Un estudio de algoritmos de selección de características para Predicción del rendimiento académico de los estudiantes

Maryam Zaffar
Departamento de Computación y Ciencias de la Información
Universidad Tecnológica PETRONAS,
32610 Seri Iskander, Malasia

KS Savita
Centro de Cómputo en la Nube de Alto Rendimiento
Departamento de Ciencias de la Computación e Información
Universiti Teknologi PETRONAS, 32610 Seri
Iskander, Malasia

Manzoor Ahmed Hashmani
Centro de computación en la nube de alto rendimiento
Departamento de Computación y Ciencias de la Información
Universidad Tecnológica PETRONAS,
32610 Seri Iskander, Malasia

Syed Sajjad Hussain Rizvi
Departamento de Telecomunicaciones
Universidad Hamdard, Karachi,
Pakistán

***Resumen—El* objetivo principal de todas las organizaciones educativas es mejorar la calidad de la educación y elevar el rendimiento académico de los estudiantes. La Minería de Datos Educativos (EDM) es una creciente campo de investigación que ayuda a las instituciones académicas a mejorar el rendimiento de sus estudiantes. Las instituciones académicas suelen ser juzgadas por las calificaciones obtenidas por los estudiantes en el examen. EDM ofrece diferentes prácticas para predecir el rendimiento académico de los estudiantes. En EDM, la selección de características (FS) juega un papel vital en la mejora de la calidad de los modelos de predicción para conjuntos de datos educativos. algoritmos de FS eliminar datos no relacionados de los repositorios educativos y por lo tanto, aumente el rendimiento de la precisión del clasificador utilizado en diferentes prácticas de EDM para respaldar la toma de decisiones en entornos educativos. La buena calidad del conjunto de datos educativos puede producir mejores resultados y, por lo tanto, las decisiones basadas en dicho conjunto de datos de calidad pueden aumentar la calidad de la educación al predecir el rendimiento de los estudiantes. A la luz de este hecho mencionado, es necesario elegir cuidadosamente un algoritmo de selección de características. Este documento presenta un análisis del rendimiento de los algoritmos de selección de características de filtro y los algoritmos de clasificación en dos conjuntos de datos de estudiantes diferentes. Los resultados obtenidos de diferentes algoritmos y clasificadores de FS en dos conjuntos de datos de estudiantes con diferente número de características también ayudarán a los investigadores a encontrar las mejores combinaciones de clasificadores y algoritmos de selección de características de filtro. Es muy necesario arrojar luz sobre la relevancia de la selección de características para la predicción del rendimiento de los estudiantes, como herramienta educativa constructiva.**

Las estrategias se pueden derivar a través del conjunto relevante de características. Los resultados de nuestro estudio muestran que existe una diferencia del 10 % en la precisión de las predicciones entre los resultados de los conjuntos de datos con un número diferente de características.

Palabras clave: minería de datos educativos; selección de características algoritmos; clasificadores; SFC; algoritmo de selección de características de relieve

I. INTRODUCCIÓN

La educación es un factor primordial para el desarrollo de una nación. La calidad de la educación es uno de los ingredientes más necesarios para crear miembros destacados de la sociedad. Los datos guardados en las bases de datos de las instituciones académicas juegan un papel importante en la mejora del proceso educativo mediante la exploración de lo oculto. información [1]. Muchas técnicas se están utilizando para evaluar el rendimiento académico de los estudiantes. Las técnicas de minería de datos se utilizan ampliamente en los datos de los estudiantes en estos días [2], [3] y están desempeñando un papel positivo en el área de la minería de datos educativos (EDM). EDM descubre los datos educativos para comprender los problemas en el rendimiento académico de los estudiantes utilizando la naturaleza fundamental de las técnicas de minería de datos [4].

La predicción del rendimiento de los estudiantes se considera un tema importante en EDM. Dado que el desempeño de los estudiantes no solo afecta la reputación de la organización, sino también el futuro del estudiante mismo, por lo tanto, los modelos de predicción del desempeño de los estudiantes están en el centro de atención frente a las partes interesadas educativas. EDM implementa datos para ayudar a las organizaciones académicas a planificar estrategias educativas y, a su vez, mejorar la calidad de la educación.

El progreso académico de los estudiantes se puede monitorear a través de los modelos de predicción. Estos modelos de predicción utilizan diferentes técnicas de EDM para analizar el rendimiento académico de los estudiantes. Es muy difícil distinguir las características que afectan el rendimiento académico de los estudiantes [5]. La predicción del rendimiento académico de los estudiantes puede ser útil para que las instituciones identifiquen a los estudiantes que necesitan asistencia financiera [6], [7], mejoren la calidad de la matrícula de la institución [7], [8], ayuden a los estudiantes a planificar mejor para el futuro y también a superar sus problemas. lucha con los estudios. El modelo de predicción del rendimiento de los estudiantes depende de las características seleccionadas del conjunto de datos. Las características más adecuadas se pueden seleccionar aplicando el algoritmo de selección de características [9].
Estos algoritmos pueden refinar los resultados de la predicción [10].
Sin embargo, los algoritmos de selección de características son los mejores para extraer las características relevantes y evitar la redundancia, sin costo de pérdida de datos [11], por lo tanto, es muy adecuado usar algoritmos FS en EDM para evitar la pérdida de datos importantes para construir estrategias con la ayuda de datos de tanta calidad.

Los algoritmos de selección de características se utilizan en el preprocesamiento paso de datos. Admite seleccionar el subconjunto apropiado de características para construir un modelo para la minería de datos. Sin embargo, los algoritmos de selección de características se utilizan para mejorar la precisión predictiva y reducir la complejidad computacional.
[4], [12], [13]. Los algoritmos de selección de características pueden aumentar

el rendimiento de los modelos de predicción del rendimiento de los estudiantes.
Hay tres tipos principales de algoritmos de selección de características, tres categorías principales: filtros, contenedores y modelos híbridos. El método de filtro se realiza en el paso de preprocesamiento y no depende de ningún algoritmo de aprendizaje, pero depende de todas las características de los datos de entrenamiento. El método Wrapper utiliza algoritmos de aprendizaje para estimar las características. Mientras que la selección de características híbridas combina las propiedades del método de filtro y contenedor [12]. En este estudio nos enfocamos principalmente en el algoritmo de selección de características de filtro.

La selección de características se ha utilizado en EDM en diferentes trabajos de investigación [5], [9], [14]. Los investigadores de EDM utilizan diferentes algoritmos de selección de funciones para obtener resultados efectivos en la predicción del rendimiento académico de los estudiantes. Pero aún se requiere mucha atención para construir modelos de predicción del rendimiento de los estudiantes con la ayuda de algoritmos de selección de características.
Nuestro artículo es un paso hacia la detección de las mejores fusiones de algoritmos de sección de características y algoritmos de clasificación en conjuntos de datos de estudiantes.

El esquema del documento es el siguiente: la Sección II proporciona la literatura relacionada con el algoritmo de selección de características utilizado en el campo de EDM. La Sección III proporciona la metodología de investigación seguida por el artículo. La Sección IV ilustra los resultados y las discusiones. La conclusión del estudio se describe en la Sección V.

II. LITERATURA RELACIONADA

Esta sección ofrece una breve revisión de la literatura sobre los algoritmos de selección de características utilizados en el campo de EDM y las diferentes combinaciones de selección de características junto con la clasificación. algoritmos utilizados en los otros estudios. El estudio en [15] propuso un árbol de decisiones mejorado para predecir los indicadores de deserción estudiantil. El estudio recopila el conjunto de datos de 240 estudiantes a través de una encuesta y aplica el algoritmo de selección de características basado en correlación (CFS) (algoritmo de selección de características de filtro) en el paso de preprocesamiento. La precisión de clasificación del modelo muestra más del 90%. Sin embargo, el estudio tomó solo un conjunto de datos en consideración. La investigación en [4] evaluó seis algoritmos de selección de características para predecir el rendimiento de estudiantes de secundaria superior. Los resultados del estudio concluyen que Voted Perceptron y One Rule (OneR) muestran un alto rendimiento predictivo con todos los subconjuntos de características obtenidos a través de algoritmos de selección de características. Además, la ganancia de información (IG) y CFS muestran un mejor valor ROC y valores de medida F en el conjunto de datos de la escuela secundaria superior.

En [1] se presentó un estudio para predecir el desempeño de los estudiantes en la escuela secundaria de Tuzla. El estudio utilizó el algoritmo de selección de características de la relación de ganancia (GR) en el conjunto de datos con 19 características. Los resultados con el algoritmo de clasificación Random Forest (RF) revelan los mejores resultados en términos de precisión de predicción.

La investigación en [16] se realizó para predecir la matriculación de estudiantes en Ciencias, Tecnología, Ingeniería y Matemáticas (STEM) en instituciones de educación superior en Kenia. Se recogieron casi 18 características a través de un cuestionario. El árbol de decisión CART muestra una mejor predicción resultados de precisión con algoritmos de selección de características Chi-Square e IG.

Se realizó un estudio para predecir las calificaciones de los estudiantes en [14], se realizó un Análisis de Componentes Principales (PCA) en el conjunto de datos de los estudiantes matriculados en la licenciatura en ciencias de la computación. El estudio utiliza PCA para construir árboles de decisión a partir de las características extraídas a través de los registros de Moodle, para predecir las calificaciones de los estudiantes.

En el estudio de [17] se llevó a cabo una comparación entre las selecciones de características Greedy, IG-ratio, Chi-Square y mRMR.
El estudio recopiló el registro de estudiantes de primer año con 15 atributos, de la base de datos de la Universidad de Tecnología de Tailandia. El estudio propuso que la selección de Greedy Forward puede dar un mejor resultado de precisión de predicción con la red neuronal artificial (ANN) en comparación con Naïve Bayes, árbol de decisión y k-NN.

Los estudios existentes en minería de datos educativos han utilizado diferentes algoritmos de selección de características de filtro en conjuntos de datos de estudiantes. En este estudio, utilizamos dos conjuntos de datos diferentes, con diferente número de características. Este estudio es una extensión de nuestro trabajo anterior [18].

tercero METODOLOGIA DE LA INVESTIGACION

Este artículo de investigación es una versión extendida del artículo [18]. Se utilizó un conjunto de datos en un estudio anterior para verificar el rendimiento de diferentes algoritmos de selección de características. El objetivo principal de esta investigación es estimar el rendimiento de diferentes algoritmos de FS junto con diferentes algoritmos de clasificación utilizando conjuntos de datos de diferentes estudiantes con un número diferente de características. La comparación entre los resultados de los algoritmos FS se basa en dos conjuntos de datos para proporcionar una nueva minería de datos educativos para el desempeño de varios algoritmos de selección de características con un número diferente de características.

Este estudio responderá a dos preguntas de investigación que son:

RQ1. ¿Cuáles son los algoritmos de selección de características importantes para predecir el rendimiento académico de los estudiantes (ya sea que aprueben o reprueben)?

RQ2. ¿Cuáles son las mejores combinaciones posibles de algoritmos de selección de características y algoritmos de clasificación para predecir el desempeño de los estudiantes (ya sea que aprueben o reprueben)?

Para lograr el objetivo de investigación y responder a las preguntas de investigación mencionadas anteriormente, se toman dos conjuntos de datos de estudiantes de fuentes válidas, después de lo cual se aplican diferentes algoritmos FS que no se usaron anteriormente en este conjunto de datos en los estudios anteriores. Como en este artículo, tratamos de evaluar diferentes algoritmos de selección de características para comprobar su rendimiento.
Se aplican varios algoritmos de clasificación utilizando diferentes algoritmos FS. Se evalúa para verificar el rendimiento entre todas las combinaciones aplicadas en el conjunto de datos de los estudiantes. La figura 1 describe un flujo básico de nuestro estudio. En este estudio se tomaron dos conjuntos de datos de estudiantes. En el segundo paso, los algoritmos de selección de características se aplican por separado en ambos conjuntos de datos, en combinación con diferentes algoritmos de clasificación. Los resultados de precisión y las instancias clasificadas correctamente se compararon en el paso final.

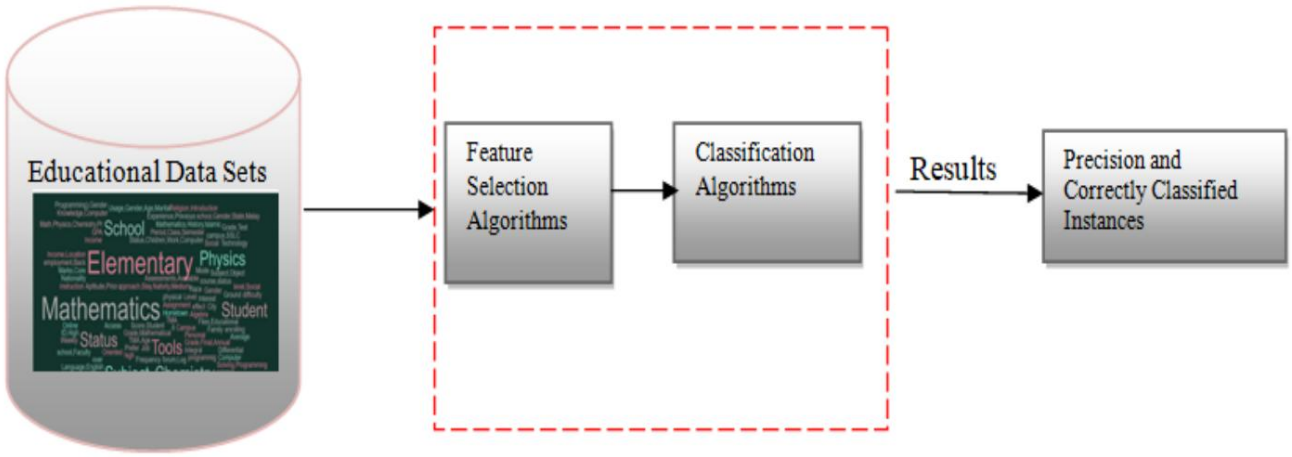

Fig. 1. Flujo de la metodología.

A. Descripción del conjunto de datos

En este estudio, hemos tomado dos conjuntos de datos de estudiantes con diferentes números de características para verificar el rendimiento del algoritmo de selección de características en diferentes números de características.
Los detalles de dos conjuntos de datos utilizados en este estudio se dan a continuación.

1) Conjunto de datos 1: el conjunto de datos 1 está compuesto por 500 estudiantes registros con 16 características. Este conjunto de datos se ha utilizado en el estudio [19] y está disponible públicamente incluso en el conjunto de datos de Kaggle.
repositorio. Se utiliza previamente para comprobar la interactividad del alumno con el sistema de gestión de e-learning. Sin embargo, solo se utiliza previamente el algoritmo de selección de características basado en la ganancia de información. Hay tres categorías de atributos en este conjunto de datos demográficos, académicos y de comportamiento. El conjunto de datos está siendo utilizado por nuestra versión anterior de este estudio.

2) Conjunto de datos 2: El conjunto de datos 2 está compuesto por 300 estudiantes con 24 características. Se recopiló de los tres collages diferentes de la India. Este conjunto de datos se utiliza en el estudio [20] . El conjunto de datos se utiliza en este documento para analizar el rendimiento académico del estudiante.

B. Configuración experimental

Waikato Environment for Knowledge Analysis (WEKA) es desarrollado por la Universidad de Waikato en Nueva Zelanda como herramienta de minería de datos. Está construido en lenguaje Java y es una rica fuente de algoritmos de minería de datos. WEKA ofrece habilidades para desarrollar técnicas de aprendizaje automático para diferentes tareas de minería de datos [21], [22]. En este experimento hemos utilizado la versión 3.9 de Weka y la aplicación Explorer.

C. Algoritmo de selección de características y clasificadores

La selección de características es una de las técnicas más recurrentes y significativas en el preprocesamiento de datos y se dice que es un elemento esencial del proceso de aprendizaje automático [23]. El enfoque principal de nuestra investigación en este documento está en seis importantes algoritmos de selección de características del algoritmo FS CfsSubsetEval, ChiSquaredAttributeEval, OneREval, FilteredAttributeEval, ReliefAttributeEval.

1) CfsSubsetEval: este enfoque identifica la capacidad predictiva de cada función. Sin embargo, el factor de redundancia también juega un papel crítico en este enfoque [24], [25]. El algoritmo CFS utiliza características homogéneas en el proceso de selección junto con pasos de preprocesamiento de discretización [26].

2) ChiSquaredAttributeEval: Chi-Squared se utiliza para comparar las pruebas de independencia y la prueba de bondad de ajuste. La prueba de independencia estima si la etiqueta de clase es dependiente o independiente de una característica. ChiSquared AttributeEval estima un atributo calculando el valor de la estadística chi-cuadrado relacionada con la clase [17], [25].

3) FilteredAttributeEval: este algoritmo de selección de características de filtro está disponible en forma de placa Weka.

4) GainRatioAttributeEval: el Gain Ratio es la medida no simétrica que se introduce para compensar el sesgo de la ganancia de información [27]. Es un algoritmo de selección de características de filtro que mide qué tan común es una característica en una clase asociada a todas las demás clases.

5) Componentes Principales: El análisis de Componentes Principales reduce la dimensionalidad del espacio, sin reducir el número de características [28].

6) ReliefAttributeEval: Relief es un algoritmo simple basado en el peso que depende totalmente de un método estadístico. Evalúa la importancia de un atributo muestreando una instancia repetidamente [25]. Detecta aquellas características que están estadísticamente relacionadas con el concepto de destino. Tiene una limitación de tamaño de conjunto de características no óptimo [29].

La precisión de la predicción de las características seleccionadas de los algoritmos de selección de características se puede evaluar a través de algoritmos de clasificación. En nuestro trabajo anterior hemos utilizado quince algoritmos de clasificación que son: Bayesian Network (BN), Naïve Bayes (NB), NaiveBayesUpdateable (NBU), MLP, Simple Logistic (SL), SMO, Decision Table (DT), OneR J rip, Decsion Stump (DS), J48, Random Forest (RF), RandomTree (RT), REPtree (RepT). Sin embargo, debido a la limitación de espacio, hemos seleccionado seis algoritmos de clasificación en este documento.

IV. RESULTADOS Y DISCUSIONES

Esta investigación informada se centra en la evaluación del rendimiento de seis algoritmos de selección de características utilizando dos conjuntos de datos de estudiantes diferentes. La efectividad de estos algoritmos se mide a través de Precisión, Recuperación, Medida F y Precisión de predicción (Instancias correctamente clasificadas). La medida F se define como la media armónica de precisión y recuperación [30]. Los resultados presentados en nuestro estudio anterior [18] y que luego se comparan con los resultados obtenidos usando el conjunto de datos 1 y el conjunto de datos 2. Los resultados de las seis técnicas de selección de características que usan el conjunto de datos 1 se informan en las Tablas I a VI mediante la aplicación de 15 clasificadores. Estas tablas ilustran los resultados obtenidos por cada uno de los algoritmos de selección de características (FS). Además, cada tabla de resultados contiene cuatro columnas que son valores de algoritmo de clasificación FS, precisión, recuperación y medida F.

A. Resultados en el conjunto de datos 1

Los resultados en la Tabla I muestran los diferentes valores de las medidas de precisión para quince clasificadores con el algoritmo de selección de características Cfssubseteval usando el conjunto de datos 1. La Fig. 2 ilustra gráficamente los resultados obtenidos con ChiSquared-AttributeEval algoritmos de selección de características. Los resultados presentados en la Tabla II y la Fig. 3 muestran que el clasificador Decision Stump (DS) tiene el rendimiento más bajo en el conjunto de datos educativos 1 con ChiSquaredAttributeEval; sin embargo, el clasificador MLP muestra resultados comparativamente mejores que otros clasificadores con la misma técnica FS.

Los resultados presentados en la Tabla III y la Fig. 4 indican que la precisión de los clasificadores utilizados en los datos educativos con Algoritmo de selección de características FilteredAttributeEval. Los resultados demuestran que los valores de precisión, recuperación y medida F son comparativamente bajos cuando se aplican los clasificadores Decision Stump y Jip. Si bien el rendimiento de MLP es relativamente mejor que el de otros clasificadores que usan FilteredAttributeEval.

TABLA I. RESULTADOS DE CFSSUBSETEVAL EN EL CONJUNTO DE DATOS 1 UTILIZANDO DIFERENTES CLASIFICADORES [18]

FS Clasificación Algoritmo	Precisión	Recuerdo	Medida F
Cfs-BN	0,724	0,743	0.742
Cfs-NB	0,73	0,729	0.728
Cfs-NBU	0,73	0,729	0.729
Cfs-MLP	0,736	0,729	0.729
Cfs-SL	0,724	0,722	0.723
Cfs-SMO	0,668	0,667	0.667
Cfs-DT	0,693	0,688	0.688
Cfs-Jrip	0,659	0,66	0.658
Cfs-OneR	0,611	0,583	0.571
Cfs-PARTE	0,713	0,708	0.71
Cfs-DS	0,373	0,528	0.437
Cfs-J48	0,708	0,701	0.702
Cfs-RF	0,64	0,632	0.633
Cfs-RT	0,627	0,618	0.621
Cfs-RepT	0,667	0,66	0.655

TABLA II. RESULTADOS DE CHISQUAREDATTRIBUTEEVAL EN EL CONJUNTO DE DATOS 1USO DE DIFERENTES CLASIFICADORES [18]

FS Clasificación Algoritmo	Precisión	Recuerdo	Medida F
Chi-BN	0,716	0,715	0.716
Chi-NB	0,66	0,66	0.654
Chi-NBU	0,66	0,66	0.654
Chi-MLP	0,769	0,764	0.764
Chi-SL	0,715	0,708	0.709
Chi-SMO	0,741	0,736	0.737
Chi-DT	0,71	0,701	0.702
Chi-Jrip	0,698	0,694	0.692
Chi-OneR	0,611	0,583	0.571
Chi-PARTE	0,64	0,639	0.639
Chi-DS	0,373	0,528	0.437
Chi-J48	0,709	0,708	0.708
Chi-RF	0,718	0,715	0.716
Chi-RT	0,674	0,674	0.674
Chi-RepT	0,651	0,653	0.651

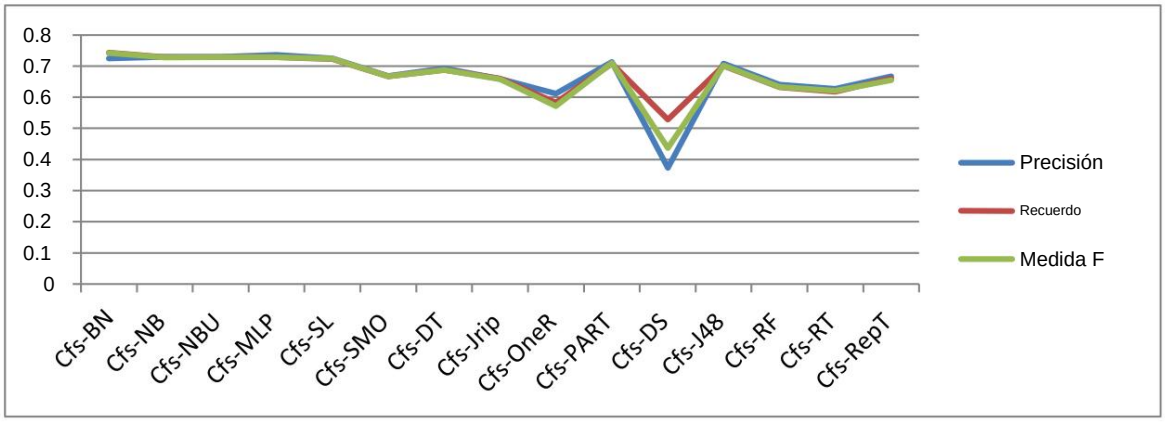

Fig. 2. Rendimiento de CfsSubsetEval utilizando el conjunto de datos 1.

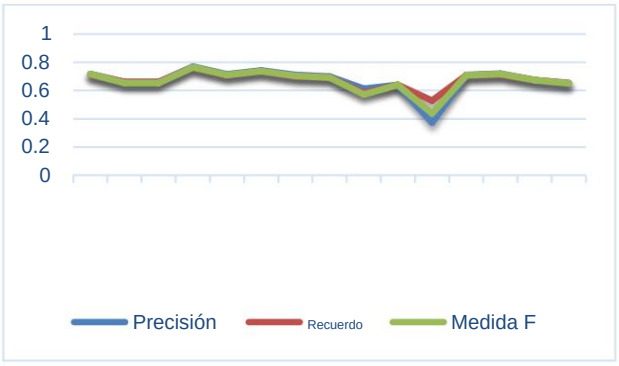

Fig. 3. Rendimiento de ChiSquredAttributeEval utilizando el conjunto de datos 1.

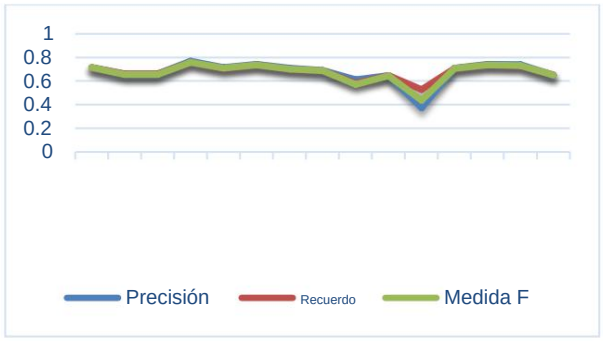

Fig. 4. Rendimiento de FilteredAttributeEval utilizando el conjunto de datos 1.

CUADRO III. EVALUACIÓN DEL RENDIMIENTO DE LA EVALUACIÓN DE ATRIBUTOS FILTRADOS UTILIZANDO LA RECUPERACIÓN DE PRECISIÓN Y LA MEDIDA F EN EL CONJUNTO DE DATOS 1 [18]

FS Clasificación Algoritmo	Precisión	Recuerdo	Medida F
Filtro-BN	0,716	0,715	0.716
Filt-NB	0,66	0,66	0.654
Filt-NBU	0,66	0,66	0.654
Filt-MLP	0,768	0,757	0.758
Filt-SL	0,715	0,708	0.709
Filtro-SMO	0,741	0,736	0.737
Filtro-DT	0,71	0,701	0.702
Filt-Jrip	0,691	0,688	0.688
Filt-OneR	0,611	0,583	0.571
Filt-PARTE	0,646	0,646	0.645
Filt-DS	0,373	0,528	0.437
Filtro-J48	0,709	0,708	0.707
Filtro-RF	0,741	0,736	0.737
Filtro-RT	0,738	0,729	0.73
Filt-RepT	0.651	0.653	0.651

Los resultados informados en la Tabla IV y la Fig. 5 exhiben los detalles de rendimiento idénticos a los ilustrados anteriormente en la Tabla III y la Fig. 4. Los resultados muestran que la disminución en el rendimiento al aplicar el clasificador GainRatioAttributeEval Jrip, sin embargo, MLP y SMO se desempeñaron comparativamente mejor que otros clasificadores.

Los resultados en la Tabla V presentan el desempeño de los Componentes Principales utilizando quince algoritmos de clasificación seleccionados. La Fig. 6 es la representación gráfica del desempeño de los Componentes Principales. El resultado en la Tabla V muestra que el clasificador SMO funcionó relativamente mejor, mientras que el rendimiento de los clasificadores Jrip y Decision Stump es contradictorio con el esperado con el componente Principal.

CUADRO IV. EVALUACIÓN DEL RENDIMIENTO DE LA RELACIÓN DE ATRIBUTOS DE GANANCIA UTILIZANDO LA RECUPERACIÓN DE PRECISIÓN Y LA MEDIDA F EN EL CONJUNTO DE DATOS 1 [18]

FS Clasificación Algoritmo	Precisión	Recuerdo	F La medida
GR-BN	0,716	0,715	0.716
GR-NB	0,66	0,66	0.654
GR-NBU	0,66	0,66	0.654
GR-MLP	0,768	0,757	0.758
GR-SL	0,715	0,708	0.709
GR-SMO	0,741	0,736	0.737
GR-DT	0,71	0,701	0.702
GR-Jrip	0,691	0,688	0.688
GR-OneR	0,611	0,583	0.571
GR-PARTE	0,646	0,646	0.645
GR-DS	0,373	0,528	0.437
GR-J48	0,709	0,708	0.707
GR-RF	0,741	0,736	0.737
GR-RT	0,738	0,729	0.73
GR-RepT	0,651	0,653	0.651

TABLA V. RESULTADOS DE LOS PRINCIPALES COMPONENTES DEL CONJUNTO DE DATOS 1 UTILIZANDO DISTINTOS CLASIFICADORES [18]

FS Clasificación Algoritmo	Precisión	Recuerdo	Medida F
PC-BN	0.643	0.632	0.633
PC-NB	0,508	0,507	0.506
PC-NBU	0,508	0,507	0.506
PC-MLP	0,694	0,694	0.693
PC-SL	0,692	0,688	0.688
PC-SMO	0,745	0,736	0.737
PC-DT	0,633	0,618	0.617
PC-Jrip	0,57	0,549	0.545
PC-OneR	0,445	0,444	0.445
PARTE DE PC	0,591	0,59	0.591
PC-DS	0,345	0,486	0.403
PC-J48	0,674	0,667	0.668
PC-RF	0,701	0,694	0.695
PC-RT	0,585	0,576	0.576
PC-Representante	0.659	0,66	0.659

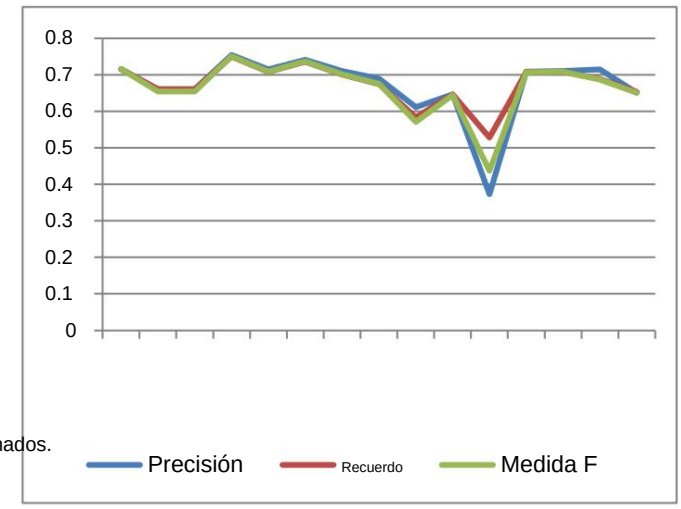

Fig. 5. Rendimiento de GainRatioAttributeEval utilizando el conjunto de datos 1.

La Tabla VI y la Fig. 7 presentan el resultado de ReliefAttributeEval(Rel) utilizando diferentes clasificadores. Se observa a través del análisis de resultados que los clasificadores de Random Forest muestran mejores resultados con ReliefAttributeEval, sin embargo, el clasificador Decision Stump (DS) muestra un rendimiento deficiente con ReliefAttributeEval utilizando el conjunto de datos 1 de los registros de los estudiantes.

B. Comparación de resultados en el conjunto de datos 1 y el conjunto de datos 2

La comparación entre las instancias clasificadas correctamente utilizando el conjunto de datos 1 y el conjunto de datos 2 se ilustra en la Tabla VII. En esta tabla se presentan solo seis clasificadores que se desempeñaron mejor en comparación con los otros clasificadores. Los resultados indican una diferencia significativa en el rendimiento utilizando ambos conjuntos de datos. Hay aproximadamente una diferencia de rendimiento y precisión del 10 al 20 % con cada uno de los algoritmos FS.

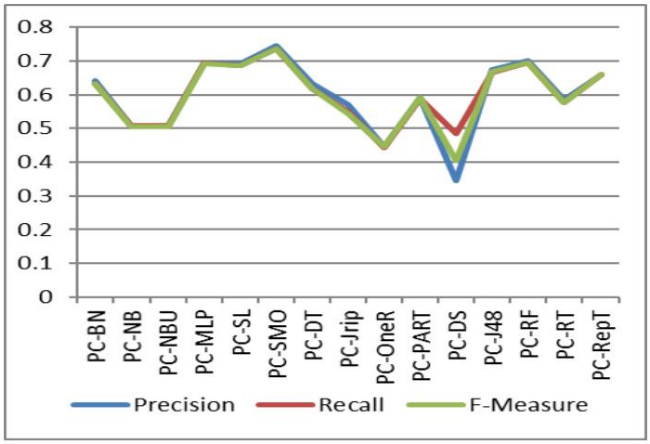

Fig. 6. Precisión, recuperación y medida F de componentes principales.

CUADRO VI. RESULTADOS DEL ATRIBUTO DE RELIEVE EN EL CONJUNTO DE DATOS 1 UTILIZANDO DIFERENTES CLASIFICADORES [18]

Clasificación FS Algoritmo	Precisión	Recuerdo	Medida F
Rel-BN	0,716	0,715	0.716
Rel-NB	0,66	0,66	0.654
Rel-NBU	0,66	0,66	0.654
Rel-MLP	0,767	0,764	0.764
Rel-SL	0,715	0,708	0.709
Rel-SMO	0,741	0,736	0.737
Rel-DT	0,71	0,701	0.702
Rel-Jip	0,713	0,708	0.708
Rel-OneR	0,611	0,583	0.571
Rel-PARTE	0,646	0,646	0.645
Rel-DS	0,373	0,528	0.437
Rel-J48	0,709	0,708	0.707
Rel-RF	0,756	0,75	0.873
Rel-RT	0,665	0,66	0.657
Rel-RepT	0,651	0,653	0.651

1) Precisión de los algoritmos de selección de características: la selección de características de relieve y el algoritmo Chi-Square con clasificador MLP proporcionan la máxima precisión utilizando el conjunto de datos 1. Mientras que el conjunto de datos 2 se usa con la técnica de selección de características chi en combinación con el algoritmo de clasificación Bayes Net (BN) ofrece la máxima precisión . La técnica de reducción de características de componentes principales en combinación con Naïve Bayes (NB), proporciona la menor precisión en el conjunto de datos 1. Aunque otros FS seleccionados Las técnicas en combinación con el algoritmo del árbol de decisión exhiben la menor precisión. Por lo tanto, el rendimiento general se degrada para el conjunto de datos 1 con la combinación de la técnica FS y los clasificadores del árbol de decisiones (DT). Del mismo modo, el algoritmo Chi-cuadrado FS con árbol de decisión da como resultado un rendimiento mínimo en el conjunto de datos 2. Se concluye a partir de las medidas de precisión ilustradas en la Tabla VII que el rendimiento es mejor con 16 características del conjunto de datos 1 que con las 24 características del conjunto de datos 2.

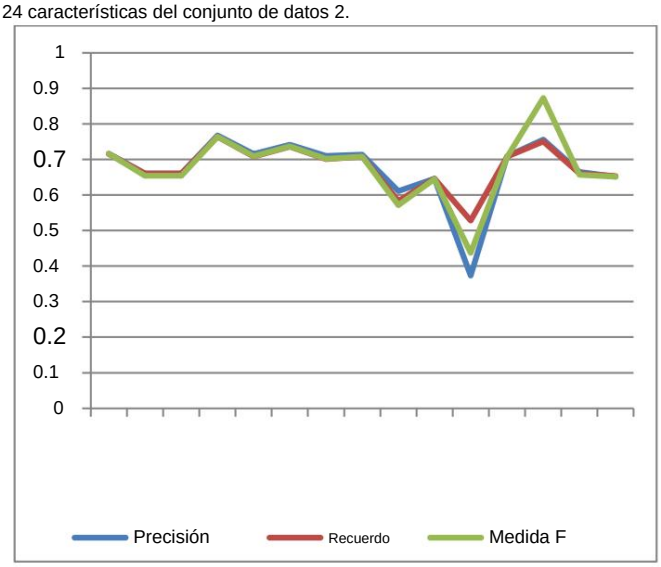

Fig. 7. Rendimiento de ReliefAttributeEval utilizando Dataset1.

En la Tabla VIII se presenta un análisis comparativo basado en el número de características seleccionadas en el conjunto de datos 1 y el conjunto de datos 2 con respecto a la precisión. La técnica chi-cuadrado FS con clasificadores Mlp da como resultado la máxima precisión utilizando el conjunto de datos 1, mientras que el algoritmo Cfs junto con Bayes Net y Naïve Bayes proporciona la máxima precisión utilizando el conjunto de datos 2. Sin embargo, el rendimiento de las técnicas FS con el algoritmo de clasificación del árbol de decisión se degrada utilizando el conjunto de datos 1 y 2. El análisis de rendimiento discutido responde a las dos preguntas de investigación discutidas en la Sección III. Estos resultados dan respuesta a dos preguntas de investigación.

RQ1. ¿Cuáles son las técnicas de selección de características importantes para predecir el desempeño de los estudiantes?

Se concluye de las Tablas VII y VIII que el desempeño de las técnicas FS ha mejorado utilizando el conjunto de datos 1 en comparación con el conjunto de datos 2. La técnica de selección de características de relieve y el algoritmo Chi cuadrado funcionan mejor en el conjunto de datos 1. Mientras que las técnicas de selección de características Chi cuadrado y Cfs funcionan mejor en el conjunto de datos 2. Por lo tanto, estas técnicas deben tenerse en cuenta al predecir el rendimiento de los estudiantes.
Según el análisis, relieve, chi-cuadrado y cfs son técnicas importantes de FS para predecir el rendimiento del estudiante.

RQ2. ¿Cuáles son las mejores combinaciones posibles de técnicas de selección de características y algoritmos de clasificación para predecir el rendimiento de los estudiantes?

Las figuras 8 y 9 muestran que existe una diferencia evidente en los resultados del conjunto de datos 1 y el conjunto de datos 2. Los resultados con el conjunto de datos 1 son mucho mejores que los resultados con el conjunto de datos 2. Ambas figuras presentan una imagen clara de los resultados. .

TABLA VII. EVALUACIÓN DEL RENDIMIENTO DE LOS ALGORITMOS DE SELECCIÓN DE CARACTERÍSTICAS EN EL CONJUNTO DE DATOS 1 Y 2 EN CONTEXTO CON EL % DE INSTANCIAS CLASIFICADAS CORRECTAMENTE

Clasificación FS Técnica	Conjunto de datos1	Conjunto de datos2
Cfs-BN	0,724	0.625
Cfs-NB	0,73	0.625
Cfs-MLP	0,736	0.561
Cfs-SMO	0,668	0.523
Cfs-DS	0,373	0.287
Cfs-RF	0,64	0.614
Chi-BN	0,716	0.616
Chi-NB	0,66	0.597
Chi-MLP	0,769	0.441
Chi-SMO	0,741	0.548
Chi-DS	0,373	0.367
Chi-RF	0,718	0.452
Filtro-BN	0,716	0,61
Filt-NB	0,66	0.614
Filt-MLP	0,768	0.496
Filtro-SMO	0,741	0.534
Filt-DS	0,373	0.287
Filtro-RF	0,741	0.438
GR-BN	0,716	0.559
GR-NB	0,66	0.555
GR-MLP	0,754	0.506
GR-SMO	0,741	0.519
GR-DS	0,373	0.287
GR-RF	0,71	0.565
PC-BN	0,643	0.367
PC-NB	0,508	0.488
PC-MLP	0,694	0.436
PC-SMO	0,745	0.495
PC-DS	0,345	0.28
PC-RF	0,701	0.363
Rel-BN	0,716	0.58
Rel-NB	0,66	0.596
Rel-MLP	0,767	0.439
Rel-SMO	0,741	0.444
Rel-DS	0,373	0.287
Rel-RF	0,756	0.499

TABLA VIII. EVALUACIÓN DEL RENDIMIENTO DE LOS ALGORITMOS DE SELECCIÓN DE CARACTERÍSTICAS EN EL CONJUNTO DE DATOS 1 Y 2 EN CONTEXTO CON % DE CORRECTAMENTE INSTANCIAS CLASIFICADAS

Clasificación FS Técnica	Conjunto de datos1	Conjunto de datos2
Cfs-BN	74,31	57.84
Cfs-NB	72,08	55.88
Cfs-MLP	72,92	57.84
Cfs-SMO	66,67	55.88
Cfs-DS	52,78	42.51
Cfs-RF	63,19	59.8
Chi-BN	71,52	61.33
Chi-NB	65,97	59.33
Chi-MLP	76,39	44.33
Chi-SMO	73,61	55
Chi-DS	52,78	42
Chi-RF	71,53	45.33
Filtro-BN	71,53	59.8
Filt-NB	65,97	59.8
Filt-MLP	75,69	48.03
Filtro-SMO	73,61	51.96
Filt-DS	52,78	42.15
Filtro-RF	73,61	42.15
GR-BN	71,53	56.33
GR-NB	65.97	55.66
GR-MLP	75	51
GR-SMO	65,97	54.3
GR-DS	52,78	42.15
GR-RF	70,83	55.88
PC-BN	63,19	45.09
PC-NB	50,69	51.96
PC-MLP	69,44	45.09
PC-SMO	73,61	49.01
PC-DS	48,61	43.13
PC-RF	69,44	47.05
Rel-BN	71,53	55.88
Rel-NB	65,97	53.92
Rel-MLP	76,39	46.07
Rel-SMO	73,61	48.03
Rel-DS	52,78	42.15

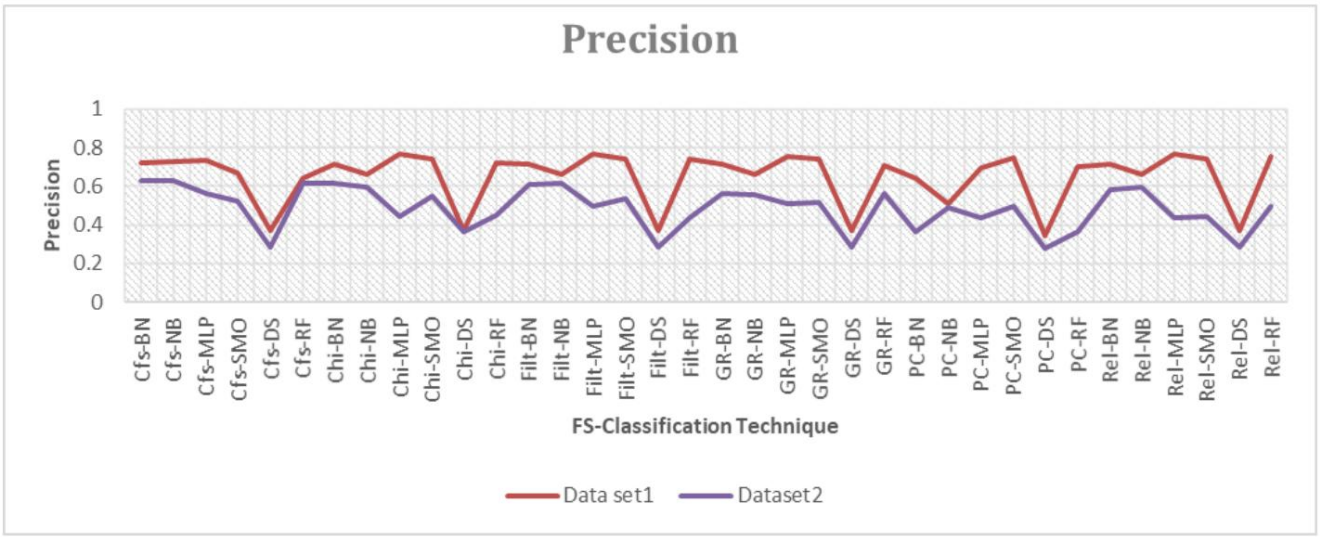

Fig. 8. Comparación de la exactitud de la precisión utilizando los conjuntos de datos 1 y 2.

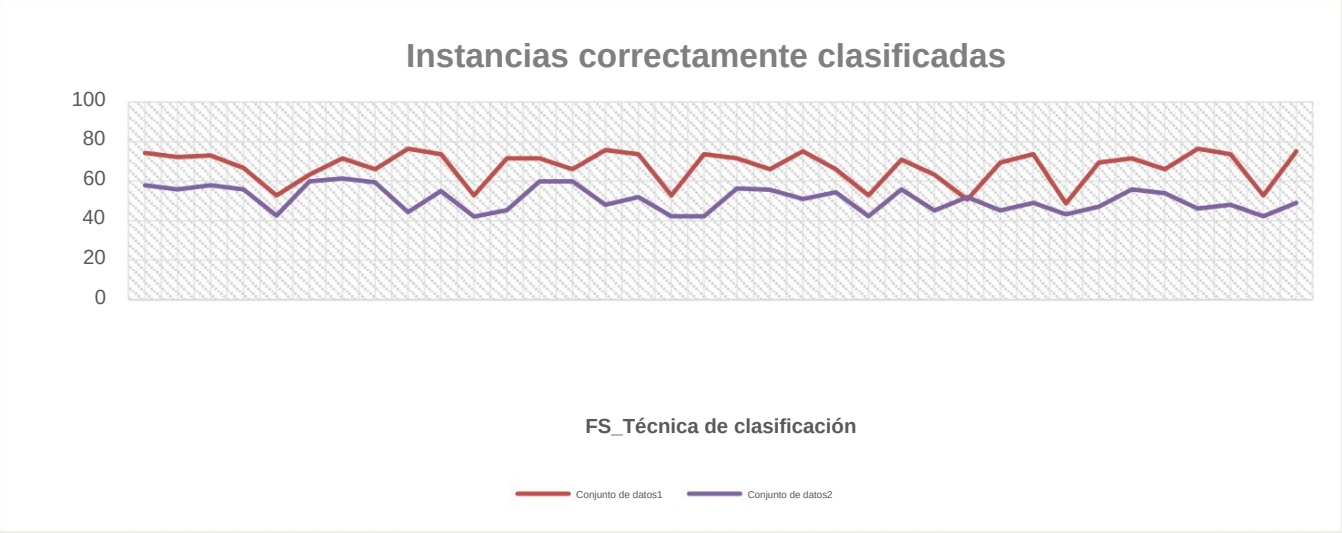

Fig. 9. Comparación de instancias correctamente clasificadas utilizando los conjuntos de datos 1 y 2.

V. CONCLUSIÓN

Este artículo presenta el estudio de varios algoritmos de selección de características y analiza su rendimiento utilizando dos conjuntos de datos diferentes. Los resultados indicaron que existe una diferencia significativa en el rendimiento de los algoritmos de selección de características que utilizan conjuntos de datos con diferentes números de características; muestra una diferencia del 10 al 20 por ciento en los porcentajes de precisión. El rendimiento de las técnicas de selección de características del filtro se reduce a medida que aumenta el número de características. Para predecir el rendimiento académico del estudiante, al tener una gran cantidad de conjuntos de características, también se pueden evaluar las técnicas de selección de características de los contenedores. En el futuro también evaluaremos los resultados de la selección de características a través de la confusión m. Además, no podemos pasar por alto las ventajas de las técnicas de selección de características de filtro. En el futuro, el estudio se puede mejorar aplicando algunos algoritmos de selección de características híbridas en conjuntos de datos de estudiantes para predecir el rendimiento del estudiante.

REFERENCIAS

[1] E. Osmanbegoviÿ, M. Suljiÿ y H. Agiÿ, "DETERMINACIÓN DEL FACTOR DOMINANTE PARA LA PREDICCIÓN DEL RENDIMIENTO DE LOS ESTUDIANTES MEDIANTE EL USO DE ALGORITMOS DE CLASIFICACIÓN DE MINERÍA DE DATOS", Tranzicija, vol. 16, págs. 147-158, 2015.

[2] AM Shahiri y W. Husain, "Una revisión sobre cómo predecir el desempeño de los estudiantes usando técnicas de minería de datos", Procedia Computer Science, vol. 72, págs. 414-422, 2015.

[3] C. Romero y S. Ventura, "Extracción de datos educativos: una revisión del estado del arte", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, págs. 601-618, 2010.

[4] M. Ramaswami y R. Bhaskaran, "Un estudio sobre las técnicas de selección de funciones en la minería de datos educativos", versión preliminar de arXiv arXiv:0912.3924, 2009.

[5] A. Mueen, B. Zafar y U. Manzoor, "Modeling and Predicting Students' Rendimiento académico utilizando técnicas de minería de datos", Revista internacional de educación moderna y ciencias de la computación, vol. 8, p. 36, 2016.

[6] M. Ramaswami y R. Rathinasabapathy, "Predicción del rendimiento de los estudiantes", Revista internacional de inteligencia computacional e informática, vol. 1, 2012.

[7] NT Nghe, P. Janecek y P. Haddawy, "Un análisis comparativo de técnicas para predecir el rendimiento académico", en Conferencia Frontiers In Education-Ingeniería global: conocimiento sin fronteras, oportunidades sin pasaportes, 2007. FIE'07. 37° Anual, 2007, págs. T2G-7-T2G-12.

[8] P. Golding y O. Donaldson, "Predicting academic performance", en Conferencia sobre las fronteras en la educación, 36th Annual, 2006, pp. 21-26.

[9] HM Harb y MA Moustafa, "Selección de un subconjunto óptimo de funciones para el modelo de desempeño de los estudiantes", Int J Comput Sci, p. 5, 2012.

[10] M. Doshi, "Técnica de selección de características basada en la correlación (Cfs) para predecir el rendimiento de los estudiantes", Revista internacional de redes informáticas y comunicaciones, vol. 6, pág. 197, 2014.

[11] W. Punlumjeak y N. Rachburee, "Un estudio comparativo de las técnicas de selección de funciones para clasificar el rendimiento de los estudiantes", en Tecnología de la información e ingeniería eléctrica (ICITEE), 2015 7th International Conference on, 2015, pp. 425-429.

[12] D. Koller y M. Sahami, "Hacia una selección óptima de características", Stanford InfoLab 1996.

[13] P. Mitra, C. Murthy y SK Pal, "Selección de características no supervisada mediante similitud de características", transacciones IEEE sobre análisis de patrones e inteligencia artificial, vol. 24, págs. 301-312, 2002.

[14] A. Figueira, "Predicción de calificaciones por análisis de componentes principales: un enfoque de minería de datos para el análisis del aprendizaje", en Advanced Learning Technologies (ICALT), 2016 IEEE 16th International Conference on, 2016, pp. 465-467.

[15] S. Sivakumar, S. Venkataraman y R. Selvaraj, "Modelado predictivo de indicadores de deserción estudiantil en la minería de datos educativos utilizando un árbol de decisión mejorado", Indian Journal of Science and Technology, vol. 9, 2016.

[16] KW Stephen, "Modelo de minería de datos para predecir la inscripción de estudiantes en cursos STEM en instituciones de educación superior", 2016.

[17] N. Rachburee y W. Punlumjeak, "Una comparación del enfoque de selección de características entre greedy, IG-ratio, Chi-square y mRMR en minería educativa", en Tecnología de la información e ingeniería eléctrica (ICITEE), 2015 7th International Conference on, 2015, págs. 420-424.

[18] M. Zaffar, MA Hashmani y K. Savita, "Análisis de rendimiento del algoritmo de selección de funciones para la minería de datos educativos", en Big Data and Analytics (ICBDA), 2017 IEEE Conference on, 2017, pp. 7-12.

[19] EA Amrieh, T. Hamtini e I. Aljarah, "Extracción de datos educativos para predecir el rendimiento académico de los estudiantes mediante métodos de conjunto" Revista internacional de teoría y aplicación de bases de datos, vol. 9, págs. 119-136, 2016.

[20] S. Hussain, NA Dahan, FM Ba-Alwi y N. RIBATA, "Extracción de datos educativos y análisis del rendimiento académico de los estudiantes mediante WEKA", Revista indonesia de ingeniería eléctrica e informática, vol. 9, 2018.

[21] K. Patel, J. Vala y J. Pandya, "Comparación de varios algoritmos de clasificación en conjuntos de datos de iris usando WEKA", Int. j adv. Ing. Res. Dev. (IJAERD), vol. 1, 2014.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann e IH Witten, "El software de minería de datos WEKA: una actualización", boletín de exploraciones ACM SIGKDD, vol. 11, págs. 10-18, 2009.

[23] A. Kalousis, J. Prados y M. Hilario, "Estabilidad de los algoritmos de selección de características: un estudio sobre espacios de alta dimensión", Sistemas de conocimiento e información, vol. 12, págs. 95-116, 2007.

[24] MA Hall y LA Smith, "Selección práctica de subconjuntos de características para aprendizaje automático", 1998.

[25] C. Anuradha y T. Velmurugan, "Técnicas de selección de funciones para analizar el rendimiento académico de los estudiantes mediante el clasificador Naïve Bayes", en la 3.ª Conferencia internacional sobre pequeñas y medianas empresas, 2016, págs. 345-350.

[26] C. Huertas y R. Juárez-Ramírez, "Comparación de rendimiento de selección de características de filtro en datos de alta dimensión: un análisis teórico y empírico análisis de los algoritmos más populares", en Information Fusion (FUSION), 2014 17th International Conference on, 2014, pp. 1-8.

[27] J. Novakoviÿ, "Hacia una selección de características óptima utilizando métodos de clasificación y algoritmos de clasificación", Yugoslav Journal of Operations Research, vol. 21, 2016.

[28] Q. Guo, W. Wu, D. Massart, C. Boucon y S. De Jong, "Selección de características en el análisis de componentes principales de datos analíticos" Chemometrics and Intelligent Laboratory Systems, vol. 61, págs. 123-132, 2002.

[29] K. Kira y LA Rendell, "El problema de selección de características: Métodos tradicionales y un nuevo algoritmo", en Aaai, 1992, pp. 129-134.

[30] T. Velmurugan y C. Anuradha, "Evaluación del rendimiento de los algoritmos de selección de características en la minería de datos educativos", Evaluación del rendimiento, vol. 5, 2016.