Modelo Predictivo Implementado en KNIME Basado en Analíticas del Aprendizajes para la Toma de Decisiones Oportunas en Entornos Virtuales de Aprendizaje

Predictive Model Implemented in KNIME Based on Learning Analytics for Timely Decision-Making in Virtual Learning Environments

Valderrama-Chauca, Enrique[1]; Apaza-Huanca, Jorge[2]; Cari-Mogrovejo, Lenin[3]; Arizaca-Machaca, Erick[4]
[1,2,3,4] Universidad Nacional de San Agustín de Arequipa-Perú

Resumen

La investigación tiene como objetivo implementar un modelo predictivo en la plataforma KNIME para analizar y comparar el rendimiento académico de los estudiantes, a partir de los datos generados en un Sistema de Gestión de Aprendizaje (LMS), identificando a los estudiantes en riesgo académico con el fin de generar intervenciones oportunas y puntuales. Se utilizó la metodología CRISP-DM en seis fases. Se toman para el análisis 22 indicadores de comportamiento en línea observados en el LMS que se distribuyen en cinco dimensiones: Rendimiento académico, número de accesos, número de tareas desarrolladas, aspectos sociales y cuestionarios. El modelo imple-mentado utiliza el algoritmo de entrenamiento Simple Regression Tree Learner. La población total lo conforman 30000 registros de estudiantes, de los cuales se ha tomado una muestra de 1000 registros mediante un muestreo aleatorio simple. Se evalúa la precisión del modelo para la predicción temprana del rendimiento académico de los estudiantes. Los 22 indicadores de comportamiento se comparan con las medias de rendimiento académico en tres asignaturas generales cuyas predicciones son satisfactorios, el error absoluto medio comparado con la media del primer curso fue de 3.81 y con una precisión del 89.7%, el error absoluto medio comparado con la media del segundo curso fue de 2.80 con una precisión del 94.2% y el error absoluto medio comparado con la media del tercer curso fue de 2.77 con una precisión del 93.8%.

Palabras clave: Modelo; Predicción; Learning; Analytics; Rendimiento; Académico; Entornos; Virtuales; Aprendizaje; Plataforma; KNIME.

Abstract

The research aims to implement a predictive model in the KNIME platform to analyze and compare the academic performance of students, based on the data stored in a Learning Management System (LMS), identifying students at academic risk in order to generate timely and timely interventions. The CRISP-DM methodology was used in six phases. Twenty-two online behavior indicators observed in the LMS are taken for analysis and distributed in five dimensions: Academic performance, number of accesses, number of tasks developed, social aspects and questionnaires. The model implemented uses the Simple Regression Tree Learner training algorithm. The total population consists of 30,000 student records, from which a sample of 1,000 records was taken by simple random sampling. The accuracy of the model for early prediction of students' academic performance is evaluated. The 22 behavioral indicators are compared with academic performance means in three general subjects whose predictions are satisfactory, the mean absolute error compared with the mean of the first course was 3.81 with an accuracy of 89.7%, the mean absolute error compared with the mean of the second course was 2.80 with an accuracy of 94.2% and the mean absolute error compared with the mean of the third course was 2.77 with an accuracy of 93.8%.

Keywords: Model; Prediction; Learning; Analytics; Performance; Academic; Environments; Virtual; Learning; Platform; KNIME.

Introducción

Actualmente los Sistemas de Gestión de Aprendizaje (LMS) almacenan una gran cantidad de datos sobre las interacciones de los estudiantes en archivos de registro, estos archivos generalmente contienen variables en los datos, como el número de inicios de sesión, el número de accesos a elementos de un curso en línea, el número de tareas completadas, el número de días en el curso en línea, las calificaciones de las actividades, la calificación del período, la calificación del curso etc. Estos datos podrían resultar interesantes para los instructores online ya que podrían contener información sobre el comportamiento de los estudiantes que podría influir en su rendimiento académico

(Bravo-Agapito, 2020).

Existe una oportunidad creciente en el ámbito educacional, donde los esfuerzos a nivel internacional buscan mejorar la calidad de la educación, para lo cual es necesario contar con sistemas de apoyo a la toma de decisiones que entreguen información de calidad y oportuna. Como menciona (Guadilla, 2000) en una esfera de la práctica institucional, los directivos y gestores académicos, muchas veces se contentan solo con el conocimiento derivado de su propia práctica. Tienden a autoabastecerse con estudios que responden a sus propósitos y a necesidades de corto plazo. Es aquí donde se busca intervenir, entregando herramientas que permitan contribuir y aportar en forma eficiente, con información relevante para estos procesos de toma decisiones, en este caso, realizando una aplicación acotada en su alcance solo a variables que impacten directamente en la retención de los alumnos, basando el análisis en estos factores y correlaciones que se presupone, que inciden en la retención, esto para poder intervenirlos oportunamente con acciones preventivas antes que correctivas (Ma-raza-Quispe et al, 2020).

Gran parte de las investigaciones en el campo de la analítica del aprendizaje ha uti-lizado datos de LMS para modelar el desempeño de los estudiantes para predecir las calificaciones de los estudiantes y para predecir qué estudiantes están en riesgo de reprobar un curso (Romero y Ventura, 2010), (Shum y Ferguson, 2012). Este es un paso importante en la analítica del aprendizaje, ya que informa la implementación de intervenciones, como la retroalimentación personalizada. Además, la pregunta es si realmente existe una única mejor manera de predecir el desempeño de los estudiantes en un conjunto diverso de cursos (Baker y Yacef, 2009).

No obstante, los estudios que han utilizado métodos y predictores similares han en-contrado resultados diferentes en los análisis correlacionales y los modelos de predic-ción. Además, la mayoría de los estudios se centran en predecir el rendimiento de los estudiantes una vez finalizado un curso, estableciendo qué tan bien se podría haber predicho el rendimiento de los estudiantes con los datos de uso de LMS, pero en un momento en el que los hallazgos ya no se pueden utilizar para una intervención opor-tuna (Campbell y Oblinger, 2007). Dado que los datos de LMS proporcionan informa-ción durante todo el curso, parece útil determinar si los datos de solo las primeras se-manas de un curso son suficientes para una predicción precisa del rendimiento de los estudiantes (Tempelaar et al, 2015). Por lo tanto, los autores argumentaron que el mejor momento para predecir el desempeño de los estudiantes es tan pronto como sea posible después de la primera evaluación, ya que este sería el mejor compromiso entre la retroalimentación temprana y el poder predictivo suficiente (Richardson et al, 2012).

Los sistemas de alerta temprana utilizan métodos de minería de datos para detectar a los estudiantes en riesgo de desaprobar cursos en diferentes niveles y contextos educativos (Howard et al, 2018). Según Knowles (Knowles y Haystacks, 2015) los in-dicadores de alerta temprana brindan a los instructores una advertencia avanzada de que los estudiantes necesitan ayuda en su proceso de aprendizaje. Estos sistemas contienen modelos predictivos con una colección de variables relacionadas con indi-cadores de alerta temprana. Estas variables generalmente contienen información so-bre datos demográficos e institucionales, características de los estudiantes, calificaciones de término o medio término y datos de interacción de LMS. El Sistema de alerta temprana de deserción escolar de Wisconsin es un desarrollo exitoso que brinda a los instructores una visión prospectiva del desempeño de los estudiantes (Knowles y Haystacks, 2015).

La mayoría de modelos de predicción del rendimiento de los estudiantes solamente se centran en la precisión de los resultados de la predicción, por esta razón lograr un modelo de predicción interpretable puede ser tan importante como obtener una alta precisión en la investigación de predicción de los aprendizajes de los estudiantes. La minería de datos educativos (MDE) es una disciplina emergente orientada al desarrollo de nuevos métodos y técnicas para explorar datos que provienen de contextos educa-tivos. Las bases de datos educativas almacenan gran cantidad de información, la misma que está siendo infrautilizada tanto por docentes, estudiantes e instituciones. Esto ocurre en vista de que los sistemas de gestión de aprendizaje (Learning Management Systems, LMS) como Moodle no disponen en su entorno de herramientas específicas de análisis de datos (Romero y Ventura, 2017).

Con el incremento de datos en los entornos virtuales de aprendizaje online, los in-vestigadores y académicos comienzan a encontrar formas de hacer que estos datos sean comprensibles y significativos (Baker, 2015). Por lo tanto, para analizar y desenterrar más información educativa potencial, los investigadores han explorado más la teoría del aprendizaje y el análisis, los marcos, las herramientas y las prácticas (De Marcos et al, 2016) y (Ruiperez-Valiente et al, 2015). En los últimos años, las personas han estudiado cada vez más la analítica de los comportamientos de aprendizaje y la predicción del rendimiento de los estudiantes ha atraído la atención de los académicos. Desde 2013, con el desarrollo continuo de la investigación y el análisis del aprendizaje, los investigadores han comenzado a utilizar el aprendizaje automático para estudiar las predicciones del aprendizaje (Greller y Hendrik, 2012). Por supuesto, esto se beneficia del desarrollo de plataformas de aprendizaje en línea como como MOOC, y un gran número de usuarios de la plataforma generan datos educativos.

Con respecto al fenómeno de que el número de usuarios registrados en la platafor-ma MOOC es alto y la tasa de abandono es extremadamente alta, los investigadores han comenzado a explorar la relación entre el comportamiento de los usuarios y si han abandonado el curso (o si pueden obtener un certificado). Al analizar la información del comportamiento del usuario y predecir los resultados del aprendizaje, esperan des-cubrir la relación, a fin de tomar

medidas tempranas para reducir la tasa de abandono de la plataforma MOOC (Baepler, 2008), (Mozs y Bowers, 2013).

En los estudios de predicción reales, la mayoría de los estudios utilizaron modelos de algoritmos incomprensibles para predecir los resultados del aprendizaje (Burgos et al, 2017), como las redes logísticas y bayesianas. Aunque tales modelos pueden predecir con precisión los resultados del aprendizaje, no se pueden interpretar. Sin duda, eso tendrá un impacto en la implementación de intervenciones específicas. Por lo tanto, para promover la construcción de un modelo de predicción y mejorar la calidad docente del aprendizaje en línea, desde la perspectiva de alta interpretación de los resultados de la predicción, es muy necesario construir un modelo de predicción del desempeño del estudiante basado en analíticas de comportamiento de aprendizaje en línea.

El modelo de análisis de aprendizaje es la base teórica para el análisis del compor-tamiento de aprendizaje en línea en el contexto de Big Data en educación. Actualmente, la analítica del aprendizaje está todavía en su fase de inicio. Sin embargo, los modelos de análisis de aprendizaje representativos existentes tienen las características comunes: Ciclo de datos. Desde la perspectiva de la analítica de enfoques de sistemas. George Siemens proporciona un modelo de analítica de aprendizaje cíclico, que incluye siete componentes: recopilación, almacenamiento, limpieza de datos, integración, análisis, representación y visualización, y acción (Iglesias-Pradas, 2014); Desde el ángulo de la mejora de la enseñanza y el aprendizaje, (Iglesias-Pradas, 2014) presenta un modelo cíclico de mejora continua para la analítica del aprendizaje, que consta de tres partes: recopilación de datos, procesamiento de información y aplicación del conocimiento, todo el proceso está respaldado por cuatro tipos de recursos tecnológicos: Computadoras, teoría, personas, organizaciones. Con el fin de explorar diferentes enfoques para el análisis de datos, (Siemens, 2013) plantea un marco de análisis de aprendizaje, que incluye diez partes, y la relación entre cada parte se convirtió en bidireccional.

Con el desarrollo continuo de la tecnología de aprendizaje y análisis (Ifenthaler y Widana Pathirana, 2014), en los últimos años han surgido más investigaciones nuevas sobre la predicción inclinada. A partir de la investigación existente, el modelo de predicción del aprendizaje se puede dividir en dos categorías, una pertenece al modelo de caja negra, es decir, para el resultado de la predicción, la razón no se puede ver directamente; el otro pertenece al modelo de caja blanca, es decir, hay una explicación directa del resultado de la predicción.

En estudios relacionados de predicción del aprendizaje, los investigadores gene-ralmente creen que la predicción de caja negra tiene mayor precisión. Especialmente cuando se trata de relaciones complejas. Los algoritmos de predicción de caja negra que se utilizan a menudo para la investigación incluyen regresión logística, máquinas de vectores de soporte (SVM) y bosque aleatorio (RF). (Baepler, 2008) utilizan algoritmos de regresión logística para predecir si los estudiantes registran cursos en sistemas asistidos inteligentes (ITS) (Baepler, 2008). Teniendo en cuenta la complejidad de la investigación y los datos difíciles recopilados del estado emocional, la motivación y el conocimiento previo, la precisión de la predicción final es cercana al 70% y el rendimiento de la predicción no es malo. Ley et al. Utilizan la regresión lineal y la regresión logística para predecir el nivel del alumno (si el alumno es un principiante, un avanza-do o un experto), los resultados de la predicción muestran que el algoritmo tiene un buen efecto de predicción. Los bosques aleatorios también se utilizan ampliamente como algoritmo de predicción.

La predicción de caja blanca tiene un mayor grado de interpretación, es decir, existe una razón específica para el resultado de la predicción. Por supuesto, cuando la inter-pretación es alta, la precisión de la predicción puede verse reducida. En el campo de la educación, los algoritmos de predicción de caja blanca que se utilizan a menudo para la investigación incluyen árboles de decisión y árboles aleatorios. Desarrollaron un sistema de predicción de caja blanca que predice el rendimiento de los alumnos en un sistema de gestión del aprendizaje (LMS) mediante el tiempo de aprendizaje dedicado a módulo de actividades y la frecuencia de uso del módulo (Rupp y Leighton, 2014). Se usaron el árbol de decisiones para desarrollar sistemas de predicción temprana usando cuatro valores propios y clasificarlos en cuatro categorías: comportamiento de aceptación, uso de materiales del curso en línea, estado de la tarea y participación en un foro de discusión. El objetivo de la predicción es la puntuación del alumno, y la precisión general de la predicción alcanzó el 95% (Harwati, 2019), después de lo cual combinaron estas técnicas con el algoritmo Adaboost. Mayor precisión al 98% (Freund).

La selección de indicadores de comportamiento de aprendizaje adecuados es una parte importante para la predicción. En la actualidad, existen muchos estudios teóricos sobre la selección de indicadores de comportamiento de aprendizaje (Macfadyen y Dawson, 2010). Estos estudios cubren indicadores que pueden estar relacionados con el efecto del aprendizaje desde diferentes perspectivas. Por ejemplo, Brown resumió tres indicadores principales de predicción: características de los estudiantes, indicado-res de comportamiento de aprendizaje y trabajos de los estudiantes. Discutió las capacidades de predicción relacionadas y los casos para diferentes tipos de indicadores. Resumió varios indicadores importantes del desempeño académico previo de los estudiantes, antecedentes de aprendizaje, participación en clase y desempeño social. Berry et al. Combinó tres indicadores que influyeron en el rendimiento académico: facto-res académicos, factores demográficos y factores culturales y sociales.

Según (Hu y Shihp, 2014), utilizó cuatro indicadores del porcentaje acumulado de conferencias en video que se pueden ver, el número de publicaciones en el foro, el número de usuarios que se basan en el foro y el número de

visualizaciones del progreso del curso como predictores; Cristóbal Romero et al. Predijo directamente el desempeño de los alumnos a partir de la participación del foro. Los indicadores incluían: el número de mensajes de los alumnos, el número de alumnos que crean nuevos temas, el número de alumnos que leen pegatinas, la concentración de alumnos y alumnos, la persistencia y otros indicadores.

Los análisis de escritura, como se ha mencionado anteriormente, se recogen de forma más eficiente por los ordenadores. Una vez almacenados como datos, pueden aplicarse en diversas circunstancias, dependiendo de la complejidad y la relevancia de las métricas. Los ejemplos más sencillos de uso de la analítica están presentes en la revisión ortográfica y gramatical en línea de Microsoft Word, que simplemente añade líneas onduladas bajo las palabras mal escritas o la gramática incorrecta para señalarlo. La herramienta de "recuento de palabras" de Word es otro ejemplo de métrica analítica sencilla (Chak y Chak, 2020).

Método

Descripción del contexto y de los participantes

La investigación se ha desarrollado en la Facultad de Educación de la Universidad Nacional de San Agustín de Arequipa. Esta universidad utiliza una plataforma de apo-yo virtual basada en el LMS Moodle. Bajo esta plataforma, las asignaturas que se im-parten en modalidad virtual permiten, por un lado, a los profesores mantener un repo-sitorio de información y registro de actividades académicas; y, por otro lado, a los alum-nos esta plataforma les permite tener una visión práctica de las actividades de apren-dizaje que se programan en los sílabos de las asignaturas. El entrenamiento del modelo implementado en la plataforma KNIME se ha implementado con 1000 registros de alumnos y 22 indicadores de comportamiento observados en el LMS en tres cursos generales durante el primer semestre de 2020, los cuales fueron seleccionados mediante un muestreo aleatorio simple de un total de 30000 registros probados de datos de asignaturas generales.

Instrumentos y procedimientos

Se ha seguido la metodología CRISP-DM, cuyo estándar incluye un modelo y una guía, estructurada en seis fases, y algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán una revisión parcial o total de las fases anteriores. Las fases o niveles que se identifican en la metodología CRISP-DM son la comprensión del negocio, la comprensión de los datos, la preparación de los datos, el modelado, la evaluación y la implementación, como se muestra en la Figura 1.

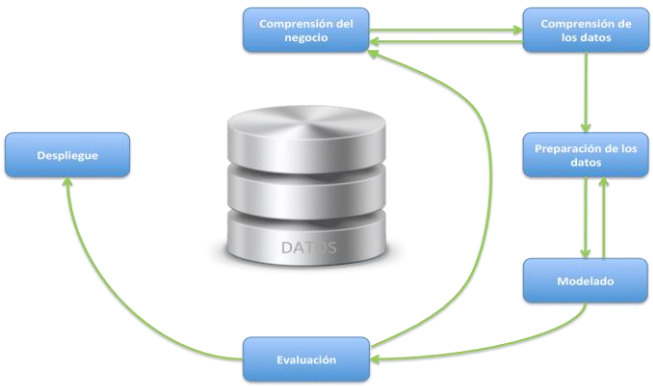

Figura 1. Fases del proceso en la metodología CRISP-DM
Fuente: (Chapman y Khabaza, 2000)

Fase 1: Comprensión del negocio

La presente investigación tiene por objetivo implementar un modelo predictivo para analizar y comparar la predicción del rendimiento académico utilizando los datos de LMS, analizamos si es posible identificar a los estudiantes en riesgo al principio de un curso y en qué medida los modelos pueden utilizarse para generar intervenciones es-pecíficas. Los indicadores de comportamientos de aprendizaje en entornos virtuales de aprendizaje afectan directamente la precisión y credibilidad de la predicción del desempeño del estudiante.

Fase 2: Comprensión de los datos

Por lo tanto, la selección científica de indicadores de comportamiento de aprendiza-je eficaces es una parte importante de la predicción del rendimiento académico del es-tudiante. Debido a la diversidad de comportamientos de aprendizaje en línea, y la complejidad de la correlación entre comportamientos, no todos los indicadores de comportamientos de aprendizaje que pueden afectar el efecto de aprendizaje pueden recopilarse de forma cuantitativa. Por lo tanto, con base en los resultados de investiga-ción existentes. Se toman como base cinco dimensiones: Predicción, accesos, tareas, aspectos sociales y cuestionarios. Los 26 indicadores de comportamiento de aprendizaje requeridos para el estudio fueron seleccionados como se muestra en la Tabla 1.

Tabla 1.
Indicadores de comportamiento de aprendizaje en línea

Dimensiones	Indicadores
Predicción	(1) Rendimiento académico
Accesos	(2) Tasa de asistencia
	(3) Número de registros en el LMS
	(4) Número de actividades realizadas por la noche
	(5) Número de inicios de sesión
	(6) Tiempo de permanencia en la sesión
	(7) Frecuencia de acceso a los foros
	(8) Número de mensajes en los foros
	(9) Frecuencia de acceso a los recursos
	(10) Frecuencia de acceso a los glosarios
	(11) Número de días transcurridos desde el último acceso al curso
	(12) Número de inicios de lecciones
Tareas	(13) Número de lecciones completadas
	(14) Número de registros en el LMS
	(15) Preparación del curso de evaluación
	(16) Número de envíos completados en total
	(17) Frecuencia de consulta de trabajos
	(18) Número de trabajos enviados
Aspectos sociales	(19) Sexo
	(20) Edad
	(21) Nivel de educación de los padres
	(22) Calidad de los alimentos que consume
Cuestionarios	(24) Promedio de 1 curso
	(25) Promedio de 2 cursos
	(26) Promedio de 3 cursos

Fase 3: Preparación de los datos

Se instaló en el servidor de la universidad la versión de Moodle 3.9 en el cual se instaló la plataforma IntelliBoard que proporciona servicios analíticos y de informes se extrajeron los datos estadísticos recopilados dentro del LMS, los datos se almacenaron en hojas de cálculo, luego fue necesario realizar un proceso de pre procesamiento de datos tal como se muestra en la Tabla 2 donde se muestra una descripción general de todas las variables predictoras y algunas estadísticas descriptivas donde haciendo un análisis de correlación para todos los cursos combinados se muestra que 15 de las 16 variables predictoras tenían una correlación estadísticamente significativa con la calificación final de los tres cursos, la excepción fue la variable edad.

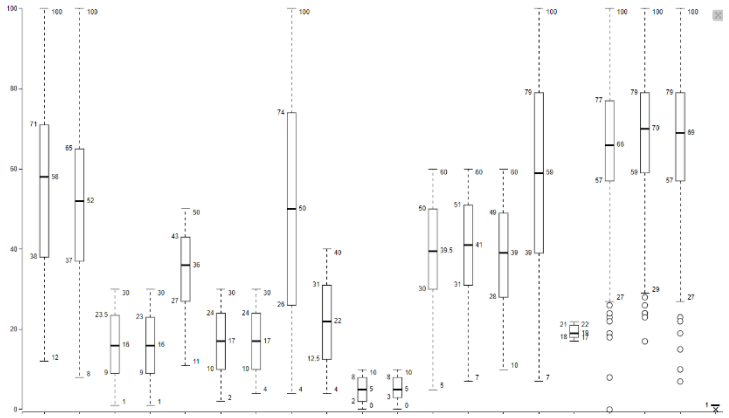

Figura 2. Preparación de los datos mediante el análisis de Box Plot.

En la tabla II se puede observar que los datos preparados no presentan valores atípicos se presenta una visión general de la simetría de la distribución de los datos ya que la mediana no está en el centro del rectángulo. En la tabla III se presenta los datos preparados para ser leídos en KNIME.

Tabla 2:
Variables predictoras y muestra estadísticas descriptivas

Columna	Min	Máximo	Significar	Desviación de Std.	Varianza	Asimetría	Curtosis	Suma total	Recuento de filas
Tasa de asistencia	12	100	56.084	20.606	424.608	-0.007	-0.966	56084	1000
Número de inicios de sesión	8	100	51.668	18.643	347.557	0.297	-0.555	51668	1000
Número de actividades realizadas a la derecha	1	30	15.896	8.355	69.801	0.084	-1.243	15896	1000
Número de lecciones iniciadas	1	30	16.077	8.289	68.708	-0.003	-1.164	16077	1000
Tiempo pasado en la sesión	11	50	35.155	9.143	83.597	-0.091	-1.114	35155	1000
Frecuencia de acceso a los foros	2	30	17.074	7.983	63.728	0	-1.242	17074	1000
Publicaciones del Foro de Números	4	30	17.15	7.761	60.238	-0.012	-1.196	17150	1000
Frecuencia de acceso a los recursos	4	100	50.713	27.663	765.236	0.044	-1.184	50713	1000
Frecuencia de acceso a los glosarios	4	40	21.678	10.59	112.146	0.011	-1.208	21678	1000
Número de días transcurridos desde el último acceso	0	10	4.915	3.141	9.866	0.049	-1.202	4915	1000
Número de registros en LMS	0	10	5.406	2.851	8.129	0.058	-1.199	5406	1000
Número de lecciones completadas	5	60	39.735	11.931	142.351	-0.003	-1.11	39735	1000

Número de envíos completados	7	60	40.787	11.664	136.05	-0.097	-1.036	40787	1000
Número de trabajos presentados	10	60	38.875	11.76	138.306	0.069	-1.175	38875	1000
Frecuencia de consultas laborales	7	100	58.922	23.399	547.527	0.04	-1.137	58922	1000
Edad	17	22	19.534	1.717	2.95	0.006	-1.307	19534	1000
Promedio 1 curso	0	100	66.089	15.163	229.919	-0.279	0.275	66089	1000
Promedio de 2 cursos	17	100	69.169	14.6	213.166	-0.259	-0.068	69169	1000
Promedio de 3 cursos	7	100	67.987	15.317	234.599	-0.333	0.122	67687	1000

Fase 4. Modelado

Técnica de Modelado

Debido a que se va a utilizar el software KNIME para realizar la minería de datos, se utiliza alguna de las técnicas de modelado que nos ofrece esta herramienta de acuerdo con los objetivos de nuestro. De los modelos que nos ofrece KNIME, el que mejor se adapta a nuestros objetivos es sería un modelo Simple Regression Tree Learner, puesto que los problemas que queremos resolver son problemas de predicción y los campos que se quieren predecir contienen valores continuos. Ver la figura 3.

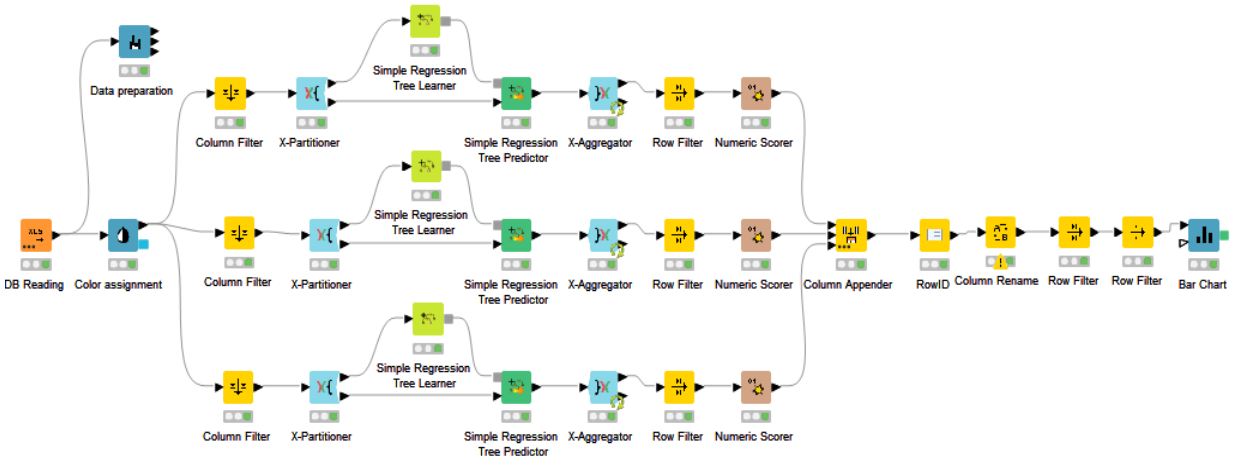

Figura 3. Modelo predictivo implementado en la plataforma KNIME

Generación del Plan de Pruebas

El procedimiento que se ha empleado para probar la calidad y validez del modelo es el de utilizar las medidas del "R al cuadrado" (R^2), el error absoluto medio (mean absolute error) y la "porcentaje del error absoluto medio" (mean absolute percentaje error). Estas medidas de error las calculan automáticamente KNIME al ejecutar el modelo.

Se hace una partición previa de los 1000 registros de los estudiantes, por un lado, está el conjunto de datos que se van a utilizar para generar el modelo, llamados datos de entrenamiento, y un segundo conjunto de datos que se empleará para realizar las pruebas y medir la calidad del modelo, llamados datos de prueba o de evaluación. Se utiliza un 70% de los datos para los datos de entrenamiento y el 30% restante para los datos de prueba.

Construir el Modelo

A continuación, se procede a ejecutar el modelo elegido sobre los datos de entrenamiento. Se describieron los ajustes de parámetros del modelo que se eligieron en la herramienta de minería de datos, así como la salida de dicho modelo y su descripción. Ver la figura 3.

Evaluación de los resultados

Resultados de la predicción

Los resultados mostraron que la precisión del modelo de predicción presenta un error absoluto medio comparado con el área de Ciencias fue de 3.813 y una precisión de 89.7%, con el área de Matemática el error absoluto medio de 2.809 y una precisión de 94.2% y con el área de letras el error absoluto medio de 2.779 y una precisión de 93.8%, estos resultados nos demuestras que en base a las 16 variables predictoras consideradas es posible realizar una predicción de rendimiento académico de los estudiantes, tal como se muestra en la figura 4.

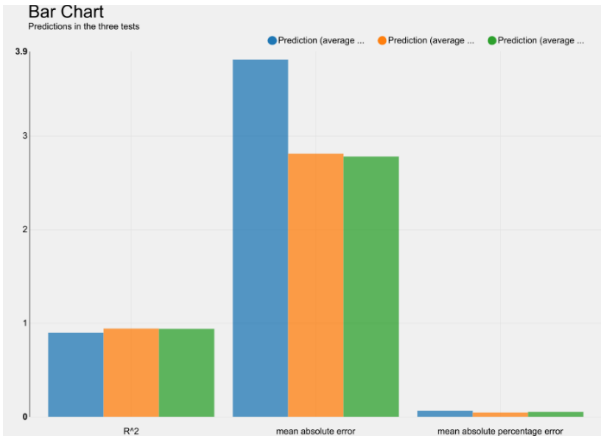

Figura 4. Precision results in the predictions made by the model

Según los resultados mostrados en el árbol de decisión en la figura 4, se realiza la predicción en base a las 16 variables y comparadas con el promedio del curso 1, se toman 900 registros de los estudiantes donde el algoritmo toma como condición la variable frecuencia de acceso a los recursos y si esta es menor o igual a 47.5 el número de instancias correctamente clasificadas es de 424 registros que corresponde al 47.1%; mientras que si la frecuencia de acceso a los recursos es mayor que 47.5 el número correcto de instancias correctamente clasificadas es 476 registros que representa un 52.9%; en general el árbol de decisión en 6 niveles considera a las 16 variables para tomar las decisiones de predicción.

Según los resultados mostrados en el árbol de decisión en la figura 5, se realiza la predicción en base a las 16 variables y comparadas con el promedio del curso 2, se toman 900 registros de los estudiantes donde el algoritmo toma como condición la variable número de lecciones culminadas y si esta es menor o igual a 14.5 el número de instancias correctamente clasificadas es de 393 registros que corresponde al 43.7%; mientras que si el número de lecciones culminadas es mayor que 14.5 el número correcto de instancias correctamente clasificadas es 507 registros que representa un 56.3%; en general el árbol de decisión en 6 niveles considera a las 16 variables para tomar las decisiones de predicción.

Según los resultados mostrados en el árbol de decisión en la figura 6, se realiza la predicción en base a las 16 variables y comparadas con el promedio del curso 3, se toman 900 registros de los estudiantes donde el algoritmo toma como condición la variable tasa de asistencia y si esta es menor o igual a 50.5 el número de instancias correctamente clasificadas es de 371 registros que corresponde al 41.2%; mientras que si la tasa de asistencia es mayor que 50.5 el número correcto de instancias correctamente clasificadas es 529 registros que representa un 58.8%; en general el árbol de decisión en 6 niveles considera a las 16 variables para tomar las decisiones de predicción.

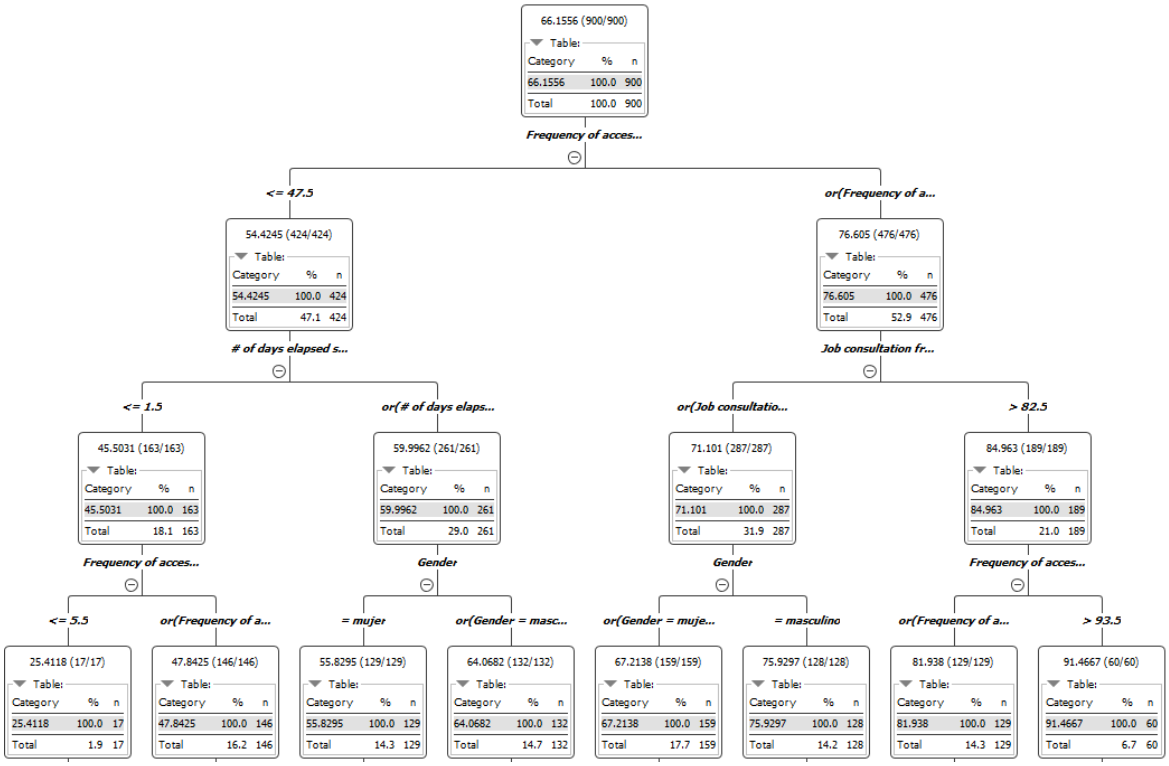

Figura 4. Diagrama de árbol de las decisiones comparadas con el promedio 1

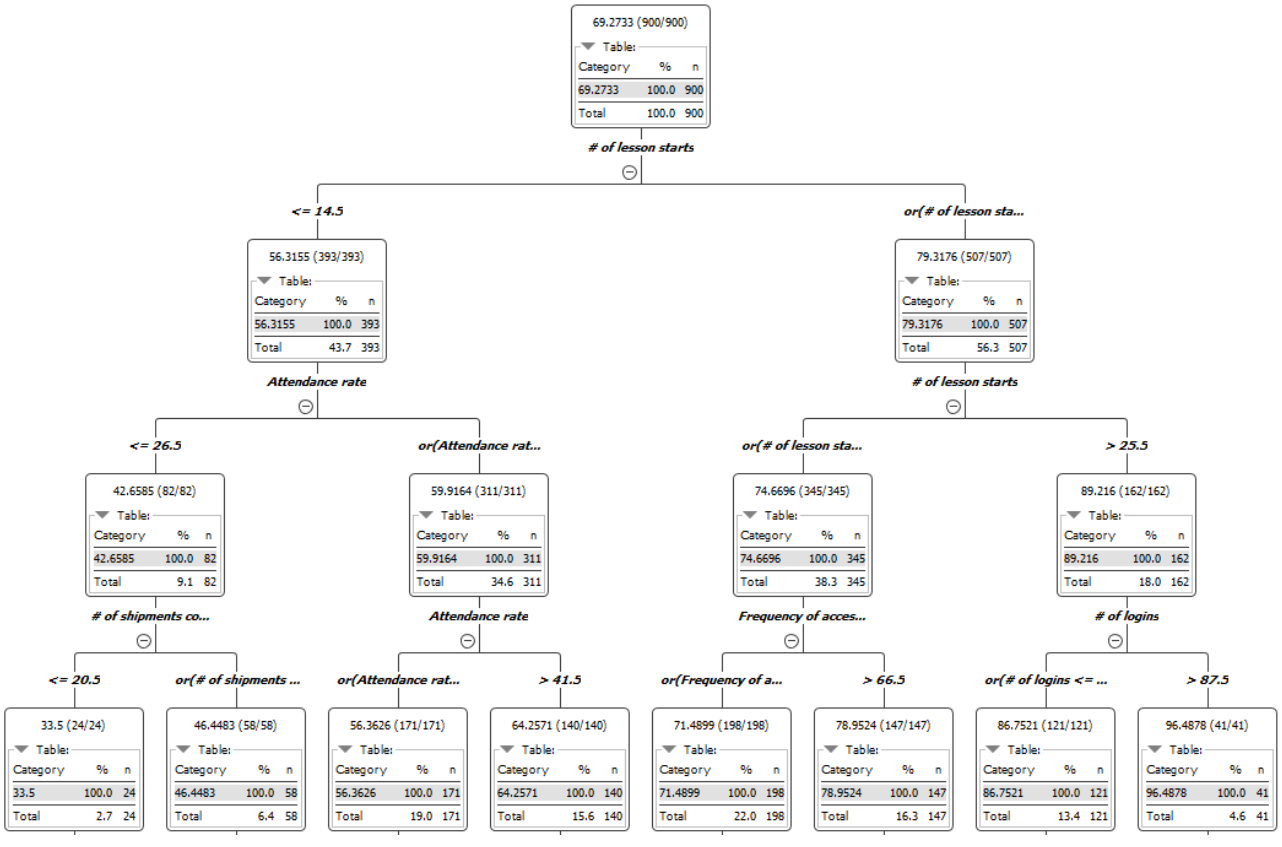

Figura 5. Diagrama de árbol de las decisiones comparadas con el promedio 2

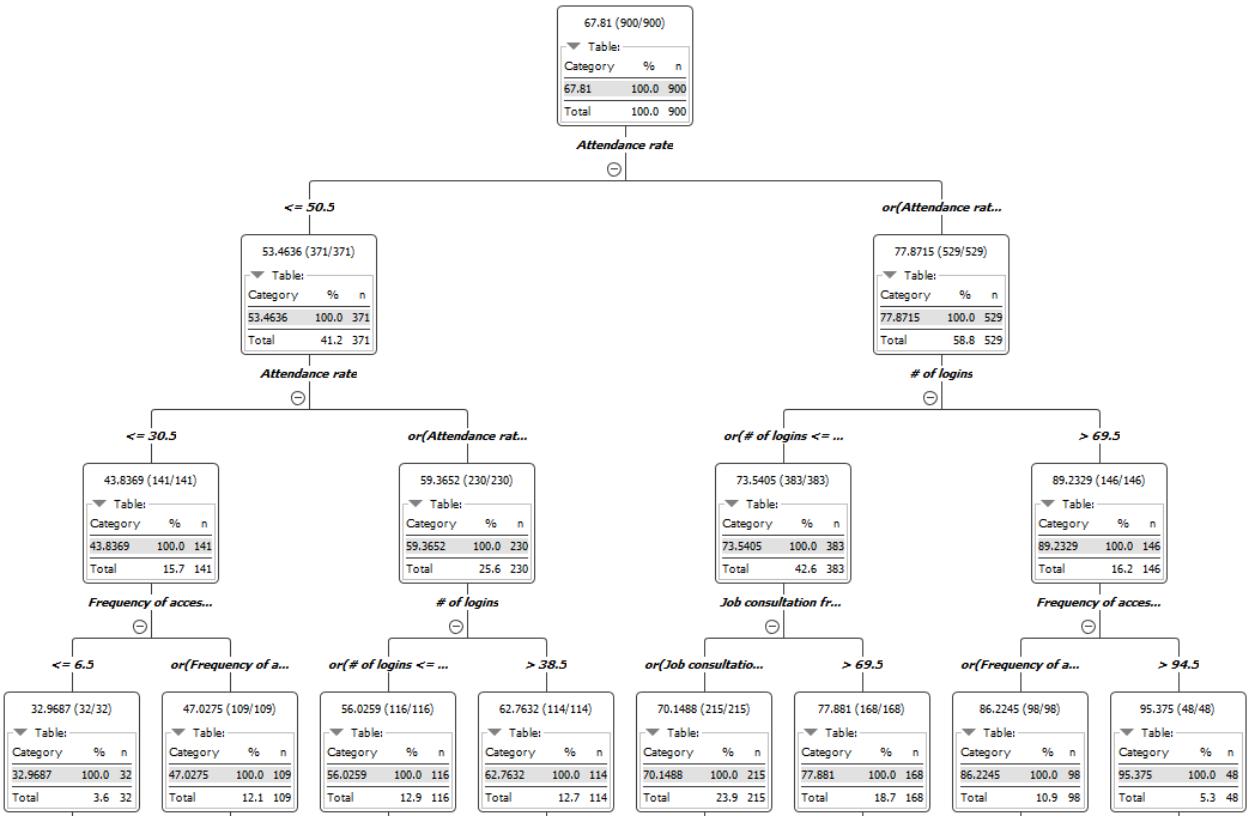

Figura 6. Diagrama de árbol de las decisiones comparadas con el promedio 3

La tabla 3 muestra 17 registros, una parte de los 1000 registros que constituyen la muestra total tomada para entrenar el modelo predictivo, los resultados son satisfactorios, donde la media ponderada del rendimiento académico en los 22 indicadores de comportamiento es de 66.71 y la media ponderada de la predicción realizada es de 65.08.

Tabla 3.

Resultado de algunas predicciones

ID	Media del primer curso	Predicción del modelo	ID	Media del segundo curso	Predicción del modelo	ID	Media del tercer curso	Predicción del modelo
16	88	89.25	10	54	52.11	4	75	73.39
53	88	84.80	14	53	56.44	6	92	94.16
61	39	19.17	17	32	24.43	34	82	83.84
64	59	47.33	19	59	61.20	46	62	64.34
65	67	65.15	20	69	63.88	49	82	79.29
72	41	39.00	23	73	74.94	51	68	69.30
111	62	60.22	24	71	76.66	67	74	74.47
137	70	65.15	28	70	67.94	71	63	58.45
139	71	70.36	34	87	86.57	74	41	45.7
148	68	67.55	35	81	76.66	78	72	74.47
160	82	82.19	46	65	63.15	79	68	70.30
175	81	81.13	48	74	71.52	81	45	45.70
176	46	47.77	89	86	82.35	83	63	65.53
180	62	70.73	97	72	71.52	106	100	94.16
183	65	63.58	99	67	67.14	114	100	99.73
201	65	70.79	122	93	90.13	124	73	73.39
210	80	82.19	123	57	57.08	128	67	67.00
Mean	66.71	65.08	Mean	68.41	67.28	Mean	72.18	72.54

Fase 5: Evaluación del modelo

El modelo implementado muestra una buena precisión en cuanto a las predicciones realizadas comparándolas con el promedio del primer curso presenta un 89.7% de aciertos, comparándolas con el promedio del segundo curso presenta un 94.2% de aciertos y comparándolos con el promedio del tercer curso presenta un 93.8% de aciertos, también se puede apreciar el menor error absoluto medio en la primera predicción de 3.813, segunda predicción de 2.809 y tercera predicción de 2.779 demostrándose que el modelo propuesto puede realizar predicciones eficientes. Ver la tabla 4.

Tabla 4.
Análisis comparativo de los resultados de las predicciones

Coeficiente de determinación	Precisión de la predicción comparada con la media del primer curso	Precisión de la predicción comparada con la media del segundo curso	Precisión de la predicción comparada con la media del tercer curso
R^2	0.897	0.942	0.938
Error medio absoluto	3.813	2.809	2.779

Fase 6: Despliegue del modelo

Para poder implantar este proyecto en el negocio real sería necesario en primer lugar tener acceso a la base de datos real de la universidad, es decir la base de datos que contiene toda la información relativa a los estudiantes de la universidad. A partir de ahí, los pasos a seguir serían los mismos que se han
seguido en la investigación desde la comprensión del negocio hasta la implantación. Si bien, cabe decir que habrá algunas fases, como la de comprensión y preparación de los datos, que en el negocio real probablemente sean más complejas y llevarán más tiempo que en este proyecto ya que se puede esperar que en la base de datos real se tengan muchos más registros y estos mismos contengan más ruido que en nuestra base de datos ficticia creada específicamente para este uso.
Como plan de supervisión y mantenimiento se podría establecer los siguientes procesos:

- Extracción y almacenamiento cuatrimestral de los datos guardando la información obtenida en formato de hoja de cálculo
- Distribución de los datos en función de los modelos de software de minería de datos a trabajar.
- Los resultados obtenidos en cada explotación de datos deberán ser llevados a formato de hoja de cálculo y generar gráficas de distintos tipos para una mejor visualización e interpretación de los resultados obtenidos en cada periodo.

Conclusiones

- La investigación implementa un modelo predictivo para analizar y comparar la predicción del rendimiento académico utilizando los datos de LMS, analizando si es posible identificar a los estudiantes en riesgo al principio de un curso y en qué medida el modelo puede utilizarse para generar intervenciones específicas. Para realizar las predicciones se implementa en el modelo el algoritmo Simple Regression Tree Learner en función de 1000 datos de registro reales tomados de un LMS, se trabaja con datos provenientes de 22 indicadores de comportamiento observados y utilizados en el LMS, los cuales al ser comparados con los promedios de rendimiento académico en tres cursos los resultados de las predicciones son satisfactorios, donde el error absoluto medio comparado con el promedio del primer curso fue de 3.813 y con una precisión de 89.7%, el error absoluto medio comparado con el promedio del segundo curso fue de 2.809 con una precisión de 94.2% y el error absoluto medio comparado con el promedio del tercer curso fue de 2.779 con una precisión de 93.8%. Estos resultados demuestran que el modelo propuesto puede ser utilizado para predecir futuros resultados del rendimiento académico de los estudiantes tomando como base un conjunto de datos provenientes de un LMS.
- Los resultados se suman a la base empírica de los hallazgos analíticos del aprendizaje y corroboran estudios previos sobre la predicción del éxito de los estudiantes, que también han mostrado resultados diferentes en correlaciones y modelos de predicción, aunque para contextos más variados que nuestra investigación.
- Un aporte muy importante del modelo propuesto es que puede ser escalable y aplicable a bases de datos grandes según los requerimientos de los usuarios.

Referencias

Baepler P, Murdoch C J. (2008). Academic Analytics and Data Mining in Higher Edu-cation. International Journal for the Scholarship of Teaching & Learning, 4(2): 267-281.

Baker, R and Yacef, K. (2009) The state of educational data mining in 2009: A review and future visions, J. Educ. Data Min., vol. 1, no. 1, pp. 3–17.

Baker, R. S. Big data and education (2nd Ed.). New York, NY: Teachers College, Co-lumbia University. (2015).

Bravo-Agapito J., Romero S.J. & Pamplona S. (2020) Early Prediction of Undergradu-ate Student's Academic Performance in Completely Online Learning: A Five-Year Study, Computers in Human Behavior, https://doi.org/10.1016/j.chb.2020.106595.

Burgos C, Campanario M L, Peña D D L. (2017). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Computers & Electrical Engineering.

Campbell, J and Oblinger, D. (2007). Academic analytics," Educause, vol. 42, pp. 40–42.

Chapman, P; Khabaza, T. and Shearer, C. CRISP-DM 1.0, Step by step data min-ing Guide. Netherlands: SPSS Inc, 2000.

De-Marcos L, García-López E, García-Cabot A. (2016). Social network analysis of a gamified e-learning course: Small-world phenomenon and network metrics as predictors of academic performance. Computers in Human Behavior, 60: 312-321.

Greller W D, Hendrik. (2012). Translating Learning into Numbers: A Generic Frame-work for Learning Analytics. Educational Technology & Society, 15(3): 42-57.

Guadilla, C. G. (2000). Investigación y Toma de Decisiones en Educación Superior. Desafíos y transformaciones de la educación en América Latina - Nueva Socie-dad 165, 97-168.

Howard, E., Meehan, M., & Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. The Internet and Higher Education, 37, 66-75.

Hu Y H, Lo C L, Shih S P. (2014). Developing early warning systems to predict stu-dents' online learning performance. Computers in Human Behavior, 36: 469-478.

Ifenthaler D, Widanapathirana C. (2014). Development and Validation of a Learning Analytics Framework: Two Case Studies Using Support Vector Machines. Tech-nology, Knowledge and Learning, 19(1): 221-240.

Iglesias-Pradas S. (2014). Can we predict success from log data in VLEs? Classifica-tion of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. Computers in Human Behavior, 31(2): 542-550.

Knowles, J. (2015). Of Needles and Haystacks: Building an Accurate Statewide Drop-out Early Warning System in Wisconsin. Journal of Educational Data Mining, 7, 3, 18-67.

Macfadyen L P, Dawson S. (2010). Mining LMS data to develop an "early warning sys-tem" for educators: A proof of concept. Computers & Education, 54(2): 588-599.

Maraza-Quispe B, Alejandro-Oviedo M, Choquehuanca-Quispe W, Caytuiro-Silva N, Herrera-Quispe J. (2020). Towards a Standardization of Learning Behavior Indica-tors in Virtual Environments. International Journal of Advanced Computer Sci-ence and Applications, Vol. 11, No. 11. DOI: 10.14569 / ISSN.2156-5570

Pedro M O Z S, Baker R S J D, Bowers A J. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. Langmuir the Acs Journal of Surfaces & Colloids, 27(11): 6897-6904.

Richardson, M; Abraham, C and Bond, R. (2012). Psychological correlates of universi-ty students' academic performance: A systematic review and meta-analysis," Psy-chol. Bull., vol. 138, no. 2, pp. 353–387.

Romero C, López M I, Luna J M (2013). Predicting students' final performance from par-ticipation in on-line discussion forums. Computers & Education, 68: 458-472.

Romero, C and Ventura, S. (2010). Educational data mining: A review of the state of the art, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 40, no. 6, pp. 601–618.

Romero, C., & Ventura, S. Educational data science in massive open online courses. WIREs: Data Mining and Knowledge Discovery, 7 (1), e1187. (2017).

Ruipérez-Valiente J A, Muñoz-Merino P J, Leony D. (2015). ALAS-KA: A learning ana-lytics extension for better understanding the learning process in the Khan Acad-emy platform. Computers in Human Behavior, 47: 139-148.

Rupp A A, Leighton J P. 16 (2014). Educational Data Mining and Learning Analytics. Springer New York, pp. 379-396.

Shum, S. B. and R. Ferguson. (2012). Social learning analytics, Educ. Technol. Soc., vol. 15, no. 3, pp. 3–26.

Siemens G. (2013). Learning Analytics. The Emergence of a Discipline. American Be-havioral Scientist, 57(10): 1380-1400.

Sinclair P M, Kable A, Levett-Jones T. (2016). The effectiveness of Internet-based e-learning on clinician

behaviour and patient outcomes: A systematic review. Inter-national Journal of Nursing Studies, 57: 70-81.

Tempelaar, D; Rienties, B and B. Giesbers (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context," Comput. Human Behavior, vol. 47, pp. 157–167.