Minería de Datos Educativos:
Predicción del rendimiento académico de los estudiantes utilizando algoritmos de aprendizaje automático

Sánchez Mosquera Joel Alexander

October 9, 2023

Abstract

This demonstrates one of the uses of Data Mining in the educational context and Machine Learning to predict the performance of university students, aided by the review of several documents that have employed various Machine Learning algorithms in identifying patterns and factors that can influence academic success. Some of the factors considered include prior grades, attendance, demographic data, and online behaviors, such as internet usage. Emphasis is placed on the importance of predictive models for identifying students who may need assistance.

Data Mining, Machine Learning, performance, students, algorithms, data, models, identification, assistance

Resumen

Se demuestra uno de los usos de la Mineria de Datos para el ámbito educativo y el Machine Learning para predecir el rendimiento de los estudiantes universitarios, ayudándose de la revisión de varios documentos que han empleado varios algoritmos de Machine Learning en la identificación de pratrones y factores que pueden influir en el éxito académico, algunos de los factores para ello son calificaciones previas, asistencia, datos demográficos y comportamientos online, como el uso de internet. Destacando la importancia de los modelos de predicción para identificar a los alumnos que pueden necesitar ayuda.

Minería de Datos, Machine Learning, rendimiento, estudiantes, algoritmos, datos, modelos, identificación, ayuda

1 Introducción

El presente trabajo es una forma resumida y traducida del documento original en inglés, en el cual se realizo un nuevo enfoque que pueda predecir cual será el

rendimiento académico de los estudiantes en un futuro, tomando datos relevantes como las calificaciones previas y datos relacionados al curso, sin tomar muy en cuenta aspectos demográficos o socioeconómicos, para buscar un desarrollo del modelo basado en algoritmos de machine learning prediciendo asi calificaciones finales, facilitando la identificación temprana de estudiantes que puedan estar en riesgo de bajar su desempeño y brindarles ayuda.

2 Literatura

En estudios acerca de **EDM** (Minería de Datos Educacional), se han analizado sistemas de **e-learning** de forma exitosa [8]. En distintas investigaciones se evidencia la clasificación de los datos educativos [5]. Al mismo tiempo, otros se enfocaron en intentar predecir el rendimiento de los jóvenes universitarios [6]. Un grupo de investigadores se centraron en dos aspectos que influyen en el desempeño de estudiantes de pregrado con el uso de metodologías de **DM** (Minería de Datos) [2]. El primero es predecir los logros académicos de los alumnos al finalizar un programa de estudio de cuatro años. El segundo es la examinación del desarrollo de los estudiantes y unirlos a los resultados predictivos. Se formaron dos grupos, los de bajas calificaciones y otros de altas calificaciones. Determinando que los docentes deben enfocarse en un pequeño número de cursos con un desempeño bueno o deficiente, ya sea para brindar apoyo a los de bajo rendimiento u orientación y oportunidades a los de alto rendimiento. Cruz-Jesus et al. (2020) utilizaron 16 características demográficas como la edad, género, asistencia, entre otros. El RF, Regresión Logística (**RL**), entre otros métodos de aprendizaje automático se usaron para predecir calificaciones de los estudiantes con una precisión aproximada del 50 al 81 %.

Varios investigadores elaboraron un modelo utilizando las características demográficas de los estudiantes y las calificaciones en trabajos durante el curso [6]. De esta manera, se predijo el rendimiento académico aplicando modelos de clasificación basados en **GBM** (Gradient Boosting Machine). Según los resultados, los mejores datos para realizar la estimación eran las calificaciones obtenidas en el año pasado y la inasistencia. Descubrieron que información como el vecindario, la escuela y la edad son posibles indicadores de éxito o fracaso. Asimismo, con datos sobre los alumnos requeridos en su inscripción y factores ambientales, se identificó a jóvenes con potencial de fracasar [7]. También se observó que este tipo de personas puede ser clasificado más específicamente usando métodos de **DM** , clasificándolos por niveles de riesgo. Concluyeron que los árboles de regresión destacan factores asociados con un rendimiento elevado, como el tamaño de la clase [10], la presión de los padres y las proporciones de género. Por otro lado, con el algoritmo **RF** (Random Forest), eran relevantes el tamaño de la escuela y el porcentaje de niñas para la precisión predictiva.

Hubo una propuesta acerca de un modelo basado en aprendizaje automático que se encarga de calcular el riesgo que los estudiantes corren en su rendimiento académico, utilizando habilidades de aprendizaje, mejorando los hábitos de estudio y la interacción académica. Consiguieron una predicción con una alta

precisión de clasificación del 85 % [1]. Los investigadores comprobaron que su modelo podría aplicarse para la identificación de los estudiantes que no son exitosos.

En otra propuesta hubo un modelo similar, el cual se basó en estrategias de aprendizaje, percepción de apoyo social, motivación, características sociodemográficas y el estado de salud para predecir el desempeño académico y deserciones [9]. Se concluyó que la variable con un mayor impacto en la **GPA** (Promedio de Puntos) son las estrategias de aprendizaje, y las de gran influencia al reconocer deserciones es la información de antecedentes.

En un modelo distinto con redes neuronales artificiales, se ayudaron de los registros de jóvenes universitarios con relación a su navegación en el **LMS** (Sistema de Gestión de Aprendizaje), con las características demográficas sobre las actividades de navegación, mostrando que poseen un impacto significativo en el rendimiento y puede ser útil para ayudar más a los alumnos [11].

Se estableció la relación que hay entre comportamientos como el uso de internet y el rendimiento académico, para predecir un futuro desempeño empleando métodos de aprendizaje automático [12]. Dicho modelo fue capaz de calcular con alta precisión, y en sus resultados sugería que características como la frecuencia de conexión a internet de los estudiantes estaban positivamente correlacionadas. Al mismo tiempo, el volumen de tráfico de internet afecta de manera positiva a ambas, lo que está relacionado con el desempeño de los estudiantes, llevando a la conclusión de que el internet juega un rol importante.

Se realizó el intento de verificar si los registros en el sistema de aprendizaje lograban predecir el rendimiento por sí solos [3]. De esta manera, se evidenció que el modelo que se basó en el comportamiento estimó con un 75 % de precisión quiénes deben repetir el curso. A su vez, afirman que dicho modelo es capaz de identificar y apoyar a jóvenes con posibilidad de ser infructuosos en semestres posteriores, también sus calificaciones y diseñaron una herramienta para los estudiantes que eran propensos a fallar [4]. Debido a ello, encontraron que el número de estudiantes no exitosos ha disminuido.

2.1 Tabla 1

Análisis comparativo

References	Variables	Objectives	Level	Dataset	Algorithms	Accuracy	
						Min	Max
Asif et al. (2017)	The marks for all the courses that are taught in the four years of the degree programme	Predicting students' performance	Undergraduate students	210	DT, 1-NN, NB, NN, RF	NN (62.50%)	NB (83.65%)
Cruz-Jesus et al. (2020)	Year of the study cycle, gender, age, number of enrolled years in high school, scholarship, internet access, class size, school size, economic level, population density, number of unit courses attended	Predicting students' performance	High schools students	110627	ANN, DT, ET, RF, SVM, kNN, LR	LR (81.1%)	SVM (51.2%)
Fernandes et al. (2019)	Class with persons with special needs, Classroom usage environment, Gender, age (mean), Student benefit, city, neighbourhood, Student with special needs, Grade (mean), Absence (mean)	Predict academic outcomes of student performance	High schools students	Dataset1:19000 Dataset2:19834	Gradient Boosting Machine	89.5%	91.9%
Hoffait and Schyns (2017)	Gender, Nationality, Studies, Prior schooling, math, scholarship, success	Predicting students at high risk of failure	secondary school students	2244	RF, LR, ANN	ANN (70.4%)	RF (90%)
Rebai et al. (2020)	Socioeconomic status, school type, school location, competition, teacher characteristic (experience, salary), class size, school size, gender, parental education, political context, parental pressure	to identify the key factors that impact school's academic performance and to explore their relationships	Secondary schools	105 schools	RT, RF		

References	Variables	Objectives	Level	Dataset	Algorithms	Accuracy	
						Min	Max
Ahmad and Shahzadi (2018)	Previous degree marks, Home environment, Study habits Learning skills, Hardworking and Academic interaction	Identification of students in the risk group	Undergraduate students	300	MPNN		95%
Musso et al., (2020)	Learning strategies, coping strategies, cognitive factors, social support, background, self-concept, self-satisfaction, use of IT and reading	Grade point average, academic retention, and degree completion	Undergraduate students	655	ANN	60.5%	80.7%
Waheed et al., (2020)	Students' demographics, clickstream events	Pass-fail, withdrawn-pass, distinction-fail, distinction-pass	Undergraduate students	32593	ANN, SVM, LR	84%	93%
Xu et al. (2019)	Internet usage behaviours comprise online time, internet connection frequency, internet traffic volume, and online time	Predicting students' performance	Undergraduate students	4000	DT, NN, SVM	71%	76%
Bernacki et al. (2020)	Log records in the learning management system	Predict achievement	Undergradeate students	337	LR, NB, J-48 DT, J-Rip DT	J-48 (53.71%)	LR (67.36%)
Burgos et al. (2018)	Historical student course grade data	Drop out of a course	Undergradeate students	100	SVM, FFNN, PESFAM, LOGIT_Act	SVM (62.50)	LOGIT_Act(97.13%)

3 Método

En este apartado se detalla el **dataset**, las técnicas de preprocesamiento y los algoritmos de aprendizaje de máquina utilizados en el estudio.

3.1 Conjunto de datos

Las instituciones educativas generalmente almacenan los datos disponibles en formato electrónico. Esta información se encontraba en una BDD (Base de Datos) para su posterior procesamiento. Los datos fueron tomados del **SIS** (Sistema de Información del Estudiante), donde suelen registrarse los datos de los estudiantes de la Universidad Estatal de Turquía. Se seleccionaron 1854 registros de estudiantes pertenecientes al curso de Lengua Turca-I en el otoño del 2019-2020, incluyendo las calificaciones de los exámenes parciales, exámenes finales, la facultad y el departamento de los estudiantes, formando así un conjunto de datos.

3.2 Preprocesamiento de datos

Las calificaciones de los exámenes parciales y finales se registraron en un intervalo de 0 a 100, lo que permitió que el sistema realizara cálculos tomando en cuenta un 40 % de la calificación del examen parcial y un 60 % de las evaluaciones finales. Aquellos con un puntaje menor a 60 se consideraron como no exitosos, mientras que aquellos que superaron dicho puntaje se consideraron exitosos.

3.3 Selección de características

Se llevó a cabo la selección de información relevante dentro del conjunto de datos con el objetivo de reducir la complejidad y aumentar la eficiencia computacional de los algoritmos de **ML**. Esto facilita la interpretación del modelo y ayuda a prevenir el sobreajuste.

3.4 Modelo de MD y Algoritmos

Se utilizaron una amplia variedad de algoritmos de aprendizaje de máquina, algunos como **RF**, **NN** (Redes Neuronales), **LR**, **SVM** (Máquinas de Vectores de Soporte), **NB** (Naive Bayes) y **kNN** (k Vecinos más Cercanos), para predecir el desempeño de los estudiantes, para la evaluación de la precisión se empleó la validación cruzada de diez pliegues.

3.5 Propósito de la MD

Se tiene dos propósitos principales al usar este método, el primero para hacer predicciones a través de un análisis de la **BDD** convirtiéndose en el modelo predictivo, mismos que son creados por el uso de resultados ya conocidos para conocer valores que se desconocen, por otro lado, en el segundo se especifican los comportamientos, este es el modelo descriptivo que permite la identificación de patrones en datos que se poseen.

3.6 Medición del Rendimiento del Modelo

Para calcular la precisión del modelo se usaron indicadores dentro de la matriz de confusión, demostrando que no existe un clasificador único que funcione completamente para todos los resultados en la predicción, viéndose en la necesidad de investigar clasificadores más adecuados para los datos analizados.

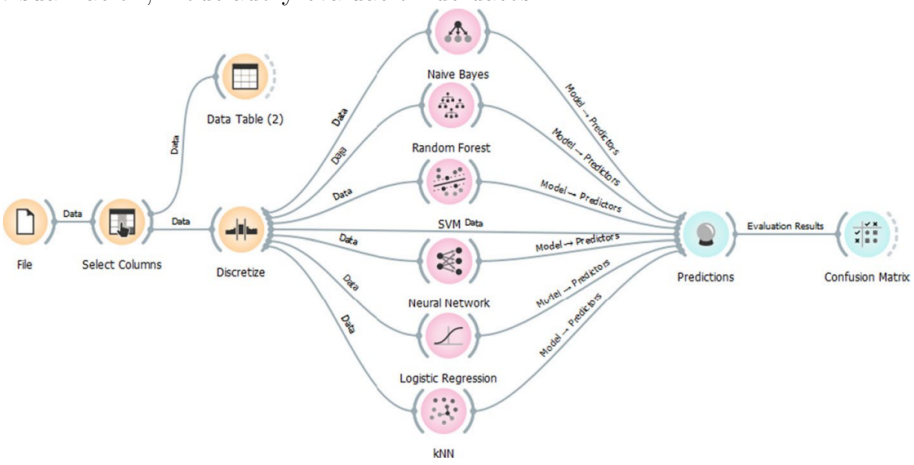
3.7 Enfoque de Algoritmos

Las técnicas de estadísticas son utilizadas para crear modelos que sean capaces de predecir exitosamente los valores de salida, usando como base los datos de entrada útiles. Por el contrario, los métodos de **ML** forman un modelo en concordancia de los valores de entrada con los valores objetivos esperados al momento de plantear un problema de optimización supervisada.

En otras palabras, para este estudio se utilizaron una diversidad de técnicas de **MD** y **ML** prediciendo el desempeño de los estudiantes con los datos de examenes y ciertas características personales de cada estudiante, los algoritmos son evaluados y se plantea hallar un enfoque más efectivo para este conjunto especifico de datos.

4 Experimentos y Resultados

Se utilizó **Orange** para la fase experimental, ya que esta herramienta de programación es adecuada para **MD** y está basada en componentes. Además, es potente y fácil de usar tanto para expertos en ciencia de datos como para principiantes en el área. Se llevó a cabo un análisis de datos mediante la combinación de widgets en flujos de trabajo, abordando tareas de recuperación, procesamiento, visualización, modelado y evaluación de datos.



Dentro del conjunto de datos se incluyó información adicional, como las calificaciones de los exámenes parciales y finales, así como la facultad y el departamento. Cada medida contiene datos relacionados con los estudiantes, y las

variables de cada puntaje se categorizaron utilizando un modelo de discretización.

4.1 Evaluación del Rendimiento del Modelo

Para conocer el rendimiento, se evaluó mediante diversas métricas, como la matriz de confusión, precisión, recuperación, puntuación F1, área bajo la curva **ROC** (Receiver Operating Characteristic) y el valor **AUC** (Área Bajo la Curva). La matriz de confusión permitió observar la situación actual del conjunto de datos y la cantidad de predicciones correctas e incorrectas realizadas por el modelo. La precisión de clasificación (**CA**) se refiere a la proporción entre predicciones correctas y el número total de instancias. También se calcularon métricas como la precisión, recuperación y puntuación F1, que ofrecen una visión completa del desempeño del modelo. La curva **ROC** y el valor **AUC** se utilizaron para evaluar el rendimiento de la clasificación. El valor de **AUC** utilizada con frecuencia para evaluar el rendimiento de los algoritmos de **ML** y mide la capacidad del modelo para predecir con precisión. En resumen, a través del estudio se consiguió alcanzar un alto nivel de precisión en la clasificación de datos, lo que demuestra la eficacia de los modelos de **ML** utilizados.

5 Discusión y Conclusión

Gracias al nuevo modelo de **ML**, se predijeron con éxito las calificaciones de exámenes finales de los jóvenes universitarios utilizando sus notas parciales como entrada. Se evaluaron diversos algoritmos de **ML**, como **RF**, **kNN**, **NN**, **SVM**, **LR** y **NB**, posteriormente se compararon sus rendimientos. Se evidenció que se logra una precisión de clasificación del 70 al 75 %, lo que sugiere que los puntajes de los exámenes son un indicador importante para la predicción deseada. Tanto **RF** como **NN** y **SVM** ofrecen una mejor precisión. En este enfoque de predicción basado en datos simples, se podría utilizar para mejorar el desempeño académico de los alumnos. Además, se destaca la importancia de utilizar datos que permitan identificar la variedad de niveles de motivación académica en los estudiantes, para poder realizar una intervención temprana y brindarles oportunidades para mejorar. Incluso se determina que para investigaciones futuras se pueden incluir parámetros y algoritmos con el fin de mejorar la precisión de las predicciones y optimizar el entorno educativo.

Referencias

[1] Z. Ahmad and E. Shahzadi. Prediction of students' academic performance using artificial neural network. *Bulletin of Education and Research*, 40(3):157–164, 2018.

[2] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider. Analyzing undergradua-

te students' performance using educational data mining. *Computers and Education*, 113:177–194, 2017.

[3] M. L. Bernacki, M. M. Chavez, and P. M. Uesbeck. Predicting achievement and providing support before stem majors begin to fail. *Computers Education*, 158:103999, 2020.

[4] C. Burgos, M. L. Campanario, D. De, J. A. Lara, D. Lizcano, and M. A. Martínez. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering*, 66:541–556, 2018.

[5] B. Chakraborty, K. Chakma, and A. Mukherjee. A density-based clustering algorithm and experiments on student dataset with noises using rough set theory. In *Proceedings of 2nd IEEE international conference on engineering and technology, ICETECH 2016*, pages 431–436, 2016.

[6] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven. Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil. *Journal of Business Research*, 94:335–343, 2019.

[7] A. Hoffait and M. Schyns. Early detection of university students with potential difficulties. *Decision Support Systems*, 101:1–11, 2017.

[8] J. A. Lara, D. Lizcano, M. A. Martínez, J. Pazos, and T. Riera. A system for knowledge discovery in e-learning environments within the european higher education area—application to student data from open university of madrid, udima. *Computers and Education*, 72:23–36, 2014.

[9] M. F. Musso, C. F. R. Hernández, and E. C. Cascallar. Predicting key educational outcomes in academic trajectories: A machine-learning approach. *Higher Education*, 80(5):875–894, 2020.

[10] S. Rebai, F. Ben Yahia, and H. Essid. A graphically based machine learning approach to predict secondary schools performance in tunisia. *Socio-Economic Planning Sciences*, 70:100724, 2020.

[11] H. Waheed, S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz. Predicting academic performance of students from vle big data using deep learning models. *Computers in Human Behavior*, 104:106189, 2020.

[12] X. Xu, J. Wang, H. Peng, and R. Wu. Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, 98:166–173, January 2019.