.921	0.58	0.507	0.638	0.485	0.514	0.460	0.884	0.58
			avg mean				0.58			avg mean				0.57
	AUC LOSS (Numerical Datasets)							**AUC LOSS (Discretized Datasets)**						
Course	**HCI**	**IS**	**ICS2**	**IP**	**PM**	**IPs**	**avg**	**HCI**	**IS**	**ICS2**	**IP**	**PM**	**IPs**	**avg**
HCI	-	0.432	0.421	0.404	0.418	0.486	0.43	-	0.148	0.201	0.352	0.200	0.220	0.22
IS	0.442	-	0.433	0.457	0.321	0.407	0.41	0.337	-	0.238	0.260	0.160	0.219	0.24
ICS2	0.270	**0.193**	-	0.283	0.215	0.271	**0.25**	0.116	**0.061**	-	0.134	0.103	0.067	**0.10**
IP	0.378	0.441	0.390	-	0.371	0.235	0.36	0.242	0.184	0.225	-	0.269	0.352	0.25
PM	0.385	0.410	0.369	0.281	-	0.290	0.35	0.219	0.305	0.271	0.200	-	0.313	0.26
IPs	0.405	0.392	0.407	0.495	0.324	-	0.40	0.377	0.246	0.400	0.370	0.424	-	0.36
			avg mean				**0.37**			avg mean				**0.24**

Table 10. AUC Results and Loss in Portability in medium level of usage of Moodle group.

	AUC (Numerical Datasets)											AUC (Discretized Datasets)										
Course	**SSCM**	**DB**	**SDC**	**PCT**	**RE**	**SE**	**InS**	**PECE**	**PESS**	**PEPI**	**avg**	**SSCM**	**DB**	**SDC**	**PCT**	**RE**	**SE**	**InS**	**PECE**	**PESS**	**PEPI**	**avg**
SSCM	0.839	0.521	0.549	0.464	0.500	0.489	0.546	0.366	0.560	0.466	0.53	0.843	0.492	0.698	0.514	0.635	0.513	0.583	0.395	0.629	0.574	0.59
DB	0.223	0.976	0.535	0.457	0.670	0.581	0.517	0.456	0.544	0.539	0.55	0.422	0.652	0.551	0.476	0.500	0.510	0.499	0.500	0.500	0.643	0.53
SDC	0.610	0.467	0.809	0.504	0.558	0.496	0.456	0.571	0.514	0.467	0.55	**0.792**	0.430	0.924	0.531	0.610	0.484	0.622	0.268	0.664	0.506	0.58
PCT	0.495	0.337	0.585	0.891	0.612	0.382	0.492	0.422	0.324	0.431	0.50	0.683	0.447	0.567	0.712	0.553	0.470	0.551	0.286	0.569	0.500	0.53
RE	0.456	0.329	0.553	0.579	0.956	0.473	0.577	0.465	0.607	0.487	0.55	0.491	0.529	0.614	0.508	0.756	0.545	0.569	0.521	0.597	0.542	0.57
SE	0.417	0.611	0.559	0.486	0.614	0.964	0.494	0.517	0.665	0.542	0.59	0.425	0.500	0.375	0.473	0.431	0.718	0.451	0.000	0.272	0.556	0.42
InS	0.605	0.671	0.583	0.486	0.610	0.533	0.704	0.564	**0.684**	0.494	0.59	0.512	0.429	0.625	0.528	0.454	0.500	0.761	0.610	0.432	0.502	0.54
PECE	0.265	0.520	0.371	0.505	0.281	0.471	0.548	0.972	0.447	0.463	0.48	0.342	0.553	0.550	0.468	0.559	0.530	0.463	0.749	0.308	0.515	0.50
PESS	0.554	0.471	0.547	0.509	0.579	0.579	0.582	0.548	0.938	0.553	0.59	0.708	0.461	0.618	0.519	0.518	0.465	0.606	0.331	0.805	0.526	0.56
PEPI	0.587	0.323	0.574	0.540	0.499	0.481	0.542	0.562	0.490	0.839	0.54	0.572	0.500	0.435	0.505	0.454	0.504	0.590	0.568	0.483	0.712	0.53
					avg mean						0.55					avg mean						0.53

	AUC LOSS (Numerical Datasets)											AUC LOSS (Discretized Datasets)										
Course	**SSCM**	**DB**	**SDC**	**PCT**	**RE**	**SE**	**InS**	**PECE**	**PESS**	**PEPI**	**avg**	**SSCM**	**DB**	**SDC**	**PCT**	**RE**	**SE**	**InS**	**PECE**	**PESS**	**PEPI**	**avg**
SSCM	-	0.318	0.290	0.375	0.339	0.350	0.293	0.473	0.279	0.373	0.34	-	0.351	0.145	0.329	0.208	0.330	0.260	0.448	0.214	0.269	0.28
DB	0.754	-	0.441	0.519	0.307	0.395	0.459	0.520	0.432	0.437	0.47	0.230	-	0.101	0.176	0.152	0.142	0.153	0.152	0.152	**0.009**	**0.14**
SDC	0.199	0.342	-	0.305	0.252	0.313	0.353	0.238	0.296	0.342	0.29	0.132	0.494	-	0.393	0.314	0.440	0.302	0.656	0.261	0.418	0.38
PCT	0.397	0.554	0.306	-	0.279	0.509	0.399	0.469	0.568	0.460	0.44	0.029	0.265	0.145	-	0.159	0.242	0.161	0.426	0.143	0.212	0.20
RE	0.500	0.627	0.403	0.377	-	0.483	0.379	0.491	0.350	0.469	0.45	0.265	0.227	0.142	0.248	-	0.211	0.187	0.235	0.160	0.214	0.21
SE	0.548	0.353	0.405	0.478	0.351	-	0.470	0.447	0.299	0.422	0.42	0.294	0.218	0.343	0.245	0.287	-	0.267	0.718	0.447	0.162	0.33
InS	0.100	0.033	0.121	0.218	0.094	0.171	-	0.140	**0.021**	0.210	**0.12**	0.249	0.332	0.136	0.233	0.307	0.261	-	0.151	0.330	0.259	0.25
PECE	0.707	0.452	0.602	0.467	0.691	0.501	0.424	-	0.526	0.509	0.54	0.408	0.196	0.200	0.281	0.191	0.219	0.286	-	0.442	0.234	0.27
PESS	0.384	0.467	0.391	0.429	0.359	0.359	0.356	0.390	-	0.385	0.39	0.097	0.344	0.187	0.286	0.287	0.340	0.199	0.474	-	0.279	0.28
PEPI	0.253	0.516	0.265	0.299	0.341	0.358	0.297	0.277	0.349	-	0.33	0.141	0.212	0.278	0.207	0.258	0.208	0.122	0.144	0.229	-	0.20
					avg mean						0.38					avg mean						0.25

Table 11. AUC Results and Loss in Portability in low level of usage of Moodle group.

	AUC (Numerical Datasets)									AUC (Discretized Datasets)								
Course	**ICS1**	**MM**	**MA1**	**MA2**	**KSCE**	**HWA**	**ICS3**	**ICS4**	**avg**	**ICS1**	**MM**	**MA1**	**MA2**	**KSCE**	**HWA**	**ICS3**	**ICS4**	**avg**
ICS1	0.917	0.524	0.523	0.512	0.653	0.498	0.491	0.404	0.57	0.761	0.480	0.485	0.448	0.531	0.597	0.470	0.591	0.55
MM	0.501	0.807	0.559	0.683	0.347	0.475	0.519	0.461	0.54	0.639	0.688	0.630	0.530	0.559	0.660	0.538	0.444	0.59
MA1	0.676	0.542	0.880	0.447	0.674	0.519	0.505	0.481	0.59	0.644	0.578	0.697	0.556	0.568	0.333	0.472	0.485	0.54
MA2	0.519	0.607	0.574	0.905	0.486	0.496	0.521	0.489	0.57	0.457	0.526	0.532	0.642	0.484	0.451	0.518	0.519	0.52
KSCE	0.594	0.554	0.563	0.354	0.705	0.663	0.545	0.522	0.56	**0.674**	0.560	0.574	0.422	0.931	0.608	0.570	0.445	0.60
HWA	0.490	0.434	0.489	0.428	0.590	1.000	0.522	0.512	0.56	0.628	0.516	0.547	0.492	0.532	0.778	0.522	0.516	0.57
ICS3	0.554	0.562	0.457	0.426	0.653	**0.758**	0.938	0.527	0.61	0.375	0.428	0.454	0.510	0.456	0.528	0.707	0.502	0.49
ICS4	0.414	0.563	0.539	0.550	0.472	0.521	0.495	0.771	0.54	0.410	0.443	0.390	0.475	0.452	0.500	0.460	0.682	0.48
					avg mean				0.57					avg mean				0.54
	AUC LOSS (Numerical Datasets)									AUC LOSS (Discretized Datasets)								
Course	**ICS1**	**MM**	**MA1**	**MA2**	**KSCE**	**HWA**	**ICS3**	**ICS4**	**avg**	**ICS1**	**MM**	**MA1**	**MA2**	**KSCE**	**HWA**	**ICS3**	**ICS4**	**avg**
ICS1	-	0.393	0.394	0.406	0.264	0.419	0.426	0.513	0.40	-	0.281	0.276	0.313	0.230	0.164	0.291	0.170	0.25
MM	0.307	-	0.249	0.125	0.460	0.332	0.288	0.347	0.30	0.048	-	0.057	0.158	0.129	**0.028**	0.150	0.244	**0.12**
MA1	0.204	0.337	-	0.433	0.206	0.361	0.374	0.399	0.33	0.053	0.119	-	0.142	0.129	0.364	0.225	0.212	0.18
MA2	0.386	0.298	0.331	-	0.419	0.409	0.384	0.416	0.38	0.185	0.116	0.110	-	0.158	0.191	0.124	0.123	0.14
KSCE	0.112	0.151	0.142	0.351	-	**0.042**	0.160	0.183	**0.16**	0.258	0.371	0.357	0.510	-	0.323	0.361	0.486	0.38
HWA	0.511	0.566	0.511	0.573	0.410	-	0.478	0.488	0.51	0.150	0.262	0.231	0.287	0.246	-	0.256	0.262	0.24
ICS3	0.384	0.376	0.481	0.513	0.285	0.180	-	0.411	0.38	0.333	0.280	0.253	0.197	0.251	0.179	-	0.205	0.24
ICS4	0.357	0.208	0.232	0.222	0.299	0.250	0.277	-	0.26	0.273	0.239	0.292	0.207	0.230	0.182	0.222	-	0.23
					avg mean				0.34					avg mean				0.22

5. Discussion

About the obtained accuracy of the student performance prediction models, as we can see in previous section tables for Experiments 1 and 2, it is noticeable that average AUC values are always a little better in the case of the numerical datasets than the discretized datasets. It is logical and expected that the models' predictive power is higher when we use numerical values. In Experiment 1, the average AUC highest values are obtained for the Physics group, having 0.68 for the numerical dataset and 0.60 for the discretized one. In Experiment 2 the highest values are found in the High group, obtaining values of 0.58 for the numerical dataset and 0.57 for the discretized dataset. Thus, in general the average AUC values are not high and only a little higher than a change (0.5). If we have a look at the maximum values for AUC, there is not a clear rule that we can obtain since we have found similar good values in both experiments: 0.89 in Computer group of experiment 1 with numerical datasets and 0.79 in medium level group of experiment 2 with discretized datasets. We can conclude that the accuracy of the prediction models when we transfer them to other different courses are not very high (but higher than a chance, AUC > 0.5), it is a little higher when using numerical values (but only slightly) and similar results are obtained in both experiments. We think that this can be in part due to the number of students vary a lot of from one course to another, ranging from 50 (minimum) to 302 (maximum) and the number of attributes vary from one dataset to another.

When assessing the models' portability, we have also used the AUC loss as an indicator of portability loss. According to Baker [10], prediction models are portable as long as their portability loss values stay around 0.1 (and AUC is kept above randomness). In general, in our two experiments, we have only obtained these good values in one group, namely, the Physics group with discrete datasets with 0.60 AUC average value and 0.09 AUC loss average. Thus, this group of courses fit the Baker's rule for model portability. However, if we look at specific cases, we also found that some specific models that applied to specific courses datasets obtain good results and fit the Baker's rule. For instance, in Experiment 1, the minimum values of portability loss was 0.008 for the numerical dataset (Computer group; DB transfer to SE) and 0.006 for discretized dataset (Computer group; PM transfer to RE). In Experiment 2, the minimum value of portability loss was 0.021 for numerical dataset (Medium group; InS transfer to PESS) and 0.009 for discretized dataset (Medium group; DB transfer to PEPI). These results indicate that some particular prediction models are applicable to some other different courses. However, we are more interested in finding if a model can be correctly transferred to all the rest of the courses in its group, and thus, we have a look at portability loss average values ("*avg*" loss column). In this regard, we have also found some good results, and the best four prediction models are described below. In Experiment 1, we have obtained good average results for the DB prediction model in the numerical dataset (average loss of 0.10) and the MM prediction model in the discretized dataset (0.04). Some similar results were obtained in Experiment 2 with InS prediction model in the numerical case (0.12) and ICS2 prediction model in the case of discretized dataset (0.10). It is important to highlight that those best four models not only present average portability loss values close to 0.10, but they all also keep average values of AUC above randomness. Thus, it indicates that those models are portables and they can be used to correctly predict in the rest of the courses in their group and we can conclude that they meet the conditions established in the portability challenge defined by Baker in The Baker Learning Analytics Prizes [10]. We also checked if these courses are very similar (number of students, number of types of activity, teachers in charge of the course, etc.), having only found some similarities in the group of Physics (which obtained the best average mean AUC Loss). In particular, we noticed that the instructors in charge of the three Mathematics courses in the Physics group were the same and they used the same methodology and evaluation approach in all their courses.

Next, we will show and comment those best four decision trees prediction models. The discovered knowledge from a decision tree can be extracted and presented in the form of classification IF-THEN rules. One rule is created for each path from the root to a leaf node. Each attribute-value pair along a given path forms a conjunction in the rule antecedent (IF part). The leaf node holds the class prediction, forming the rule consequent (THEN) part. In our case, we will show the J48 pruned tree that Weka

Appl. Sci. **2020**, *10*, x FOR PEER REVIEW 18 of 23

provides when training a classification prediction model. We have added the word "**THEN**" to the output of Weka in order to make easier the reading of each rule.

5.1. Best Models of Experiment 1

In Figure 3 we can see the best decision tree for Computer group with numerical datasets that is the prediction model of DB course. It is a big tree (27 nodes in total) that consists of eight leaf nodes or rules for the *Pass* class and six rules for the *Fail* class. We can see that all the attributes or Moodle events counts are about assignment, choice, forum, page, and resource. In most of the branches that lead to *Pass* leaf nodes, we can see "greater than" conditions over attributes and "less or equal than" condition in the attributes of branches that lead to *Fail* classification. Thus, we can conclude that in this prediction model to have a minimum threshold number of events in these activities seem to be much related with students' success in the course.

```
J48 pruned tree
------------------
assignment_view_all <= 0
| choice_view_all <= 1
| | forum_add_post > 0.45473
| | | forum_add_discussion <= 0
| | | | forum_subscribe <= 0.201712
| | | | | choice_choose <= 0.209779: THEN Fail
| | | | | choice_choose > 0.209779
| | | | | | assignment_view > 9.111531
| | | | | | | forum_view_forums <= 1.379725
| | | | | | | | forum_view_discussion <= 30
| | | | | | | | | page_view > 1.469242
| | | | | | | | | | choice_view <= 5.482719
| | | | | | | | | | | resource_view > 1.448263
| | | | | | | | | | | | choice_view <= 3.43581: THEN Fail
| | | | | | | | | | | | choice_view > 3.43581: THEN Pass
| | | | | | | | | | | resource_view <= 1.448263: THEN Fail
| | | | | | | | | | choice_view > 5.482719: THEN Pass
| | | | | | | | | page_view <= 1.469242: THEN Fail
| | | | | | | | forum_view_discussion > 30: THEN Pass
| | | | | | | forum_view_forums > 1.379725: THEN Pass
| | | | | | assignment_view <= 9.111531: THEN Fail
| | | | forum_subscribe > 0.201712: THEN Pass
| | | forum_add_discussion > 0: THEN Pass
| | forum_add_post <= 0.45473: THEN Fail
| choice_view_all > 1: THEN Pass
assignment_view_all > 0: THEN Pass

Number of Leaves:    14
Size of the tree:    27
```

Figure 3. Best model of the Computer group with numerical dataset—Subject DB.

Figure 4 shows the best decision tree in the Physic group with discretized dataset, that is, the prediction model of the MM course. It is a small decision tree (11 nodes in total) with five leaf nodes labeled with the *Pass* value and only one leaf node with the label *Fail*. The attributes or events that appear in the tree are about page, resource, and forum. Thus, thanks to the little number or rules and the high comprehensibility of the two labels (HIGH and LOW) the tree is very interpretable and usable

Appl. Sci. **2020**, *10*, x FOR PEER REVIEW 19 of 23

by an instructor. For example, if we analyze the branch leading to that *Fail* leaf node, we can see that students that showed a low number of events with pages, resources, and forums are quite likely to fail the course.

```
J48 pruned tree
------------------
page_view = Low
| resource_view = Low
| | forum_view_forums = Low
| | | resource_view_all = Low
| | | | resource_view = Low: THEN Fail
| | | | resource_view = High: THEN Pass
| | | resource_view_all = High: THEN Pass
| | forum_view_forums = High: THEN Pass
| resource_view = High: THEN Pass
page_view = High: THEN Pass

Number of Leaves:     6
Size of the tree:     11
```

Figure 4. Best Model of the Physics group with discretized dataset—Subject MM.

5.2. Best Models of Experiment 2

In Figure 5, we show the best decision tree in the medium level group with numerical datasets, that is, the prediction model of the Ins course. It is a medium size tree (15 nodes in total) that has three rules or leaf nodes for *Pass* class and five rules for *Fail*. The attributes or events that appear in this tree are about forum, page, and choice. Most of the branches that lead to *Pass* show that students must have a greater number of events in these attributes than a specific threshold. The rest of paths lead to students' fail.

```
J48 pruned tree
------------------
forum_view_forums <= 0.937213
| page_view <= 1.866007: THEN Fail
| page_view > 1.866007
| | choice_view_all <= 0.496039
| | | forum_view_forum <= 4: THEN Fail
| | | forum_view_forum > 4: THEN Pass
| | choice_view_all > 0.496039: THEN Pass
forum_view_forums > 0.937213
| choice_view <= 2.576183: THEN Fail
| choice_view > 2.576183
| | forum_add_discussion <= 0: THEN Fail
| | forum_add_discussion > 0
| | | choice_view_all <= 0.496039: THEN Fail
| | | choice_view_all > 0.496039: THEN Pass

Number of Leaves:     8
Size of the tree: 15
```

Figure 5. Best model of the Medium group with numerical dataset—Subject InS.

Figure 6 show the best decision tree obtained in the high level group with discretized dataset, which is the prediction model of ICS2 course. It is a small tree (only nine nodes in total) that has

Figure 6 show the best decision tree obtained in the high level group with discretized dataset, which is the prediction model of ICS2 course. It is a small tree (only nine nodes in total) that has three leaf nodes or rules for predicting when the students *Pass* and two rules for *Fail*. In this model, the attributes or Moodle events that appear in the rules are about forum, resource, and choice activities. Again, most of the branches that lead to *Pass* show that student must have a greater number of events in these attributes than a specific threshold. The rest of paths lead to students' fail.

J48 pruned tree
----
forum_view_forum = Low
| forum_view_discussion = Low
| | resource_view= Low: **THEN** Fail
| | resource_view= High: **THEN** Pass
| forum_view_discussion = High: **THEN** Pass
forum_view_forum = High
| choice_view = Low: **THEN** Fail
| choice_view = High: **THEN** Pass

Number of Leaves: 5
Size of the tree: 9

Figure 6. Best Model of the High group with discretized dataset—Subject ICS2.

6. Conclusions and Future Research

This paper presents a detailed study about the portability of predictive models between universities courses. To our knowledge, this work is one of the first exhaustive studies about portability of performance prediction models with blended university courses, and thus, we hope that it can be of help to other researchers who are also interested in developing models for portability solutions in their educational institutions.

In order to answer to our two research questions, we have carried out two experiments executing the J48 classification algorithms over 24 courses in order to obtain the AUC and AUC loss of the models when applying to different courses of the same group by using numerical and discretized datasets. Starting of the results obtained in our experiments, the answers to our two research questions are:

a. How feasible it is to directly apply predictive models between courses belonging to the same degree. By analyzing the results shown in Tables 5–8, we can see that the average AUC values are not very high (both in numerical and discretized datasets), but when we used discretized datasets, the obtained models are better in terms of AUC loss or portability loss, in spite of the fact that numerical datasets present the best AUC values, which is something that we expected in advance. In fact, portability loss values are inside the interval from 0.09 to 0.28 for the discretized datasets and we obtained good portability loss results in the Computer group and in the Physics group.

b. How feasible it is to directly apply predictive models between courses that make a similar level of usage of Moodle. By analyzing the results shown in Tables 9–11, we can see that again, the best AUC values are obtained with the numerical datasets but they are not very high. However, the best lowest portability loss values are obtained with the discretized datasets in the range from 0.22 to 0.25. In this experiment, we did not find results as good as in the first one, but nevertheless, the results obtained are inside an acceptable range.

In conclusion, the results obtained in our experiments with our 24 university courses show that it is only feasible to directly transfer predictive models or apply them to different courses with an acceptable accuracy and without losing portability under some circumstances. In our case, only when we have used discretized datasets and the transfer is between courses of the same degree, although only in two specific degrees of the four degrees tested, the loss portability is feasible. Additionally, we have shown the four best prediction models obtained in each experiment (1 and 2) and type of dataset (numerical and discretized). We have obtained that the most important attributes or Moodle events that appear in the decision trees are about forums, assignments, choices, resource, and page. However, it is important to remark that prediction models when using discretized datasets not only provide the lowest AUC loss values, that is, the best portability, but they also provide smaller decision trees than numerical ones and they only use two comprehensible values (HIGH and LOW) in their attributes (instead of continues values with threshold) that make them much easier to interpret and transfer to other courses.

A limitation of this work is the fact that the best obtained models (decision trees) might not be directly actionable by the teachers of the other courses since those models may include activities or actions that their courses do not have. We have technically solved this problem by executing J48 as Wrapper classifier that addresses incompatible training and test data by building a mapping between the training data that a classifier has been built with and the incoming test instances' structure. Model attributes that are not found in the incoming instances receive missing values. We have to do it because there are some cases when the source course and the target course do not exactly use the same attributes (they do not have the same events in their logs). We also think that this issue can be one of the reasons why we have obtained low accuracy values when applying a model to other courses that use different activities.

Finally, this work is a first step in our research. The experimental results obtained show that new strategies must be explored in order to get more conclusive results. In the future, we want to carry out new experiments by using much more additional courses and other degrees in order to check how generalizable our results can be. We are also very interested in the next potential lines or future research lines:

- To use a low number of higher-level attributes proposed by pedagogues and instructors (such as ontology-based attributes) in order to analyze whether using only few high level semantic sets that remain same in all the course datasets has a positive influence on portability results.
- To use other factors (apart from the degree and the level of Moodle usage) that can be used to group different courses and analyze how portable the models are inside those groups, for example, the number of students, the number of assessment tasks, the methodology used by the instructor, etc. Furthermore, if we have a higher number of different courses, we can do groups inside groups, for example, for each degree, to group the course by the level of Moodle usage and the same used activities.

Author Contributions: Conceptualization, C.R., J.A.L., and J.L.-Z.; methodology, C.R. and J.A.L.; software, J.L.-Z.; validation, C.R., J.A.L. and J.L.-Z.; formal analysis, C.R.; investigation, J.L.-Z.; resources, C.R., and J.L.-Z.; data curation, J.L.-Z.; writing—original draft preparation, J.A.L. and J.L.-Z.; writing—review and editing, C.R., and J.A.L.; visualization, J.L.-Z.; supervision, C.R. and J.A.L.; project administration, C.R.; funding acquisition, C.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by projects of the Spanish Ministry of Science and Technology TIN2017-83445-P.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dougiamas, M.; Taylor, P. Moodle: Using learning communities to create an open source course management system. In *EdMedia: World Conference on Educational Media and Technology*; The LearnTechLib: Waynesville, NC, USA, 2003; pp. 171–178.
2. Romero, C.; Ventura, S.; García, E. Data mining in course management systems: Moodle case study and tutorial. *Comput. Educ.* **2008**, *51*, 368–384. [CrossRef]
3. Romero, C.; Ventura, S. 2020. Educational Data mining and Learning Analytics: An updated survey. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*; John Wiley and Sons: Hoboken, NJ, USA, 2011.

4. Romero, C.; Ventura, S. Data mining in education. *Wiley Interdiscip. Rev. Data Min. Knowl. Dis.* **2013**, *3*, 12–27. [CrossRef]
5. Romero, C.; Ventura, S. Educational data mining: A review of the state of the art. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2010**, *40*, 601–618. [CrossRef]
6. Romero, C.; Espejo, P.G.; Zafra, A.; Romero, J.R.; Ventura, S. Web usage mining for predicting final marks of students that use Moodle courses. *Comput. Appl. Eng. Educ.* **2013**, *21*, 135–146. [CrossRef]
7. Luna, J.M.; Castro, C.; Romero, C. MDM tool: A data mining framework integrated into Moodle. *Comput. Appl. Eng. Educ.* **2017**, *25*, 90–102. [CrossRef]
8. Tayebinik, M.; Puteh, M. Blended Learning or E-learning? *Int. Mag. Adv. Comput. Sci. Telecommun.* **2013**, *3*, 103–110.
9. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
10. Baker, R.S. Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes. *J. Educ. Data Min.* **2019**, *11*, 1–17.
11. Boyer, S.; Veeramachaneni, K. Transfer learning for predictive models in massive open online courses. In *Artificial Intelligence in Education. AIED 2015. Lecture Notes in Computer Science*; Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M., Eds.; Springer: Cham, Switzerland, 2015; Volume 9112.
12. Boyer, S.; Veeramachaneni, K. Robust predictive models on MOOCs: Transferring knowledge across courses. In Proceedings of the 9th International Conference on Educational Data Mining, Raleigh, NC, USA, 29 June–2 July 2016; pp. 298–305.
13. Ding, M.; Wang, Y.; Hemberg, E.; O'Reilly, U.-M. Transfer learning using representation learning in massive open online courses. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK19), Empe, AZ, USA, 4–8 March 2019; ACM: New York, NY, USA, 2019; pp. 145–154.
14. Gašević, D.; Dawson, S.; Rogers, T.; Gasevic, D. Learning Analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Int. High. Educ.* **2016**, *28*, 68–84. [CrossRef]
15. Kidzinsk, L.; Sharma, K.; Boroujeni, M.S.; Dillenbourg, P. On generalizability of MOOC models. In Proceedings of the 9th International Conference on Educational Data Mining (EDM), Raleigh, NC, USA, 29 June–2 July 2016; International Educational Data Mining Society: Buffalo, NY, USA, 2016.
16. Conijn, R.; Snijders, C.; Kleingeld, A.; Matzat, U. Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS. *IEEE Trans. Learn. Technol.* **2017**, *10*, 10–29. [CrossRef]
17. Gardner, J.; Yang, Y.; Baker, R.S.; Brooks, C. Enabling end-to-end machine learning replicability: A case study in educational data mining. In Proceedings of the 1st Enabling Reproducibility in Machine Learning Workshop, Stockholmsmässan, Sweden, 25 June 2018.
18. Hunt, X.J.; Kabul, I.K.; Silva, J. Transfer learning for education data. In Proceedings of the ACM SIGKDD Conference, El Halifax, NS, Canada, 17 August 2017.
19. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey on transfer learning. *J. Big Data* **2016**, *3*, 9. [CrossRef]
20. Csurka, G.A. Comprehensive survey on domain adaptation for visual applications. A comprehensive survey on domain adaptation for visual applications. In *Domain Adaptation in Computer Vision Applications. Advances in Computer Vision and Pattern Recognition*; Csurka, G., Ed.; Springer: Cham, Switzerland, 2017.
21. Zeng, Z.; Chaturvedi, S.; Bhat, S.; Roth, D. DiAd: Domain adaptation for learning at scale. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK19), Empe, AZ, USA, 4–8 March 2019; ACM: New York, NY, USA, 2019; pp. 185–194.
22. Sun, S.; Shi, H.; Wu, Y. A survey of multi-source domain adaptation. *Inform. Fusion* **2015**, *24*, 84–92. [CrossRef]
23. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
24. López-Zambrano, J.; Martinez, J.A.; Rojas, J.; Romero, C. A tool for preprocessing moodle data sets. In Proceedings of the 11th International Conference on Educational Data Mining, Buffalo, NY, USA, 15–18 July 2018; pp. 488–489.
25. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann Publishers: San Francisco, CA, USA, 2016.
26. Quinlan, R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Mateo, CA, USA, 1993.

27. Hamoud, A.; Hashim, A.S.; Awadh, W.A. Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *Int. J. Interact. Multimed. Artif. Intell.* **2018**, *5*, 26–31. [CrossRef]
28. Wu, X.; Kumar, V.; Quinlan, J.R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Philip, S.Y.; et al. Top 10 algorithms in data mining. *Knowl. Inform. Syst.* **2008**, *14*, 1–37. [CrossRef]
29. Thai-Nghe, N.; Busche, A.; Schmidt-Thieme, L. Improving academic performance prediction by dealing with class imbalance. In Proceedings of the 9th International Conference on Intelligent Systems Design and Applications, Pisa, Italy, 30 November–2 December 2009; pp. 878–883.
30. Käser, T.; Hallinen, N.R.; Schwartz, D.L. Modeling exploration strategies to predict student performance within a learning environment and beyond. In Proceedings of the 7th International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, 13–17 March 2017; pp. 31–40.
31. Santoso, L.W. The analysis of student performance using data mining. In *Advances in Computer Communication and Computational Sciences*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 559–573.
32. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
33. Gamulin, J.; Gamulin, O.; Kermek, D. Comparing classification models in the final exam performance prediction. In Proceedings of the 37th IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 26–30 May 2014; pp. 663–668.
34. Whitehill, J.; Mohan, K.; Seaton, D.; Rosen, Y.; Tingley, D. MOOC Dropout Prediction: How to Measure Accuracy? In Proceedings of the Fourth Association for Computing Machinery on Learning @ Scale, Cambridge, MA, USA, 20–21 April 2017.

Journal of Computing in Higher Education
https://doi.org/10.1007/s12528-021-09273-3

Improving the portability of predicting students' performance models by using ontologies

Javier López-Zambrano[1,3] · Juan A. Lara[2] · Cristóbal Romero[3]

Accepted: 9 March 2021

Abstract
One of the main current challenges in Educational Data Mining and Learning Analytics is the portability or transferability of predictive models obtained for a particular course so that they can be applied to other different courses. To handle this challenge, one of the foremost problems is the models' excessive dependence on the low-level attributes used to train them, which reduces the models' portability. To solve this issue, the use of high-level attributes with more semantic meaning, such as ontologies, may be very useful. Along this line, we propose the utilization of an ontology that uses a taxonomy of actions that summarises students' interactions with the Moodle learning management system. We compare the results of this proposed approach against our previous results when we used low-level raw attributes obtained directly from Moodle logs. The results indicate that the use of the proposed ontology improves the portability of the models in terms of predictive accuracy. The main contribution of this paper is to show that the ontological models obtained in one source course can be applied to other different target courses with similar usage levels without losing prediction accuracy.

Keywords Educational data mining · Predictive modelling · Student performance · Transfer learning · Model portability · Ontology

✉ Cristóbal Romero
cromero@uco.es

Javier López-Zambrano
jlopez@outlook.com; jlopez@espam.edu.ec

Juan A. Lara
juanalfonso.lara@udima.es

1 Escuela Superior Politécnica Agropecuaria de Manabí (ESPAM MFL), Faculty of Computing, SISCOM Group, 131106 Calceta, Ecuador

2 Department of Computer Science, Madrid Open University, UDIMA, Carretera de La Coruña, KM. 38500, Vía de Servicio, No. 15, 28400 Collado Villalba, Madrid, Spain

3 Department of Computer Science and Numerical Analysis, University of Córdoba (UCO), 14071 Córdoba, Spain

Published online: 24 March 2021

Introduction

In recent decades, one of the main educational milestones is the advent of a new form of learning called e-learning (electronic learning), based on the use of the internet and technology to support students' online education. Nowadays, this form of learning is becoming particularly important due to the limitations defined by the authorities to restrain the spread of pandemics such as the one caused by Covid-19. The use of e-learning poses important advantages including the enabling of a more flexible temporal and spatial interaction than other forms of learning. Besides, vast amounts of learning process data can be collected, since it is based on the use of Learning Management Systems (LMS). Moodle (Dougiamas & Taylor, 2008) is one of the most used LMS overall, because, among other advantages, it is free, open and there is an important community of users who support its development. Data recorded by Moodle, in particular those that reflect students' interactions with educational resources, can be of great interest and applicability for building student behavior models. To analyze these data, approaches such as Educational Data Mining (EDM) and Learning Analytics (LA) are useful (Romero et al., 2008). In EDM, a field whose purpose is the extraction of knowledge from educational data, there are well-defined problems that have been addressed by the scientific community, such as the prediction of students' performance (Romero & Ventura, 2013, 2020). Recently, it is more frequent to find works that propose new approaches to analyzing educational data for a particular course. However, one of the due challenges is creating models for a particular course that can be useful when used in other courses (Baker, 2019). These are what we call transferable or portable models (Boyer & Veeramachaneni, 2015).

In our previous work (López-Zambrano et al., 2020), we obtained models generated from Moodle's logs data and we studied the degree of portability of the models between subjects, grouped by area of knowledge and by the usage level of platform resources. We used Moodle's native raw attributes which, in certain combinations of courses, led us to a certain loss in the portability of models since these low-level attributes are very dependent on each particular course. To overcome this limitation from our previous research, in this paper we present a new approach based on the use of resources from the semantic web area, in particular, ontologies (Fong et al., 2011; Tang & Fong, 2010). One of the most promising lines in this respect, particularly when analyzing logs of students' interactions with the LMS, is the categorization or taxonomy of attributes. In this regard, Bloom's taxonomy plays an important role. Bloom's taxonomy is a multi-tiered model of classifying thinking according to six cognitive levels of complexity which in this new version are: Remembering, Understanding, Applying, Analysing, Evaluating and Creating (Forehand, 2005). Based on this idea, some works have even defined correspondence between the levels of Bloom's taxonomy and the different actions conducted by students in Moodle (Rollins, 2010). Some authors (Cerezo et al., 2020) proposed a categorization of low-level attributes into different higher-level codifications, such as Executing, Planning, Learning, and Reviewing. Precisely, our research aims to evaluate the degree of portability of models built by using ontologies of interaction-with-the-platform attributes. To do so, we defined an ontology inspired by Bloom's taxonomy and based on the work by Cerezo et al. (2020), with the purpose of conducting a comprehensive study to measure the degree of portability of the models built based on that ontology (denoted as ontological models), compared with a previous similar study conducted by the authors López-Zambrano et al. (2020) in which we did not use ontologies but instead employed low-level Moodle attributes (denoted as non-ontological models). The models have been built from students' interactions with Moodle logs and the class attribute to predict is binary and represents whether or not the student will pass the course (Pass/Fail). In this work, the courses have been grouped according to the usage level of Moodle activities/resources. This approach has already been used in previous studies with satisfactory results (López-Zambrano et al., 2020). Taking all this into consideration, the global objective of this paper is to provide an answer for the research question below:

- Can the ontological models obtained in one (source) course be applied in other different (target) courses with a similar usage level without losing prediction accuracy?

The rest of the paper is organized as follows: Sect. "Background" reviews the literature related to this research. Section "Materials and methods" describes the data and the experiments. Section "Results" includes and discusses the results obtained. Finally, Sect. "Conclusions" presents the conclusions and future lines of research.

Background

Achieving generalizable and portable models is still an important challenge in the area of EDM, in spite of the important advances made in the last few years (Boyer & Veeramachaneni, 2015; Ding et al., 2019; Gašević et al., 2016; Hunt et al., 2017). In fact, Baker (2019) has considered what he calls the "Generalizability" or "New York City and Marfa" problem as one of the main challenges for the future of EDM, which is explained in detail in López-Zambrano et al. (2020).

To address this challenge, the resource of resources from the semantic web seems to be a promising line. The semantic web is an extension of the current web in which information is provided with a certain meaning, which makes cooperation and portability easier (Dhuria & Chawla, 2014). Fundamental resources from the semantic web are the ontologies, because they provide a common understanding of a domain. In particular, they may be interesting resources in the e-learning field (Al-Yahya et al., 2015).

In this regard, several particular works should be highlighted. In Octaviani et al. (2015) they present a tool, called RDB2Onto, for creating ontologies from Moodle logs, but this work does not validate the utility of such an ontology. In Castro and Alonso (2011) they propose a general architecture for EDM in which there is an educational ontology, but they do not define or develop the ontology, only providing a general statement of it as a part of a higher-level architecture. There are even some works such as the one presented in Chang et al. (2020) where they utilize data mining techniques (association in this case) to build ontology-driven tutoring models for intelligent tutoring systems (this is precisely the opposite process to ours since we use the ontology for a further data mining analysis).

These previous works present general approaches. Other more specific works bear greater similarity to our study because they define particular ontologies to facilitate the EDM process. We found some works where the ontology created is not focused on attributes of students' interaction with the LMS. In Marinho et al. (2010) they propose an ontology to model EDM tasks, techniques, and parameters. In Grivokostopoulou et al. (2014) they propose an educational system that utilizes ontologies and semantic rules to enhance the quality of educational content (curriculum) and the learning activities delivered to each student. In Nouira et al. (2019), they propose an ontological model for assessment analytics. And finally; in Dorça et al. (2017), they present an approach for the automatic and dynamic analysis of learning object repositories in which ontology models the relationships between the attributes and learning styles of the learning objects.

Other related works are those that define ontologies to model data of students' interactions with LMS resources. In El-Rady (2020), they propose an ontology where the student is the main class from which a series of associations arise that are connected to other classes that model the students' data (education, profile, social activities, etc.). That ontology is used as a part of a validation process to predict student dropout rates. Other related works are based on the idea of organizing the interaction attributes as part of a kind of taxonomy. It is worth highlighting the work presented in Cerezo et al. (2020), where they propose a process mining method for a self-regulated learning assessment, and make use of an ontology inspired by Bloom's taxonomy. In Montenegro-Marin et al. (2011), they also propose an ontology based on the idea of taxonomy, but not restricted to interaction attributes, as they consider many other features, such as the curriculum design, productivity, management, and so on. However, they do not validate the utility of the ontology.

Considering all the previous works, and to the best of the authors' knowledge, our work presented in this paper is the first that analyses the power of ontologies as a resource that makes the portability of EDM models easier and, in particular, it is also the only one for that purpose which is based on data from the students' interactions with the LMS. Furthermore, it is the first research that depicts a comparative study against a previous non-ontological similar approach, with the purpose of demonstrating the performance improvement obtained when using ontologies. Both of these innovative aspects are the core contributions of this paper.

Materials and methods

In this section, we describe both the data used and the preprocessing tasks we applied to them in order to transform the raw data gathered from the Moodle logs to the high-level attributes of the proposed ontology. We also describe the experimentation that we carried out in order to address our research question.

Table 1 Information of all subjects

Subject	Code	Degree	Year	#Users	Moodle usage
Introduction to programming (group 1)	IP1	Computer	1	144	Medium
Introduction to programming (group 2)	IP2	Computer	1	145	High
Programming methodology (group 1)	PM1	Computer	1	114	Medium
Programming methodology (group 2)	PM2	Computer	1	119	High
Professional computer tools	PCT	Computer	1	124	Medium
Databases	DB	Computer	2	58	Medium
Human computer interfaces	HCI	Computer	2	260	High
Information systems	InS	Computer	2	188	Medium
Software engineering	SE	Computer	2	58	Medium
Interactive systems	IS	Computer	3	84	High
Requirement engineering	RE	Computer	3	36	Medium
Software design and construction	SDC	Computer	3	50	Medium
Introduction to computer science	ICS1	Electrical engineering	1	100	Low
Introduction to computer science	ICS2	Electronic engineering	1	198	High
Introduction to computer science	ICS3	Civil engineering	1	85	Low
Introduction to computer science	ICS4	Mining engineering	1	77	Low

Table 2 List of groups by Moodle usage

No	Group	No. of subjects
1	High	5
2	Medium	8
3	Low	3

Data and preprocessing

We have used the log data of 1840 Cordoba University students from 16 different courses taught by the Computer Science Department. Table 1 summarises these courses. For each course, it shows the subject or name of the course (Subject), our own identification Code, name of the Degree, Year in the degree/curriculum, number of students (#Users), and the level of Moodle Usage (Low, Medium or High). To accomplish the ethical and privacy issues about using these data, we have used informed consent with all the instructors and we have also anonymized all information about students (Pardo & Siemens, 2014).

We divided or grouped our 16 different courses (see Table 1) into three usage levels of Moodle activities in courses (see Table 2). Moodle provides us resources (text and web page, link to files or websites, and label) and different types of activities (assignments, chat, choice, database, forum, glossary, lesson, quiz, survey, wiki, workshop, etc.). We have defined three levels of usage by the number of activities used in the course:

- *Low level* The course only has one or no activity.
- *Medium level* The course has two different types of activities.
- *High level* The course has three or more different types of activities.

Moodle provides a wide range of activities such as Assignments, Databases, Chats, Choice, Questionnaires, Quiz, Surveys, Forums, Glossaries, Lessons, SCORM packages, Workshops, Wikis, etc.). The most frequent activities in our courses are Assignments, Forums, and Quizzes. So, normally low-level courses only use one of these three activities, medium level two of them, and high level three or more activities. Table 2 shows the number of courses in each group grouped by usage level.

We also propose our ontology for defining 5 high-level attributes starting from 58 low-level attributes or actions provided by Moodle logs (see Table 3).

As depicted in Table 3, our ontology generalizes the 58 raw/low-level events provided by the Moodle logs into only five attributes or high-level categories. The first category references all the actions about consulting resources (LEARNING/READING/VIEWING), the second groups the students' communication events (COMMUNICATING), the third deals with the students' work (WORKING/DOING), the fourth is about students' evaluation (EXAMINING/EVALUATING) and the last is about the students' general ENGAGEMENT in the course. The first four attributes of our ontology are a number (from 0 to 100) that is the percentage of events of each type that each student has done in Moodle. The last attribute is the most general and is also a number (between 0 and 100) obtained from the total number of interactions/events and the number of days connected.

Finally, we have created two different datasets or data files: one with the original previously-described numerical data, and the other discretizing the attributes in two labels (HIGH and LOW) by using the equal width discretization method.

In both cases, we added a new attribute or class to predict at the end of our 5 attributes. This class is the final mark obtained by the students in the course, which is the value to predict in a classification task. The final mark (value between 0 and 10) has been discretized into two values or labels: FAIL if the student's final mark is lower than 5 or PASS if the students' final mark is higher than 5.

Methodology for experimentation

The methodology used in our experimentation consisted of these steps (see Fig. 1):

- Firstly, we downloaded and preprocessed the Moodle log in order to obtain both the numerical and discretized datasets for each course. We used a specific Java tool that we developed for doing this specific transformation task (López-Zambrano et al., 2020).
- Secondly, we executed the well-known J48 classification algorithm provided by the WEKA data mining environment for each one of the previous numerical and

Table 3 Ontology and Moodle low-level actions of each category

Learning/reading/viewing	Communicating	Working/doing	Evaluating/examining	Engagement
Blog view	Forum add discussion	Assignment upload	Hotpot submit	Number of total interactions
Book view all	Forum add post	Assignment view	Hotpot view	
Course enrol	Forum search	Assignment view all	Hotpot view all	Number of days connected
Course recent	Forum subscribe	Assignment view submission	Questionnaire submit	
Course user report	Forum subscribe all	Choice choose	Questionnaire update	
Course view	Forum update	Choice choose again	Questionnaire view	
Folder view	Forum update post	Choice view	Questionnaire view all	
Folder view all	Forum user report	Choice view all	Quiz attempt	
Imscp view all	Forum view discussion	Teamwork update	Quiz close attempt	
Page view	Forum view forum	Teamwork view	Quiz continue attempt	
Page view all	Forum view forums	Teamwork view all	Quiz continue attempt	
Resource view	Wiki edit		Quiz preview	
Resource view all	Wiki update		Quiz review	
Url view	Wiki view		Quiz view	
Url view all	Wiki view all		Quiz view all	

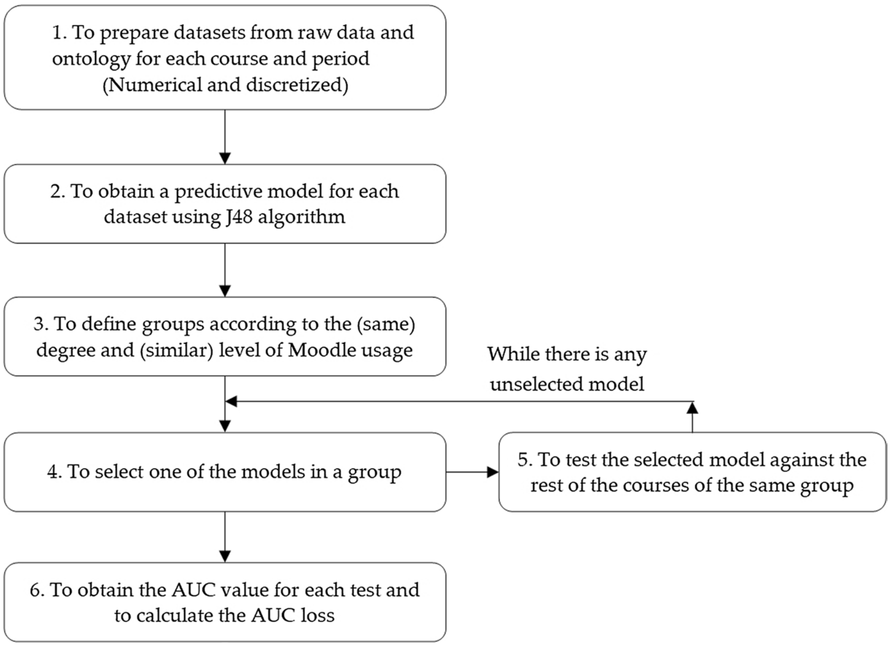

Fig. 1 Methodology used in our experimentation

categorical datasets of 16 subjects or different courses. In this step, we obtained one prediction model for each course.

- Then, we grouped our 16 subjects/courses into 3 groups depending on their level of usage of Moodle activities (see Table 1).
- Next, we repeated the next two actions. We selected each prediction model obtained in one course one by one and we applied it to testing the datasets of all the other courses in the same group. We repeated this process with all the models and with all the datasets for each group.
- Finally, we obtained the values of the two evaluation metrics that we used (the area under the ROC Curve and AUC loss) when applying the prediction model for one course/subject over the other datasets in the same group. And we compared the results obtained when using the original raw low-level

Results

The results of these three groups are set out below (summarised in Table 3). Two experiments were conducted for each group, applying the J48 algorithm with balanced numerical and discretized datasets. These experiments consisted of having a first set of experiments for which high-level datasets were constructed (ontology) and a second experiment with datasets built with low-level attributes.

For each experiment (within the same group), we conducted an analysis of the best AUC obtained and the lowest error rate, or loss of portability, of the model.

Thus, the results consist of two tables. At the top, a matrix is shown with the results of the AUC metric, obtained from the list of the general model for each subject (rows), compared to the average AUC for the individual datasets from each period for a subject (columns). The values of the main diagonal represent the testing of the general model for subjects over their own datasets, where this value is the reference AUC value (highest value), with regard to the AUCs from the other subjects. The second matrix (bottom) displays the difference between the highest AUC (reference) by row, with regard to each individual AUC. These values tell us how much precision is lost in the AUC when this model is tested with other subjects (portability), aiming to highlight the lowest values, as they indicate the lowest error rate or loss in the process of model portability or transferability.

Group of courses with high-level usage

For the high-level group, we can see in Table 4 that of the two tests, the best general results (averages) are in the datasets with ontology, revealing that the AUC average for numerical datasets is 0.62 and the average for discretized datasets is 0.61, higher than their equivalents in the tests without ontology. While there is only a small difference, the loss rate or difference in transferability does denote a greater difference, and within the same test group, the difference between numerical and discretized datasets is highly significant, where the tests with discretized data are much better.

If we focus on the tests with the best results, we can see that the best value for the AUC metric (0.675) is in the ICS2 subject obtained with discretized data and tested with the subject HCI. This is not concordant with the general average of the AUCs, whose highest value is for the numerical sets (0.62), with a tiny difference of one one-hundredth. However, it is concordant with the fact that the best rate of loss or difference is with the discretized data (0.10). We can also see that the generalized model obtained with the aforesaid subject (ICS2) has very good results, which is proven in the general averages (row) in both tests (with and without ontology).

With regard to the model that obtained the best average in the precision loss rate, we can see in Fig. 2, the decision tree, defining the attribute COMMUNICATING (from the five ontology attributes—Table 1) as the attribute with the highest increase in information, which would define the prediction for a student passing the course.

Group of courses with medium-level usage

For the medium-level group, we can see in Table 5 that of the two tests, the best general results (averages) are in the dataset tests with ontology, revealing that the AUC average for numerical datasets is 0.60 and the average for discretized datasets is 0.59, higher than their equivalents in the tests without ontology. There is a small difference, although the loss rate or difference in transferability does denote a greater difference, and within the same test group, the difference between numerical and discretized datasets is highly significant, where the tests with discretized data are much better.

Table 4 AUC results and loss of transferability (difference) with J48—high-level group

Course	With ontology											
	HCI	IS	ICS2	IP2	PM2	Avg	HCI	IS	ICS2	IP2	PM2	Avg
	AUC (numerical datasets)						*AUC (discretized datasets)*					
HCI	0.890	0.511	0.592	0.535	0.528	0.61	0.710	0.672	0.608	0.614	0.624	0.65
IS	0.488	0.886	0.498	0.555	0.629	0.61	0.509	0.672	0.512	0.576	0.629	0.58
ICS2	0.602	0.600	0.799	0.639	0.661	0.66	0.675	0.633	0.717	0.632	0.662	0.66
IP2	0.483	0.484	0.589	0.849	0.550	0.59	0.536	0.651	0.512	0.704	0.630	0.61
PM2	0.501	0.591	0.483	0.544	0.909	0.61	0.501	0.560	0.562	0.562	0.666	0.57
					Avg mean	0.62					Avg mean	0.61
	AUC loss (numerical datasets)						*AUC loss (discretized datasets)*					
HCI	–	0.379	0.298	0.355	0.362	0.35	–	0.038	0.102	0.095	0.085	0.08
IS	0.398	–	0.388	0.331	0.257	0.34	0.163	–	0.159	0.096	0.043	0.12
ICS2	0.197	0.199	–	0.160	0.138	0.17	0.042	0.084	–	0.085	0.055	0.07
IP2	0.366	0.364	0.260	–	0.298	0.32	0.169	0.053	0.192	–	0.074	0.12
PM2	0.408	0.318	0.426	0.364	–	0.38	0.166	0.107	0.104	0.104	–	0.12
					Avg mean	0.31					Avg mean	0.10
Course	Without ontology											
	HCI	IS	ICS2	IP2	PM2	Avg	HCI	IS	ICS2	IP2	PM2	Avg
	AUC (numerical datasets)						*AUC (discretized datasets)*					
HCI	0.943	0.510	0.522	0.538	0.524	0.61	0.769	0.621	0.569	0.417	0.570	0.59
IS	0.485	0.927	0.494	0.470	0.606	0.60	0.479	0.816	0.577	0.555	0.656	0.62
ICS2	0.514	0.590	0.783	0.500	0.569	0.59	0.503	0.558	0.619	0.485	0.516	0.54
IP2	0.484	0.420	0.472	0.862	0.490	0.55	0.519	0.576	0.535	0.761	0.491	0.58
PM2	0.514	0.489	0.530	0.618	0.899	0.61	0.574	0.488	0.522	0.592	0.793	0.59
					Avg mean	0.59					Avg mean	0.58

Table 4 (continued)

Course	Without ontology											
	HCI	IS	ICS2	IP2	PM2	Avg	HCI	IS	ICS2	IP2	PM2	Avg
	AUC loss (numerical datasets)						*AUC loss (discretized datasets)*					
HCI	–	0.432	0.421	0.404	0.418	0.42	–	0.148	0.201	0.352	0.200	0.23
IS	0.442	–	0.433	0.457	0.321	0.41	0.337	–	0.238	0.260	0.160	0.25
ICS2	0.270	0.193	–	0.283	0.215	0.24	0.116	0.061	–	0.134	0.103	0.10
IP2	0.378	0.441	0.390	–	0.371	0.40	0.242	0.184	0.225	–	0.269	0.23
PM2	0.385	0.410	0.369	0.281	–	0.36	0.219	0.305	0.271	0.200	–	0.25
					Avg mean	0.37					Avg mean	0.21

```
J48 pruned tree
------------------

COMMUNICATING = LOW: Fail
COMMUNICATING = HIGH: Pass

Number of Leaves:     2
Size of the tree:     3
```

Fig. 2 The best model for the high-level group with discretized dataset—subject ICS2

If we focus on the tests with the best results, we see that the best value for the AUC metric (0.718) is for the subject SDC, obtained with numerical data and tested with the subject RE, which is concordant with the general AUC average, whose highest value is in the numerical tests (0.60), with is a small difference of one one-hundredth. However, it is not concordant with the fact that the best loss or difference rate is for discretized data (0.18). We also see that in the generalized model within the tests with discretized data, the subject PM1 has a good result in the general average for the loss rate (row) in the tests without ontology, although in the tests with ontology (employing a generalized model of high-level attributes), there are also good results in the IP1 and INS subjects, which share the same value of 0.13, six one-hundredths more, but still within the ideal value for good transferability of the model.

With regard to the subjects with the best average loss rate, Fig. 3 shows that the decision tree defines the attributes LEARNING/READING/VIEWING and COMMUNICATING (from the five ontology results—Table 1) as the attributes with the greatest gain in information, defining that if there is a high level of LEARNING/READING/VIEWING, the student will pass or, conversely, if it is low, but with a high level of interaction in COMMUNICATING, the student will also pass.

Concerning the decision tree shown in Fig. 4, it defines the attributes LEARNING/READING/VIEWING, COMMUNICATING, WORKING/DOING and EVALUATING/EXAMINING (from the five ontology attributes—Table 1) as the attributes with the greatest increase in information, once again defining that if there is a high level of LEARNING/READING/VIEWING, the student will pass or, conversely, if it is low, but with a high level of interaction in COMMUNICATING, the student will also pass. If the COMMUNICATING level is low, but the level of WORKING/DOING is high, then the student would pass, but if it is not high, then the student will only pass if the EVALUATING/EXAMINING level is high.

Group of courses with low-level usage

For the low-level group, we can see in Table 6 that of the two tests, the best general results (averages) are in the dataset tests with ontology, revealing that the AUC averages for numerical datasets is 0.63 and the average for discretized datasets is 0.61, higher than their equivalents in the tests without ontology. There is a small difference, although the loss rate or difference in transferability does denote a greater difference, and within the same test group, the difference between numerical and discretized datasets is highly significant, where the tests with discretized data are much better.

Table 5 AUC results and loss of transferability (difference) with J48—medium-level group

Course	With ontology								
	IP1	PM1	DB	SDC	PCT	RE	SE	InS	Avg
AUC (numerical datasets)									
IP1	0.835	0.567	0.589	0.552	0.508	0.589	0.620	0.582	0.61
PM1	0.519	0.821	0.540	0.520	0.530	0.510	0.550	0.567	0.57
DB	0.670	0.623	0.980	0.590	0.571	0.566	0.521	0.640	0.65
SDC	0.502	0.596	0.516	0.788	0.469	0.718	0.549	0.504	0.58
PCT	0.633	0.621	0.611	0.610	0.911	0.641	0.572	0.670	0.66
RE	0.494	0.519	0.497	0.643	0.476	0.869	0.527	0.512	0.57
SE	0.540	0.510	0.520	0.560	0.530	0.511	0.962	0.523	0.58
InS	0.608	0.580	0.591	0.563	0.508	0.560	0.562	0.815	0.60
								Avg mean	0.60
AUC loss (numerical datasets)									
IP1	–	0.267	0.246	0.283	0.327	0.246	0.215	0.253	0.26
PM1	0.302	–	0.281	0.301	0.291	0.311	0.271	0.254	0.29
DB	0.310	0.357	–	0.390	0.409	0.414	0.459	0.340	0.38
SDC	0.286	0.192	0.272	–	0.319	0.070	0.239	0.284	0.24
PCT	0.278	0.290	0.300	0.301	–	0.270	0.339	0.241	0.29
RE	0.375	0.350	0.372	0.226	0.393	–	0.342	0.357	0.34
SE	0.422	0.452	0.442	0.402	0.432	0.451	–	0.439	0.43
InS	0.207	0.235	0.224	0.252	0.307	0.255	0.253	–	0.25
								Avg mean	0.31
AUC (discretized datasets)									
IP1	0.772	0.637	0.621	0.601	0.688	0.643	0.643	0.652	0.66
PM1	0.634	0.763	0.532	0.604	0.562	0.510	0.521	0.602	0.59
DB	0.612	0.583	0.775	0.555	0.616	0.567	0.543	0.551	0.60
SDC	0.474	0.562	0.628	0.696	0.505	0.590	0.480	0.551	0.56
PCT	0.592	0.564	0.577	0.582	0.812	0.567	0.581	0.582	0.61
RE	0.589	0.591	0.520	0.583	0.572	0.801	0.563	0.571	0.60
SE	0.527	0.562	0.550	0.588	0.504	0.614	0.694	0.548	0.57
InS	0.648	0.635	0.549	0.640	0.471	0.369	0.529	0.677	0.56
								Avg mean	0.59
AUC loss (discretized datasets)									
IP1	–	0.135	0.151	0.172	0.084	0.129	0.129	0.120	0.13
PM1	0.129	–	0.231	0.159	0.201	0.253	0.242	0.162	0.20
DB	0.163	0.192	–	0.220	0.159	0.208	0.232	0.224	0.20
SDC	0.222	0.134	0.068	–	0.191	0.107	0.216	0.145	0.15
PCT	0.220	0.248	0.235	0.230	–	0.245	0.231	0.230	0.23
RE	0.212	0.210	0.281	0.218	0.229	–	0.238	0.230	0.23
SE	0.167	0.132	0.144	0.107	0.190	0.080	–	0.146	0.14
InS	0.029	0.042	0.128	0.038	0.206	0.309	0.148	–	0.13
								Avg mean	0.18

Table 5 (continued)

Course	Without ontology								
	IP1	PM1	DB	SDC	PCT	RE	SE	InS	Avg
AUC (numerical datasets)									
IP1	0.938	0.588	0.542	0.545	0.610	0.493	0.579	0.523	0.60
PM1	0.496	0.689	0.589	0.478	0.567	0.624	0.484	0.486	0.55
DB	0.495	0.491	0.976	0.535	0.457	0.670	0.581	0.517	0.59
SDC	0.492	0.518	0.467	0.809	0.504	0.558	0.496	0.456	0.54
PCT	0.459	0.496	0.337	0.585	0.891	0.612	0.382	0.492	0.53
RE	0.439	0.524	0.329	0.553	0.579	0.956	0.473	0.577	0.55
SE	0.526	0.581	0.611	0.559	0.486	0.614	0.964	0.494	0.60
InS	0.484	0.495	0.671	0.583	0.486	0.610	0.533	0.704	0.57
								Avg mean	0.57
AUC loss (numerical datasets)									
IP1	–	0.350	0.396	0.393	0.328	0.446	0.359	0.415	0.38
PM1	0.193	–	0.100	0.211	0.122	0.065	0.205	0.203	0.16
DB	0.481	0.485	–	0.441	0.519	0.307	0.395	0.459	0.44
SDC	0.317	0.291	0.342	–	0.305	0.252	0.313	0.353	0.31
PCT	0.432	0.395	0.554	0.306	–	0.279	0.509	0.399	0.41
RE	0.517	0.432	0.627	0.403	0.377	–	0.483	0.379	0.46
SE	0.438	0.383	0.353	0.405	0.478	0.351	–	0.470	0.41
InS	0.221	0.209	0.033	0.121	0.218	0.094	0.171	–	0.15
								Avg mean	0.34
AUC (discretized datasets)									
IP1	0.811	0.441	0.496	0.535	0.500	0.500	0.414	0.510	0.53
PM1	0.476	0.585	0.458	0.550	0.515	0.564	0.512	0.559	0.53
DB	0.551	0.500	0.652	0.551	0.476	0.500	0.510	0.499	0.53
SDC	0.532	0.593	0.430	0.924	0.531	0.610	0.484	0.622	0.59
PCT	0.494	0.500	0.447	0.567	0.712	0.553	0.470	0.551	0.54
RE	0.568	0.543	0.529	0.614	0.508	0.756	0.545	0.569	0.58
SE	0.487	0.500	0.500	0.375	0.473	0.431	0.718	0.451	0.49
InS	0.526	0.500	0.429	0.625	0.528	0.454	0.500	0.761	0.54
								Avg mean	0.54
AUC loss (discretized datasets)									
IP1	–	0.370	0.315	0.277	0.311	0.311	0.397	0.301	0.33
PM1	0.108	–	0.127	0.035	0.070	0.021	0.073	0.025	0.07
DB	0.101	0.152	–	0.101	0.176	0.152	0.142	0.153	0.14
SDC	0.392	0.331	0.494	–	0.393	0.314	0.440	0.302	0.38
PCT	0.218	0.212	0.265	0.145	–	0.159	0.242	0.161	0.20
RE	0.188	0.213	0.227	0.142	0.248	–	0.211	0.187	0.20
SE	0.231	0.218	0.218	0.343	0.245	0.287	–	0.267	0.26
InS	0.235	0.261	0.332	0.136	0.233	0.307	0.261	–	0.25
								Avg mean	0.23

Fig. 3 Best model for the medium-level group with discretized dataset—subject IP1

```
J48 pruned tree
------------------

LEARNING/READING/VIEWING = LOW
|   COMMUNICATING = LOW: Fail
|   COMMUNICATING = HIGH: Pass
LEARNING/READING/VIEWING = HIGH: Pass


Number of Leaves:        3
Size of the tree:        5
```

Fig. 4 Best model for the medium-level group with discretized dataset—subject InS

```
J48 pruned tree
------------------

LEARNING/READING/VIEWING = LOW
|   COMMUNICATING = LOW
|   |   WORKING/DOING = LOW
|   |   |   EVALUATING/EXAMINING = LOW: Fail
|   |   |   EVALUATING/EXAMINING = HIGH: Pass
|   |   WORKING/DOING = HIGH: Pass
|   COMMUNICATING = HIGH: Pass
LEARNING/READING/VIEWING = HIGH: Pass


Number of Leaves:        5
Size of the tree:        9
```

If we now focus on the tests with the best results, we can see that the best value for the AUC metric (0.683) is in the ICS1 subject obtained with discretized data and tested with the subject ICS4. This is not concordant with the general average of the AUCs whose highest value is for the numerical sets (0.63), with a small difference of two one-hundredths. However, it is concordant with the fact that the best rate of loss or difference is with the discretized data (0.13). We can also observe that the generalized model obtained with the aforesaid subject (ICS1) has very good results, which is proven in the general averages (row) for the matrix of discretized data with ontology. The value is 0.07, which is below the ideal for determining a good transfer of the model, in this case for a general model with high-level attributes.

With regard to the subject with the best average loss rate, we can see in Fig. 5, the decision tree, defining the attribute COMMUNICATING (from the five attributes of ontology—Table 1) as the attribute with the highest increase in information, which would define the predictability for a student passing the course.

Conclusions

This paper aims to improve the portability or transferability of predictive models of students' performance by using an ontology that uses a taxonomy of actions on students' interactions with the Moodle learning management system. We compare the results of this new proposed approach against our previous results when we used low-level raw attributes directly obtained from Moodle logs. The results obtained

Table 6 AUC results and loss of transferability (difference) with J48—low-level group

Course	With ontology							
	ICS1	ICS3	ICS4	Avg	ICS1	ICS3	ICS4	Avg
	AUC (numerical datasets)				*AUC (discretized datasets)*			
ICS1	0.860	0.592	0.500	0.65	0.722	0.615	0.683	0.67
ICS3	0.506	0.820	0.560	0.63	0.512	0.750	0.565	0.61
ICS4	0.510	0.531	0.832	0.62	0.500	0.500	0.600	0.53
			Avg mean	0.63			Avg mean	0.61
	AUC loss (numerical datasets)				*AUC loss (discretized datasets)*			
ICS1	–	0.268	0.360	0.31	–	0.107	0.039	0.0
ICS3	0.314	–	0.260	0.29	0.239	–	0.186	0.21
ICS4	0.322	0.301	–	0.31	0.100	0.100	–	0.10
			Avg mean	0.30			Avg mean	0.13
Course	Without ontology							
	ICS1	ICS3	ICS4	Avg	ICS1	ICS3	ICS4	Avg
	AUC (numerical datasets)				*AUC (discretized datasets)*			
ICS1	0.917	0.491	0.404	0.60	0.761	0.470	0.591	0.61
ICS3	0.554	0.938	0.527	0.67	0.375	0.707	0.502	0.53
ICS4	0.414	0.495	0.771	0.56	0.410	0.460	0.682	0.52
			Avg mean	0.61			Avg mean	0.55
	AUC loss (numerical datasets)				*AUC loss (discretized datasets)*			
ICS1	–	0.426	0.513	0.47	–	0.291	0.170	0.23
ICS3	0.384	–	0.411	0.40	0.333	–	0.205	0.27
ICS4	0.357	0.277	–	0.32	0.273	0.222	–	0.25
			Avg mean	0.39			Avg mean	0.25

Fig. 5 Best model for the low-level group with discretized dataset—subject ICS1

```
J48 pruned tree
------------------

COMMUNICATING = LOW: Fail
COMMUNICATING = HIGH: Pass

Number of Leaves:          2
Size of the tree:          3
```

show that the use of the proposed ontology significantly improves the portability of the models in terms of their predictive accuracy. So, the answer to our initial research question is yes, the ontological models obtained in one source course can be applied to other different target courses with similar usage levels without losing prediction accuracy.

One of the limitations of this work is the specific attributes/variables used in our proposed ontology.

For example, it is also important to discuss if the “number of total interactions” are truly showing engagement when learning using LMS. The number of actions includes the behavior of supposed relevant activity in the LMS and were are assuming that all of these actions could indicate that the student is properly involved in his learning process. As traditionally happens with study time, however, this variable by itself is very tricky. It may seem that the more time those students spend studying, the better grades they should receive, but it is not that simple; it mainly depends on the quality of the study time, and something similar could be occurring with the relevant actions; more activity in the LMS does not assure better results (Cerezo et al., 2016).

Regarding the application of the results obtained in this work and the potential for using them within other domains; it is important to notice that currently there is an increasing interest in the generalization and portability of prediction models and specifically with Moodle LMS (Monllao-Olive et al., 2019). In this line, our proposal can be applied not only to Learning Management Systems as Moodle but also to other different domains or data sources such as Intelligent Tutoring Systems (ITSs), Massive Online Open Courses (MOOCs), Traditional face-to-face educational environments, Blended Learning and Multimodal Learning environments, and so on.

Finally, as a future study, we are currently working on:

- Using a higher number of courses with much more data/students from different areas/domains, not only engineering and computer science, but also fields such as science, biology, medicine, philosophy, and literature, in order to generalize the good results that we obtained in this study.
- Discovering predictive models that can be portable/transferable as soon as possible in the early stages of the course. This means we would not have to wait until the end of the course to have all Moodle usage data available, and the obtained models could be used as general early warning prediction models for different similar courses (Romero & Ventura, 2019).

Acknowledgements This work would not be possible without the funding from the Ministry of Sciences and Innovation I+D+I TIN2017-83445-P.

References

Al-Yahya, M., George, R., & Alfaries, A. (2015). Ontologies in E-learning: Review of the literature. *International Journal of Software Engineering and Its Applications, 9*(2), 67–84.

Baker, R. S. (2019). Challenges for the future of educational data mining: The baker learning analytics prizes. *Journal of Educational Data Mining, 11*, 1–17.

Boyer, S., & Veeramachaneni, K. (2015). Transfer learning for predictive models in massive open online courses. In C. Conati, N. Heernan, A. Mitrovic, & M. Verdejo (Eds.), *Artificial intelligence in education. AIED 2015. Lecture notes in computer science.* (Vol. 9112). Springer.

Castro, F. A., & Alonso, M. A. (2011). Learning objects and ontologies to perform educational data mining. In *International conference on frontiers in education: Computer science and computer engineering*.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education, 96*, 42–54.

Cerezo, R., Boragín, A., Esteban, M., & Romero, C. (2020). Process mining for self-regulated learning assessment in E-learning. *Journal of Computing in Higher Education, 32*, 74–88.
Chang, M., D'Aniello, G., Gaeta, M., Orciuoli, F., Sampson, D., & Simonelli, C. (2020). Building ontology-driven tutoring models for intelligent tutoring systems using data mining. *IEEE Access, 8*, 48151–48161.
Dhuria, S., & Chawla, S. (2014). Ontologies for personalized E-learning in the semantic web. *International Journal of Advanced Engineering and Nano Technology, 1*(4), 13–18.
Ding, M., Wang, Y., Hemberg, E., & O'Reilly, U.-M. (2019). Transfer learning using representation learning in massive open online courses. In *Proceedings of the 9th international conference on learning analytics & knowledge (LAK19), Empe, AZ, USA, 4–8 March 2019* (pp. 145–154). ACM.
Dorça, F. A., Carvalho, V. C., Mendes, M. M., Araújo, R. D., Ferreira, H. N., & Cattelan, R. G. (2017). An approach for automatic and dynamic analysis of learning objects repositories through ontologies and data mining techniques for supporting personalized recommendation of content in adaptive and intelligent educational systems. In *IEEE 17th international conference on advanced learning technologies (ICALT)* (pp. 514–516).
Dougiamas, M., & Taylor, P. (2008). Moodle: Using learning communities to create an open source course management system. In *EdMedia: World conference on educational media and technology* (pp. 171–178). The LearnTechLib.
El-Rady, A. A. (2020). An ontological model to predict dropout students using machine learning techniques. In *3rd international conference on computer applications & information security (ICCAIS)* (pp. 1–5).
Fong, S., Ip, A., & Mohammed, S. (2011). *A multi-level biomedical classification model by using aggregation and abstraction technique*s. Biomedical Engineering (Biomed 2011).
Forehand, M. (2005). Bloom's taxonomy: Original and revised. In M. Orey (Ed.), *Emerging perspectives on learning, teaching, and technology (E-Book)*. Retrieved January 15, 2021, from https://textbookequity.org/Textbooks/Orey_Emergin_Perspectives_Learning.pdf
Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education, 28*, 68–84.
Grivokostopoulou, F., Perikos, I., & Hatzilygeroudis, I. (2014). Utilizing semantic web technologies and data mining techniques to analyze students learning and predict final performance. In *2014 IEEE international conference on teaching, assessment and learning for engineering (TALE)* (pp. 488–494).
Hunt, X. J., Kabul, I. K., & Silva, J. (2017). Transfer learning for education data. In *Proceedings of the ACM SIGKDD conference*.
López-Zambrano, J., Lara, J. A., & Romero, C. (2020). Towards portability of models for predicting students' final performance in university courses starting from Moodle Logs. *Applied Science, 10*(354), 1–23.
Marinho, T., Costa, E. B., Dermeval, D., Ferreira, R., Braz, L. M., Bittencourt, I. I., & Luna, H. P. L. (2010). An ontology-based software framework to provide educational data mining. In *Proceedings of the 2010 ACM symposium on applied computing (SAC '10). Association for computing machinery* (pp. 1433–1437).
Monllao-Olive, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2019). A quest for a one-size-fits-all neural network: Early prediction of students at risk in online courses. *IEEE Transactions on Learning Technologies, 12*(2), 171–183.
Montenegro-Marin, C. E., Cueva-Lovelle, J. M., Sanjuan, O., & Nuñez-Valdez, E. R. (2011). Towards an ontology to describe the taxonomy of common modules in learning management systems. In *International journal of interactive multimedia and artificial intelligence, 1, special issue on computer science and software engineering* (pp. 47–53).
Nouira, A., Cheniti-Belcadhi, L., & Braham, R. (2019). An ontology-based framework of assessment analytics for massive learning. *Computer Applications on Engineering Education, 27*, 1343–1360.
Octaviani, D., Pranolo, A., & Othman, S. (2015). RDB2Onto: An approach for creating semantic metadata from relational educational data. In *International conference on science and information technology (ICSITech)* (pp. 137–140).
Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology, 45*(3), 438–450.
Rollins, M. (2010). *Moodle and Bloom's taxonomy*. . Lulu Enterprise.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*, 368–384.
Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*, 12–27.
Romero, C., & Ventura, S. (2019). Guest editorial: Special issue on early prediction and supporting of learning performance. *IEEE Transactions on Learning Technologies, 12*(2), 145–147.
Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. In *Wiley interdisciplinary reviews: Data mining and knowledge discovery*. Wiley.
Tang, A., & Fong, S. (2010). A taxonomy-based classification model by using abstraction and aggregation. In *2010 6th international conference on advanced information management and service (IMS)* (pp. 448–454).

Javier López-Zambrano is Assistant Professor at Escuela Superior Politécnica Agropecuaria de Manabí (ESPAM MFL), Ecuador. He is currently Ph.D. student at University of Córdoba, UCO, Spain, where he is conducting research on early prediction of students' academic performance by using data mining techniques. His research interests in computer science include data mining, knowledge discovery in databases, and e-learning.

Juan A. Lara is Associate Professor and Research Scientist at Madrid Open University, UDIMA, Spain. He is currently member of Department of Computer. He holds a Ph.D. in Computer Science and two Post Graduate Masters in Information Technologies and Emerging Technologies to Develop Complex Software Systems from Technical University of Madrid, Spain. He is author of more than 30 papers published in international impact journals. He has served as guest editor in a number of special issues of high impact factor journals and organized one international conference. His research interests in computer science include data mining, knowledge discovery in databases, data fusion, artificial intelligence and e-learning.

Cristóbal Romero is Full Professor at the University of Córdoba in Spain and member of the KDIS (Knowledge Discovery and Intelligent Systems) research group. He has published more than 150 papers in books, journals and conferences, 50 of which have been published in Thomson-Reuters Impact Factor (IF) journals and some of them are important EDM (Educational Data Mining) surveys/reviews. He was also the co-editor of several special issues and two books regarding EDM specific topics. He was a founding officers of the international EDM society and he has served in the program committee of a great number of international conferences about education, personalization, data mining, AI and he has organized as Program Chair and Conference Chair several editions of EDM conference. His main research interests are the application of data mining/learning analytics and artificial intelligence techniques to educational data/environment/domain.

Psicothema

Psicothema 2021, Vol. 33, No. 3, pp. 1-10
doi: 10.7334/psicothema2021.62

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Early Prediction of Student Learning Performance Through Data Mining: A Systematic Review

Javier López-Zambrano[1], Juan Alfonso Lara Torralbo[2], and Cristóbal Romero[3]
[1] Escuela Superior Politécnica Agropecuaria de Manabí, [2] Madrid Open University, and University of Córdoba[3]

Abstract

Background: Early prediction of students' learning performance using data mining techniques is an important topic these days. The purpose of this literature review is to provide an overview of the current state of research in that area. **Method:** We conducted a literature review following a two-step procedure, looking for papers using the major search engines and selection based on certain criteria. **Results:** The document search process yielded 133 results, 82 of which were selected in order to answer some essential research questions in the area. The selected papers were grouped and described by the type of educational systems, the data mining techniques applied, the variables or features used, and how early accurate prediction was possible. **Conclusions:** Most of the papers analyzed were about online learning systems and traditional face-to-face learning in secondary and tertiary education; the most commonly-used predictive algorithms were J48, Random Forest, SVM, and Naive Bayes (classification), and logistic and linear regression (regression). The most important factors in early prediction were related to student assessment and data obtained from student interaction with Learning Management Systems. Finally, how early it was possible to make predictions depended on the type of educational system.

Keywords: Educational Data Mining; Learning Analytics; Early prediction of academic performance; Early Warning Systems; Detection of students at-risk of Dropping-out.

Resumen

Predicción Temprana del Rendimiento Académico con Minería de Datos: una Revisión Sistemática. **Antecedentes:** la predicción temprana del rendimiento académico mediante técnicas de minería de datos es un campo de estudio emergente, que se pretende analizar por medio de este artículo de revisión. **Método:** se ha revisado la literatura existente por medio de un proceso de búsqueda de artículos en los principales motores de búsqueda, y de selección de los mismos de acuerdo con ciertos criterios. **Resultados:** el proceso de búsqueda reportó 133 resultados, de los cuales 82 fueron seleccionados para dar respuesta a las preguntas de investigación planteadas. Se han agrupado los trabajos encontrados para poder dar respuesta a las preguntas por tipo de sistema educativo, técnicas de minería de datos aplicadas, variables empleadas y grado de anticipación con el que se puede predecir. **Conclusiones:** la mayor parte de los trabajos publicados corresponden a sistemas de aprendizaje en línea y presenciales-tradicionales en educación secundaria y terciaria; los algoritmos más utilizados el J48, Random Forest, SVM, Naive Bayes (clasificación), y la regresión logística y lineal (regresión); los datos de evaluación y los obtenidos de la interacción del estudiante con el entorno de aprendizaje son las variables más relevantes; finalmente, la anticipación en la predicción varía según el tipo de sistema educativo.

Palabras clave: Data Mining Educativo; Analítica de Aprendizaje; predicción temprana del rendimiento académico; sistemas de detección temprana; estudiantes en riesgo de abandono.

Predicting studentsí learning performance is a challenging but essential task in education (Romero & Ventura, 2013). The prediction of academic performance is important not only to help students take control of their own learning and become self-regulated learners but also to allow educators to identify at-risk students and reduce the chances of failure (Bogarín et al., 2018). This is a difficult task because of the many possible factors that can influence student performance. In order to shed some light on this problem, EDM (Educational Data Mining) and Learning Analytics (LA) techniques have been successfully applied, mainly in e-learning environments (LMS -Learning Management Systems-, MOOC -Massive Open Online Courses-; etc.), where the volume of generated data is especially large and the studentsí activity reflects their learning processes (Castro et al., 2007). Data with information about students can also be gathered from traditional face-to-face education environments and from blended learning environments (B-learning).

The use of EDM and LA techniques to analyze these large amounts of data has produced interesting, interpretable, useful and novel information about learners (Fayyad et al., 1996). The application of Data Mining (DM) techniques to information about learning activities produced in educational environments is known as EDM (Barnes et al., 2009). EDM uses the same DM techniques with certain adaptations depending on the specific problems to be solved (Romero & Ventura, 2020). One of its main tasks is to predict student learning performance (failure, success, school dropout, etc.). LA can be defined as the measurement, collection,

Received: February 19, 2021 • Accepted: May 20, 2021
Corresponding author: Juan Alfonso Lara Torralbo
Department of Computer Science
Madrid Open University (UDIMA)
28400 Collado Villalba (Spain)
e-mail: juanalfonso.lara@udima.es

analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs (Siemens, 2013). Hence, EDM and LA are deeply related fields, and share the common objective of predicting and guiding student learning.

Early prediction can be defined as the application of predictive models that use key variables to accurately predict student failure or dropout as early as possible (Berens et al., 2018; Yu et al., 2018). It also refers to the technological information in the management of studentsí academic work for the early detection of their potential or real academic problems (Wang et al., 2018). It is necessary to detect at-risk students as early as possible and thus provide early intervention or care to help students succeed and to prevent them from quitting or failing. A wide range of student information can be used to make early predictions of student performance. Examples include student-completed questionnaires (Krotseng, 1992), lessons and activities in the early stages of courses (Costa et al., 2017), student performance and demographic data (Berens et al., 2018), activities and comments on evaluations to analyze feelings (Yu et al., 2018), records from online environments (Howard et al., 2018), and affective and emotive variables (Mújica et al., 2019) among others.

Early prediction is a challenging task for the EDM field due to the many factors that can influence a studentís final status. It is a critical issue in education because it concerns many students at all stages (primary education, secondary education, and tertiary or higher education) and in schools and universities all over the world. Early prediction is also essential in order to identify at-risk students as early as possible in order to implement programs that provide appropriate, effective prevention strategies, give advice or recommendations, and carry out remedial actions or interventions (Romero & Ventura, 2019).

Although there are some review papers about the prediction of academic performance (Ameen et al., 2019; Felix et al., 2018), the identification of at-risk students in general (Nik Nurul Hafzan et al., 2019), the use of exclusively LMS course data for prediction (Na & Tasir, 2018), and the application of Early Warning Systems óEWSó (McMahon & Sembiante, 2020) (Liz-Domínguez et al., 2019), none of them focus on early prediction through data mining techniques. This is the main reason that the current survey is necessary.

In this paper, rather than only analyzing studies about early prediction, an analysis was also carried out looking at different aspects related to early prediction, such as the education systems considered, the most commonly-used techniques and algorithms, how early it is possible to predict, and which are the most commonly-used variables or attributes.

The purpose of this survey is to conduct a systematic review of the literature about early prediction of academic performance in order provide readers with an introduction to the application of EDM/LA for early prediction and thus answer the following research questions: In what type of educational system has early prediction been applied most often? What techniques have been used most often? Which specific algorithms are the most used, and which have produced the best prediction results? How early can academic performance be predicted with acceptable accuracy? What specific variables or attributes have been used and demonstrated better performance?

The major original scientific contributions of this paper are:

- We present and summarize the most important scientific literature about the use of data mining techniques for early prediction of student performance.
- We have taxonomized those references and grouped them by the type of educational system.
- We have discovered and presented a series of research niches and opportunities in the area by analyzing aspects such as the most-used techniques, the attributes used, and how early the predictions of academic performance can be made.

This paper is organized as follows: The procedure section describes the process used for the systematic review. The results and discussion sections describe the studies selected, and the answers to the five research questions. Finally, the conclusions and future lines of research are presented.

Method

Procedure

Search strategy

We followed the systematic literature review procedure by Tranfield et al. (2003). Systematic reviews begin by defining a review protocol that specifies the research questions and the methods that will be used to perform the review. Following that, we defined the keywords and the explicit inclusion and exclusion criteria for searching for and selecting papers about early prediction. A double filter process was applied to discard papers that did not meet the inclusion criteria after reading the abstract (first filter) and the full paper (second filter).

We used Google Scholar, Web of Science, and Scopus search engines in order to search for all academic papers about early prediction published up to November 2020. The search used the following search terms:

1. "Early prediction" AND "Data Mining" AND ("academic performance" OR "at-risk students" OR dropouts)
2. "Early prediction" AND "Learning Analytics" AND ("academic performance" OR "at-risk students" OR dropouts)
3. "Early detection" AND "Data Mining" AND ("academic performance" OR "at-risk students" OR dropouts)
4. "Early detection" AND "Learning Analytics" AND ("academic performance" OR "at-risk students "OR dropouts)
5. "Early warning systems" AND ("academic performance" OR "at-risk students" OR dropouts)

Selecting papers

The papers were selected by reading both the abstract and full content of the papers initially downloaded from the search and applying the following inclusion and exclusion rules:

- Inclusion: articles focused exclusively on the topic of early prediction of student performance through EDM techniques.
- Exclusion: articles that did not actually perform early prediction of students' performance through EDM techniques despite containing some of the search keywords.

Results

Starting from the search using the keywords noted above, a total of 133 papers were downloaded. There were 97 journal articles, 29 articles from international conferences, and 7 items corresponding to types such as books, reports, and doctoral theses.

As Figure 1 shows, the preliminary search identified 133 papers published up to November 2020 whose titles included the defined keywords. The abstract of each paper was read, leading to 17 papers being discarded for not doing early prediction. The remaining 116 papers were read in full, and 34 additional papers were discarded for the same reason. Many papers contained early prediction in the titles, but in reality they described classical prediction by using all the information provided at the end of the courses. The remaining 82 papers were used to answer the five research questions.

After reading the final selection of 82 articles, an analysis was carried out from various perspectives in order to answer each of the 5 research questions. In the sections, we describe and discuss the results and give an overview of the literature about the topic.

Discussion

Figure 2 shows that the first papers were published in the 1990s, which indicates that early prediction is not a new concern. However, it was not until 2008 when further research in this regard began,

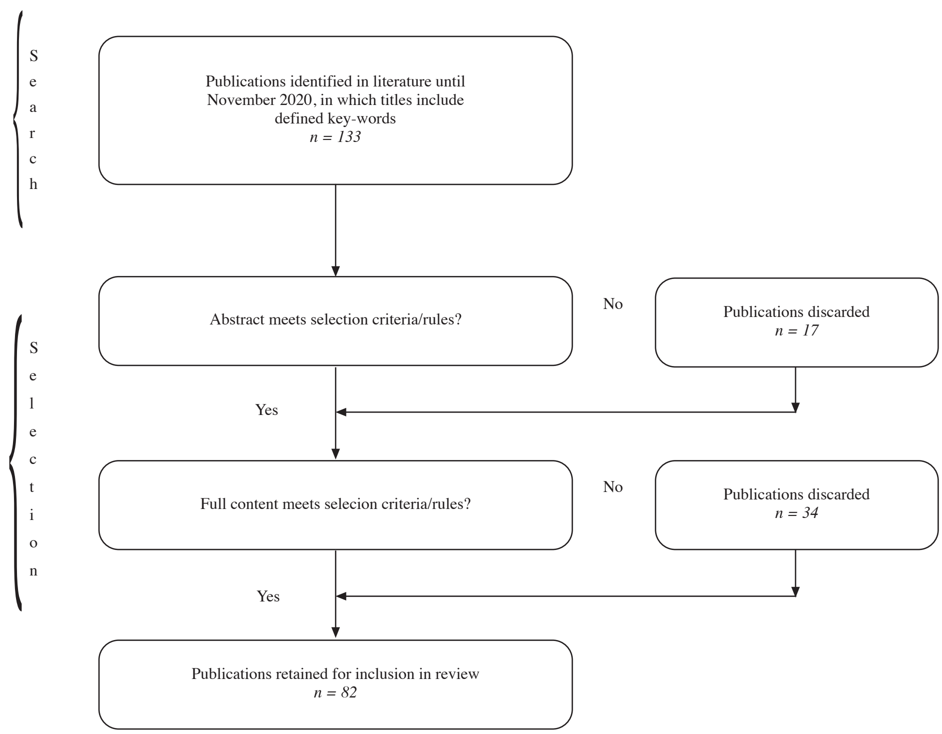

Figure 1. Procedure flowchart

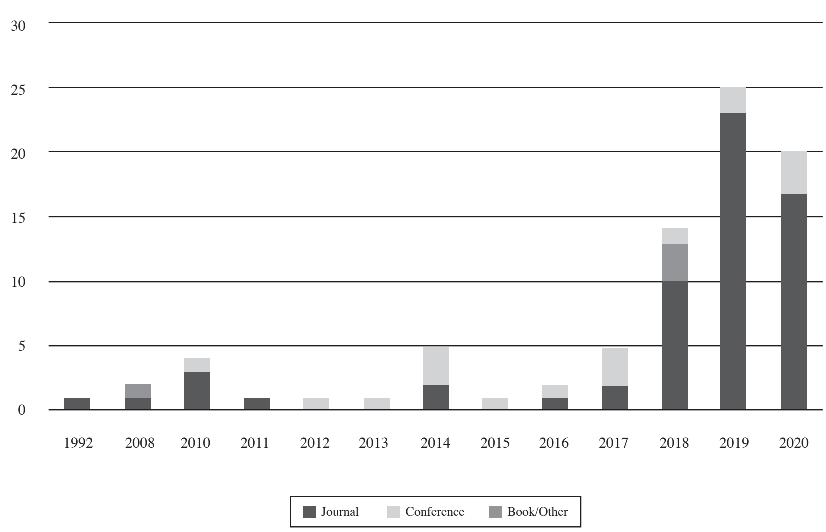

Figure 2. Number of papers published per year

and the most significant contributions came in the last decade. In addition, we have noticed that in the last 4 years (2017-2020) there have been a significant number of contributions.

Table 1 shows the 5 most-cited papers about early prediction of student learning performance. The first ranked paper affirms that LMS-generated student data can be used for identifying at-risk students and can allow more timely pedagogical interventions (Macfadyen & Dawson, 2010). The second describes the goals and objectives of the Open Academic Analytics Initiative (OAAI), and describes the process and challenges of collecting, organizing and mining student data to predict academic risk and the results of interventions with at□risk students (Jayaprakash et al., 2014). The third paper explores the socio-demographic variables and study environment that may influence student persistence or dropout and examines the extent to which these factors help us in pre-identifying successful and unsuccessful students (Kovačić, 2010). The fourth paper seeks to identify significant behavioral indicators of learning using LMS data regarding online course achievement (You, 2016). The fifth paper in the ranking presents a comparative study on the effectiveness of educational data mining techniques for early prediction of students likely to fail in introductory programming courses (Costa et al., 2017).

What type of educational system has early prediction been applied to most often?

Early prediction can be applied to various types of educational systems and levels. These include: Traditional education, referring to long-established practices traditionally used in schools (in-person); E-learning, which is a form of distance learning completely virtualized through digital channels (mainly the internet); and Blended learning, in which e-learning is combined with in-person classes (Romero & Ventura, 2013). The different educational levels are: Primary education, the first stage in formal compulsory education; Secondary education, the final stage of basic education and the phase before tertiary level; and Tertiary education, which refers to education provided mainly at universities, for example leading to academic or professional degrees.

To answer this question, we classified the selected papers by the type of educational system and education level. As Figure 3 shows, the studies used data mostly from online learning (47 papers – 57.3%) followed by traditional in-person environments (30 papers – 36.6%), while very few studies were conducted in hybrid or B-learning environments (5 papers – 6.1%). Figure 3 also shows that most of the 82 papers described studies done with students in tertiary education (76 papers – 86.6%), a few with secondary level students (6 papers – 7.3%), and none with primary level students. This indicates that most of the effort to date has been in early prediction with university students, which is also in accordance with the accessibility of the data. Student data from learning environments is easier to collect, manage and analyse, and in the authors' experience, higher education is much more computerized than primary and secondary education.

Table 2 shows a summary of the 82 selected papers grouped by type of educational environment and education level.

Table 1
Top 5 most cited papers in Google Scholar

#	Title	Reference	#Cites
1	Mining LMS data to develop an "early warning system" for educators: A proof of concept	(Macfadyen & Dawson, 2010)	1028
2	Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative	(Jayaprakash et al. 2014)	332
3	Early Prediction of Student Success: Mining Students Enrolment Data	(Kovačić, 2010)	262
4	Identifying significant indicators using LMS data to predict course achievement in online learning	(You, 2016)	245
5	Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses	(Costa et al., 2017)	199

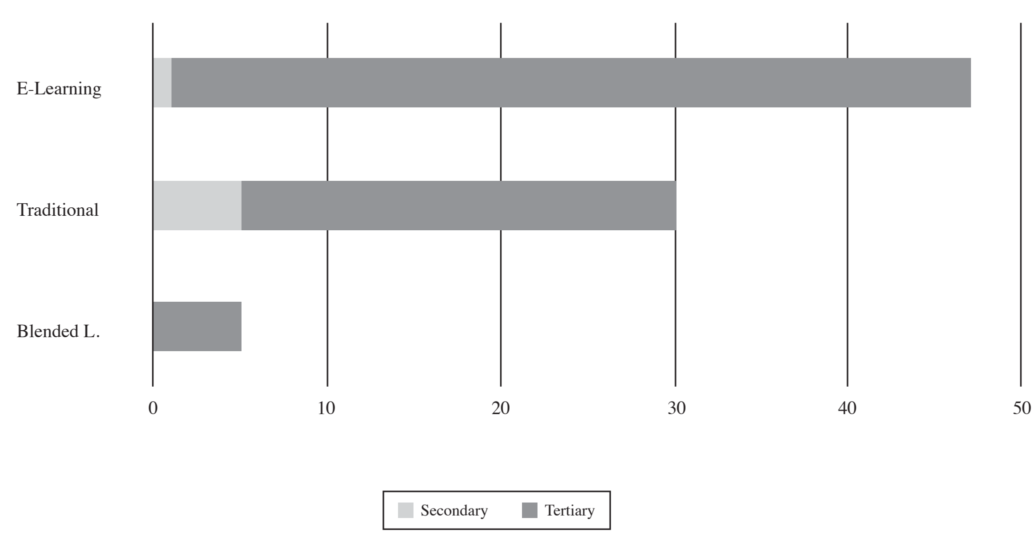

Figure 3. Education level data by type of learning environment

Table 2
Summary of all selected papers by type of educational environment and education level

Educational Environment	Education Level	# Papers	%
Face-to-face	Secondary	5	6.1
	Tertiary	25	30.5
E-Learning	Secondary	1	1.2
	Tertiary	46	56.1
B-Learning	Tertiary	5	6.1

What EDM techniques have been most used to date?

There are different data mining techniques for early prediction of student performance, both supervised (classification and regression) and unsupervised (clustering and association). Classification tries to predict a categorical or nominal value whereas regression tries to predict a numerical value. Clustering puts similar objects into groups and association finds associations or relationships.

Figure 4 shows the frequency of use of techniques in the 82 selected papers in order to determine the most widely-used techniques in EDM. Classification is the most commonly-used technique with 50 papers (42.4%), followed by regression with 33 papers (28%). Clustering, with 13 papers (11%), and association, with 2 papers (1.7%), were used much less often, along with other techniques that were not specified (16.9% noted Machine Learning / Data Mining generically). Hence, the two main DM techniques that have traditionally been applied to early prediction of student academic performance are classification and regression, both supervised techniques. Regression techniques have been used to predict the specific numerical value of a student's performance, and classification has been used to predict the class to which the student belongs, such as Pass/Fail, Success/Failure, or Retain/Dropout.

Which specific algorithms are the most used, and which have produced the best prediction results?

There is a wide range of specific data mining algorithms for doing early prediction. In classification, the most popular were Decision Tree, Random Forest, Support Vector Machine, Naive Bayes, K-Nearest-Neighbour, Boosted Tress, Adaptive Boosting, Gradient Boosting. Popular regression algorithms included Logistic Regression, Linear Regression, and Bayesian Additive Regressive Trees. In Clustering, the popular algorithms were K-Means, Balanced Iterative Reducing, and Clustering using Hierarchies, while in Association, they were Class Association Rule and Random Guess.

Table 3 shows a summary giving the type of DM method, the name of the specific algorithm, and the number of times each algorithm was used in the papers in absolute and percentage terms. The most widely-used algorithms were Naive Bayes, Decision Tree, Support Vector Machine and Logistic Regression.

In terms of algorithm accuracy, the best results were obtained by Miguéis et al. (2018), who achieved 96.1% prediction accuracy with Random Forest, and Razak et al. (2018), who achieved 96.2% with linear regression and 82% with decision tree (J48). Jiang et al. (2014) achieved 92.6% accuracy with logistic regression. Costa

Table 3
Most used algorithms and best results if authors provide them

Method	Algorithm	#	%
Classification	Decision Tree (J48)	31	38%
	Random Forest	25	30%
	Support Vector Machine	21	26%
	Naive Bayes	14	17%
	K-Nearest-Neighbor	10	12%
	Boosted Trees	7	9%
	Adaptive Boosting	7	9%
	Gradient Boosting (XGBoost)	3	4%
	Other	5	6%
Regression	Logistic Regression	23	28%
	Linear Regression	12	15%
	Bayesian Additive Regressive Trees	1	1%
	Other	12	15%
Clustering	K-Means clustering	2	2%
	Balanced Iterative Reducing and Clustering using Hierarchies	1	1%
Association	Class Association Rule	1	1%
	Random Guess	1	1%

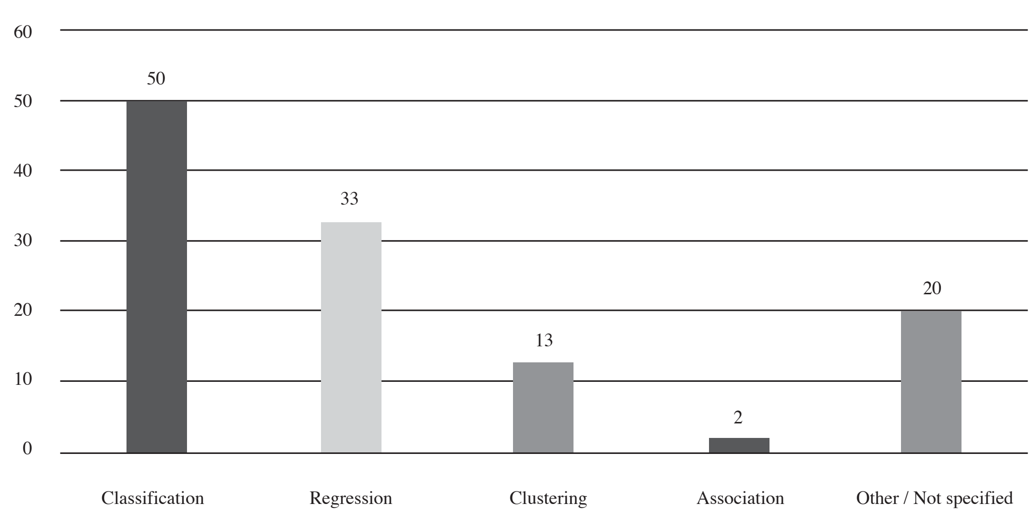

Figure 4. Frequency of use of EDM techniques

et al. (2017) achieved 92% accuracy. However, they also applied naive Bayes and decision tree algorithms as did Casey (2017), who achieved 69% prediction accuracy. In contrast, Chung & Lee (2018) achieved 95% accuracy with their best model applying random forest, while Wang et al. (2018) applied naive Bayes and achieved 85% accuracy.

How much earlier can academic performance be accurately predicted through EDM techniques?

Course length varies depending on the educational environment. For example, traditional education courses can last from four months to a semester or a year. The B-learning modality is similar because the system generally fits the times of an in person (traditional) course, while in e-learning, a course can last from a few weeks to several months. This means that there are different timespans for early prediction, therefore, the answer to this question is addressed for each type of educational environment separately. Early prediction times will depend on the modality of the course.

Traditional Environment

Within the traditional in-person educational environment, most papers do not explicitly indicate how early they can predict academic performance, very few provide that information. Berens et al. (2018) conducted a study over several semesters of bachelor's degrees at two universities (state and private). They showed that the prediction accuracy significantly improved as the semesters went by. At the time of the students' enrolment, they achieved 68% prediction accuracy for the public university and 67% for the private. After obtaining student performance data at the end of the first semester, they achieved 79% accuracy for the public university and 85% for the private, and after the fourth semester, the prediction accuracy reached 90% for the public and 95% for the private. In contrast, Wang et al. (2018) only indicated that success or failure can be predicted in the first semester with good accuracy. Bursać et al. (2019) used models that were, in the second week of a 13-week course, able to determine whether some of the students needed assistance in learning and assimilating learning materials in order to achieve a good grade at the end of the educational process.

E-Learning Environment

One of the most notable of the papers about e-learning courses was from Kuzilek et al. (2015). They managed to increase prediction accuracy by approximately 50% at the beginning of the semester and more than 90% at the end of a high school course. In a 16-week course, Han et al. (2016) produced a model in which the area under the curve, AUC (an indicator of the goodness of the prediction that represents the relationship between the sensitivity and specificity of a predictive model), was in the 0.62-0.83 range, predicting a week ahead. Howard et al. (2018) predicted students' final grades at week 6 (out of 12), based on a mean absolute error up to 6.5 percentage points. Vitiello et al. (2018) achieved 0.8 Accuracy when considering the active time of 10% of the users or the first five days after the initial user interaction. According to Hlosta et al. (2017), it is important for evaluations to be performed in the first few days of a course. If the score is over 50%, there is a high probability of students' academic success. Aljohani et al. (2019) Predicted pass/fail classes with around 90% accuracy within the first 10 weeks of student interaction in a virtual learning environment. Queiroga et al. (2020) predicted at-risk students with an AUC above 0.75 in the initial weeks of a course. Li et al. (2020), reported an AUC score of 0.8262 in the task of next-day prediction while the performance fell to 0.7430 in a next-two-week prediction task.

B-Learning Environment

In papers about B-learning, Costa et al. (2017) achieved an accuracy that varied between 0.50 and 0.82 in a distance education course and from 0.50 to 0.79 for a course on the learning environment. These results indicate that after the first week of these courses, it was possible to identify students who were likely to fail with at least 50% effectiveness. Lu et al. (2018) showed that the final academic performance of students in a blended course could be predicted with high stability and accuracy between weeks 1-6 of the course (out of 18). Macarini et al. (2019) detected at-risk students in the first week of a course with an AUC value from 0.7 to 0.9.

What specific variables or attributes have been used and produced better performance?

The variables and student attributes used for prediction vary depending on the educational environment, and even within the

Table 4
Most used variables classified by educational environment and source of data

FACE-TO-FACE	E-LEARNING	B-LEARNING
DEMOGRAPHICS: AGE, NATIONALITY, SEX, CITY, FAMILY INCOME LEVEL, HAVING A SCHOLARSHIP, HAVING A JOB OR BABY, LIVING WITH PARENTS, LEGAL GUARDIANS' EDUCATIONAL ATTAINMENT ACTIVITY: HOMEWORK GRADE, HOMEWORK CLICKS, ATTENDANCE, DISCUSSION, POSITIVE VALENCE, NEGATIVE VALENCE, NEUTRAL VALENCE, AVERAGE OF VALENCE, EPORTFOLIO ENGAGEMENT FEATURES PERFORMANCE: TOTAL CREDITS, CREDITS GAINED, FAILING CREDITS, PASSING RATE, ARITHMETIC MEAN SCORE, WEIGHTED AVERAGE CREDIT SCORE, AVERAGE CREDIT SCORE POINT, CREDIT SCORE POINT, FAILING SCORE	Interaction: Videos watched, problems attempted; total number of activities; total number of active days; total number of sessions, number of successful compilations, ratio between on-campus and off-campus connections, number of connections, time spent on the platform, time spent on slides within the platform, time spent typing in the platform, time idle in the platform, slides covered, number of slides visited, number of slides opened, number of transactions, number of mail messages read, number of mail messages sent, number of discussion messages read, number of files viewed, number of web links viewed, number of clicks. Performance: number of assessments started, number of assessments finished, time spent on assessments, number of assignments read, number of assignments submitted, time spent on assignments	On-campus: age, gender, civil status, income, number of homework exercises, participation in class, performance in weekly activities and final exam Distance education: time and number of accesses and messages in communication tools (blog, glossary, wiki, and forums), video-viewing behaviour, out-of-class practice behaviour, number of clicks and time with other course resources, quiz scores and virtual tutoring

same environment, the variables vary between studies. Researchers have used different groups of variables in each paper, which makes it hard to tabulate the variables by frequency of use. In general, these variables come from the same data sources, such as student demographics, student activities and student interactions. Table 4 shows the most commonly-used variables in the selected papers grouped by the type of educational system and source of data.

As Table 4 shows, in Traditional education, there are three main sources of variables: demographics, performance, and activity. In E-learning environments there are only two: variables related to student interactions and performance. Finally, on-campus and distance education related variables were found to be used in B-learning systems. In order to see which variables produced the most accurate predictions, we examine each type of educational environment separately below.

Traditional Environments

In traditional in-person educational environments, there are a group of variables that were used most. Berens et al. (2018), Cano & Leonard (2019), and Araújo et al. (2019) used academic performance data and student demographic data to achieve a 79% prediction accuracy at the end of the first semester for a public university and 85% for a private university in applied sciences. Along similar lines, Aguiar et al. (2014) used similar data, supplemented with ePortfolio engagement features, where the highest AUROC value (0.929) was obtained by the dataset with the highest academic participation, and the academic performance was worst with an AUROC value of 0.654. Kovačić (2010) used student demographic data and the study environment to achieve a general classification percentage of 60.5%. Yu et al. (2018) considered the relative variables of tasks, assistance, and discussion. They also considered a variable called courage, which is obtained by applying sentiment analysis to identify affective information within self-evaluations based on written text, comments that reflect learning attitudes towards the lesson, comprehension of the course content, and learning difficulties, which produced a prediction accuracy of 76%.

E-Learning Environments

In e-learning education systems, most of the studies used attributes related to interaction with the learning environment. Kuzilek et al. (2015) used these types of attributes to achieve 93.4% accuracy. Similarly, Chui et al. (2018) used these same types of attributes, among others related to module presentation, and achieved between 92.2% and 93.8% accuracy predicting at-risk students. Among the papers that focused more on the attributes of interaction with the study courses, Han et al. (2016) used attributes such as time of interaction with resources, the interaction of students with problems and submissions, and study habits to achieve an AUC between 0.62 and 0.83. Other studies used attributes such as the number of emails sent, and the number of evaluations made. Macfadyen & Dawson (2010) and Nistor & Neubauer (2010) achieved significant prediction results and they indicated that quiz marks were a very important predictive factor. Olivé et al. (2019) used neural networks to predict which students were likely to submit their assignments on time using data from student and peer activity, student activity and peer activity separated from course info, and student activity, peer activity, and course information trained separately (the networks with the greatest predictive power). Mbouzao et al. (2020) identified failure patterns of up to 60% of students who would dropout or fail the course based on the first week student interaction with MOOC videos in a thirteen-week course, and were able to identify 78% of successful students. Kuzilek et al. (2015), Ortigosa et al. (2019), Kostopoulos et al. (2019), and Waheed et al. (2020) used demographic and variable data from the LMS. Choi et al. (2018), Aljohani et al. (2019), Villa-Torrano et al. (2020), Chen & Cui (2020), and Cui et al. (2020) used the number of clicks as a predictive attribute.

B-Learning Environments

The most used variables for B-Learning environments came from on-campus traditional in-person and distance or e-learning sources. Costa et al. (2017) used attributes such as gender, marital status, age, exam, forums, access, messages, wiki, and transfers, producing predictions that were 92% accurate. Lu et al. (2018) used attributes such as video visualization, out-of-class practice behaviour, homework and questionnaire marks, and after-school tutoring assistance, achieving accuracy between 82-83%. Macarini et al. (2019) used data linked to three different aspects of student interactions (cognitive presence, teaching presence, and social presence) aiming to predict students at risk of failing based on an existing theory about how interactions work inside Virtual Learning Environments. Gitinabard et al. (2019) found that the most important features were total time spent in both types of sessions, total number of actions performed in both browser and study sessions, number of study and browser sessions, number of homogeneous sessions between study and browser sessions.

Research Directions

In this paper, we have described the current state of the art in early prediction of student performance through data mining techniques by means of a systematic review of the literature. We also defined five research questions whose answers can provide important findings for the scientific educational community:

- With regard to the first research question, we have shown that most of the published papers were about online learning systems and traditional in-person secondary and tertiary education. However, very little research has been conducted on early prediction in primary education, which is an open research area. According to the results published in some recent papers, one very promising field is the application of data mining techniques for early prediction of student performance in blended learning environments.
- In relation to the second question, we have shown that the most commonly-used techniques were classification and regression. However, it should be noted that the application of association and clustering in conjunction with the first two may imply a certain trend. At the very least, the clustering technique was shown to be able to be used to make a prediction without using any other techniques (Chau et al., 2018).
- In terms of the third question, we have shown that within each technique, there were some specific algorithms that were widely used and which have produced very good

prediction results. In the classification technique, the stand outs were J48, Random Forest, SVM, and Naive Bayes stand out, while in the regression technique, logistic regression and linear regression stood out. These algorithms are recommended for new researchers when dealing with an early prediction problem.

- With regard to the fourth question, we have shown that how early the prediction can be done varies based on the type of educational system. Within traditional in-person education, Berens et al., (2018) achieved an accuracy of between 78%-84% predicting dropout, with data from the first semester by using average grade (avg. Grade/semester) as the most important predictor. In e-learning environments, an evaluation test should be performed in the first few days of the course, such that if the test score is over 50%, there will be a high probability of a student's academic success (Hlosta et al., 2017).
- In relation to the fifth question, we have shown that most studies used student assessment data when doing early prediction. Within traditional environments, most of the papers also used demographic data to make predictions (Aguiar et al., 2014). Meanwhile, in virtual environments (e-learning and B-learning), most of the variables were gathered from students' interaction with the system and there is an increasing interest in sentiment analysis data (Yu et al., 2018).

Finally, we would like to highlight some future lines that we consider important research opportunities for the EDM research community:

- Selecting and evaluating what are the most important very early factors or indicators that affect student performance in each type of educational system and at each level: More research is needed on selecting the best features to use according to the type of educational system in order to be able to provide earlier predictions (for example in the first day or week, or even before starting the course, when the student registers). This can be dealt with as a multi-view problem, in which the huge amounts of data used for making predictions come from multiple sources and different data sources and we need to select the best attributes.
- Generalizing early prediction models in order to apply them or transfer them to other courses. There is a need to generalize and reuse these models but providing good accuracy is a challenge because they are specific to the courses. The problem is that each study uses different features according to the characteristics of each course, which creates difficulties in adapting any one of the existing plethora of models to any course. More work is necessary to produce good models that are transferable to different courses from the original.
- Developing and testing Early Warning Systems (EWS) and Response to Intervention (RtI) in a real education environment. Real early warning environments should be integrated to close the circle so that following prediction, actions or mitigation measures should be taken for at-risk students at risk: show results, send reports, make recommendations, provide feedback to different stakeholders, etc. More research is necessary in EDM to develoo frameworks, early warning systems and apply real-time intervention strategies in educational environments to work together with educational science (Romero & Ventura, 2019).

Acknowledgments

The authors gratefully acknowledge the financial support provided by the Spanish Ministry of Science and Innovation in the project: "Improving Data Science User's Experience with Computational Intelligence (INTENSE)" and the University of Cordoba, project UCO-FEDER 18 REF.1263116 MOD.A.s

References

Aguiar, E., Ambrose, G. A. A., Chawla, N. V., Goodrich, V., & Brockman, J. (2014). Engagement vs Performance: Using Electronic Portfolios to Predict First Semester Engineering Student Persistence. *Journal of Learning Analytics, 1*(3), 7-33. https://doi.org/10.18608/jla.2014.13.3

Aljohani, N. R., Fayoumi, A., & Hassan, S.-U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability, 11*(24), 7238. https://doi.org/10.3390/su11247238

Ameen, A. O., Alarape, M. A., & Adewole, K. S. (2019). Students' academic performance and dropout prediction. *Malaysian Journal of Computing, 4*(2), 278-303. https://doi.org/10.24191/mjoc.v4i2.6701

Araújo, A., Leite, C., Costa, P., & Costa, M. J. (2019). Early identification of first-year students at risk of dropping out of high-school entry medical school: The usefulness of teachers' ratings of class participation. *Advances in Health Sciences Education, 24*, 251-268.

Barnes, T., Desmarais, M., Romero, C., & Ventura, S. (2009, July). Educational Data Mining 2009 [Conference presentation]. 2nd International Conference On Educational Data Mining, Córdoba, Spain.

Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early detection of students at risk-predicting student dropouts using administrative student data from German universities and machine learning methods. *Journal of Educational Data Mining, 11*(3), 1-41. https://doi.org/10.5281/zenodo.3594771

Bogarín, A., Cerezo, R., & Romero, C. (2018). Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs). *Psicothema, 30*(3), 322-329. http://dx.doi.org/10.7334/psicothema2018.116

Bursać, M., Blagojević, M., & Milošević, D. (2019). Early prediction of student success based on data mining and artificial neural network. In D. Milošević, Y. Tang, & Q. Zu (Eds.), Lecture Notes in Computer Science: Vol. 11956. Human Centered Computing (pp. 26-31). Springer. https://doi.org/10.1007/978-3-030-37429-7_3

Cano, A., & Leonard, J. D. (2019). Interpretable Multiview Early Warning System Adapted to Underrepresented Student Populations. *IEEE Transactions on Learning Technologies, 12*(2), 198-211. https://doi.org/10.1109/TLT.2019.2911079

Casey, K. (2017). Using Keystroke Analytics to Improve Pass-Fail Classifiers. *Journal of Learning Analytics, 4*(2), 189-211. https://doi.org/10.18608/jla.2017.42.14

Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying Data Mining Techniques to e-Learning Problems. In L. C. Jain, R. A. Tedman, & D. K. Tedman (Eds.), Studies in Computational Intelligence: Vol. 62. Evolution of Teaching and Learning Paradigms in Intelligent Environment (pp. 183-221). Springer. https://doi.org/10.1007/978-3-540-71974-8_8

Chau, L. M., Chau, V. T. N., & Phung, N. H. (2018). On Temporal Cluster Analysis for Early Identifying In-trouble Students in an Academic Credit System. In H. T. Bao, L. A. Cuong, H. Van Khuong, & B. T. Lam (Eds.), Proceedings of the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS) (pp. 171-176). IEEE. https://doi.org/10.1109/nics.2018.8606827

Chen, F., & Cui, Y. (2020). Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics, 7*(2), 1-17. https://doi.org/10.18608/JLA.2020.72.1

Choi, S. P. M., Lam, S. S., Li, K. C., & Wong, B. T. M. (2018). Learning analytics at low cost: At-risk student prediction with clicker data and systematic proactive interventions. *Educational Technology and Society, 21*(2), 273-290. https://doi.org/10.2307/26388407

Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2018). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior, 107*, 105584. https://doi.org/10.1016/j.chb.2018.06.032

Chung, J. Y., & Lee, S. (2018). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review, 96*, 346-353. https://doi.org/10.1016/j.childyouth.2018.11.030

Cui, Y., Chen, F., & Shiri, A. (2020). Scale up predictive models for early detection of at-risk students: A feasibility study. *Information and Learning Sciences, 121*(3-4), 97-116. https://doi.org/10.1108/ILS-05-2019-0041

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *Artificial Intelligence Magazine, 17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Felix, I., Ambrósio, A. P., LIMA, P. D. S., & Brancher, J. D. (2018, October). Data Mining for Student Outcome Prediction on Moodle: A systematic mapping [Conference presentation]. Anais Do XXIX Simpósio Brasileiro de Informática Na Educação (SBIE 2018), Fortaleza, Brasil. https://doi.org/10.5753/cbie.sbie.2018.1393

Gitinabard, N., Xu, Y., Heckman, S., Barnes, T., & Lynch, C. F. (2019). How Widely Can Prediction Models be Generalized? An Analysis of Performance Prediction in Blended Courses. *ArXiv, 12*(2), 184-197.

Han, W., Jun, D., Xiaopeng, G., Qiaoye, Y., & Kangxu, L. (2017). Predicting Performance in a Small Private Online Course. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), Proceedings of the 10th International Conference on Educational Data Mining (pp. 384-385). Education Resources Information Center.

Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: Early identification of at-risk students without models based on legacy data. In A. Wise, P. H. Winne, & G. Lynch (Eds.), Proceedings of the Seventh International Learning Analytics & Knowledge Conference - LAK í17 (pp. 6-15). ACM. https://doi.org/10.1145/3027385.3027449

Howard, E., Meehan, M., & Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. *Internet and Higher Education, 37*, 66-75. https://doi.org/10.1016/j.iheduc.2018.02.001

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics, 1*(1), 6-47.

Jiang, S., Williams, A. E., Schenke, K., Warschauer, M., & Dowd, D. O. (2014). Predicting MOOC Performance with Week 1 Behavior. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), Proceedings of the 7th International Conference on Educational Data Mining (pp. 273-275). IEDMS.

Kostopoulos, G., Karlos, S., & Kotsiantis, S. (2019). Multiview Learning for Early Prognosis of Academic Performance: A Case Study. *IEEE Transactions on Learning Technologies, 12*(2), 212-224. https://doi.org/10.1109/TLT.2019.2911581

Kovačić, Z. (2010). Early Prediction of Student Success: Mining Students Enrolment Data. Proceedings of Informing Science & IT Education Conference (pp. 647-665). https://doi.org/10.28945/1281

Krotseng, M. V. (1992). Predicting persistence from the student adaptation to college questionnaire: Early warning or siren song? *Research in Higher Education, 33*, 99-111. https://doi.org/10.1007/BF00991974

Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z., & Wolff, A. (2015). OU Analyse: Analysing at-risk students at The Open University. Learning Analytics Review, LAK15-1, 1-16.

Li, H., Ding, W., & Liu, Z. (2020). Identifying at-risk k-12 students in multimodal online environments: A machine learning approach. arXiv preprint, 2003.09670.

Liz-Domínguez, M., Caeiro-Rodríguez, M., Llamas-Nistal, M., & Mikic-Fonte, F. (2019, June). Predictors and early warning systems in higher education ó A systematic literature review [Conference presentation]. CEUR Workshop at LASI-SPAIN, Vigo, Spain.

Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Lin, A. J. Q., & Yang, S. J. H. (2018). Applying Learning Analytics for the Early Prediction of Students í Academic Performance in Blended Learning. *Educational Technology and Society, 21*(2), 220-232. https://doi.org/10.2307/26388400

Macarini, L. A. B., Cechinel, C., Machado, M. F. B., Ramos, V. F. C., & Munoz, R. (2019). Predicting students success in blended learning-Evaluating different interactions inside learning management systems. *Applied Sciences, 9*(24), 5523. https://doi.org/10.3390/app9245523

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an ìearly warning systemî for educators: A proof of concept. *Computers and Education, 54*(2), 588-599. https://doi.org/10.1016/j.compedu.2009.09.008

Mbouzao, B., Desmarais, M. C., & Shrier, I. (2020). Early Prediction of Success in MOOC from Video Interaction Features. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), Lecture Notes in Computer Science: Vol. 12164. Artificial Intelligence in Education (pp. 191-196). Springer. https://doi.org/10.1007/978-3-030-52240-7_35.

McMahon, B. M., & Sembiante, S. F. (2020). Re-envisioning the purpose of early warning systems: Shifting the mindset from student identification to meaningful prediction and intervention. *Review of Education, 8,* 266-301. https://doi.org/10.1002/rev3.3183

Miguéis, V. L., Freitas, A., García, P. J. V, & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36-51. https://doi.org/10.1016/j.dss.2018.09.001

Mújica, D., Pérez Villalobos, A., Bernardo Gutiérrez, M. V., Cervero Fernández-Castañón A. B., & González-Pienda García, J. A. (2019). Affective and cognitive variables involved in structural prediction of university dropout. *Psicothema, 31*(4), 429-436. https://doi.org/31. 10.7334/psicothema2019.124

Na, K. S., & Tasir, Z. (2018). Identifying at-risk students in online learning by analysing learning behaviour: A systematic review. In A. Selamat (Ed.), Proceedigns of the 2017 IEEE Conference on Big Data and Analytics (pp. 118-123). IEEE. https://doi.org/10.1109/ICBDAA.2017.8284117

Nik Nurul Hafzan, M. Y., Safaai, D., Asiah, M., Mohd Saberi, M., & Siti Syuhaida, S. (2019). Review on Predictive Modelling Techniques for Identifying Students at Risk in University Environment. In L. M. Hee (Ed.), Proceedings of the 2019 MATEC Web of Conferences 255 (03002). EDP Sciences. https://doi.org/10.1051/matecconf/201925503002

Nistor, N., & Neubauer, K. (2010). From participation to dropout: Quantitative participation patterns in online university courses. *Computers and Education, 55*(2), 663-672. https://doi.org/10.1016/j.compedu.2010.02.026

Olivé, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2019). A Quest for a One-Size-Fits-All Neural Network: Early Prediction of Students at Risk in Online Courses. *IEEE Transactions on Learning Technologies, 12*(2), 171-183. https://doi.org/10.1109/TLT.2019.2911068

Ortigosa, A., Carro, R. M., Bravo-Agapito, J., Lizcano, D., Alcolea, J. J., & Blanco, Ó. (2019). From Lab to Production: Lessons Learnt and Real-Life Challenges of an Early Student-Dropout Prevention System. *IEEE Transactions on Learning Technologies, 12*(2), 264-277. https://doi.org/10.1109/TLT.2019.2911608

Queiroga, E. M., Lopes, J. L., Kappel, K., Aguiar, M., Araújo, R. M., Munoz, R., Villarroel, R., & Cechinel, C. (2020). A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course. *Applied Sciences, 10*(11), 3998. https://doi.org/10.3390/app10113998

Razak, R. A., Omar, M., Ahmad, M., & Mara, P. (2018). A Student Performance Prediction Model Using Data Mining Technique. *International Journal of Engineering & Technology, 7*(2.15), 61-63.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*, 12-27. https://doi.org/10.1002/widm.1075

Romero, C., & Ventura, S. (2019). Guest Editorial: Special Issue on Early Prediction and Supporting of Learning Performance. *IEEE Transactions on Learning Technologies, 12*(2), 145-147. https://doi.org/10.1109/TLT.2019.290810

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

Costa, E. B., Fonseca, B., Almeida-Santana, M., Ferreira de Araújo, F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of studentsí academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247-256. https://doi.org/10.1016/j.chb.2017.01.047

Siemens, G. (2013). Learning Analytics: The Emergence of a Discipline. *American Behavioral Scientist, 57*(10), 1380-1400. https://doi.org/10.1177/0002764213498851

Tranfield, D., Denyer, D. & Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management, 14*, 207-222. https://doi.org/10.1111/1467-8551.00375

Villa-Torrano, C., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., & Gómez-Sánchez, E. (2020). Early prediction of studentsí efficiency during online assessments using a long-short term memory architecture. In A. Martínez-Monés, A., Álvarez, M. Caeiro-Rodríguez, & Y. Dimitriadis (Eds.), Proceedings of Learning Analytics Summer Institute Spain 2020 (pp. 39-46). CEUR Workshop Proceedings.

Vitiello, M., Walk, S., Helic, D., Chang, V., & Guetl, C. (2018). User behavioral patterns and early dropouts detection: Improved users profiling through analysis of successive offering of MOOC. *Journal of Universal Computer Science, 24*(8), 1131-1150.

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior, 104*, 106-189. https://doi.org/10.1016/j.chb.2019.106189

Wang, Z., Zhu, C., Ying, Z., Zhang, Y., Wang, B., Jin, X., & Yang, H. (2018). Design and Implementation of Early Warning System Based on Educational Big Data. In W. Tu, L. Wang, C. Ji, N. Chen, Q. Sun, X. Song, & X. Wang (Eds.), Proceedigns of the 5th International Conference on Systems and Informatics (pp. 549-553). IEEE. https://doi.org/10.1109/icsai.2018.8599357

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education, 29*, 23-30. https://doi.org/10.1016/j.iheduc.2015.11.003

Yu, L. C., Lee, C. W., Pan, H. I., Chou, C. Y., Chao, P. Y., Chen, Z. H., Tseng, S. F., Chan, C. L., & Lai, K. R. (2018). Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. *Journal of Computer Assisted Learning, 34*, 358-365. https://doi.org/10.1111/jcal.12247