

Vea discusiones, estadísticas y perfiles de autor para esta publicación en: <https://www.researchgate.net/publication/335979954>

Modelo predictivo de graduados a tiempo utilizando algoritmos de aprendizaje automático

Capítulo · Septiembre 2019

DOI: 10.1007 / 978-981-15-0399-3_11

CITACIÓN

1

LEE

1,010

5 autores, incluyendo:



Nurafifah Mohammad Suhaimi

2 PUBLICACIONES 15 CITAS

[VER EL PERFIL](#)



Shuzlina Abdel Rahman

Mara Tecnológica Universitaria

89 PUBLICACIONES 380 CITAS

[VER EL PERFIL](#)



Sofianita Mutalib

Mara Tecnológica Universitaria

63 PUBLICACIONES 235 CITAS

[VER EL PERFIL](#)



Nurzeatul Abdul Hamid

Mara Tecnológica Universitaria

26 PUBLICACIONES 138 CITAS

[VER EL PERFIL](#)

Algunos de los autores de esta publicación también están trabajando en estos proyectos relacionados:



Proyecto UiTM AV [Ver proyecto](#)



Inteligencia Artificial [Ver proyecto](#)



Modelo predictivo de graduados a tiempo

Uso de algoritmos de aprendizaje automático

Nurafifah Mohammad Suhaimi³, Shuzlina Abdul Rahman^{1,2,3} (&),
Sofianita Mutalib^{1,2,3}, Nurzeatul Hamimah Abdul Hamid^{1,2,3},
y Ariff Md Ab Malik^{1,3,4}

¹ Grupo de Iniciativa de Investigación de Sistemas Inteligentes, Universiti Teknologi MARA,
40450 Shah Alam, Selangor, Malasia

{shuzlina,sofi,nurzea}@tmsk.uitm.edu.my,
ariff215@puncakalam.uitm.edu.my Facultad de Ciencias

² Informáticas y Matemáticas,

Universidad Tecnológica MARA, 40450 Shah Alam, Selangor, Malasia

³ Universidad Tecnológica MARA, 40450 Shah Alam, Selangor, Malasia
afifahsuhaimi01@gmail.com Facultad de

⁴ Administración y Negocios, Universiti Teknologi MARA,
42300 Puncak Alam, Selangor, Malasia

Abstracto. En la mayoría de las universidades, el número de estudiantes que se graduaron a tiempo reflejan tremendamente en sus costos de operación. En tales casos, el elevado número de Los logros de los estudiantes graduados a tiempo o GOT reducirán indirectamente el costo operativo anual de la universidad por estudiante. No es tan trivial como parece, para asegurar la mayoría de los estudiantes capaces de GOT es un reto. Puede variar en la perspectiva. de prácticas universitarias, programas académicos y antecedentes de los estudiantes. En el nivel universitario, los datos de los estudiantes se pueden utilizar para identificar el logro y capacidad de los estudiantes, intereses y debilidades. Para construir una predicción precisa modelo, requiere un estudio extenso sobre los factores significativos que pueden contribuir a capacidad de los estudiantes para graduarse a tiempo. En consecuencia, este estudio tiene como objetivo construir un modelo predictivo que puede predecir el estado de graduación de los estudiantes. aplicamos cinco diferentes algoritmos de aprendizaje automático (clasificadores), a saber, Árbol de decisión, Bosque aleatorio, Naïve Bayes, Máquina de vectores de soporte (PolyKernel) y Soporte Máquina vectorial (RBFKernel). Estos clasificadores se evaluaron con cuatro pliegues k diferentes de 5, 10, 15 y 20. El rendimiento de estos clasificadores fue comparado basado en diferentes mediciones sujetas a exactitud, precisión, recuperación, y puntuación F. Los resultados indicaron que Support Vector Machine (PolyKernel) superó a otros clasificadores y los mejores números de k pliegues para este experimento son 5 y 20. Se espera que este modelo predictivo de GOT sea beneficioso para administración universitaria y académicos para diseñar sus estrategias para ayudar y mejorar la debilidad del rendimiento académico de los estudiantes y asegurar pueden graduarse a tiempo.

Palabras clave: Minería de datos Graduate-On-Time Machine learning
Modelo predictivo Algoritmo supervisado

1. Introducción

El patrón decreciente de los estudiantes graduados a tiempo (GOT) se ha convertido en un problema importante para el Ministerio de Educación de Malasia, profesores, académicos y partes relacionadas a pesar del número creciente de estudiantes matriculados. Asegurar a los estudiantes GOT se ha convertido en el mayor desafío para la universidad ya que la posición de una universidad en la industria de la educación se basa en este indicador que actúa como una de las métricas para medir la efectividad institucional [1]. La medición de la productividad académica se mide de muchas maneras, dependiendo del rango de parámetros de entrada y resultados académicos. Los dos indicadores de productividad utilizados por las universidades de Malasia son la Intake Graduation on Time (iGOT) y el costo anual por estudiante de tiempo completo equivalente (Cost per FTSE) [2]. El iGOT mide la productividad en función del número de estudiantes (basado en la cohorte de admisión) que se gradúan dentro de la duración estipulada del programa. En lugar de observar la tasa de graduación anual, iGOT destaca la cantidad de estudiantes que se graduaron en comparación con las cifras de admisión. Con base en esta medida, aunque un estudiante puede tomar un período más largo para terminar su carrera, la universidad asumirá un mayor costo. La mejora de la tasa iGOT también se vincula con el Coste por FTSE [2]. Según las estadísticas de 2013/2014, el iGOT promedio de las universidades públicas es del 74 %. Las estadísticas también muestran una gran brecha entre las universidades del 27%.

Aunque en algunos casos, la graduación retrasada es inevitable, el análisis de datos históricos podría proporcionar información beneficiosa [1]. Por lo tanto, la aplicación de Education Data Mining (EDM) en tales áreas podría allanar el camino para comprender mejor los problemas en cuestión.

EDM diseña un nuevo paradigma de exploración para analizar el comportamiento de aprendizaje de los estudiantes para obtener información. Utiliza técnicas de aprendizaje automático para explorar datos de entornos educativos, como registros en línea, enfoques de enseñanza, recursos de enseñanza, pruebas intermedias y resultados de exámenes para predecir y aprender los patrones que caracterizan los comportamientos de los estudiantes que afectan su desempeño. Posteriormente, el objetivo de la exploración es comprender y mejorar los resultados educativos. Por ejemplo, el modelo de predicción del rendimiento de los estudiantes mediante la previsión de la calificación de los estudiantes permite que la dirección académica de la universidad diseñe un mecanismo de advertencia adecuado para los estudiantes que están en riesgo. Por lo tanto, capaz de ayudarlos a superar las dificultades en su estudio. Por lo tanto, el modelo de predicción podría proporcionar información útil para que los programas estratégicos planifiquen una medida adecuada para mejorar el rendimiento de los estudiantes [3]. De manera similar, los hallazgos de la investigación revelan que entre los factores que contribuyeron al bajo rendimiento de los estudiantes se encuentran el género [4], el estado civil y la edad [5], los problemas familiares [6], el historial académico previo [7], la demografía, antecedentes personales, educativos, psicológicos, progreso académico y otras variables ambientales [8]. Aunque se han identificado la mayoría de los factores significativos, la predicción del rendimiento de los estudiantes es dinámica, ya que varía según las universidades, los programas y los antecedentes de los estudiantes [9]. En el método EDM, el modelado predictivo anticipa el tiempo de graduación de los estudiantes. Para construir el modelo predictivo, se utilizan varias tareas, que son clasificación, regresión y categorización. Por ejemplo, la tarea de clasificación crea modelos predictivos para la predicción de variables objetivo en función de varias variables de entrada [10]. Las técnicas de clasificación se aplican con frecuencia para facilitar el proceso de toma de decisiones [11]. Este trabajo tiene como objetivo descubrir el clasificador ideal que funciona mejor para predecir la

estado de graduación o simplemente GOT. Hay varios clasificadores bajo la tarea de clasificación que se han aplicado para predecir el tiempo de graduación de los estudiantes. Entre los clasificadores utilizados se encuentran el árbol de decisión, las redes neuronales artificiales, Naive Bayes, K-Nearest Neighbor y Support Vector Machine [10, 12–14]. Sin embargo, el rendimiento de estos clasificadores es variado, ya que funcionan de manera diferente según el tipo de datos. Es necesario realizar varios experimentos para encontrar qué clasificador funciona mejor. Presentamos este documento de la siguiente manera: la primera sección resume la descripción general de esta investigación, mientras que la sección dos describe los trabajos relacionados. La tercera sección revisa la metodología involucrada para llevar a cabo esta investigación. La cuarta sección presenta los resultados y las discusiones. Por último, concluimos este trabajo en la sección cinco.

2. Trabajo relacionado

El creciente número de estudiantes que no pueden graduarse a tiempo o GOT afecta significativamente a la institución para producir resultados de calidad cada año y contribuye al bajo puntaje en las tasas de graduación [2]. En consecuencia, tiene un impacto adverso en la productividad de la universidad y, por lo tanto, afecta la clasificación de la universidad. El iGOT mide si los estudiantes terminan sus estudios en el tiempo requerido. Shariff et al. [14] definieron iGOT como un estado donde los estudiantes completaron sus estudios en un tiempo determinado establecido por la universidad. En el que la mayoría de las instituciones en otros países de Europa han establecido que el tiempo para que los estudiantes de pregrado se gradúen es de cuatro años, pero lamentablemente, muchos estudiantes se retrasan en terminar su carrera (el 40% lo completa en cuatro años mientras que el 60% en seis años) [15].

Los estudiantes que tardan más en graduarse afectan el presupuesto de la universidad, ya que la universidad tiene que gastar más dinero para proporcionar recursos adicionales, como aulas adicionales para atender a la cantidad de estudiantes [2]. Hay alrededor del 25% de los estudiantes en las universidades de Malasia que se quedan más allá de la duración del curso [16]. Entre las sugerencias que se proponen está la de penalizar a los alumnos que no aprueben GOT. Esta acción podría actuar como un recordatorio para que los demás siempre tomen su estudio en serio. Este escenario desafiante preocupa a muchas partes interesadas, especialmente a la administración de la universidad, ya que tienen que pensar fuera de la caja y elaborar un plan sólido para resolver tal T y mejorar la cantidad de tasas de graduación. Deben manejar este problema de manera brillante y proactiva, ya que el rendimiento de la universidad depende en gran medida de la tasa de graduación. Una de las soluciones para manejar este problema es analizar el desempeño de los estudiantes, ya que puede ser el indicador para predecir el tiempo de graduación de los estudiantes. Sin embargo, analizar el desempeño de los estudiantes es muy complicado y tedioso ya que involucra muchos datos que aumentan continuamente año tras año [17]. Alternativamente, la minería de datos es aplicable para realizar análisis para resolver este problema.

La minería de datos (DM) es un proceso que puede convertir una cantidad masiva de datos y convertirlos en información o conocimiento significativo. Se trata de analizar y categorizar la información, así como resumir el conocimiento de diferentes tipos de datos almacenados en bases de datos y almacenes de datos [9, 18]. Hoy en día, la aplicación de la minería de datos está muy extendida en el sistema educativo y es beneficiosa para analizar los resultados de los estudiantes.

desempeño, pronosticar el tiempo de graduación de los estudiantes y otros temas relacionados [19–21].

[14] realizó un modelo predictivo para predecir el estado de graduación de los estudiantes de doctorado. El resultado muestra que hay un total de 79 estudiantes que se pronostican GOT.

Con base en su investigación, señalaron que las mujeres tienden a GOT en comparación con los hombres, ya que la cantidad de estudiantes mujeres que se predice que GOT es un 56% más alta en comparación con los hombres. Los hallazgos de [14] pueden ser intensificados por [22], ya que subrayaron que, a partir de su investigación, el nivel de rendimiento de las estudiantes en educación de software fue más alto que el de los estudiantes varones. La precisión del modelo DM se puede mejorar y realizar un análisis más profundo de los datos extraídos con el uso de algoritmos de aprendizaje automático.

Los algoritmos de aprendizaje automático son los métodos que suelen emplear los investigadores para descubrir patrones a partir de conjuntos de datos permitiéndoles aprender por sí mismos [23]. Además del avance de las amplias capacidades informáticas, ahora parece posible aprender a través de una gran cantidad de datos.

3. Metodología

Hay seis fases involucradas en esta metodología de investigación siguiendo el Proceso intersectorial para la minería de datos (CRISP-DM) [24] donde cada fase está asociada con las actividades necesarias. El detalle de cada fase se discute brevemente en las siguientes secciones.

Comprensión empresarial. Business Understanding es la primera fase de esta investigación.

En esta fase, el área principal que se examinará son las cuestiones relacionadas con el estado de graduación de los estudiantes y los factores que contribuyen a la graduación oportuna de los estudiantes.

Comprensión de datos. La segunda etapa del proceso CRISP-DM es la comprensión de datos, que requiere que el investigador obtenga los datos requeridos y los transforme en un formato que se pueda extraer utilizando la herramienta de minería de datos. En esta investigación, la técnica aplicada en la recopilación de datos fue a través de documentos y registros, mediante el examen de la base de datos de estudiantes de la UiTM que contiene los datos históricos de los estudiantes de la UiTM de la cohorte 2013. Luego, estos datos se examinaron cuidadosamente para identificar la distribución y su rango. valores. Según el examen, los datos sin procesar de la base de datos de UiTM constan de 31 atributos con 74 670 instancias.

Preparación de datos. Esta es la fase crucial en el proceso CRISP-DM, ya que la competencia del modelo de investigación depende en gran medida de la calidad del conjunto de datos. Varias actividades involucradas en esta fase que incluyen selección y limpieza de datos, construcción de datos e integración y formateo de datos. Se seleccionaron varios atributos del conjunto de datos sin procesar en función de la relevancia para los objetivos de esta investigación. El resto de los atributos no seleccionados fueron descartados ya que no otorgaba ningún significado y aporte para lograr los objetivos de esta investigación. La Tabla 1 muestra la lista de los atributos seleccionados, 13 atributos con estado GOT como destino o etiqueta de clase.

Tabla 1. Selección de atributos

Atributo	Valor	Descripción
Prog_desc	Programa de Ciencias, Ingeniería programa	Programa tomado por los estudiantes
Modo de estudio	Tiempo completo, tiempo completo extendido	El modo de estudio
Patrocinador	Sí No	préstamo o beca tomado por el estudiante
Discapacidad	Sí No	discapacidad del estudiante
Género	Macho femenino	sexo del estudiante
Estado civil	Soltero Casado	matrimonio del estudiante estado
Raza	Dusun, Iban/Sea Dayak, Jawa, Bidayuh, malayo	carrera de estudiante
Edad	18, 20, 25, 30	edad del estudiante
Permanent_state_desc	Johor, Kedah, Kelantan, Perak, Perlis, Penang, Sabah, Sarawak	permanente del estudiante habla a
Requisitos de entrada	Diploma, Matrícula KPM 0–1.99,	Modo de admisión del estudiante
CGPA	2.00–2.49, 2.50–2.99, 3.00–3.49, 3.5–4,0	CGPA del estudiante
Ingresos familiares	Ingreso cero 1 € - 499,99 € 500yRMN - 999,99yRMN 1000yRMN - 1999,99yRMN \$2000 - \$2999.99 \$3000 - \$3999.99 \$4000 - \$4999.99 5000yRMN - 7999,99yRMN \$8000 - \$9999.99 RM10000 y superior	familia del estudiante ingreso
GOTstatus	GOT, no GOT	graduación del estudiante estado

Esta investigación atendió los datos de los estudiantes del Programa de Ciencias e Ingeniería. Por utilizando la tarea de "filtro" en Microsoft Excel, para el atributo prog_desc, todos los demás programas que el Programa de Ciencias y el Programa de Ingeniería fueron descartados. el total de instancias del Programa de Ingeniería es 1160 instancias mientras que el Programa de Ciencias es 2575 instancias.

Modelado. La fase de modelado es la parte para buscar patrones útiles en los datos. Dentro proceso de aprendizaje automático, el conjunto de datos debe someterse al proceso de modelado para identificar el patrones de los conjuntos de datos. En la minería de datos, hay un número de algoritmos de modelado o comúnmente conocido como clasificador, pero no todos ellos encajan con el proyecto de esta investigación. El La lista de clasificadores se redujo, en base a pocas reglas, particularmente basadas en el preguntas de negocio y el tipo de variables involucradas. Los datos preparados derivados de El proceso de preparación de datos se entrenó en cinco clasificadores diferentes, a saber, árbol de decisión, Random Forest, Naïve Bayes, SVM con PolyKernel y SVM con RBFKernel.

Evaluación. En la fase de evaluación del modelo, estos clasificadores se evaluaron en función de cuatro medidas de rendimiento que son el valor de puntuación de precisión, precisión, recuperación y medida F. El clasificador que obtuvo la mejor puntuación es designado como el mejor clasificador para esta investigación.

Despliegue. La fase de implementación es la fase en la que todo el progreso de la investigación, los resultados, los resultados y los hallazgos, así como cualquier problema o limitación, se implementan en un formulario de informe.

Los modelos predictivos se desarrollan utilizando una herramienta de minería de datos, llamada Waikato. Entorno para el Análisis del Conocimiento (WEKA) [25].

4 Resultados y Discusiones

Esta sección presenta el desempeño de los cinco clasificadores como se mencionó en la sección anterior. Estos clasificadores se evaluaron en base a la validación cruzada con 5, 10, 15 y 20 pliegues. El rendimiento de cada clasificador se interpreta utilizando diferentes medidas de rendimiento, como la puntuación de exactitud, la precisión, la recuperación y la puntuación F. Las siguientes subsecciones presentan y analizan los resultados obtenidos de estos clasificadores.

4.1 Análisis de puntuación de precisión

La puntuación de precisión es el número correcto de predicciones realizadas dividido por el número total de predicciones realizadas, multiplicado por 100 para convertirlo en un porcentaje. En otras palabras, la precisión es la posibilidad de que el clasificador pueda predecir correctamente las instancias positivas y negativas. La Figura 1 muestra el puntaje de precisión general para los cinco clasificadores.

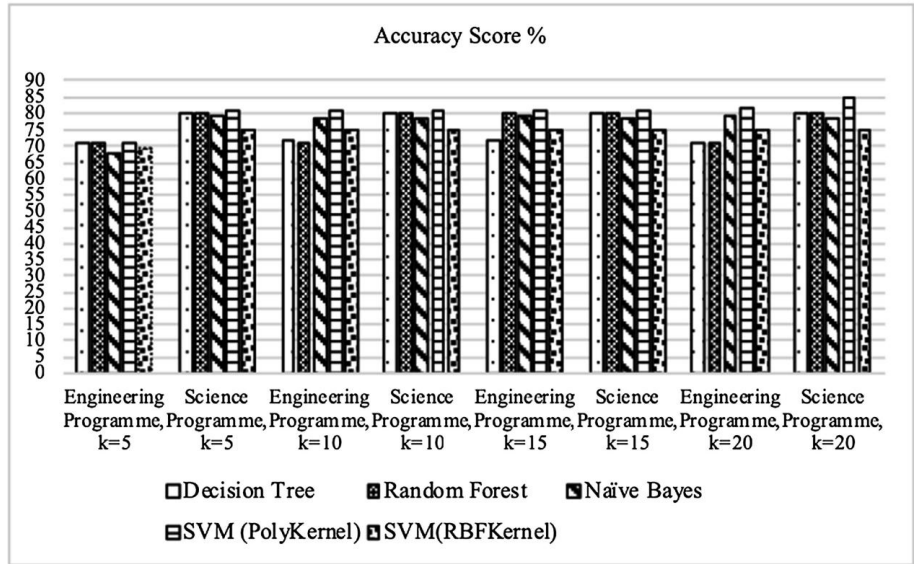


Fig. 1. Puntuación de precisión de diferentes clasificadores

De acuerdo con el gráfico de la Fig. 1, se ilustran los resultados de cada clasificador, según el tipo de datos y el pliegue de validación cruzada. En el gráfico, el gráfico de barras más alto que se puede ver en Science Program es cuando $k = 20$, que es de SVM (PolyKernel). Este clasificador ha superado a los demás con una puntuación de precisión del 84,78 %. Por otro lado, el puntaje de precisión más bajo lo logró Naïve Bayes con $k = 5$, en el Programa de Ingeniería, con solo 68.13%. Cuando $k = 5$, podemos ver claramente que los clasificadores se desempeñaron mejor en el Programa de Ciencias, en comparación con el Programa de Ingeniería. Sin embargo, en promedio, la puntuación de precisión obtenida de todos los clasificadores en el Programa de Ingeniería aumentó cuando $k = 10$, pero no se produjeron grandes diferencias en cuanto a la puntuación de precisión en el Programa de Ciencias. La puntuación de precisión de Random Forest aumentó drásticamente cuando $k = 15$ para el Programa de Ingeniería, pero para los otros clasificadores, la puntuación se mantuvo constante para ambos tipos de datos. Sin embargo, el puntaje de precisión para Random Forest disminuyó cuando el valor k se cambió a 20. SVM (PolyKernel), por otro lado, aumentó constantemente en el término del puntaje de precisión cuando el valor k se cambió de 15 a 20. Además, podemos ver una clara tendencia al alza en la puntuación de precisión de SVM (PolyKernel) para ambos tipos de datos, cuando se probaron en $k = 5, 10, 15$ y 20 y alcanzó la puntuación máxima cuando el valor de k es 20 . En general, a partir del gráfico anterior, podemos concluir que, en promedio, todos los clasificadores dieron un mejor resultado de precisión para el programa de Ciencias en el rango de 75 a 85% y el valor ideal de k para determinar el puntaje de precisión es cuando $k = 20$.

4.2 Análisis de puntuación de precisión

En el aprendizaje automático, la precisión es la cantidad de valores positivos que se pronostican correctamente hasta el total de valores positivos pronosticados. Aquí, la predicción positiva correcta se destaca de todas las predicciones positivas. Por ejemplo, en esta investigación, queremos resaltar el número real de estudiantes que realmente GOT de todos los estudiantes que se predice que GOT. La puntuación de alta precisión indica que hay un número bajo de falsos positivos.

Falso positivo es cuando el clasificador etiqueta datos que en realidad son negativos con positivos. En esta investigación, los estudiantes que no son GOT están etiquetados con GOT. El gráfico de barras de la Fig. 2 muestra la comparación de la puntuación de precisión para todos los clasificadores.

En general, a partir del gráfico anterior, podemos ver que la puntuación de precisión de estos cinco clasificadores se mantuvo constante a lo largo de los experimentos cuando $k = 5, 10, 15$ y 20 .

Aquí, cuando $k = 5$, el rendimiento de SVM (PolyKernel) cuando se probó en el Programa de Ciencias logró el puntaje de precisión más alto en comparación con otros clasificadores.

La puntuación de precisión de SVM (PolyKernel) disminuyó alrededor de un 1 % cuando $k = 10, 15, 20$ y la puntuación se mantuvo sin cambios a lo largo de los experimentos, que es 0,82, con el Programa Científico. Desde Science Programme, se puede ver que SVM (PolyKernel) y SVM (RBFKernel) compartieron el mismo puntaje de precisión cuando $k = 15$ y $k = 20$, lo que muestra que estos dos clasificadores se desempeñaron de manera idéntica con esos valores de k . De manera similar, para Science Programme, Decision Tree y Random Forest también lograron la misma puntuación de precisión en todos los experimentos, que es 0,8. Por otro lado, para el Programa de Ciencias, la tendencia del puntaje de precisión para todos los clasificadores es la misma, excepto para Random Forest. Cuando $k = 5$ y $k = 10$, Random Forest obtuvo una puntuación de 0,71, pero la puntuación aumentó ligeramente a 0,72 para los siguientes pliegues de validación cruzada, que son $k = 15$ y $k = 20$. Además, SVM (PolyKernel) obtuvo la puntuación de precisión más baja, en comparación con

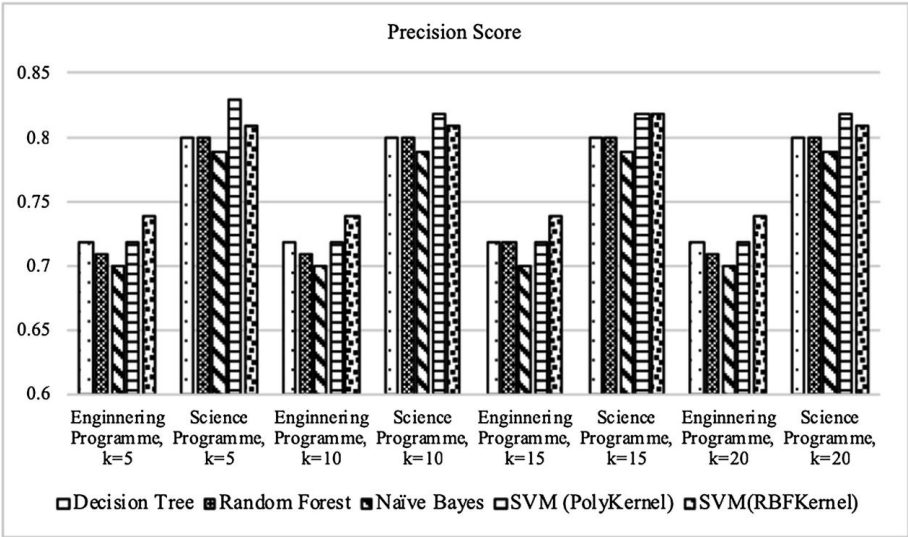


Fig. 2. Puntuación de precisión de diferentes clasificadores

los otros clasificadores, que es 0.70. Sin embargo, este clasificador se desempeñó mejor en el Programa de Ciencias, lo que demuestra que este clasificador fue bueno para clasificar datos en el Programa de Ciencias, en comparación con el Programa de Ingeniería. En general, a partir del gráfico, se puede concluir que todos los clasificadores capaces de producir resultados más precisos en la predicción de GOT con programas Science y, en promedio, el mejor número de pliegues de validación cruzada para medir la puntuación de precisión de todos los clasificadores es 5. En la siguiente sección, se analiza en detalle el análisis del rendimiento en cada pliegue.

4.3 Análisis de puntuación de recuperación

El recuerdo o la sensibilidad es la proporción del valor positivo real que se predice correctamente como positivo. También se puede definir como la relación entre los valores positivos pronosticados correctamente y los valores positivos reales. Recall destaca la sensibilidad del algoritmo, es decir, de todos los positivos reales de cuántos fueron capturados por el clasificador. La puntuación de recuperación se calcula dividiendo el número de positivos reales entre el número de positivos reales más el número de negativos falsos. Los datos positivos reales son aquellos en los que el clasificador que en realidad es positivo clasifica los datos como positivos, o en una palabra más simple, son correctos. Los falsos negativos, por otro lado, son datos que el clasificador marca como negativos y que en realidad son positivos o incorrectos. En esta investigación, los positivos reales son GOT correctamente predichos y los falsos negativos son los estudiantes que los clasificadores etiquetan como no GOT que en realidad eran GOT.

La recuperación se puede considerar como la capacidad de un clasificador para encontrar todos los puntos de interés de un conjunto de datos. El gráfico de barras de la Fig. 3 muestra la comparación de la puntuación de recuerdo para todos los clasificadores.

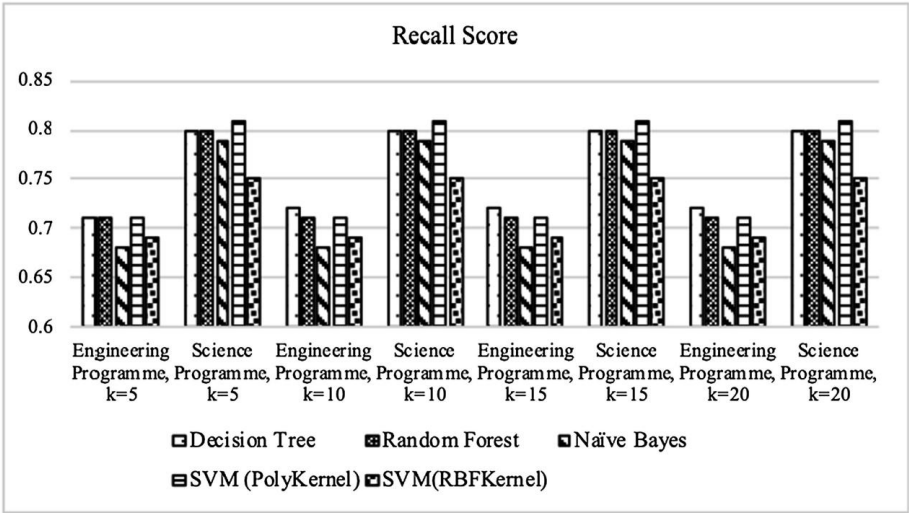


Fig. 3. Puntuación de recuerdo de diferentes clasificadores

En el gráfico, para el Programa de ingeniería, se puede ver que el rendimiento de todos los clasificadores fue ligeramente bajo cuando los pliegues de validación cruzada son iguales a 5. Aquí, Árbol de decisión, Bosque aleatorio y SVM (PolyKernel) compartieron el mismo nivel de recuperación que es del 0,71 %, y la puntuación de recuerdo más baja la alcanza Naïve Bayes, con solo el 0,68 %. Sin embargo, el rendimiento de Decision Tree aumentó a 0,72% y se mantuvo constante para k = 10, 15 y 20. Además, el rendimiento de todos los clasificadores también se mantuvo constante para k = 10, 15 y 20. Se puede concluir que cuando k = 5, el desempeño de los clasificadores fue levemente menor, en comparación con el resto de los pliegues de validación cruzada. Para Science Programme, el patrón de la puntuación de recuerdo para todos los clasificadores es similar cuando k = 5, 10, 15 y 20. Aquí, Decision Tree y Random Forest compartieron una puntuación similar, que es del 0,80 %, a lo largo de los experimentos. SVM (PolyKernel) superó a los demás clasificadores con una puntuación del 0,81 %, mientras que SVM (RBFKernel) logró la puntuación de recuperación más baja, que es del 0,75 %. En general, a partir del gráfico, se puede concluir que todos los clasificadores predijeron mejor el estado GOT con Science Programme y que el diferente número de pliegues no afectó la puntuación de recuerdo de los clasificadores.

4.4 Análisis de rendimiento

La puntuación F, también llamada medida F, es una medida de la precisión de una prueba. La puntuación F se define como la media armónica ponderada de la precisión y recuperación de la prueba. La puntuación de la puntuación F tiene en cuenta la precisión y la recuperación de una prueba. Como se explicó anteriormente, la precisión es la proporción de resultados positivos que son realmente positivos. Esto también se conoce como el valor predictivo positivo. La recuperación, por otro lado, o más conocida como sensibilidad, es la capacidad de una prueba para identificar correctamente los resultados positivos para obtener la verdadera tasa positiva. Cuando un clasificador logró una puntuación F más cercana a 1, eso significa que el clasificador tiene un equilibrio perfecto

de precisión y recuerdo. El puntaje F más cercano a 0 es el peor, lo que significa que tiene un puntaje bajo de precisión y recuperación. La siguiente Fig. 4 ilustra la tendencia de la puntuación F entre el Programa de Ingeniería y el Programa de Ciencias para todos los clasificadores.

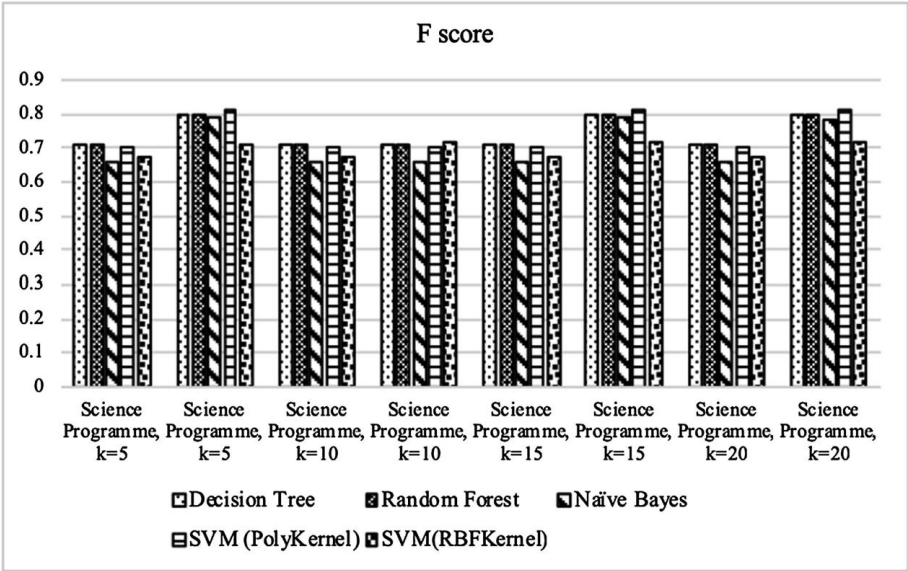


Fig. 4. Puntuación F de diferentes clasificadores

Del gráfico, la tendencia de la puntuación F para todos los clasificadores con el Programa de Ingeniería y el Programa de Ciencias, cuando $k = 5, 15, 20$, es similar. Aquí, se puede interpretar que Decision Tree, Random Forest y SVM (PolyKernel) han dominado la puntuación F para ambos tipos de datos, si se comparan según el tipo de datos.

Específicamente, estos clasificadores se desempeñaron mejor con el Programa de Ciencias, en comparación con el Programa de Ingeniería. Cuando $k = 10$, la puntuación F de Decision Tree, Random Forest, Naïve Bayes y SVM (PolyKernel) se redujo significativamente cuando se probó con Science Programme. Sin embargo, en contra del desempeño de esos clasificadores, el puntaje F para SVM (RBFKernel) aumentó levemente cuando $k = 10$ y se mantuvo constante cuando $k = 15$ y 20 . Esto muestra que este clasificador tuvo dificultades para clasificar datos cuando $k = 5$, pero mejora cuando los valores de k aumentan. Para el Programa de Ingeniería, la tendencia de todos los clasificadores es la misma a lo largo de los experimentos. Decision Tree y Random Forest lograron el mismo nivel de puntuación F cuando $k = 5, 10, 15$ y 20 , que es la puntuación F más alta para este tipo de datos (0,71 %). Naïve Bayes, por otro lado, logró la puntuación más baja, que es de 0,66 %, donde la diferencia en términos de puntuación F entre este clasificador con Árbol de decisión y Random Forest es de 0,05 %, que es pequeña. En general, en promedio, se puede decir que los clasificadores se desempeñaron mejor con el Programa de Ciencias, cuando los pliegues de validación cruzada son iguales a 15 y SVM (PolyKernel) es el mejor clasificador que se puede usar para predecir el estado de graduación de los estudiantes.

5. Conclusión

Este documento demostró el uso de cinco clasificadores de algoritmos de aprendizaje automático para producir un modelo predictivo de Graduate-On-Time (GOT). Estos clasificadores se modelaron en el Programa de Ingeniería y el Programa de Ciencias usando validación cruzada con 5, 10, 15 y 20 pliegues. El rendimiento de cada clasificador se interpretó utilizando diferentes medidas de rendimiento, como la puntuación de precisión, la precisión, la recuperación y la puntuación F. Los resultados mostraron que SVM (PolyKernel) superó a otros clasificadores. Sin embargo, Naïve Bayes tuvo dificultades para predecir el estado de graduación de los estudiantes ya que este clasificador había producido la precisión promedio más baja y el puntaje sujeto a precisión, recuerdo y puntaje F.

El modelo generado a partir del Programa de Ciencias ha contribuido a un mejor rendimiento de los clasificadores en comparación con el Programa de Ingeniería debido a su distribución de datos de equilibrio. Los valores ideales de k para los pliegues de validación cruzada para este experimento de investigación fueron 5 y 20. Esta investigación es beneficiosa para muchas partes, como la dirección académica de la universidad, los académicos y los estudiantes, ya que puede alertar sobre el rendimiento de los estudiantes que tienen más probabilidades de fallar en GOT y las acciones que se pueden tomar para resolver este problema. Además, este enfoque también puede mejorar la calidad académica de la universidad, ya que la cantidad de estudiantes que no pueden GOT puede reducirse significativamente. En el futuro, este trabajo se puede mejorar utilizando más conjuntos de datos de diferentes programas, en particular de campos no científicos y tecnológicos.

Reconocimiento. El conjunto de datos para esta investigación fue proporcionado por el Centro de Planificación e Información Estratégica (CSPI) Universiti Teknologi MARA. Los autores agradecen a la Facultad de Ciencias Informáticas y Matemáticas e Instituto de Calidad y Avance del Conocimiento (InQKA), UiTM por todo el apoyo brindado.

Referencias

1. Ojha, T., Heileman, GL, Martinez-Ramon, M., Slim, A.: Predicción del retraso en la graduación basado en el desempeño de los estudiantes (2017)
2. Enhancing Academic Productivity and Cost Efficiency (Libro de Plata del Programa de Transformación Universitaria), Ministerio de Educación de Malasia (2016). http://mohe.gov.my/muat_turun/awam/penerbitan/university-transformation-programme/188-the-unitp-silver-book
3. Ibrahim, Z., Rusli, D.: Predicción del rendimiento académico de los estudiantes: comparación de redes neuronales artificiales, árboles de decisión y regresión lineal. En: Proceedings 21st Annual SAS Malaysia Forum, págs. 1 a 6 (2007)
4. Dayioglu, M., Türüt-Asik, S.: Diferencias de género en el rendimiento académico en una gran universidad pública de Turquía. Alto. Educ. 53(2), 255–277 (2007). <https://link.springer.com/article/10.1007/s10734-005-2464-6>
5. Amuda, BG, Bulus, AK, Joseph, HP: El estado civil y la edad como predictores del rendimiento académico de los estudiantes de las facultades de educación en el noreste de Nigeria. Soy. J. Educ. Res. 4, 896–902 (2016). <https://doi.org/10.12691/EDUCACION-4-12-7>
6. Asif, R., Merceron, A., Ali, SA, Haider, NG: Análisis del rendimiento de los estudiantes universitarios mediante la minería de datos educativos. computar Educ. 113, 177–194 (2017). <https://doi.org/10.1016/j.compedu.2017.05.007>
7. Herath, D.: Minería de datos educativos para investigar comportamientos de aprendizaje: una revisión de la literatura (2018). <https://doi.org/10.13140/RG.2.2.20919.01446>

8. Agrawal, RS, Pandya, MH: Minería de datos con redes neuronales para predecir los logros académicos de los estudiantes. En t. J.Comp. ciencia Tecnología 7(2) (2016). <http://www.ijcst.com/vol72/1/19-richa-shambhulal-agrawal.pdf> _ 9. Mohammad Suhaimi, N., Abdul-Rahman, S., Mutalib, S., et al.: Revisión sobre la predicción del tiempo de graduación de los estudiantes mediante algoritmos de aprendizaje automático. En t. Mod. J. Educ. computar ciencia 11, 1–13 (2019). <https://doi.org/10.5815/ijmecs.2019.07.01>
10. Wah, YB, Ibrahim, N., Hamid, HA, et al.: Métodos de selección de características: caso de enfoques de filtro y envoltura para maximizar la precisión de la clasificación. Pertanika J. Sci. Tecnología 26, 329–340 (2018)
11. Akmal, E., Zaman, K., Farhan, A., et al.: Soft computing en ciencia de datos. Cómputo suave. Datos ciencia 545, 387–401 (2015). <https://doi.org/10.1007/978-981-287-936-3>
12. Cahaya, L., Hiryanto, L., Handhayani, T.: Predicción del tiempo de graduación de los estudiantes usando algoritmo inteligente K-Medoids, págs. 263–266 (2017)
13. Pang, Y., Judd, N., O'Brien, J., Ben-Avie, M.: Predecir los resultados de graduación de los estudiantes a través de máquinas de vectores de soporte. En: Actas - Conferencia Fronteras en Educación FIE, octubre de 2017, pp. 1–8 (2017). <https://doi.org/10.1109/FIE.2017.8190666>
14. Shariff, SSR, Rodzi, NAM, Rahman, KA, et al.: Predicción del "graduado a tiempo (GOT)" de Ph.D. estudiantes utilizando el modelo de regresión logística binaria. En: Actas de la conferencia AIP (2016)
15. Mujani, WK, Muttaqin, A., Khalid, KA: Desarrollo histórico de las instituciones públicas de educación superior en Malasia. Oriente Medio J. Sci. Res. 20, 2154–2157 (2014). <https://doi.org/10.5829/idosi.mejsr.2014.20.12.21113>
16. Graduarse a tiempo es el objetivo de Malasia. En: Afterschool.my (2015). <https://afterschool.my/articulos/graduarse-a-tiempo-es-el-objetivo-de-malasia>. Consultado el 14 de noviembre de 2018
17. Ogwoka, TM, Cheruiyo, W., Okeyo, G.: Un modelo para predecir el rendimiento académico de los estudiantes utilizando un híbrido de K-medias y algoritmos de árboles de decisión. En t. J. Cómputo. aplicación Tecnología Res. 4, 693–697 (2015)
18. Jing, L.: Minería de datos y sus aplicaciones en la educación superior. nuevo directorio Inst. Res. 2002, 17 (2002)
19. Ma, X., Zhou, Z.: Predicción de las tasas de aprobación de los estudiantes utilizando una máquina de vectores de soporte optimizada y un árbol de decisiones. En: 2018 IEEE 8th Annual Computing Communication Workshop Conference CCWC 2018, enero de 2018, págs. 209–215 (2018). <https://doi.org/10.1109/CCWC.2018.8301756>
20. Athani, SS, Kodli, SA, Banavasi, MN, Hiremath, PGS: predictor del rendimiento académico y comportamiento social de los estudiantes mediante técnicas de minería de datos. En: Proceeding - IEEE International Conference Computing Communication Automation ICCCA 2017, enero de 2017, págs. 170–174 (2017). <https://doi.org/10.1109/CCAA.2017.8229794>
21. Al-Shehri, H., Al-Qarni, A., Al-Saati, L., et al.: Predicción del rendimiento de los estudiantes utilizando la máquina de vectores de soporte y el vecino K-Nearest. En: Conferencia canadiense sobre ingeniería eléctrica e informática, págs. 17–20 (2017). <https://doi.org/10.1109/CCECE.2017.7946847>
22. Lee, SJ, Kim, JM, Lee, WG: Análisis de los factores que afectan el rendimiento en la educación de programación de creadores en la era de la comunicación inalámbrica. Wirel. pers. común 93, 187–209 (2017). <https://doi.org/10.1007/s11277-016-3450-2>
23. Asyraf, AS, Abdul-Rahman, S., Mutalib, S.: Minería de términos textuales para el análisis de predicción del mercado de valores utilizando noticias financieras. común computar información ciencia 788, 293–305 (2017). https://doi.org/10.1007/978-981-10-7242-0_25
24. Marbán, Ó., Mariscal, G., Segovia, J.: A data mining & Knowledge Discovery Process Model. En: Ponce, J., Karahoca, A. (eds.) Data Mining and Knowledge Discovery in Real Life Applications, febrero de 2009, págs. 438–453. I-Tech, Viena, Austria (2009). ISBN 978-3-902613-53-0
25. Frank, E., et al.: El banco de trabajo WEKA. Apéndice en línea para minería de datos: herramientas y técnicas prácticas de aprendizaje automático, 4.ª ed. Prensa de Morgan Kaufmann, Burlington (2016)