Predicción de desempeño de estudiantes usando Naive Bayes

Edú David Moreno Ccama

Lima, Perú

[U201517035@upc.edu.pe](mailto:U201517035@upc.edu.pe)

Renato Mercado Luna

Lima, Perú

[U201510738@upc.edu.pe](mailto:U201510738@upc.edu.pe)

*Abstract*— Actualmente con el aumento de centros educativos junto con el número de asistentes y el incremento de la cantidad de centros de educación en el negocio, por lo cual las empresas que ofrecen este servicio requieren ofrecer un mejor servicio que sus competidoras y mejorar la experiencia de sus usuarios, lo cual resulta complicado ya que no se cuentan con un medio para sustraer información de las necesidades del usuario de forma rápida. En vista ello este texto académico propone la implementación de un algoritmo de inteligencia artificial usando naive bayes para determinar en base a patrones sobre el comportamiento de estudiantes su desempeño, para así brindar a las instituciones educativas información sobre el desempeño de sus alumnos respecto a sus comportamientos. [PASOS][TECNIA USADA][RESAULTADOS]300 PALABRAS

1. INTRODUCCIÓN

En la actualidad, con el incremento del número de centros educativos y los esfuerzos por parte del gobierno para mejorar la educación el Perú mantiene un nivel educativo bajo respecto a otros países ocupando el puesto 127 de 138.[1] Asimismo, el nivel de abandono escolar prematuro sigue teniendo un porcentaje alto de 14%, de igual forma se tiene un alto grado de alumnos que repiten cursos por motivos familiares, económicos, falta de interés o de salud.[2] Además, el país sigue entre los países que realizan una menor inversión en el ámbito educativo lo cual se refleja en su pobre desempeño en evaluaciones de educación internacionales como PISA donde ocupa los puesto 64,62,63 en la área de ciencia, matemática y lectura de un total de 72 países.[3] Para casos como estos la inteligencia artificial ofrece grandes ventajas para adaptar los contenidos y metodología de enseñanza al grupo de alumnos que se esté enseñando lo cual reduciría la desmotivación de los alumnos y permitiría una mejor atención de sus necesidades. De este modo las materias se pueden ir adaptando en función a los intereses y resultados de los alumnos. [4]

Para poder alcanzar el objetivo descrito en el párrafo anterior, nosotros proponemos el diseño de un sistema para la detección del rendimiento de estudiantes tomando como base sus comportamientos. Para lo cual haremos uso de un algoritmo de inteligencia artificial basado en el teorema de bayes para la predicción del desempeño de un estudiante basándose en múltiples variables sobre el comportamiento de los estudiantes. Asimismo, se hará uso de los datos proporcionados por la universidad de california en Irvine(UCI) a través de un repositorio para machine learning de la universidad [5].

Con la aplicación de esta propuesta se espera obtener como resultados la posible note de un alumno en base a las variables establecidas lo cual ayuda a poder establecer estrategias reducir el nivel de deserción y de alumnos desaprobados. En los siguientes párrafos, se explicará el estado del arte basado en informes acerca de trabajos relacionados.

1. ESTADO DEL ARTE

Se realiza una exhaustiva búsqueda a documentos de investigación científica relacionados la minería de datos en temas educativos, presentados a continuación.

1. *Prerocesamiento de la información para el uso algoritmos de Data Mining y aplicaciones de WEKA*

Data Mining es el proceso de descubrir conocimiento en bases de datos, ayuda a encontrar patrones escondidos en un largo conjunto de datos. Hoy en día, medir el rendimiento de los estudiantes es dificultoso, ya que estos dependen de diversos factores como personal, socioeconómico, psicológico y otras variables del ambiente. Educational Data Mining es una disciplina en crecimiento que ayuda a desarrollar métodos que exploren tipos de datos únicos de una base de datos educacional para predecir el rendimiento académico de los estudiantes.

En el primer estudio se consideran estudiantes que cursan el grado de Master of Computer Application (MCA) de Pune University [6]. En primer lugar, se usa la técnica de redes neuronales para seleccionar los atributos de un conjunto y con base en la precisión de los datos correctamente clasificados, se identifican atributos importantes y se generan reglas [6]. En segundo lugar, se usa la herramienta de software libre WEKA, implementado en lenguaje java, este software orece una colección de machine learning y algoritmos de data mining para pre-procesamiento de datos, clasificación, regresión, agrupamiento, y reglas asociativa. Ahora, se comparan los resultados entre la precisión obtenida por las redes neuronales en todos los atributos y la precisión obtenida aplicando la técnica de redes neuronales en atributos seleccionados [7]. En tercer lugar, se aplica la Association Rule Mining para encontrar los conjuntos de ítems más frecuentes y generar reglas de asociaciones fuertes para los conjuntos de ítems frecuentes, también, se hace uso del algoritmo Apriori que está basado en el hecho de que el algoritmo usa conocimiento anterior de propiedades de conjunto de elementos [6].

Para el segundo estudio se tomó datos de colegios rurales y urbanos del distrito de Betul en India [7]. Entre sus objetivos se tiene: Identificar las variables predictivas más influentes en el rendimiento académico de estudiante de la escuela primaria, encontrar el mejor árbol de decisiones, algoritmo de Naive Bayes, algoritmos Zero R Classification Algorithms en los datos de los estudiantes, predecir el resultado de rendimientos en el estudio y aplicar técnicas de Data Mining usando datos de estudiantes de áreas rurales y urbanas de nivel primario en el distrito de Betul en la India [7].

Antes de correr las pruebas en los datos es necesario limpiar y preparar los datos para usarlos en WEKA workbench, La selección de atributos implica la búsqueda a través de todas las posibles combinaciones de atributos en la data para encontrar los mejores subconjuntos de atributos para la predicción [7]. Para realizar esto se crean dos objetos, un atributo evaluados y un método de búsqueda. El evaluador determina que método se usa para asignar un valor a cada subconjunto de atributos, mientras que el método de búsqueda determina que estilo de búsqueda es usado [7]. De este modo, haciendo uso de árbol de deciciones, Naive Bayes y Zero R, se puede predecir el rendimiento de un estudiante de nivel primario. La precisión de este modelo es de 85.53% que significa que el modelo es exitoso prediciendo el grado final de estudiantes de alto nivel fuera de los 600 que fue clasificados exitosamente. El profesor, estudiantes y sus padres puede mejorar el resultado de un estudiante que probablemente aprobará en un nivel bajo a través de un asesoramiento adecuado [7].

Para el tercer estudio se busca crear un modelo que ayude a los estudiantes a poder encontrar un campo de especialización en ciencias de la computación en la universidad de Suan Sunandha Rajabhat en Tailandia, para lo cual se hace uso de minado de datos para encontrar patrones y relaciones de información como información de registro, información del curso entre otros [8].

Para llevar a cabo el proyecto se extrajo data del programa de ciencias de la computación de la universidad entre los periodos de 2006 a 2012, el set de datos estaba compuesto por información de 312 récords personales de estudiantes [8]. Asimismo, los estudiantes de ciencias de la computación habían recibidos preguntas sobre sus habilidades en base de datos, ingeniería de software, multimedia y redes, y comunicaciones [8]. Los datos fueron pre procesados y transformados para ser adecuados al formato necesario y eliminación de información que no era requerida, se particiono el valor de los campos en 5 niveles muy pobre, pobre, medio, bueno y muy bueno. Luego de preparar la información fue analizada por WEKA una colección de algoritmos de machine learning para analizar sets de datos par atareas de minado de datos [8]. Los algoritmos utilizados fueron redes de bayes y J48 una variante de los árboles de decisiones, para medir el resultado se usó el método de validación cruzada [8]. Los resultados obtenidos fueron que las redes de bayes fueron más precisos que el J48 con un nivel de 92.13% de precisión [8].

En conclusión, de los tres estudios se puede concluir que la información inicial tiene que pasar por un proceso de preprocesamiento, en el caso del primer y segundo estudio se usó redes neuronales para el preproceso de datos. Asimismo, el tres estudios se hace uso de algoritmos de Data Mining como árbol de decisiones y el algoritmo de Nave Bayes, que con la herramienta WEKA resulta fácil de implementar. Por ello, en los tres estudios WEKA forma parte importante de la investigación, siendo un software importante.

1. *Uso de algoritmos inteligencia artificial en el minado de datos*

Para poder llevar a cabo el minado de datos como es el caso de la búsqueda de patrones se requiere de un profundo análisis de los datos. Para esto es necesario el uso de algoritmos de inteligencia artificial, ya que permiten un rápido y profundo análisis de los datos. En los siguientes se explican los principales métodos utilizados para minar datos en campos de la educación.

En el estudio presente se toman los datos del segundo año de bachillerato de tecnología de la base de datos de la United College of Engineering and Research Naini Allahabad y se usó un cuestionario para recoger datos reales que describan la relación entre el comportamiento de aprendizaje de los estudiantes y su rendimiento académico. [9]

Como la data ya se encuentra organizada no será necesario un preproceso, se realizarán los procesos de asociación, clasificación, y agrupamiento seguido de evaluación de resultado. [9]

Asociación: Analizar los datos para identificar la ocurrencia consolidada de eventos y utiliza criterios de apoyo y confianza. En el caso presente se aplica al comportamiento del estudiante y busca relaciones entre ítems. [9]

Clasificación: Se usa para clasificar cada ítem de un conjunto de datos en uno de los conjuntos de clases o grupos predefinidos. Se usan un conjunto de reglas IF-THEN para la clasificación. Las reglas de clasificación son reglas de predicción para describir una situación futura. [9]

Agrupamiento: Dividir los datos en grupo de objetos similares. El algoritmo K-means es uno de los más conocidos de los algoritmos de agrupamiento, está basado en una simple idea: Dado un conjunto de grupos iniciales, asigne cada punto a uno de ellos y luego cada centro del grupo se reemplaza por el punto medo del grupo respectivos, esto realiza hasta lograr la convergencia. [9]

En conclusión: Se aplican técnicas de Data mining para descubrir el conocimiento, reglas de asociación, reglas de clasificación para predecir el desempeño de los estudiantes, asimismo, se agrupa a los estudiantes en grupos usando k-means.

Para el segundo estudio, se presentan las técnicas más usadas para el minado de datos para la predicción del desempeño de estudiantes. Primero el estudio presenta el uso de árboles de decisión, los cuales cuenta con seis pasos los cuales son: primero, se establece un set de datos inicial, se convierte los valores de los datos a una forma discreta, se incorpora los valores del set de data a un árbol de nodos, si el árbol es homogéneo se terminar el proceso, de lo contrario se selecciona un individuo para hacer más homogéneo el árbol y así hasta que el árbol se acerca a la homogeneidad.[10] Luego con los datos preparados se hace uso del algoritmo J48 y se obtuvo los datos identificado a los estudiantes que podrían tener un mal desempeño y a los que tendrían un buen desempeño. Asimismo, se presento el uso de la múltiple regresión para predicción de promedios, calculando un promedio en base a los datos con los que se cuenta en la base de datos, Finalmente se presentó la clasificación de estudiantes para puestos de trabajo. [10]

En conclusión, las técnicas de regresión y arboles de decisión pueden predecir de forma acertada le desempeño de alumnos y el agrupamiento de alumnos resulta útil cunado se busca crear grupos de alumnos basándose en su comportamiento y desempeño. [10]

Para el tercer estudio, se utiliza la data de los estudiantes y los métodos de filtro incluyen técnicas para valorar los atributos de evaluación que dependen de la heurística en función de las características generales de los datos [11]. Para la minería de datos, los métodos de filtro son una solución más práctica por ciertas razones: la elección y evaluación de atributos es más corta, la independencia del algoritmo de estudio de máquina permite el uso en combinación con la técnica de modelado de data [11].

Los algoritmos usados fueron los siguientes:

* Algoritmos Rules-based: JRip, NNge, PART y Ridor.
* Algoritmos Trees-based: ADTree, J48, LAD Tree y RandomForest.
* Algoritmos Functions-based: Logistic, MultilayerPerceptron, RBFNetwork y SMO. [11]

Todos los algoritmos obtuvieron una precisión sólida con valores muy similares (65% - 75%). Los resultados indican que la mayoría de los algoritmos mejoran cuando se usan solo once y nueve atributos [11]. Los resultados más altos los obtiene J48 cuando usa solo nueve atributos y Random Forest cuando usa todos los atributos [11].

El desempeño de los estudiantes en la actualidad se determina ene base a las notas de los estudiantes dividas entra las obtenidas en clases y los exámenes finales del semestre, cada estuante debe obtener unas notas como mínimo para poder pasar los cursos. Para el caso presentado primero se preparó la data en este caso la data se obtuvo de la universidad de Purvanchal de la India desde el 2007 hasta el 2010, la data proveniente de múltiples tablas fue fusionada en una sola tabla, en el siguiente paso de selección de data, se eliminó la información que no fuese necesaria para el minado, en el tercer paso se seleccionó el árbol de decisión como técnica a utilizar, en el cuarto paso se determinó usar el algoritmo ID3 diseñado por Ross Quinlan, el cual realiza una búsqueda voraz de arriba hacían abajo en el árbol, para poder entortar el mejor nodo se estableció como métrica la información obtenida calculado usando la homogeneidad de los atributos de un nodo, en el quinto paso se midieron el nivel de homogeneidad de las tablas basando en medidores como el índice Gini o la entropía, sexto se divide entre los atributos de un nodo para seleccionar el mejor usando el criterio de la información granada calculado usando el nivel de homogeneidad, esto se repite en cada nodo hasta que cada atributo del árbol ha pasado por este proceso. Finalmente, después de aplicar los pasos se logró clasificar a los alumnos según su nivel de desempeño.

[Conclusión]

En síntesis, los tres estudios presentan múltiples formas de abordar el minado de datos en el caso de alumnos y centros educativos. Sin embargo, resaltan como mentados eficaces el uso de árboles de decisión por su alto nivel de eficacia en los resultados que ofrece y de naive bayes por la rapidez con la que puede realizar el minado.

1. ESTUDIO AUTODIRIGIDO

* Aprendizaje profundo

El aprendizaje profundo o también conocido como *Deep Learning* es un campo de la inteligencia artificial cuyo objetivo es el estudio y la construcción de sistemas basados en la emulación del aprendizaje que los seres humanos realizan, para lo cual hace uso de una gran variedad de algoritmos usualmente utilizado en redes neuronales. El aprendizaje profundo forma parte del campo más grande machine learning por lo cual esta basado en algoritmos de múltiples niveles no lineales de representación y abstracción, donde cada nuevo nivel usa como input el output del nivel anterior. Los sistemas que hacen uso de aprendizaje profundo deben de ser entrenados par poder llevar sus objetivos a cabo mediante la repetición de múltiple pruebas con los objetos para los que fueron diseñados donde hacen uso de las múltiples capas para poder reconocer los patrones que caracterizan el objeto con el que son probados, para así poder llegar a generalizar su funcionamiento y de manera automática poder identificar casos no visto previamente sin necesidad de reglas predefinidas, ejemplo de esto es el reconocimiento de imágenes.[12]

El aprendizaje profundo es ampliamente utilizado en redes neuronales, sistemas de reconocimiento de voz, procesamiento del lenguaje natural, bioinformática entre otros.

* Web mining

Con el gran aumento de información disponible en la web, la world wide web se ha convertido en un área muy conveniente para la investigación de minería de datos. La internet constituye un medio interactivo y muy popular para compartir información. La red es inmensa, diversa y por tanto esto incrementa su escalabilidad.

Web mining es el concepto que agrupa todas las técnicas, métodos y algoritmos utilizados para extraer información y conocimiento desde los datos originados en la Web [13] Parte de estas técnicas apuntan a analizar el comportamiento de los usuarios. De esta manera podemos definir web mining en tres diferentes categorías según su aplicación.

* Web usage mining:

Son los registros de web logs, que contienen toda la interacción entre los usuarios y la web [13].

* Web content mining:

Son los objetos que aparecen dentro de una página web, por ejemplo, las imágenes, textos libres, sonidos, etc [13].

* Web structure mining:

Se refiere a la estructura de hipervínculos presentes en una página [13].

Esta técnica tiene muchas ventajas para las empresas y marcas, incluidas las instituciones del estado, que la aplican para detectar amenazas y lucha contra el terrorismo.

* Análisis de sentimientos

Existe una gran cantidad de textos en los que el contenido subjetivo es lo más relevante y cuyo procesamiento no debería limitarse únicamente a las técnicas de clasificación de documentos. Frente a esta necesidad de clasificar la orientación u opinión que se expresa en los documentos surge el área de análisis de sentimientos.

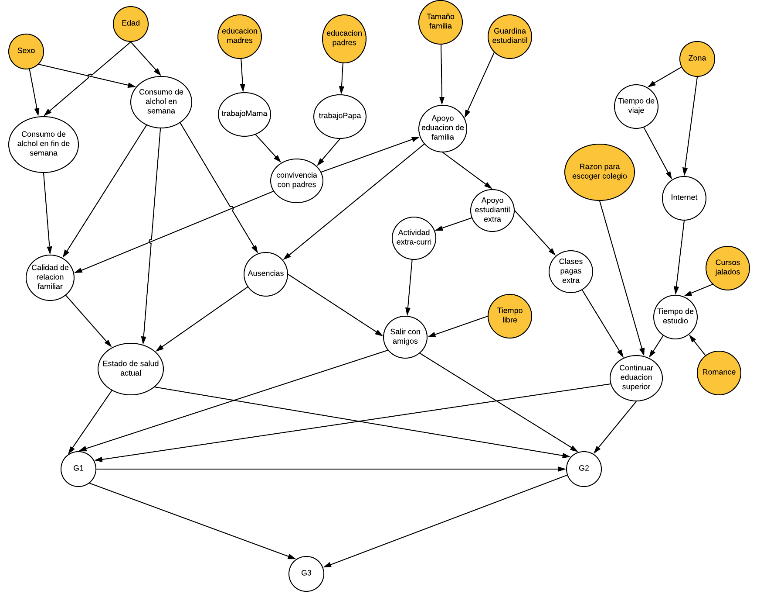
El análisis de sentimiento se refiere a los diferentes métodos de lingüística computacional que ayudan a identificar y extraer información subjetiva del contenido existente en el mundo digital [14]. Gracias al análisis de sentimiento se puede extraer un valor tangible y directo, como pude ser determinar si un texto extraído de la red internet contiene algo negativo o positivo [14].

Gracias a toda la información que se recopila de esta forma se le denomina minería de opinión y gracias a ella; las empresas tienen una inmediata disponibilidad de la información deseada [14]. Asimismo, gracias a este análisis podrimos recopilar información suficiente para que conocer lo qué piensa o qué opinan otros usuarios en la internet [14].

Gracias a este proceso se consigue obtener datos de calidad, se evita tener multitud de datos que carecen de valor para la toma de decisiones, tomar decisiones en tiempo real, desarrollar mejores estrategias empresariales y facilitar la gestión de la reputación online y ayuda a saber que acciones llevar a cabo en el plan estratégico de marketing online [14].

1. ARQUITECTURA DE LA SOLUCION

Para la elaboración se plantea la creación de un aplicativo con interfaz gráfica para determinar la probabilidad de que un alumno pase un curso basado en unos pocos datos como input. Para esto, se plantea el uso del teorema de Bayes para poder detener la probabilidad del alumno. Primero se elaboró una red bayesiana tomando como nodos las variables presentes en el set de datos y las relaciones entre los nodos fueron extraídas en base al juicio de expertos sobre la relación de como las variables afectan al promedio de un alumno. La red bayesiana será usada para según el input que introduzca el usuario realizar las respectivas asignaciones de los valores que le corresponderían de cada nodo para tener como resultado final si el desempeño del alumno a lo largo del curso daría una nota probatoria. En la siguiente grafica se encuentra representada en la siguiente gráfica, cuenta con 31 nodos, de los cuales se identificaron 11 de ellos como input del usuario.



Una vez definida la red bayesiana, se procedió con el cálculo de las probabilidades asociadas a cada nodo, para esto se realizo un filtrado de los datos para calcular sus porcentajes y haciendo uso del teorema de Bayes y la probabilidad condicional.

[IMAGEN PROABABILIDA CONDICONAL ¿?]

Tal como se mención anteriormente este algoritmo usa el teorema de bayes para realizar sus inferencias y lo realiza en base al input del usuario. A continuación, se desarrollará una inferencia para ilustrar el funcionamiento del algoritmo.

[EJEMPLO]

[EXPLICACION EJEMPLO]

Una vez agregados las frecuencias asociadas a cada nodo de la red se continuo con la implementación de la solución.

1. Implementación

Para la implementación de la solución se hizo uso de R studio tanto como para la creación de la red bayesiana y la resolución de consultas como para la creación de la interfaz gráfica. La interfaz pide la introducción de las 11 variables definidas como inputs en el punto anterior.

* Sexo
* Edad
* Educación de su madre
* Educación de su padre
* Tamaño de familia
* Guardian estudiantil
* Razon para escoger colegio
* Tiempo libre
* Zona
* Cursos jalados
* Romance

[Imagen formulario]

Estas variables serán ingresadas en la interfaz entre un conjunto de opciones predefinidas. Una vez ingresados los datos, son procesados a través de la red bayesiana y se determina un resultado.

1. Evaluación de resultados

[Comparación de fidelidad con papers]

Conclusiones

En síntesis, …

1. REFERENCIAS

[1] Perú sigue en la cola respecto a calidad educativa a nivel mundial [online]. Diario Gestión, 2016 Disponible en: <https://gestion.pe/economia/peru-sigue-cola-respecto-calidad-educativa-nivel-mundial-148579>

[2] Así está el Perú 2016: la deserción escolar y la calidad educativa [online]. RPP, 2016 Disponible en: <http://rpp.pe/politica/elecciones/asi-esta-el-peru-2016-la-desercion-escolar-y-la-calidad-educativa-noticia-938483>

[3] Perú entre los países que menos invierten en educación, por debajo de los US$ 50,000 [online]. Diario Gestión, 2018 Disponible en: <https://gestion.pe/economia/peru-paises-invierten-educacion-debajo-us-50-000-229121>

[4] Cómo la Inteligencia Artificial puede revolucionar la educación [online]. blog.Boostup.io, 2017 Disponible en: <https://blog.boostup.io/c%C3%B3mo-la-inteligencia-artificial-puede-revolucionar-la-educaci%C3%B3n-9748a1f45fd>

[5] P. Cortez, Student Performance Data Set [online]. California: University of California, Irvine, 2014 Disponible en: <https://archive.ics.uci.edu/ml/datasets/student+performance>

[6] S. Borkar and K. Rajeswari, "Attributes Selection for Predicting Students’ Academic Performance using Education Data Mining and Artificial Neural Network," IJCA, vol. 86, no. 6, pp. 25-29, January 2014.

[7] M. Singh, H. Nagar and A. Sant, "Using Data Mining to Predict Primary School Student Performance," IJARIIE, vol. 2, no. 1, pp. 43-46, January 2016.

[8] K. Kularbphettong. C. Tongsiri “Mining Educational Data to Support Students’ Major Selection”, *IJEPD*, vol 8, pp. 21-23, 2014

[9] M. Tiwari, R. Singh and N. Vimal, "An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education", *IJCSMC, vol.*2, no. 2, pp. 53-57, 2013.

[10] J. John & J. Kavya , K. Paarth , P. Shubha “Educational Data Mining techniques and their applications”, *IEEE*, vol. 1, pp. 1344-1348., Noviembre 2015

[11] E. Osmanbegović, H. Agić and M. Suljić, “Determining dominant factor for students performance prediction by using data mining”, HRCAK, Disponible en: <https://hrcak.srce.hr/file/197395>

CLASSIFICATION ALGORITHMS

[12] Aprendizaje profundo (deep learning) [online].TechTarget, 2017 Disponible en: <https://searchdatacenter.techtarget.com/es/definicion/Aprendizaje-profundo-deep-learning>

[13] J. Velásquez and L. Donoso, “Aplicación de Técnicas de Web Mining sobre los Datos Originados por Usuarios de Páginas Web. Visión Crítica desde las garantías fundamentales, especialmente la Libertad, la Privacidad y el Honor de las Personas”, RIS, vol. 14, junio 2010. Disponible: <http://www.dii.uchile.cl/~ris/RISXXIV/Velasquez47.pdf>

[14] Análisis de sentimiento, ¿qué es, cómo funciona y para qué sirve? [online]. ITELLIGENT, 2017 Disponible en: <http://www.itelligent.es/es/analisis-de-sentimiento/>

Anexo 1

