UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

MINADO PRELIMINAR DE DATOS DE LOS INTERESES ACADÉMICOS DE ESTUDIANTES DE LICENCIATURA

TESIS PARA OBTENER EL GRADO DE:

MAESTRA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

FÁTIMA GUADALUPE MONTEJO COLLADO

BAJO LA DIRECCIÓN DE:

DR. JOSÉ HERNÁNDEZ TORRUCO

EN CODIRECCIÓN:

DR. OSCAR ALBERTO CHÁVEZ BOSQUEZ

CUNDUACÁN, TABASCO, A: AGOSTO 2025

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

MINADO PRELIMINAR DE DATOS DE LOS INTERESES ACADÉMICOS DE ESTUDIANTES DE LICENCIATURA

TESIS PARA OBTENER EL GRADO DE:

MAESTRA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

FÁTIMA GUADALUPE MONTEJO COLLADO

BAJO LA DIRECCIÓN DE:

DR. JOSÉ HERNÁNDEZ TORRUCO

EN CODIRECCIÓN:

DR. OSCAR ALBERTO CHÁVEZ BOSQUEZ

CUNDUACÁN, TABASCO, A: AGOSTO 2025

Declaración de Autoría y Originalidad

En la Ciudad de Cunduacán el día 12 del mes de agosto del año 2025, el que suscribe **Fátima Guadalupe Montejo Collado**, alumna del Programa de la **Maestría en Ciencias de la Computación** con número de matrícula **222H21001**, adscrito a la **División Académica de Ciencias y Tecnologías de la Información**, de la Universidad Juárez Autónoma de Tabasco, como autor de la Tesis presentada para la obtención de Grado y maestría y titulada **Minado preliminar de datos de los intereses académicos de estudiantes de licenciatura**, dirigida por el Dr. José Hernández Torruco y el Dr. Oscar Alberto Chávez Bosquez.

DECLARO QUE: La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita. Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente.

Cunduacán, Tabasco a 12 de agosto de 2025.

Estudiante: Fátima Guadalupe Montejo Collado

UJAT
UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
"ESTUDIO EN LA DUDA, ACCIÓN EN LA FE"

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

2025 AÑO DE LA Mujer Indígena

Cunduacán, Tabasco, a 11 de agosto de 2025
Oficio No. 1388/2025/DACYTI/D

Asunto: Autorización de impresión de Tesis

C. Fátima Guadalupe Montejo Collado
Egresada de la Maestría en Ciencias de la Computación

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional ***"Minado preliminar de datos de los intereses académicos de estudiantes de licenciatura"***, para presentar examen y obtener el Grado de Maestra en Ciencias de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

Atentamente

Dr. Óscar Alberto González Gonzalez
Director

C.c.p. Dr. Eddy Arquímedes García Alcocer. - Encargado del Despacho de la Coordinación de Posgrado DACYTI
Archivo.
Consecutivo.

M.T.E. OAGG/EAGA

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690.
Cunduacán, Tabasco, México.
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx

Carta de Cesión de Derechos

Villahermosa, Tabasco a 12 de agosto de 2025.

Por medio de la presente manifiesto haber colaborado como AUTOR en la producción, creación y/o realización de la obra denominada: **Minado preliminar de datos de los intereses académicos de estudiantes de licenciatura**.

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y/o realización de la obra antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entendemos y aceptamos el alcance del artículo en mención de que tenemos el derecho al reconocimiento como autores de la obra, y a la Universidad Juárez Autónoma de Tabasco mantendrá en un 100 % la titularidad de los derechos patrimoniales por un período de 20 años sobre la obra en la que colaboramos, por lo anterior, cedemos el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADOR

Estudiante: Fátima Guadalupe Montejo Collado

TESTIGOS

Dr. José Hernández Torruco

Dr. Oscar Alberto Chávez Bosquez

Índice general

Índice de figuras

Índice de tablas

Resumen

Esta investigación se enfoca en identificar los intereses académicos de los estudiantes universitarios de la División Académica de Ciencias y Tecnologías de la Información (DACYTI) de la Universidad Juárez Autónoma de Tabasco (UJAT). Esta División cuenta con cuatro carreras activas, por lo que se diseñó un cuestionario como instrumento de recolección de datos por carrera, basados en el mapa curricular. Los cuestionarios fueron elaborados con la finalidad de recabar información precisa y contextualizada sobre las preferencias, motivaciones y áreas de interés académico de los estudiantes. Los instrumentos fueron enviados y aprobados por el Comité de Ética de la Universidad. La muestra estuvo conformada por 152 estudiantes, quienes debían tener al menos el 40 % de avance curricular. Para el análisis de la información se aplicaron técnicas de minería de datos, y para tareas de clasificación se utilizaron algoritmos como: *Random Forest*, J48, JRip, OneR, *K-Nearest Neighbors* (KNN), Máquinas de Vectores de Soporte (SVM) con diferentes kernels. También se aplicaron filtros de selección de variables relevantes: CFS (*Correlation-based Feature Selection*), *Chi Squared*, *Information Gain* y *Random Forest*. Para evaluar cada modelo se tomaron en cuenta las métricas: *Accuracy*, Kappa, Sensibilidad, Especificidad y *Balanced Accuracy*. Con base en los experimentos realizados, el clasificador con mejores resultados en todos los experimentos fue *Random Forest* y el filtro que menos variables relevantes generó fue CFS. Algunas de las variables más frecuente en los experimentos que mostró mayor relevancia fue la calificación del área de interés y la experiencia previa que los estudiantes tenían antes de ingresar a la carrera. Esto sugiere que tanto el rendimiento académico en asignaturas como la experiencia previa a temas del área influyen significativamente en los intereses académicos de los alumnos. Para futuras investigaciones, se propone emplear metaheurísticas y ensamble de clasificadores, así como también realizar un análisis profundo por carrera balanceando los datos de las clases.
Palabras clave: Inteligencia Artificial, Minería de datos, Intereses académicos.

Abstract

This research focuses on identifying the academic interests of university students from the Academic Division of Information Sciences and Technologies (DACYTI) at the Juárez Autonomous University of Tabasco (UJAT). This Division offers four active degree programs, and therefore, a questionnaire was designed as a data collection instrument for each program, based on the corresponding curriculum. The questionnaires were developed with the aim of gathering precise and contextualized information about students' preferences, motivations, and areas of academic interest. These instruments were submitted to and approved by the University's Ethics Committee. The sample consisted of 152 students who had completed at least 40 % of their academic coursework. Data analysis techniques from the field of data mining were applied, and for classification tasks, the following algorithms were used: Random Forest, J48, JRip, OneR, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) with different kernels. Feature selection filters were also applied, including CFS (Correlation-based Feature Selection), Chi-Squared, Information Gain, and Random Forest. To evaluate each model, the following metrics were considered: Accuracy, Kappa, Sensitivity, Specificity, and Balanced Accuracy. Based on the experiments conducted, the classifier that performed best across all experiments was Random Forest, while the filter that produced the fewest relevant variables was CFS. Some of the most frequent variables that showed higher relevance in the experiments were the grade in the area of interest and the prior experience students had before entering the program. This suggests that both academic performance in specific subjects and prior exposure to topics related to the field significantly influence students' academic interests.

For future research, it is proposed to use metaheuristic techniques and classifier ensembles, as well as to conduct an in-depth analysis by degree program, balancing class data accordingly.

Keywords: Artificial Intelligence, Data Mining, Academic Interests.

Capítulo 1

Generalidades

1.1. Introducción

En los últimos años, la minería de datos (*Data mining*) ha experimentado un gran crecimiento en el ámbito académico. Las instituciones se han dado cuenta que la gran cantidad de datos almacenados en sus sistemas se pueden analizar y utilizar para obtener nuevos conocimientos a partir de ellos (Moine et al., 2011). Comprender los intereses académicos de los estudiantes universitarios representa un factor clave para mejorar la orientación vocacional, el diseño curricular y las estrategias pedagógicas dentro de las instituciones educativas. Identificar de manera temprana estos intereses permite no solo optimizar la permanencia y el rendimiento estudiantil, sino también alinear los perfiles de egreso con las demandas del entorno profesional.

La minería de datos implica la identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles, a partir de conjuntos de datos (Ballesteros Román et al., 2013). Su objetivo principal consiste en descubrir información implícita u oculta que mediante los métodos estadísticos tradicionales no es posible obtener, encontrar patrones que en ocasiones no son visibles para el ser humano (Moine et al., 2011).

En este trabajo se recopilaron y analizaron datos relacionados con los intereses académicos de estudiantes de licenciatura de la DACyTI, mediante la aplicación de técnicas de Minería de datos. El objetivo principal fue identificar los intereses de los estudiantes en torno a asignaturas específicas, áreas de especialización (asignaturas optativas), temas de tesis, campos laborales

vinculados al ejercicio profesional y preferencias en estudios de posgrado.

La investigación se centró en las etapas iniciales del proceso, que incluyeron el diseño y validación de un instrumento de encuesta, su aplicación para la recolección de datos, el análisis exploratorio y el preprocesamiento de la información, así como la obtención de resultados estadísticos preliminares.

Este proyecto aporta múltiples beneficios en el ámbito computacional, destacando principalmente en el desarrollo y aplicación de técnicas de minería de datos para el análisis educativo. En primer lugar, promueve el uso de algoritmos de clasificación supervisada y métodos de selección de variables relevantes en un contexto real, permitiendo validar su eficacia en la identificación de patrones relacionados con los intereses académicos de los estudiantes.

Además, fomenta la integración de procesos computacionales como el preprocesamiento de datos, la reducción de dimensionalidad y la evaluación de modelos mediante métricas estándar, lo cual fortalece las competencias técnicas en ciencia de datos aplicada a la educación. El flujo de trabajo propuesto, desde la recolección de datos hasta el análisis exploratorio y la modelación, puede ser replicado en otros contextos educativos.

Asimismo, este trabajo contribuye a la generación de conocimiento sobre cómo variables académicas, curriculares y personales influyen en la toma de decisiones estudiantiles, lo que puede ser aprovechado en el desarrollo de sistemas inteligentes de recomendación académica o plataformas de apoyo a la orientación vocacional.

1.2. Planteamiento del problema

1.2.1. Definición del problema

En el ámbito universitario, comprender los intereses académicos reales de los estudiantes resulta esencial para diseñar programas educativos relevantes, fomentar el aprendizaje y disminuir los índices de deserción escolar. No obstante, dichos intereses tienden a ser variados, cambiantes, y en muchas ocasiones no se expresan de forma directa, lo que complica su detección mediante métodos tradicionales de recolección y análisis de información.

A pesar de contar con mecanismos como entrevistas o cuestionarios directos, muchas veces

los patrones subyacentes en las preferencias estudiantiles pasan desapercibidos. Por ello, se requiere un enfoque que permita analizar grandes volúmenes de datos de forma automática para detectar correlaciones y tendencias no evidentes a simple vista.

Este proyecto tiene como objetivo la creación de un conjunto de datos a partir de encuestas aplicadas a estudiantes de licenciatura de la DACyTI, con la finalidad de implementar técnicas de minería de datos que permitan identificar patrones ocultos en sus intereses académicos. El propósito es proporcionar a la institución educativa una herramienta analítica que facilite la toma de decisiones estratégicas en relación con la oferta de cursos, diplomados y programas de posgrado, alineándolos con los intereses reales de los estudiantes.

De esta forma, se espera que el aprovechamiento escolar aumente, que se diseñen programas más atractivos y que se reduzca la tasa de abandono escolar, contribuyendo así a una educación más personalizada y eficaz.

1.2.2. Delimitación de la investigación

- Los datos serán recopilados de los alumnos de licenciatura que tengan al menos el 40 % de avance curricular de las siguientes carreras: Ingeniería en Informática Administrativa (IIA), Ingeniería en Sistemas Computacionales (ISC), Licenciatura en Sistemas Computacionales (LSC) y Licenciatura en Tecnologías de la Información (LTI), de la División Académica de Ciencias y Tecnologías de la información.
- Los datos de las calificaciones y la veracidad de los datos depende de las respuestas de los alumnos en la encuesta.

1.3. Preguntas de investigación

1. ¿Cuáles son los patrones preliminares en el dataset construido a partir de los intereses académicos de los alumnos de licenciatura?
2. ¿Que técnicas de Minería de datos permiten encontrar los patrones antes mencionados eficazmente?

1.4. Objetivo general

Realizar un análisis preliminar de los intereses académicos de estudiantes de licenciatura aplicando minería de datos.

1.5. Objetivos específicos

1. Diseñar el instrumento y aplicar la encuesta.
2. Crear un dataset de los intereses académicos de estudiantes de licenciatura.
3. Descubrir patrones preliminares en el dataset creado.

1.6. Justificación

Contar con un modelo basado en los intereses académicos de los estudiantes de licenciatura permitirá a la institución ampliar y optimizar su oferta educativa. Esto incluye la creación de nuevos cursos optativos, talleres, diplomados y programas de posgrado que respondan a las áreas de mayor interés identificadas. Asimismo, este modelo podrá ser útil para revisar y actualizar los contenidos de asignaturas existentes, incorporar nuevas líneas de especialización, o fortalecer aquellas que resulten estratégicas según las preferencias del alumnado. En conjunto, estas acciones contribuirán a una mejor alineación entre la formación académica y las expectativas profesionales de los estudiantes, así como a una oferta más atractiva, pertinente y flexible.

Darles la importancia pertinente a los intereses académicos de los alumnos permitirá que ellos tomen la iniciativa por estudiar un diplomado o un posgrado de su preferencia y esto impactará positivamente a la sociedad, ya que el formar buenos profesionales en especialidades de su interés hará que sean buenos en sus trabajos, que puedan desarrollarse plenamente en el ámbito laboral y social.

Para la construcción del modelo fue necesario desarrollar un instrumento de recolección de datos en forma de encuesta, con el cual se obtuvo un conjunto de datos (dataset) representativo. Posteriormente, se llevó a cabo un proceso de preparación y depuración de los datos, así como

la identificación de las variables más relevantes, lo cual permitió estructurar adecuadamente la información para su análisis mediante técnicas de minería de datos.

1.7. Metodología utilizada

A continuación, se describen los métodos y técnicas empleados durante el proceso de investigación para alcanzar los objetivos propuestos.

1. **Revisión y análisis de la literatura.** Se realizará una revisión sistemática de la literatura para conocer y entender los factores en el interés académico de los estudiantes de licenciatura. La información será consultada en artículos científicos recientes y artículos considerados clásicos en la literatura.

2. **Diseñar y validar una encuesta.** Para recopilar los datos relacionados con los intereses académicos de estudiantes de licenciatura, se diseñará la encuesta aplicando las metodologías y normativas adecuadas para el caso, así como seleccionando las preguntas adecuadas. La encuesta se validará con los instrumentos correspondientes.

3. **Recopilación de los datos mediante la encuesta.** Se aplicará la encuesta en la población especificada respetando los protocolos correspondientes.

4. **Comprensión de los datos recopilados.** Se aplicarán técnicas estadísticas descriptivas y de visualización de datos para un análisis exploratorio y comprensión de los datos.

5. **Preparación de los datos.** Se aplicarán técnicas de limpieza de datos, tratamiento de datos faltantes, transformación de datos e ingeniería de atributos con el objetivo de preparar el dataset para la etapa de modelado.

6. **Descubrimiento preliminar de patrones.** Se aplicarán técnicas para descubrir asociaciones entre variables del dataset.

7. **Análisis de los resultados.** Se describirán los resultados obtenidos.

Capítulo 2

Marco teórico

2.1. Conceptos y teorías fundamentales de la investigación

- ***Minería de datos***. Se puede definir originalmente como el proceso de examinar grandes cantidades de datos para descubrir nuevas relaciones, patrones y tendencias importantes (César Pérez López, 2007). La minería de datos emplea una combinación de técnicas semiautomáticas de inteligencia artificial, análisis estadístico, manipulación de bases de datos y representación gráfica, para descubrir información que no se encuentra limpiamente declarada en los datos (Martínez, 2001).

- ***Aprendizaje automático* (Machine Learning)**. Es un área de la inteligencia artificial enfocada en crear algoritmos y sistemas capaces de mejorar su desempeño de manera autónoma. Su propósito principal es detectar regularidades en la información del pasado para construir modelos que faciliten la predicción de eventos o comportamientos futuros (LLevot, 2020).

  En la Figura 2.1 se muestra una imagen de la relación entre aprendizaje profundo, aprendizaje automático y la inteligencia artificial.

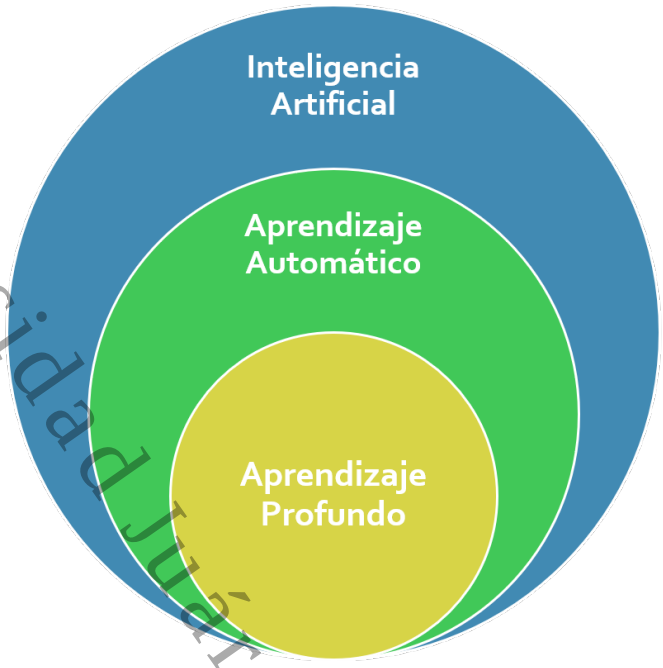

Figura 2.1. Representación gráfica de la relación del Aprendizaje profundo, Aprendizaje automático y la Inteligencia artificial (**?**).

- **Redes neuronales**. Una red neuronal se define como un conjunto de elementos de procesamiento de la información altamente interconectados, que son capaces de aprender con la información que se alimenta. Su principal característica es que puede aplicarse a un gran número de problemas que pueden ir desde problemas complejos reales a modelos teóricos sofisticados, tales como: reconocimiento de imágenes, reconocimiento de voz, análisis y filtrado de señales, clasificación, discriminación, análisis financiero, predicción dinámica, entre otros (César Pérez López, 2007).

- ***Clustering***. Consiste en encontrar grupos de datos que tengan características similares o estén relacionadas entre sí (Tan et al., 2006). Algunos tipos de agrupamiento o clustering son:

  - Agrupamiento particional: Es la división de un conjunto de datos en subconjuntos no superpuestos, es decir se hacen diferentes grupos con los elementos de características similares (Tan et al., 2006).
  - Agrupación jerárquica: Es un método de análisis de un conjunto de clústeres anidados organizados como un árbol jerárquico. Existen dos tipos de agrupamiento jerárquico:

las aglomerativas y las divisivas. Las aglomerativas consisten en un acercamiento ascendente, cada observación inicia en su propio grupo y los pares de grupo son unidos mientras uno sube en la jerarquía, hasta que al final del proceso todos los casos tratados están englobados en un mismo conglomerado. Por otro lado, las divisivas consisten en un acercamiento descendente, todas las observaciones inician en un grupo y se realizan divisiones mientras uno baja en la jerarquía (Berzal, 2025).

- **Modelo predictivo**. Un modelo predictivo es el resultado de un algoritmo que es entrenado con datos para predecir lo que se está buscando en ese proyecto. Una vez entrenado el modelo es probado varias veces con los datos y da como resultado un pronóstico, diagnostico o clasificación basada en los datos que entrenaron al modelo (Rojas, 2020).

- **Dataset** Son un conjunto de datos recolectados durante la ejecución de un proyecto de investigación. Son utilizados para entrenar al sistema que identifica patrones (Aliende, 2017).

- ***Random forest***. El método *random forest* consiste en un modelo de predicción formado por varios árboles de decisión del tipo CART(Classification and Regression Trees), construidos a partir de la técnica bootstrap. El algoritmo toma un conjunto de entrenamiento con *n* observaciones y *m* variables, y de él genera distintas muestras aleatorias con reemplazo. Para cada muestra se entrena un árbol de decisión, considerando únicamente un subconjunto aleatorio de las variables. Como resultado, se obtiene un conjunto de árboles entrenados de forma independiente, lo que permite disminuir la varianza y aumentar la capacidad de generalización del modelo (Liu et al., 2012).

- **J48**. El algoritmo J48 es una implementación en Java del algoritmo C4.5, desarrollado por Ross Quinlan, y ampliamente utilizado en el campo del aprendizaje automático para resolver problemas de clasificación. Este algoritmo construye árboles de decisión basándose en la teoría de la información, utilizando la ganancia de información como criterio principal para la selección de atributos en cada nodo del árbol. J48 es eficaz tanto con datos categóricos como continuos, permite manejar valores faltantes, realiza poda posterior al entrenamiento (post-pruning) para evitar el sobreajuste, y es capaz de generar reglas de decisión interpretables. Su aplicabilidad abarca diversos campos como la medicina, la educación, el comercio

electrónico y el gobierno digital (Quinlan, 1993).

Pseudocódigo de C4.5 / J48

Algoritmo C4.5(D, A)
Entrada:
D: conjunto de datos de entrenamiento
A: conjunto de atributos disponibles

1. Si todas las instancias en D pertenecen a la misma clase: devolver una hoja con esa clase.
2. Si A está vacío o D está vacío: devolver una hoja con la clase mayoritaria en D.
3. Para cada atributo $a \in A$: calcular la ganancia de información normalizada (gain ratio) al dividir D por a.
4. Seleccionar el atributo a_{best} con la mayor gain ratio.
5. Crear un nodo de decisión que divida D según los valores de a_{best}.
6. Para cada valor $v \in$ Valores(a_{best}):
   - Crear el subconjunto $D_v \subseteq D$ donde $a_{\text{best}} = v$
   - Llamar recursivamente a C4.5(D_v, $A - \{a_{\text{best}}\}$)
   - Agregar el resultado como hijo del nodo actual.
7. Aplicar poda al árbol resultante (opcional, incluida en J48).

- **JRip**. JRip (*RIPPER*) es uno de los algoritmos más básicos y populares. Las clases se analizan en orden creciente de tamaño, y se genera un conjunto inicial de reglas para cada clase utilizando el método de reducción de errores incremental. JRip funciona tratando todos los ejemplos de una determinada categoría en los datos de entrenamiento como una clase, y busca un conjunto de reglas que cubra a todos los miembros de dicha clase. Posteriormente, pasa a la siguiente clase y repite el mismo proceso, hasta que todas las clases hayan sido cubiertas (Rajput et al., 2011), la regla tiene la siguiente forma:

$$\textbf{si}\Big(atributo_1 < \text{operador relacional} > valor_1 < \text{operador lógico} > atributo_2 < \text{operador relacional} > valor_2 < \cdots > \Big) \textbf{ entonces } valor\text{-}decisin$$

- ***OneR***. Está basado en el algoritmo ID3, donde la meta principal consiste en adquirir las reglas de clasificación directamente desde el conjunto de datos de entrenamiento, es simple y efectivo, de uso frecuente en el aprendizaje de automático, utiliza un único atributo para la clasificación, el cual es el de menor porciento de error y se obtiene un conjunto de reglas (Cardoso García & Arza Valdés, 2017).

- ***Chi-squared***. La prueba de Chi-cuadrado de independencia, o Chi-cuadrado de Pearson, es una prueba que utiliza tablas de contingencia para determinar si el patrón observado entre dos variables categóricas en la tabla es lo suficientemente fuerte como para considerar que dichas variables están relacionadas (dependen entre sí) (Translational Interventional Radiology, 2023).

- ***Information Gain***. La Ganancia de Información (IG) está basada en la entropía utilizada en aprendizaje automático para evaluar qué tan relevante es una característica respecto a la clase objetivo. Las características con alta IG aportan más información y suelen mejorar la clasificación. Sin embargo, IG no elimina características redundantes, por lo que es necesario aplicar filtros adicionales. Su valor depende de la diferencia de información antes y después de conocer un atributo, y su valor máximo en problemas multiclase es 1 (Win & Kham, 2019).

$$H(Class)+H(Attribute)\text{-}H(Class, Attribute)$$

  donde:
  H(Class): es la entropía de Shannon para una variable;
  H(Class, Attribute): es una articulación Entropía de Shannon para una variable X con una condición a Y.

- **CFS**. La Selección de Características Basada en Correlación (CFS) es una técnica de preprocesamiento que selecciona subconjuntos de atributos evaluando qué tan bien están correlacionados con la clase objetivo y qué tan poco están correlacionados entre sí. Su objetivo es identificar características relevantes y no redundantes para mejorar el rendimiento de los modelos de aprendizaje automático (Hall, 1999).

- **Validación**. Se utiliza una técnica de *cross-validation* o validación cruzada para probar la eficacia de un modelo de *Machine Learning*. Este es un proceso de *re-sampling* (remuestreo) que permite estimar modelos incluso con datos limitados. Se utiliza validación cruzada para comparar los resultados de los algoritmos de clasificación y seleccionar de forma certera el modelo que realice la predicción con mayor precisión para un problema específico. Es una

de las técnicas fácil de comprender, fácil de implementar y genera un mínimo de sesgo que otros métodos (L. Laura & Baluarte, 2017). Algunas técnicas de validación cruzada son:

- *Train-Test*: Consiste en dividir aleatoriamente la serie de datos. Una parte se usa para entrenar el modelo ML y la otra parte se usa para probarlo para su validación. Normalmente, del 70 % al 80 % del conjunto de datos se reserva para entrenamiento. El 20-30 % restante se utiliza para la validación cruzada (scientest, 2025).
- *K-Folds*: La técnica *K-Folds* es fácil de comprender, con respecto a otros enfoques de validación cruzada, resultando en un modelo con menos sesgo. Esto ayuda a garantizar que todas las observaciones en el conjunto de datos original tengan la oportunidad de aparecer tanto en el conjunto de entrenamiento como en el de prueba. Este es uno de los mejores enfoques cuando los datos de entrada son limitados (scientest, 2025).

- **Matriz de confusión**. Una matriz de confusión o tabla de contingencia es una herramienta utilizada para evaluar el desempeño de un modelo de clasificación. Se representa como una tabla que compara las clases reales con las clases predichas por el modelo, permitiendo identificar aciertos y errores en las predicciones. En la tabla se muestra la matriz de confusión creado con un modelo de clasificación binaria.

Tabla 2.1. Matriz de confusión para una clasificación binaria

	Valor	**Real**
Valor Predicho	TP (Verdadero Positivo) FN (Falso Negativo)	FP (Falso Positivo) TN (Verdadero Negativo)

donde:

TP = Verdadero positivo.

FP = Falso positivo.

FN = Falso negativo.

TN = Verdadero negativo.

- ***Accuracy***. El accuracy (exactitud o precisión global) es una métrica utilizada para evaluar el desempeño de un modelo de clasificación. Indica qué proporción de las predicciones del modelo son correctas, considerando tanto las clases positivas como negativas.

$$\text{Precisión} = \frac{TP + TN}{TP + FN + FP + TN}$$

donde:

TP = Verdadero positivo.

FP = Falso positivo.

FN = Falso negativo.

TN = Verdadero negativo.

Un valor cercano a 1 (o 100 %) indica que el modelo clasifica correctamente la mayoría de las instancias. Sin embargo, no siempre es una métrica suficiente, especialmente cuando hay clases desbalanceadas (por ejemplo, 95 % de una clase y 5 % de otra).

- **Especificidad**. La especificidad mide la proporción de los verdaderos negativos que se identificaron correctamente en el modelo.

$$\text{Especificidad} = \frac{TN}{TN + FP}$$

donde:

TN = Verdadero negativo.

FP = Falso positivo.

- **Sensibilidad**. También conocida como recall o verdadero positivo rate, es una métrica que evalúa la capacidad de un modelo de clasificación para identificar correctamente las instancias positivas. Es decir, mide el porcentaje de verdaderos positivos sobre el total de casos que realmente son positivos.

$$\text{Sensibilidad} = \frac{TP}{TP + FN}$$

donde:

TP = Verdadero positivo.

FN = Falso negativo.

- **Kappa**. El índice Kappa de Cohen, mide la coincidencia de la predicción con la clase real (1 significa que ha habido coincidencia absoluta)

$$\textit{Kappa} = \frac{P(O) - P(E)}{1 - P(E)}$$

donde:

P(O): proporción de concordancia observada.

P(E): proporción esperada de veces que k coincidencia es evaluada

- ***Accuracy balanceado***. *Balanced Accuracy* (Precisión Balanceada), es una métrica de evaluación que promedia la sensibilidad (también conocida como recall o true positive rate) de cada clase. Es especialmente útil cuando las clases están desbalanceadas, ya que evita que una clase mayoritaria domine la métrica.

$$\text{Precisión balanceada} = \frac{TN}{TN + FP} + \frac{TP}{TP + FN}$$

donde:

TP = Verdadero positivo.

FP = Falso positivo.

FN = Falso negativo.

TN = Verdadero negativo.

- **Validación de un instrumento de investigación** La validación de un instrumento es el proceso mediante el cual se evalúa si un cuestionario, prueba, escala u otra herramienta de medición realmente mide lo que pretende medir y lo hace de forma precisa, coherente y adecuada para el contexto en el que se va a aplicar. En términos simples, busca comprobar que el instrumento es válido y confiable. Existen diferentes tipos de validez:

  - Validez de contenido: Evalúa si los ítems cubren de manera completa y representativa todas las dimensiones del concepto que se quiere medir.
  - Validez de criterio: Compara los resultados del instrumento con un criterio externo o una medida ya validada.

- Validez de constructo: Verifica si el instrumento realmente refleja el concepto teórico subyacente, usando análisis estadísticos como el análisis factorial.

La validación incluye:

1. Revisión de expertos (validez de contenido).
2. Prueba piloto para ajustar ítems.
3. Análisis de confiabilidad (ej. Alfa de Cronbach).
4. Análisis factorial para verificar la estructura interna.

2.2. Literatura relacionada

La minería de datos y el ML en el ámbito académico ha sido aplicado a diferentes áreas, destancando el rendimiento académico, la deserción estudiantil, entre otros. A continuación, se muestran algunos de los estudios relacionados al presente trabajo.

Los autores de “Técnicas de Machine Learning aplicadas a la evaluación del rendimiento y a la predicción de la deserción de estudiantes universitarios” (Cruz et al., 2022) recopilaron datos de múltiples fuentes que contienen información sobre 54 atributos distintos de cada estudiante. Estos datos muestran que muchos factores, como los académicos, demográficos, psicológicos, de salud, entre otros, influyen significativamente en el abandono escolar, algunas de las técnicas que analizaron son Naïve Bayes, Máquina de Vectores de Soporte (SVM), Árboles de Decisión, Redes Neuronales, *K-Nearest Neighbors* (k-NN), Regresión Logística, Bagging y Boosting.

Por otro lado, (Alsayed et al., 2021) aplicaron técnicas de Aprendizaje supervisado para la selección de la especialidad de pregrado adecuada por parte de los estudiantes. En este artículo se muestran varias técnicas de ML, Árbol de decisión, Clasificadores de árboles adicionales (*Extra tree classifiers* (ETC)), Bosques aleatorios (*Random Forest*), Clasificadores de aumento de gradiente (*Gradient Boosting Classifiers* (GBC)) y Máquina de vectores de soporte (*Support Vector Machine* (SVM)), estos fueron probados para predecir la especialidad de pregrado correcta de los estudiantes antes de la admisión en el nivel de pregrado basado en los mercados laborales actuales y la experiencia. En este estudio, la base de datos contiene 216 registros de estudiantes y 19 características de entrada, este conjunto de datos fue publicado en Kaggle, después convirtieron

esos datos en un formato que fuera aceptable para los modelos de ML. Posteriormente aplicaron varias técnicas de aprendizaje supervisado a el conjunto de datos y las evaluaron utilizando un método de validación cruzada de 10 pliegues. Los hallazgos mostraron que RF y GBC predicen los campos de estudio de los estudiantes con una precisión de 0,75 % y 0,61 %, respectivamente. Los resultados indican que RF y GBC son los clasificadores más apropiados para integrar en el sistema de recomendación de campo inteligente para predecir campos adecuados para los estudiantes de acuerdo con el mercado laboral, porque el rendimiento de estos clasificadores es bueno con menos datos de capacitación.

Por su parte (Rodrıguez-Maya et al., 2017) presentan un modelo de clasificación para predecir la deserción de los alumnos de universidades en México, a partir de la información recolectada del estudiante, (como sus calificaciones) y los puntos obtenidos en el examen de ingreso a la institución EXANI-II, el cual evalúa diferentes factores, personales, académicos, sociales, entre otros. Se realizó un caso de estudio para evaluar este modelo, cuya precisión fue del 86 %. El modelo de clasificación utiliza una metodología de Minería de datos, para aprender modelos basados en Árboles de decisión. Se utilizó un caso de estudio del Instituto Tecnológico de Zitácuaro con información de generaciones del 2010-2015 y 2011-2016.

En otros países también se han utilizado técnicas de ML para abordar el problema, tal como (Zumárraga et al., 2017) se muestran resultados alcanzados por la Universidad Politécnica Salesiana de Ecuador, donde construyeron y validaron una herramienta propia para la evaluación de preferencias e intereses profesionales. La muestra fue de 1027 aspirantes, 48,31 % de mujeres y un 51,69 % de hombres, entre los 17 a 25 años. Para este estudio se realizó un Cuestionario de Intereses Profesionales (CIPRO-UPS), y la técnica estadística de análisis de varianza (ANOVA), para conocer el rendimiento académico. Para conocer los intereses profesionales, se realizó la CIPRO-UPS y se aplicó Análisis Factorial Confirmatorio (AFC).

Finalmente, (Romero et al., 2015) utilizan técnicas de ML para predecir el rendimiento académico de los alumnos de Ingeniería Civil en Informática de la Universidad Bío de Chile, donde desarrolló una herramienta de software que utiliza la técnica k-Nearest Neigbors para clasificar variables nominales. Asimismo, lograron alcanzar niveles de acierto aceptables en el mejor de los casos, con un acierto del 60 % e índices de error cuadrático medio del 0.4.

Capítulo 3

Diseño del instrumento y conjunto de datos obtenidos

3.1. Instrumento

Debido a que actualmente no se encontró un instrumento completo y específico para conocer los intereses académicos de estudiantes de licenciatura de la DACyTI, se diseñó una encuesta para cada una de las 4 carreras activas de la División, las cuales constan de 24 a 27 reactivos dependiendo del área de interés que el alumno seleccione.Las cuatro carreras activas en la DACYTI son las siguientes:

1. Licenciatura en Tecnologías de la Información (LTI).
2. Licenciatura en Sistemas Computacionales (LSC).
3. Ingeniería en Sistemas Computacionales (ISC).
4. Ingeniería en Informática Administrativa (IIA).

La encuesta está diseñada en diferentes secciones con base en el mapa curricular de cada una de las carreras. En primer lugar consta de una sección general con 15 reactivos, donde se cuestiona porque eligió la carrera que está cursando, su formación a nivel bachillerato, si actualmente se encuentra laborando, por cual modalidad le gustaría titularse, si esta interesado en

estudiar una maestría, diplomados, certificaciones, entre otros. La siguiente sección diseñada fue la específica, donde puede variar de 4 a 6 secciones más ya que el mapa curricular de las carreras son diferentes, esta sección se centró en los intereses académicos específicos en el área que los estudiantes seleccionaran. En este apartado se cuestiona su lenguaje de programación favorito, calificación académica en el área de interés seleccionada, si ha reprobado alguna asignatura relacionada al área seleccionada, porqué le interesa esa área, entre otras.
Las secciones adicionales en el área especifica que se mencionó anteriormente son las siguientes:

1. Redes.
2. Programación.
3. Ingeniería de software.
4. Tratamiento de la información (Diseño, Modelado, Programación y Administración de Base de datos).
5. Entorno Social (Contabilidad, Normatividad informática, Derechos Humanos, Presupuestos, entre otras).
6. Interacción hombre-máquina.

La mayoría de las preguntas son de opción múltiple, en donde las respuestas se enumeraron, para así poder extraer ese digito y ser analizado en los experimentos.

Una vez diseñado el cuestionario, se procedió a validar su contenido mediante juicio de expertos. Para ello, dos especialistas en ciencias computacionales evaluaron la relevancia, claridad y pertinencia de cada ítem.
Posteriormente la validación, estos se enviaron al Comité de Ética de la UJAT para que fuera revisado y asegurarse así de que no se comprometen datos de información personal de los estudiantes. El resultado fue Aprobado con recomendaciones acerca de que al crearse en Google Forms, no se recopilarán los correos electrónicos, para que la participación de los estudiantes estuviera consentida bajo anonimato.

En este sentido, se crearon las 4 encuestas en Google Forms respetando cada una de las observaciones del comité de ética, para no comprometer información personal de los estudiantes. La encuesta se aplicó a 152 estudiantes de la DACyTI correspondientes a las carreras de IIA, ISC, LSC y LTI, específicamente a los estudiantes que contaran a partir del 40 % de avance curricular durante el periodo 2023-A, esto debido a que, en esta fase de su formación académica, los estudiantes poseen una experiencia más consolidada dentro de su plan de estudios, lo que les brinda una perspectiva más clara y definida respecto a sus intereses académicos, áreas de afinidad y posibles trayectorias profesionales.

3.2. Datos obtenidos

Los datos obtenidos por las encuestas fueron guardados en 4 archivos CSV, uno por cada carrera. Todas las respuestas eran obligatorias por lo cual no se podía omitir ninguna pregunta. Los nombres de los atributos obtenidos eran las preguntas completas del cuestionario. Debido a que la encuesta estaba dividida en secciones y en la pregunta especifica solo podían seleccionar un área de interés, las demás preguntas de las otras áreas quedaban en blanco.

En la Tabla siguiente tabla se muestran los nombres de los atributos del conjunto de datos original.

Tabla 3.1. Tabla de la relación de preguntas incluidas en la encuesta.

Nombre del atributo:
Marca temporal
Género
Edad
Semestre
1. ¿Por qué elegiste esta carrera?
2. ¿Qué formación académica o experiencia profesional previa posees de la carrera que estás cursando?
3. ¿Qué capacitación o equivalente cursaste en el bachiller?
4. ¿Qué serie o bloque cursaste en el bachiller?

5. ¿Qué tanta información tenías antes de elegir tu carrera?
6. ¿Actualmente en qué área estás laborando?
7. Incluso si no te pagaran por lo que estás haciendo, ¿podrías seguir haciéndolo por gusto?
8. ¿Qué método de aprendizaje te funciona mejor?
9. ¿En qué tipo de actividades extracurriculares te gustaría participar en el futuro?
10. ¿Por qué modalidad te gustaría titularte?
11. ¿Qué tipo de actividades académicas te resultan más interesantes?
12. De las siguientes opciones, ¿qué preferirías obtener adicional al título como parte de tus estudios?
13. ¿En qué tipo de organización te gustaría trabajar en el futuro?
14. ¿Cuántas veces has solicitado convenio?
15. ¿Qué maestría te gustaría estudiar en la DACYTI?
16 ¿Qué área de sistemas computacionales te interesa más, de acuerdo con el mapa curricular correspondiente?
1. Explica brevemente las razones de tu respuesta anterior.
2. ¿Cuál es tu experiencia previa en redes?
3. ¿Te interesa la investigación en redes?
4. ¿Cuáles de las siguientes áreas te resultan interesantes o te gustaría aprender más?
5. ¿Te gustaría seguir estudiando redes al concluir tu licenciatura?
6. ¿Qué aspecto del área de redes te motiva?
7. ¿Cuál es tu calificación promedio en redes? (Número entero entre 5 y 10).
8. ¿Cuántas veces has reprobado asignaturas del área de redes?
1. Explica brevemente las razones de tu respuesta anterior.
2. ¿Cuál es tu experiencia previa en programación?
3. ¿Te interesa la investigación en programación?
4. ¿Cuáles de las siguientes áreas te resultan interesantes o te gustaría aprender más?
5. ¿Cuál es tu lenguaje de programación favorito?

6. ¿Te gustaría seguir estudiando programación al concluir tu licenciatura?
7. ¿Qué te motiva del área de programación?
8. ¿Cuál es tu calificación promedio en programación? (Número entero entre 5 y 10).
9. ¿Cuántas veces has reprobado asignaturas del área de programación?
1. Explica brevemente las razones de tu respuesta anterior.
2. ¿Cuál es tu experiencia previa en Ingeniería de software?
3. ¿Te interesa la investigación en Ingeniería de software?
4. ¿Cuáles de las siguientes áreas te resultan interesantes o te gustaría aprender más?
5 . ¿Cuál es tu lenguaje de programación favorito para desarrollar software?
6. ¿Te gustaría seguir estudiando Ingeniería de software en el futuro?
7. ¿Qué te motiva del área de Ingeniería de software?
8. ¿Qué metodología de desarrollo de software te interesa más?
9. ¿Qué tipo de herramientas de software te gustaría aprender a utilizar?
10. ¿Cuál es tu calificación promedio en Ingeniería de software? (Número entero entre 5 y 10).
11. ¿Cuántas veces has reprobado asignaturas del área de Ingeniería de software?
1. Explica brevemente las razones de tu respuesta anterior.
2. ¿Cuál es tu experiencia previa en el área de Tratamiento de información?
3. ¿Te interesa la investigación en el área de Tratamiento de información?
4. ¿Cuáles de las siguientes áreas te resultan interesantes o te gustaría aprender más?
5. ¿Qué tipo de bases de datos te interesa más?
6. ¿Qué lenguaje de programación te gusta más para programar bases de datos?
7. ¿Te gustaría seguir estudiando bases de datos en el futuro?
8. ¿Qué te motiva del área de Tratamiento de Información?
9. ¿Qué metodología de modelado y diseño de bases de datos te parece más interesante?
10. ¿Cuál es tu calificación promedio en el área de Tratamiento de Información? (Número entero entre 5 y 10).

11. ¿Cuántas veces has reprobado asignaturas del área de Tratamiento de Información?
1. Explica brevemente las razones de tu respuesta anterior.
2. ¿Cuál es tu experiencia previa en el área de entorno social?
3. ¿Te interesa la investigación en el área de entorno social?
4. ¿Cuáles de las siguientes áreas te resultan interesantes o te gustaría aprender más?
5. ¿Qué herramienta digital te gustaría aprender a usar para trabajar en el área del entorno social en tecnologías de la información?
6. ¿Te gustaría seguir estudiando el entorno social en tecnologías de la información en el futuro?
7. ¿Qué te motiva del área de entorno social?
8. ¿Qué metodología de trabajo en el área del entorno social en tecnologías de la información te parece más interesante?
9. ¿Cuál es tu calificación promedio en el área de entorno social? (Número entero entre 5 y 10).
10. ¿Cuántas veces has reprobado asignaturas del área de entorno social?
1. Explica brevemente las razones de tu respuesta anterior.
2. ¿Cuál es tu experiencia previa en el área de Interacción hombre-máquina?
3. ¿Te interesa la investigación en el área de Interacción hombre-máquina?
4. ¿Cuáles de las siguientes áreas te resultan interesantes o te gustaría aprender más?
5. ¿Te gustaría seguir estudiando la Interacción hombre-máquina en el futuro?
6. ¿Qué tecnologías te gustaría aprender a usar para trabajar en el área de la Interacción hombre-máquina?
7. ¿Qué te motiva del área de la Interacción hombre-máquina?
8. ¿Qué áreas de la inteligencia artificial te parecen más interesantes?
9. ¿Qué tipo de proyectos te gustaría desarrollar en el futuro en el área de la inteligencia artificial?
10. ¿Qué lenguajes de programación te gustaría aprender para trabajar en el área de la inteligencia artificial?

11. ¿Cuál es tu calificación promedio en el área de Interacción hombre-máquina? (Número entero entre 5 y 10).
12. ¿Cuántas veces has reprobado asignaturas del área de Interacción hombre-máquina?
¿Qué área de la Informática Administrativa te interesa más, de acuerdo con el mapa curricular correspondiente?
¿Qué área de la tecnología de la información te interesa más, de acuerdo con el mapa curricular correspondiente?

Capítulo 4

Preprocesamiento y descripción del conjunto de datos

4.1. Conjunto de Datos

El conjunto de datos obtenido en las encuestas aplicadas a los estudiantes de la DACyTI, tiene como objetivo principal el minado preliminar para conocer los intereses académicos de estudiantes de licenciatura, en donde se toman en cuenta variables muy importantes como su capacitación y bloque previo a la licenciatura, genero, sexo, entre otras.

El conjunto de datos tiene un total de 81 variables y 152 instancias (estudiantes), el tipo de datos de las variables es entero y carácter. Para que el nombre de las variables no fuera tan extenso se modificaron los nombres de manera que se entienda la pregunta, pero resumida. En la tabla 4.1 se puede observar el nombre de las variables ya modificadas y su tipo de datos.

Tabla 4.1. Tabla de los atributos del dataset y su tipo de datos.

No. de atributo	Nombre del atributo	Tipo de dato
2	Genero	character
3	Edad	integer
4	Semestre	integer
5	RazonEleccion	character

6	FormacionPrevia	character
7	CapacitacionPrevia	character
8	BloquePrevio	character
9	InfoDisponible	character
10	AreaLabora	character
11	PorGusto	character
12	MetodoAprende	character
13	ActExtraCur	character
14	ModTit	character
15	ActAcadInteres	character
16	AdicionalTitulo	character
17	TipoOrg	character
18	NumConvenios	character
19	MaestriaPrefiere	character
20	AreaInteres	character
21	1Razones	character
22	2ExperienciaPRedes	character
23	3TeInteresaInvestigacionRedes	character
24	4AprenderMas	character
25	5ContinuarEstRedes	character
26	6QueTeMotivadeRedes	character
27	7CalificacionRedes	character
28	8VecesReprobadasRedes	character
29	1Razones	character
30	2ExperienciaPreviaProgramacion	character
31	3TeInteresaInvestigacionProgramacion	character
32	4AprenderMas	character
33	5LenguajesProgramFav	character

34	6ContinuarEstProgramacion	character
35	7QueTeMotivaProgramacion	character
36	8CalificacionProgramacion	character
37	9VecesReprobadasProgramacion	character
38	1Razones	character
39	2ExperienciaPreviaIngSoftware	character
40	3TeInteresaInvestigacionEnElAreaDeIngSoftware	character
41	4AprenderMas	character
42	5LenguajePrgramFavPDesarrollarSoftware	character
43	6ContinuarEstIngSoftware	character
44	7QueTeMotivaIngSoftware	character
45	8QueMetodologiaDesarrolloSoftwareTeInteresaMas	character
46	9QueTipoDeHerramientasDeSoftwareAprenderias	character
47	10CalificacionIngSoftware	character
48	11VecesReprobadasIngSoftaware	character
49	1Razones	character
50	2ExperienciaPreviaTratamientoDeLaInfo	character
51	3TeInteresaInvestigacionEnElAreaTratamientoDeInfo	character
52	4AprenderMas	character
53	5QueTipoDeBDTeInteresaMas	character
54	6LenguajePrgramFavPProgramarBD	character
55	7ContinuarEstBD	character
56	8QueTeMotivaTratamientoDeInfo	character
57	9QueMetodologiaDeModeladoYDisenoBDTeInteresaMas	character
58	10CalificacionTratamiendoDeInfo	character
59	11VecesReprobadasTratamientoDeInformacion	character
60	1Razones	character
61	2ExperienciaPreviaEntornoSocial	character

62	3TeInteresaInvestigacionEnElAreaEntornoSocial	character
63	4AprenderMas	character
64	5QueHerramientaDigitalTeInteresaMasPTrabajar EnAreaEntornoSocialEnTI	character
65	6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo	character
66	7QueTeMotivaEntornoSocial	character
67	8QueMetodologiaDeTrabajoEnElAreaEntornoSocial EnTITeInteresaMas	character
68	9CalificacionEntornoSocial	character
69	10VecesReprobadasEntornoSocial	character
70	1Razones	character
71	2ExperienciaPreviaInteraccionHombreMaquina	character
72	3TeInteresaInvestigacionEnElAreaInteraccionHombre Maquina	character
73	4AprenderMas	character
74	5ContinuarEstLaInteraccionHombreMaquina	character
75	6QueTecnologiasTeInteresaMasPTrabajarEnArea InteraccionHombreMaquina	character
76	7QueTeMotivaDelAreaInteraccionHombreMaquina	character
77	8QueAreasDeInteligenciaArtificialTeInteresaMas	character
78	9TipoProyectosPDesarrollarEnIA	character
79	10LenguajePrgramFavPTrabajarEnIA	character
80	11CalificacionInteraccionHombreMaquina	character
81	12VecesReprobadasInteraccionHombreMaquina	character

Como se puede observar la mayoría de las variables son de tipo carácter.

4.2. Preprocesamiento

Para poder utilizar el conjunto de datos obtenido mediante las encuestas, primero se realizó un análisis exploratorio, posteriormente, se unieron los cuatro archivos en un solo conjunto de datos para iniciar con los experimentos. Es importante mencionar que las encuestas estaban divididas por secciones, lo que generaba la presencia de celdas en blanco dependiendo de la elección que los estudiantes hicieron respecto a su área de interés. Estas celdas en blanco no eran valores faltantes, ya que los estudiantes solo podian elegir un área de interés. Sin embargo para que no afectara a los experimentos y se pudieran trabajar los datos, se le asigno el número 99 como valor centinela, y de esta manera poder identificar esas celdas como no aplicable, evitando así que estas observaciones fueran interpretadas erróneamente durante el análisis.

En el nuevo conjunto de datos se agregó un atributo que fue llamado como `Lic` de tipo numérico, donde almacena la licenciatura que está cursando el estudiante, en este caso son: 1=IIA, 2=ISC, 3=LSC, 4=LTI. Todo esto para poder realizar experimentos en rstudio, utilizar el dataset correctamente y poder identificar los datos analizados por carrera. De igual manera se extrajo el primer dígito de cada respuesta para quedarnos con un *dataset* numerico, y se cambiaron a tipo *factor* los atributos `AreaInteres` y `Genero`, en donde `AreaInteres` es nuestra variable predictora, el atributo `Lic` se mantiene con el tipo numérico y los atributos restantes se cambiaron a entero. Se omite el atributo `MarcaTemporal`, ya que este almacena la fecha y hora en que el estudiante respondió la encuesta.

De igual forma se omitieron las preguntas abiertas, ya que algunos modelos requieren que todos los valores sean númericos, en este caso eran 6 preguntas abiertas, cada una corresponde al área de interés seleccionada, en la que se solicitaba explicar la razón de porqué le interesaba esa área. Por lo tanto, el dataset quedó constituido con 75 atributos o variables y 152 instancias.

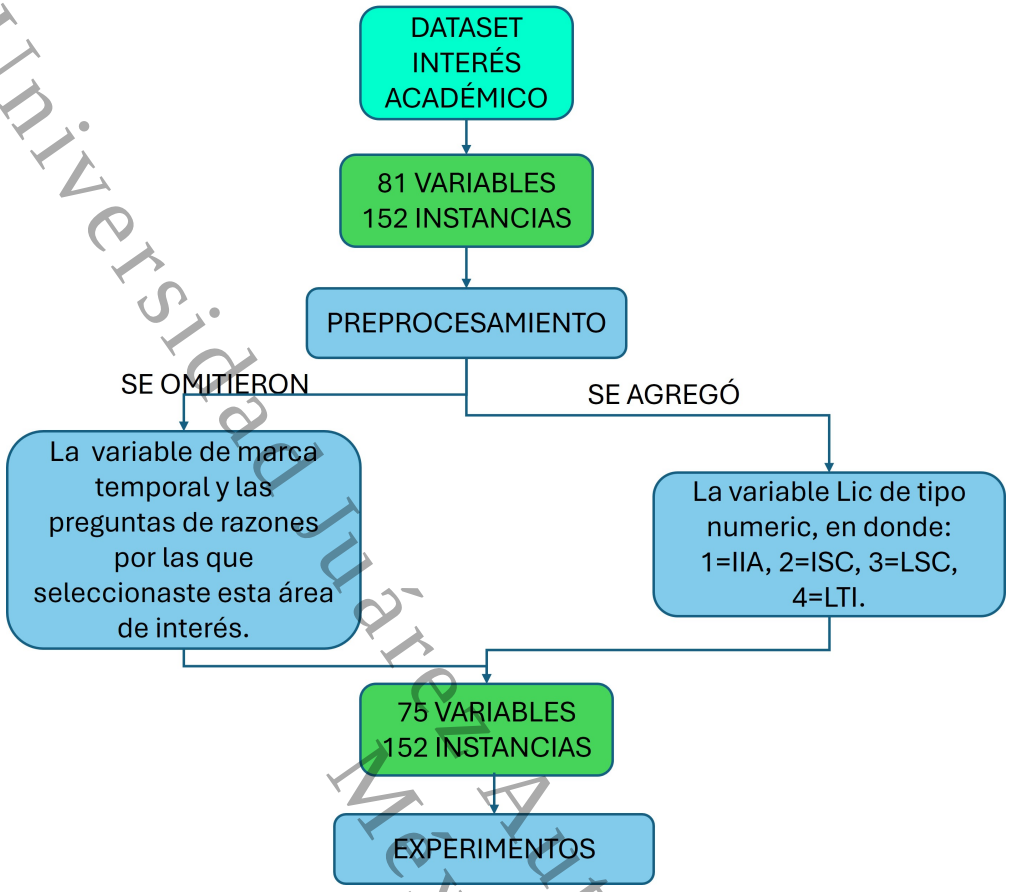

Figura 4.1. Diagrama de flujo para ilustrar el tratamiento que se le aplicó a los datos y el dataset final para los experimentos.

Los resultados obtenidos en la encuesta con respecto a la pregunta del área de interés conforme el mapa curricular se muestra en las gráficas siguientes.

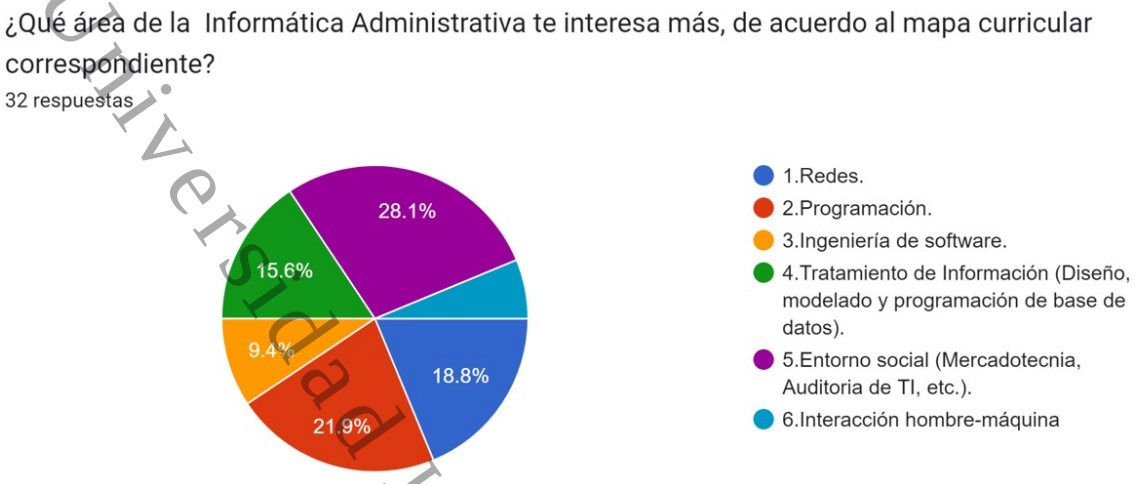

Figura 4.2. Gráfica de los resultados obtenidos del área de interés de los estudiantes de la IIA.

En la gráfica 4.2 se puede observar que el área que más les interesa a los estudiantes de la licenciatura en informática administrativa es el entorno social, es decir todo lo relacionado con mercadotecnia, publicidad, administración, contaduría, entre otros.

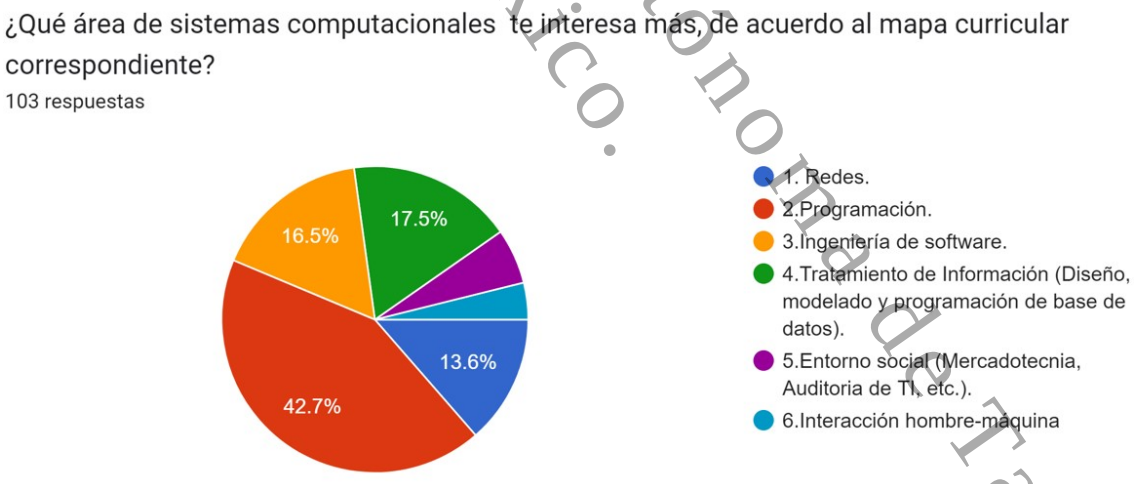

Figura 4.3. Gráfica de los resultados obtenidos del área de interés de los estudiantes de la ISC.

En la gráfica 4.3 se muestra la respuesta de los estudiantes de la ingeniería en sistemas computacionales, donde el área de mayor interés es el de la programación y el de menor interés es el de Interacción hombre-máquina.

Figura 4.4. Gráfica de los resultados obtenidos del área de interés de los estudiantes de la LSC.

En la Figura 4.4 se puede observar que el área de mayor interés de los estudiantes de la licenciatura en sistemas computacionales es programación y la que menos porcentaje de interés muestra es Tratamiento de la información, todo lo relacionado con las Bases de datos.

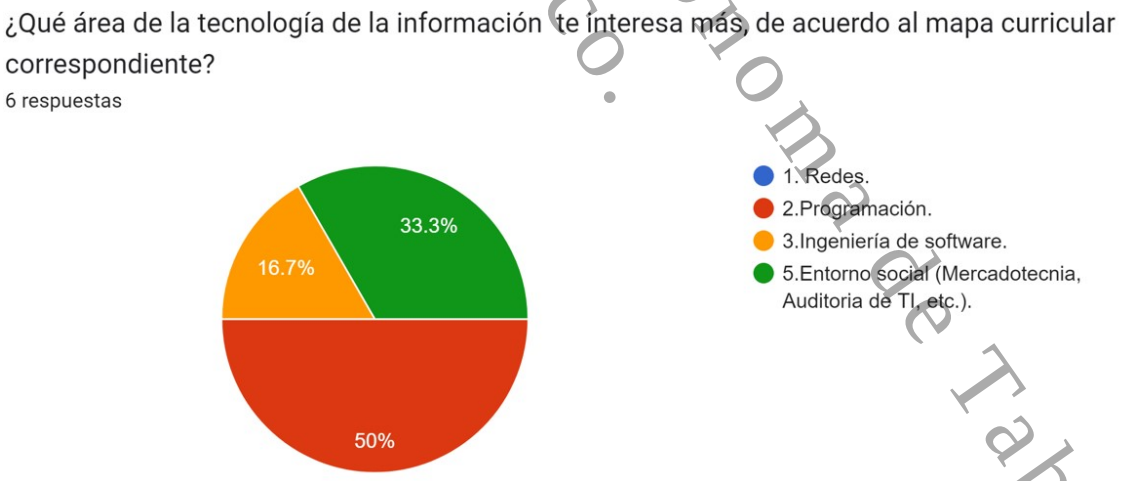

Figura 4.5. Gráfica de los resultados obtenidos del área de interés de los estudiantes de la LTI.

En la Figura 4.5 se puede observar que el área de mayor interés de los estudiantes de la licenciatura en tecnologías de la información es programación.

Finalmente podemos observar en una gráfica general con los resultados de las 4 carreras que la mayor área de interés es la programación, seguida por el Entorno social.

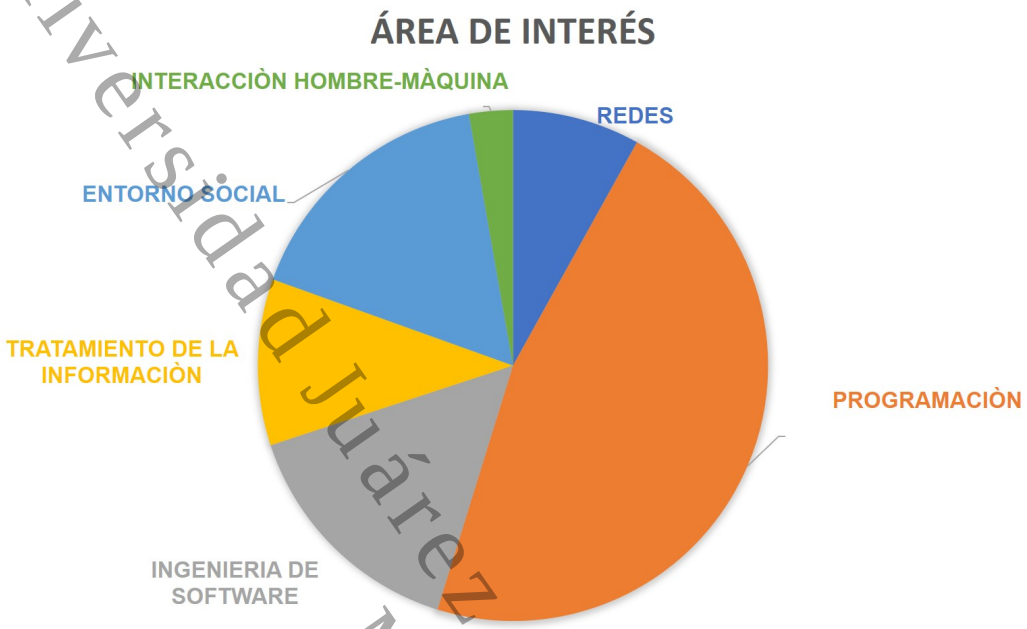

Figura 4.6. Gráfica de los resultados obtenidos del área de interés académico de los estudiantes de la DACyTI.

Capítulo 5

Experimentos y Resultados

En este capítulo se presentan los experimentos realizados como parte del proceso de análisis de datos para la identificación de los intereses académicos de los estudiantes de licenciatura. Con base en el conjunto de datos previamente construido a partir de encuestas aplicadas, se aplicaron diversas técnicas de minería de datos que incluyeron tanto métodos de selección de atributos como modelos de clasificación.

El objetivo principal de estos experimentos fue evaluar el desempeño de diferentes algoritmos y métodos de selección de atributos tipo filtros, con el fin de determinar cuáles permiten identificar de manera más efectiva los patrones ocultos en las preferencias académicas de los estudiantes. Por lo tanto, se utilizaron métodos filtros como: *Chi-squared*, *Information gain*, CFS y *Random forest importance*, así como los clasificadores *Support Vector Machines (SVM)* (con distintos kernels, *Random forest*, J48, JRip y OneR.

A lo largo del capítulo se describe la metodología experimental, los criterios de evaluación utilizados, y se analizan los resultados obtenidos con cada técnica. Finalmente, se discuten los hallazgos más relevantes y su posible aplicación en la mejora de la oferta educativa de la institución.

5.1. Experimento #1: Utilizando las primeras 20 variables del conjunto de datos

El objetivo de este experimento fue identificar las áreas de interés de los estudiantes de licenciatura utilizando las respuestas de la sección del área general de la encuesta.

5.1.1. Datos

El conjunto de datos utilizado en está sección es el dataset obtenido mediante las encuestas, el cual contiene 152 instancias. Se utilizaron para este experimento únicamente los primeros 20 atributos del dataset obtenido, ya que estos corresponden a las preguntas del área general de las encuestas.

La variable predictora es la pregunta especifica #20 “área de interés”. Para este experimento la variable `áreaInteres` se divide en 2 clases: la clase 0 para el área de interés de Entorno social y la clase 1 para las áreas tecnológicas que son: Programación, Redes, Interacción hombre-máquina, Ingeniería de Software y Tratamiento de la información.

Tabla 5.1. Tabla de los atributos utilizados en el experimento 1 y su tipo de datos.

Nombre del atributo	Tipo de dato
Genero	Factor
Edad	Integer
Semestre	Integer
Lic	Numeric
RazonEleccion	Integer
FormacionPrevia	Integer
CapacitacionPrevia	Integer
BloquePrevio	Integer
InfoDisponible	Integer
AreaLabora	Integer
PorGusto	Integer
MetodoAprende	Integer
ActExtraCur	Integer
ModTit	Integer
ActAcadInteres	Integer
AdicionalTitulo	Integer
TipoOrg	Integer
NumConvenios	Integer
MaestriaPrefiere	Integer
AreaInteres	Factor

5.1.2. Diseño experimental

Para este experimento se utilizaron Árboles de decisión con diferentes algoritmos, tales como: JRIP, J48, OneR, y *Random forest*, con la finalidad de predecir el área de interés de los estudiantes y por su facilidad de interpretación.

En la Figura 5.1 se puede observar el proceso que se llevó a cabo para la creación de cada uno de los modelos. Como se mencionó anteriormente, se utilizaron los primeros 20 atributos del dataset para crear cada modelo, se utilizaron dos terceras partes de los datos para entrenamiento de cada modelo y una tercera parte para evaluar el rendimiento del modelo. Se aplicó cada clasificador de manera independiente y se evaluó el rendimiento de los modelos, para evaluarlos se tomaron en cuenta las métricas de rendimiento: *accuracy*, sensibilidad, especificidad, kappa y *accuracy* balanceado.

Posteriormente por cada modelo realizamos 50 ejecuciones utilizando diferentes semillas por cada ejecución para obtener diferentes resultados, identificando el valor mínimo, máximo y la

mediana.

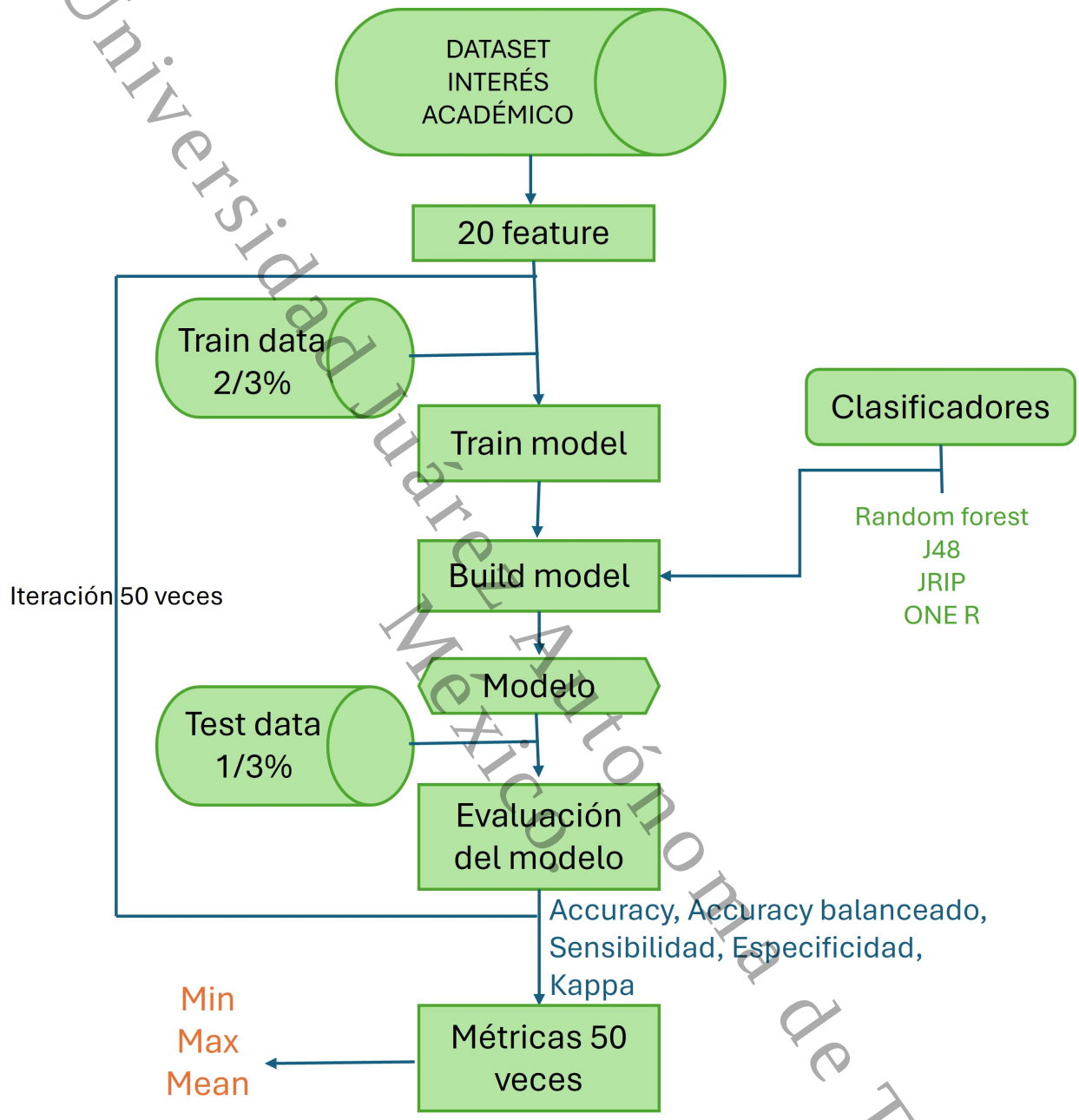

Figura 5.1. Diagrama de flujo del proceso de los datos en el experimento 1, clasificadores y métricas.

En la Tabla 5.2 se muestra el rendimiento de los modelos utilizando las métricas ya mencionadas, en donde se puede observar que el mejor rendimiento en el *accuracy* balanceado lo obtiene el clasificador J48. *Random forest*, JRip y OneR predicen muy bien la clase positiva, mientras que la clase negativa no muestra resultado. Sin embargo, J48 si predice ambas clases y tiene un

accuracy balanceado de 71.30 %.

Tabla 5.2. Medidas de rendimiento de los clasificadores usando las primeras 20 variables.

Clasificador	Accuracy	Especificidad	Sensibilidad	Kappa	Accuracy balanceado
Random forest	0.9	1.0	0.0	0.0	0.5
J48	**0.8833**	**0.9259**	**0.5000**	**0.396**	**0.7130**
JRIP	0.9	1.0	0.0	0.0	0.5
OneR	0.9	1.0	0.0	0.0	0.5

En este experimento también se utilizaron los métodos filtros CFS, *Chi-squared*, *Information gain* y *Random forest*, para obtener variables relevantes. En la Figura 5.2 se puede observar los resultados de los filtros aplicados al conjunto de datos, en donde una de las variables relevantes es el Género.

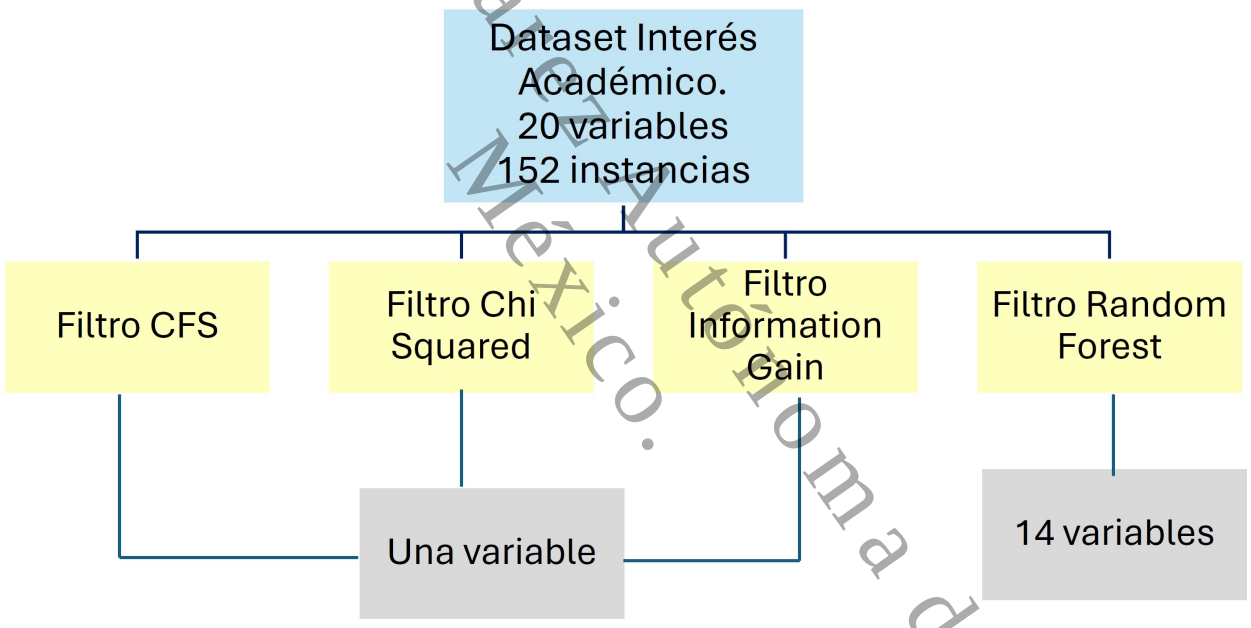

Figura 5.2. Variables relevantes obtenidas mediante los métodos filtro.

Las variables relevantes encontradas por los métodos filtro son las siguientes:

- **Variables relevantes encontradas por el filtro CFS:**
  1. Genero
- **Variables relevantes encontradas por los filtro *chi-squared e information gain:***
  1. Genero

- **Variables relevantes encontradas por el filtro *random forest:***

  1. AreaLabora
  2. BloquePrevio
  3. RazonEleccion
  4. Lic
  5. MaestriaPrefiere
  6. ActAcadInteres
  7. ActExtraCur
  8. InfoDisponible
  9. Edad
  10. AdicionalTitulo
  11. PorGusto
  12. CapacitacionPrevia
  13. TipoOrg
  14. Genero
  15. NumConvenios

Como podemos observar en el listado anterior, la variable relevante encontrada por los métodos filtros en común es la de Genero, esto sugiere que, en el contexto analizado las diferencias de género podrían estar asociadas a variaciones en el rendimiento o en la elección de trayectorias académicas, lo que abre la posibilidad de explorar desigualdades o patrones diferenciados en la población estudiada. Por otra parte, el método *random forest* seleccionó un conjunto más amplio de variables, como Área laboral, Bloque previo, Razón de elección, Edad, entre otras. Desde una perspectiva pedagógica, estas variables pueden interpretarse como indicadores que reflejan tanto factores personales, como formativos y motivacionales. Su relevancia en la predicción indica que el rendimiento estudiantil y la elección de programas no solo dependen de aspectos académicos, sino también de experiencias previas y preferencias individuales.

Tabla 5.3. Valor máximo, mínimo y la mediana del *accuracy* obtenido al ejecutar cada modelo 50 veces.

Clasificador	Máximo	Mínimo	Mediana
Random forest	0.9166667	0.8833333	0.907
J48	0.9166667	0.7333333	0.8523333
JRIP	0.9	0.7166667	0.8683333
OneR	0.9	0.7833333	0.8843333

En la tabla 5.3 se puede observar que *random forest* obtuvo la mediana de accuracy más alta (0.907), seguido por *OneR*, lo que indica que incluso modelos simples pueden generar predicciones competitivas si se entrenan con variables bien seleccionadas. Esta información es valiosa para el diseño de sistemas de apoyo educativo, ya que permite implementar modelos interpretables (como OneR) en contextos donde la transparencia es prioritaria, o modelos más complejos (como Random Forest) cuando se busca maximizar la precisión predictiva.

5.2. Experimento # 2: Utilizando todas las variables del conjunto de datos

5.2.1. Datos

Para este experimento se utilizaron todas las variables del conjunto de datos, el cual tiene 75 variables y 152 instancias. Como se había mencionado anteriormente las celdas en blanco por cada sección fueron sustituidas por el número 99, para que no afectara a los experimentos y se pudieran trabajar los datos, se le asigno el número 99 como valor centinela, y de esta manera poder identificar esas celdas como no aplicable, evitando así que estas observaciones fueran interpretadas erróneamente durante el análisis.

Tabla 5.4. Tabla donde se muestra que el área de interés se dividió en 2 clases: el 1 para las áreas de interés en informática y 0 para el área del entorno social.

0	1
17	135

En la Tabla 5.4 se puede observar que a 17 estudiantes les interesa el área de Entorno social, mientras que a 135 estudiantes les interesa más las áreas de Informática.

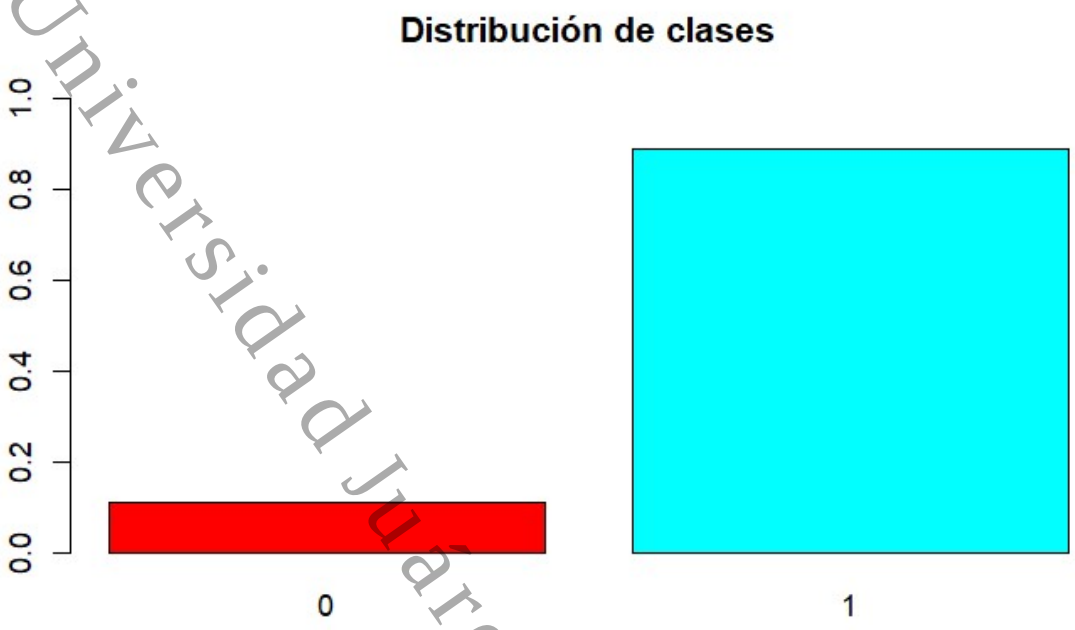

Figura 5.3. Proporción de la clase 0 y 1.

En la Figura 5.3 se observa la proporción de cada una de las clases, se puede observar que la clase 1 tiene alrededor del 88.82 % de las observaciones y la clase 0 el 11.18 %.

5.2.2. Diseño experimental

Para este experimento se crearon modelos con los clasificadores *Random forest*, J48, Jrip, OneR, SVM con kernel lineal (SVKL), SVM kernel polinomial (SVMKP), SVM kernel radial (SVMKR), SVM kernel sigmoidal (SVMKS), k-NN. En la Figura 5.4 se observa el proceso que se llevó a cabo para la creación de cada uno de los modelos. Como se mencionó anteriormente, se utilizaron todos los atributos del dataset, es decir 75 atributos y 152 instancias. Para este experimento se tomó el 60 % de los datos para entrenamiento y el 40 % para prueba; se aplicó cada clasificador de manera independiente y se evaluó el rendimiento de los modelos. Para evaluar los modelos se tomaron en cuenta las métricas de rendimiento: *accuracy*, sensibilidad, especificidad, kappa y *accuracy* balanceado.

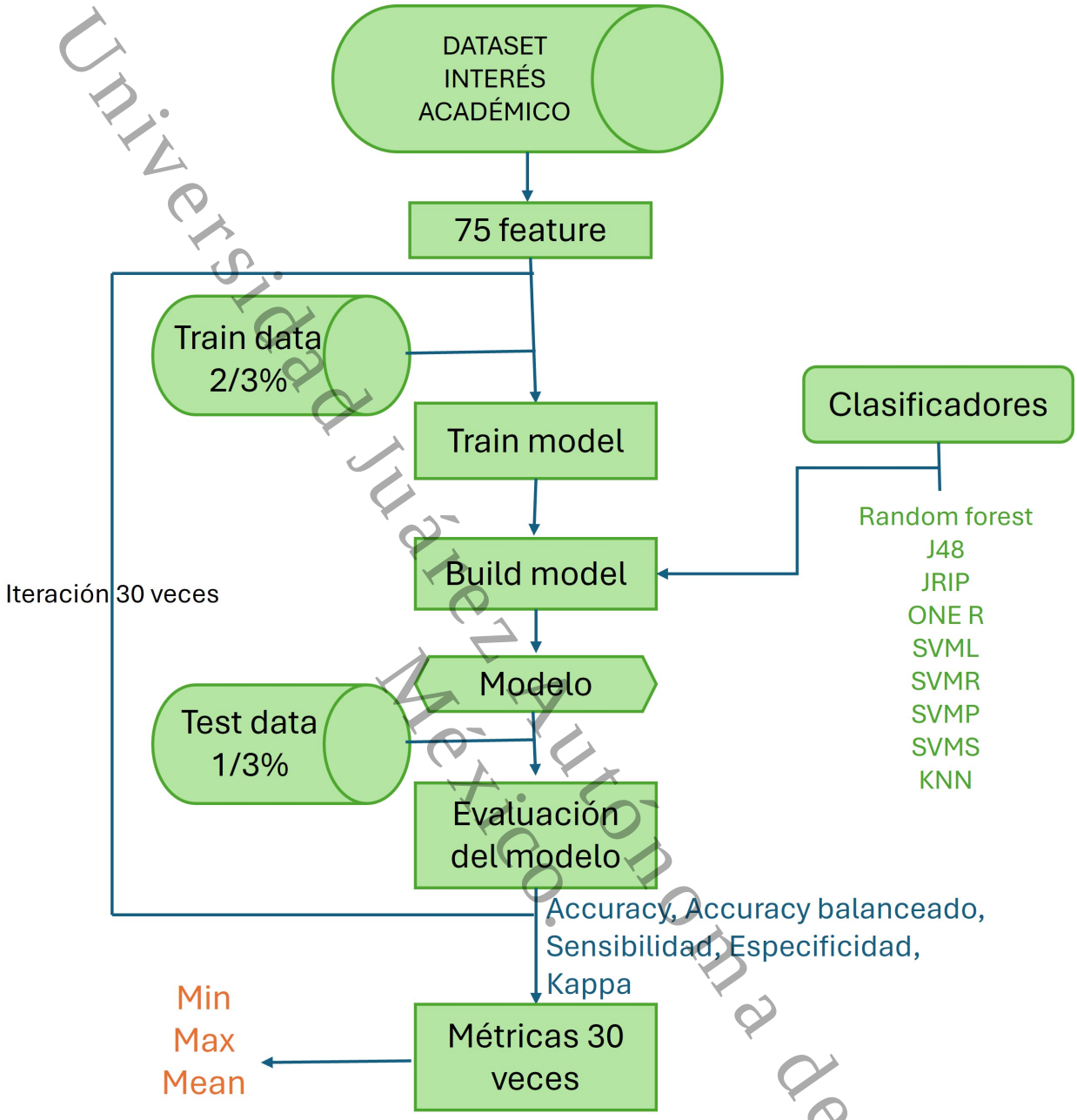

Figura 5.4. Diagrama de flujo del proceso de los datos utilizando todas las variables del dataset.

En la Tabla 5.5 se muestra el rendimiento de los modelos utilizando las métricas ya mencionadas. Se puede observar que en todos los clasificadores se obtiene el 100 % en todas las métricas, esto significa que los modelos están haciendo predicciones perfectas sobre el conjunto de datos de prueba y que no hay errores en las predicciones. De igual forma se aplicaron los filtros CFS, *Chi-squared*, *Information gain*, *Random forest*.

Tabla 5.5. Medidas de rendimiento de los clasificadores utilizando todas las variables.

Clasificador	Accuracy	Especificidad	Sensibilidad	Kappa	Accuracy balanceado
Random forest	1	1	1	1	1
J48	1	1	1	1	1
JRIP	1	1	1	1	1
OneR	1	1	1	1	1
SVMKL	1	1	1	1	1
SVMKP	1	1	1	1	1
SVMKR	1	1	1	1	1
SVMKS	1	1	1	1	1
k-NN	1	1	1	1	1

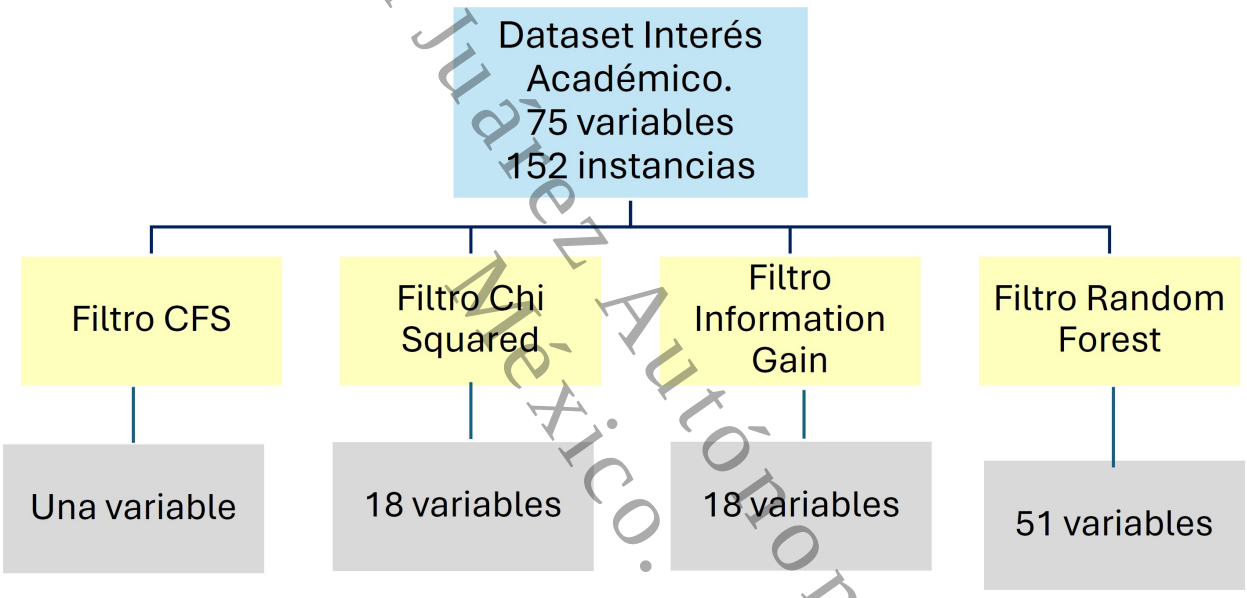

Figura 5.5. Resultados de los métodos filtro utilizando todas las variables del conjunto de datos.

Como podemos observar en la Figura 5.5 el filtro *Chi-squared* y el filtro *Information gain* dan como resultado las mismas variables relevantes, mientras que con *Random forest* se obtienen 51 variables relevantes. Cabe mencionar que se ejecutaron 30 veces los modelos con diferentes semillas, para corroborar si los resultados son consistentes a lo largo de varias ejecuciones.
Las variables relevantes encontradas por los métodos filtro son las siguientes:

- **Variable relevante encontrada por el filtro CFS:** X9CalificacionEntornoSocial
- **Variables relevantes encontradas por los filtros *Chi-squared* e *information gain*:**

1. X9CalificacionEntornoSocial
2. X2ExperienciaPreviaEntornoSocial
3. X3TeInteresaInvestigacionEnElAreaEntornoSocial
4. X5QueHerramientaDigitalTeInteresaMasPTrabajarEnAreaEntornoSocialEnTI
5. X6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo
6. X7QueTeMotivaEntornoSocial
7. X8QueMetodologiaDeTrabajoEnElAreaEntornoSocialEnTITeInteresaMas
8. X10VecesReprobadasEntornoSocial
9. X4AprenderMas.4
10. X8CalificacionProgramacion
11. X3TeInteresaInvestigacionProgramacion
12. X6ContinuarEstProgramacion
13. X7QueTeMotivaProgramacion
14. X9VecesReprobadasProgramacion
15. X2ExperienciaPreviaProgramacion
16. X5LenguajesProgramFav
17. X4AprenderMas.1
18. Genero

- **Variables relevantes encontradas por el filtro *random forest:***

1. X9CalificacionEntornoSocial
2. X6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo
3. X8QueMetodologiaDeTrabajoEnElAreaEntornoSocialEnTITeInteresaMas
4. X10VecesReprobadasEntornoSocial
5. X2ExperienciaPreviaEntornoSocial
6. X5QueHerramientaDigitalTeInteresaMasPTrabajarEnAreaEntornoSocialEnTI

7. X3TeInteresaInvestigacionEnElAreaEntornoSocial
8. X7QueTeMotivaEntornoSocial
9. X4AprenderMas.4
10. X9VecesReprobadasProgramacion
11. X7QueTeMotivaProgramacion
12. X3TeInteresaInvestigacionProgramacion
13. X8CalificacionProgramacion
14. X2ExperienciaPreviaProgramacion
15. X6ContinuarEstProgramacion
16. X4AprenderMas.1
17. X5LenguajesProgramFav
18. X8QueTeMotivaTratamientoDeInfo
19. X4AprenderMas.3
20. ActExtraCur
21. X5QueTipoDeBDTeInteresaMas
22. X3TeInteresaInvestigacionEnElAreaTratamientoDeInfo
23. X2ExperienciaPRedes
24. X7CalificacionRedes
25. X2ExperienciaPreviaTratamientoDeLaInfo
26. X9QueTipoDeHerramientasDeSoftwareAprenderias
27. X8QueMetodologiaDesarrolloSoftwareTeInteresaMas
28. X9QueMetodologiaDeModeladoYDisenoBDTeInteresaMas
29. X3TeInteresaInvestigacionRedes
30. X8VecesReprobadasRedes
31. X7ContinuarEstBD
32. X10CalificacionTratamiendoDeInfo

33. X2ExperienciaPreviaIngSoftware
34. PorGusto
35. X3TeInteresaInvestigacionEnElAreaDeIngSoftware
36. X4AprenderMas
37. X4AprenderMas.2
38. AreaLabora
39. Genero
40. X10CalificacionIngSoftware
41. TipoOrg
42. NumConvenios
43. X11VecesReprobadasTratamientoDeInformacion
44. X5ContinuarEstLaInteraccionHombreMaquina
45. CapacitacionPrevia
46. MaestriaPrefiere
47. InfoDisponible
48. BloquePrevio
49. ActAcadInteres
50. X6LenguajePrgramFavPProgramarBD
51. X6QueTeMotivadeRedes

Tabla 5.6. Resultado de las 30 ejecuciones de cada modelo, tomando en cuenta la métrica *accuracy*.

Nombre del clasificador	Máximo	Mínimo	promedio
Random forest	**1**	**1**	**1**
J48	1	0.95	0.9905
JRIP	1	0.95	0.9966
OneR	**1**	**1**	**1**
SVMKL	**1**	**1**	**1**
SVMKP	1	0.9833	0.9983
SVMKR	1	0.9833	0.9955
SVMKS	1	0.95	0.9961

En la Tabla 5.6 se presentan los resultados obtenidos tras 30 ejecuciones de diversos modelos de clasificación, evaluados mediante la métrica de accuracy. Los clasificadores *Random Forest*, *OneR* y SVM con kernel lineal (SVMKL) alcanzaron un desempeño perfecto en todas sus ejecuciones, logrando valores máximos, mínimos y promedio iguales a 1, lo que indica una capacidad óptima para predecir correctamente los intereses académicos de los estudiantes de licenciatura.

5.3. Experimento # 3: Utilizando las variables relevantes obtenidas mediante el filtro *Random Forest*

Este experimento tiene el objetivo de crear modelos predictivos utilizando las variables relevantes encontradas por el filtro *Random forest*.

5.3.1. Datos

Los datos utilizados en este experimento son los atributos encontrados mediante el filtro *Random forest*, es decir 51 variables.

Tabla 5.7. Variables relevantes encontradas mediante el filtro *RF*.

N°	Nombre del atributo	N°	Nombre del atributo
1	X9CalificacionEntornoSocial	2	X6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo
3	X8QueMetodologiaDeTrabajoEnElAreaEntornoSocialEnTITeInteresaMas	4	X10VecesReprobadasEntornoSocial
5	X2ExperienciaPreviaEntornoSocial	6	X5QueHerramientaDigitalTeInteresaMasPTrabajarEnAreaEntornoSocialEnTI
7	X3TeInteresaInvestigacionEnElAreaEntornoSocial	8	X7QueTeMotivaEntornoSocial
9	X4AprenderMas.4	10	X9VecesReprobadasProgramacion
11	X7QueTeMotivaProgramacion	12	X3TeInteresaInvestigacionProgramacion
13	X8CalificacionProgramacion	14	X2ExperienciaPreviaProgramacion
15	X6ContinuarEstProgramacion	16	X4AprenderMas.1
17	X5LenguajesProgramFav	18	X8QueTeMotivaTratamientoDeInfo
19	X4AprenderMas.3	20	ActExtraCur
21	X5QueTipoDeBDTeInteresaMas	22	X3TeInteresaInvestigacionEnElAreaTratamientoDeInfo
23	X2ExperienciaPRedes	24	X7CalificacionRedes
25	X2ExperienciaPreviaTratamientoDeLaInfo	26	X9QueTipoDeHerramientasDeSoftwareAprenderias
27	X8QueMetodologiaDesarrolloSoftwareTeInteresaMas	28	X9QueMetodologiaDeModeladoYDisenoBDTeInteresaMas
29	X3TeInteresaInvestigacionRedes	30	X8VecesReprobadasRedes
31	X7ContinuarEstBD	32	X10CalificacionTratamiendoDeInfo
33	X2ExperienciaPreviaIngSoftware	34	PorGusto
35	X3TeInteresaInvestigacionEnElAreaDeIngSoftware	36	X4AprenderMas
37	X4AprenderMas.2	38	AreaLabora
39	Genero	40	X10CalificacionIngSoftware
41	TipoOrg	42	NumConvenios
43	X11VecesReprobadasTratamientoDeInformacion	44	X5ContinuarEstLaInteraccionHombreMaquina
45	CapacitacionPrevia	46	MaestriaPrefiere
47	InfoDisponible	48	BloquePrevio
49	ActAcadInteres	50	X6LenguajePrgramFavPProgramarBD
51	X6QueTeMotivadeRedes		

5.3.2. Diseño experimental

En este experimento se crearon modelos con los clasificadores *Random forest*, J48, Jrip, OneR, SVM con kernel lineal, SVM kernel polinomial, SVM kernel radial, SVM kernel sigmoidal, k-NN. En la Figura 5.6 se muestra el proceso que se llevó a cabo para la creación de cada uno de los modelos. Como se mencionó anteriormente, se utilizaron las variables relevantes obtenidas mediante el método filtro *Random forest*, es decir 51 atributos y 152 instancias. Para este experimento se tomaron dos terceras partes de los datos para entrenamiento y una tercera parte para prueba, se aplicó cada clasificador de manera independiente y se evaluó el rendimiento de los modelos. Para evaluar los modelos se tomaron en cuenta las métricas de rendimiento: *accuracy*, sensibilidad, especificidad, kappa y *accuracy* balanceado.

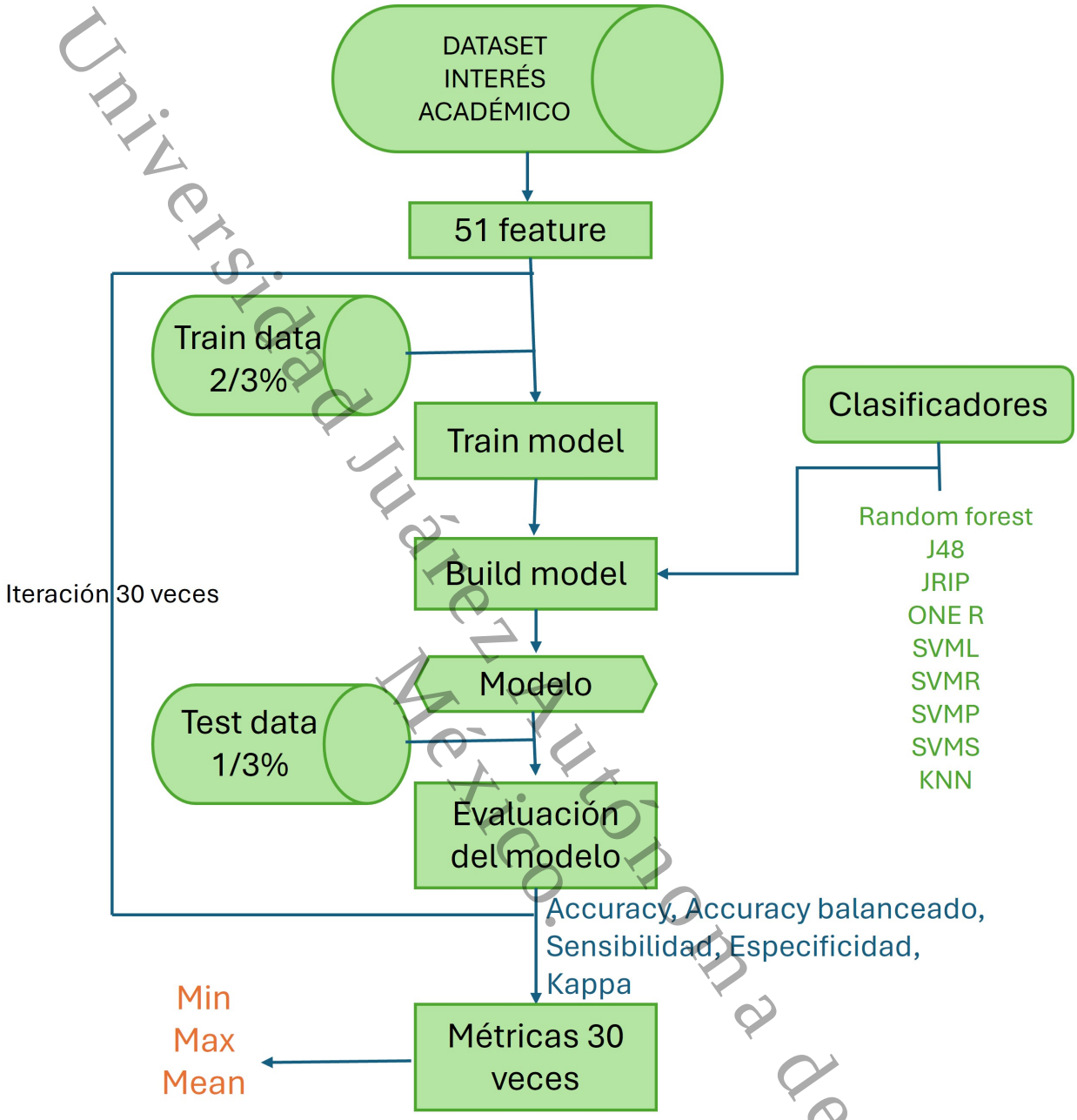

Figura 5.6. Diagrama de flujo del proceso de los datos, utilizando las variables relevantes obtenidas por el método filtro *random forest*.

En la Tabla 5.8 se muestra el rendimiento de los modelos utilizando las métricas ya mencionadas. Como se puede observar en todos los clasificadores se obtiene el 100 % en todas las métricas.

Tabla 5.8. Medidas de rendimiento de los clasificadores.

Clasificador	Accuracy	Especificidad	Sensibilidad	Kappa	Accuracy balanceado
Random forest	1	1	1	1	1
J48	1	1	1	1	1
JRIP	1	1	1	1	1
OneR	1	1	1	1	1
SVMKL	1	1	1	1	1
SVMKP	1	1	1	1	1
SVMKR	1	1	1	1	1
SVMKS	1	1	1	1	1
k-NN	1	1	1	1	1

Se ejecutaron 30 veces los modelos con diferentes semillas, para corroborar si los resultados son consistentes a lo largo de varias ejecuciones.

Tabla 5.9. Resultado de las 30 ejecuciones de cada modelo, mostrando los valores máximo, promedio y mínimo de cada uno de los clasificadores. En donde *Random forest*, *OneR*, SVM con kernel lineal, SVM con kernel polinomial, SVM con kernel radial y k-NN, tienen los mejores resultados en cuanto a la métrica de *accuracy* en un 100 %.

Nombre del clasificador	Máximo	Mínimo	promedio
Random forest	**1**	**1**	**1**
J48	1	0.95	0.9933
JRIP	1	0.95	0.9966
OneR	**1**	**1**	**1**
SVMKL	**1**	**1**	**1**
SVMKP	**1**	**1**	**1**
SVMKR	**1**	**1**	**1**
SVMKS	1	0.95	0.9972
k-NN	**1**	**1**	**1**

Como se puede observar J48, JRIP y SVMKS también muestran buenos resultados con un *accuracy* promedio de 0.99.

5.4. Experimento # 4: Utilizando las variables relevantes obtenidas mediante el filtro CFS

Este experimento tiene el objetivo de crear modelos predictivos utilizando la variables relevantes encontradas por el filtro CFS.

5.4.1. Datos

Los datos utilizados en este experimento son los atributos encontrados mediante el filtro CFS, en este caso el filtro CFS encontró como variable relevante `X9CalificacionEntornoSocial`.

5.4.2. Diseño experimental

En este experimento se crearon modelos con los clasificadores, *Random forest*, J48, Jrip, OneR, SVM con kernel lineal, SVM kernel polinomial, SVM kernel radial, SVM kernel sigmoidal y k-NN. En la Figura 5.7 se muestra el proceso que se llevó a cabo para la creación de cada uno de los modelos. Se utilizó la variable encontrada por el filtro CFS. Para este experimento se tomaron dos terceras partes de los datos para entrenamiento y una tercera parte para prueba, se aplicó cada clasificador de manera independiente y se evaluó el rendimiento de los modelos. Para evaluar los modelos se tomaron en cuenta las métricas de rendimiento: *accuracy*, sensibilidad, especificidad, kappa y *accuracy* balanceado.

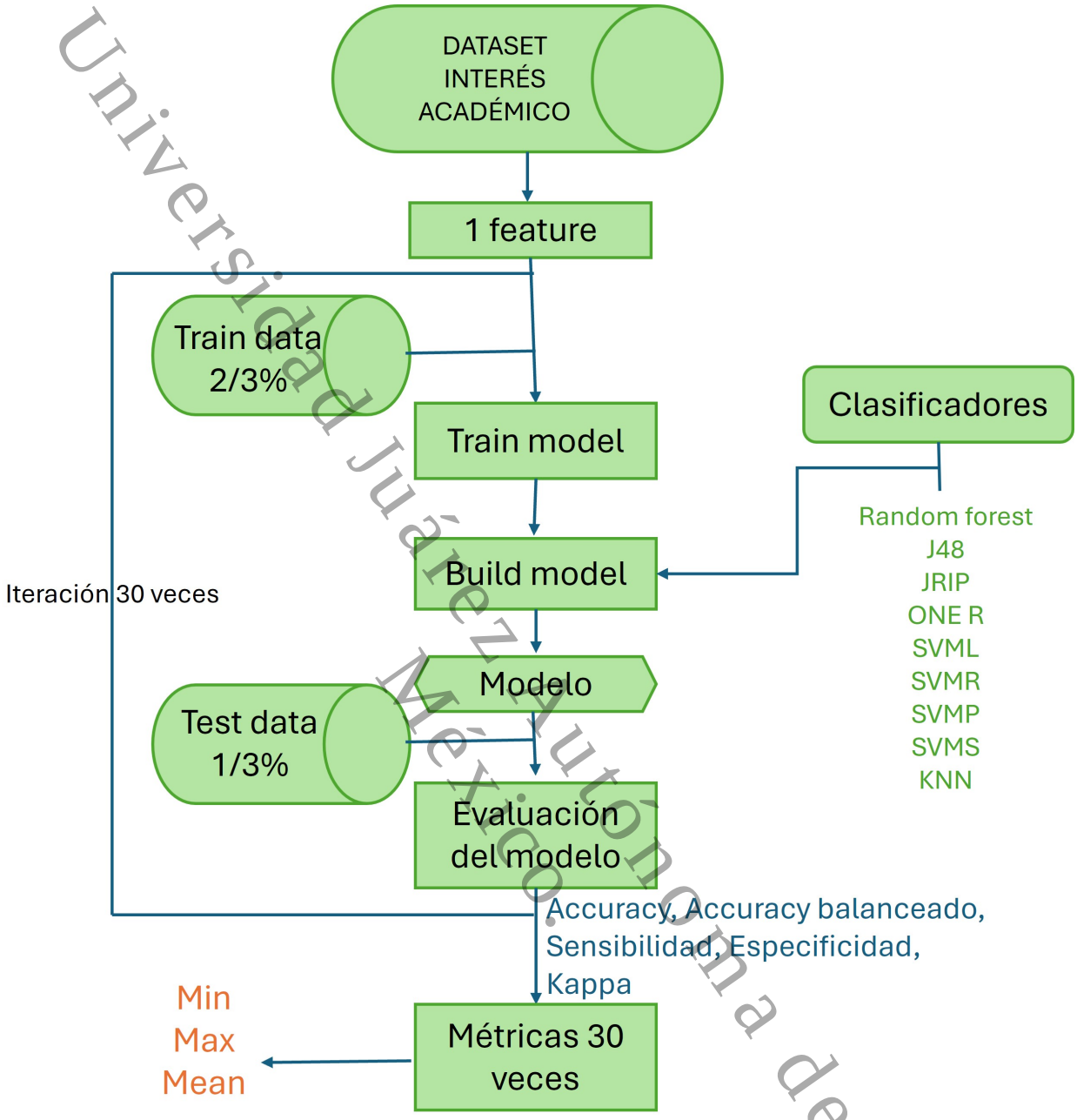

Figura 5.7. Diagrama de flujo del proceso de los datos, utilizando la variable relevante obtenida por el método filtro CFS.

En la Tabla 5.10 se muestra el rendimiento de los modelos utilizando las métricas ya mencionadas. Como se puede observar en todos los clasificadores se obtiene el 100 % en todas las métricas.

Tabla 5.10. Medidas de rendimiento de los clasificadores.

Clasificador	Accuracy	Especificidad	Sensibilidad	Kappa	Accuracy balanceado
Random forest	1	1	1	1	1
J48	1	1	1	1	1
JRIP	1	1	1	1	1
OneR	1	1	1	1	1
SVMKL	1	1	1	1	1
SVMKP	1	1	1	1	1
SVMKR	1	1	1	1	1
SVMKS	1	1	1	1	1

Para corroborar si los resultados eran consistentes a lo largo de varias ejecuciones, se ejecutaron 30 veces los modelos con diferentes semillas, dando como resultado la Tabla 5.11.

Tabla 5.11. Resultado de las 30 ejecuciones de cada modelo, se puede observar que los clasificadores J48 y Jrip tienen un menor rendimiento en comparación con los otros clasificadores, ya que los otros clasificadores muestran un rendimiento en un 100 % en la métrica de *accuracy*.

Nombre del clasificador	Máximo	Mínimo	promedio
Random forest	**1**	**1**	**1**
J48	1	0.95	0.9966
JRIP	1	0.95	0.9966
OneR	**1**	**1**	**1**
SVMKL	**1**	**1**	**1**
SVMKP	**1**	**1**	**1**
SVMKR	**1**	**1**	**1**
SVMKS	**1**	**1**	**1**
k-NN	**1**	**1**	**1**

5.5. Experimento # 5: Utilizando las variables relevantes obtenidas mediante el filtro *Chi-Squared*

Este experimento tiene el objetivo de crear modelos predictivos utilizando las variables relevantes encontradas por el filtro *Chi-squared*.

5.5.1. Datos

Los datos utilizados en este experimento son los atributos encontrados mediante el filtro *Chi-squared*, es decir 18 variables, las que se muestran a continuación.

Tabla 5.12. Variables relevantes mediante el filtro *Chi-squared*.

N°	Nombre del atributo	N°	Nombre del atributo
1	X9CalificacionEntornoSocial	2	X6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo
3	X8QueMetodologiaDeTrabajoEnElAreaEntornoSocialEnTITeInteresaMas	4	X10VecesReprobadasEntornoSocial
5	X2ExperienciaPreviaEntornoSocial	6	X5QueHerramientaDigitalTeInteresaMasPTrabajarEnAreaEntornoSocialEnTI
7	X3TeInteresaInvestigacionEnElAreaEntornoSocial	8	X7QueTeMotivaEntornoSocial
9	X4AprenderMas.4	10	X9VecesReprobadasProgramacion
11	X7QueTeMotivaProgramacion	12	X3TeInteresaInvestigacionProgramacion
13	X8CalificacionProgramacion	14	X2ExperienciaPreviaProgramacion
15	X6ContinuarEstProgramacion	16	X4AprenderMas.1
17	X5LenguajesProgramFav	18	Genero

5.5.2. Diseño experimental

En este experimento se crearon modelos con los clasificadores *Random forest*, J48, Jrip, OneR, SVM con kernel lineal, SVM kernel polinomial, SVM kernel radial, SVM kernel sigmoidal, k-NN. En la Figura 5.8 se muestra el proceso que se llevó a cabo para la creación de cada uno de los modelos. Como se mencionó anteriormente, se utilizaron las variables relevantes obtenidas mediante el método filtro *Chi-squared*, es decir 18 atributos. Para este experimento se tomó dos terceras partes de los datos para entrenamiento y una tercera parte para prueba, se aplicó cada clasificador de manera independiente y se evaluó el rendimiento de los modelos. Para evaluar los modelos se tomaron en cuenta las métricas de rendimiento: *accuracy*, sensibilidad, especificidad, kappa y *accuracy* balanceado.

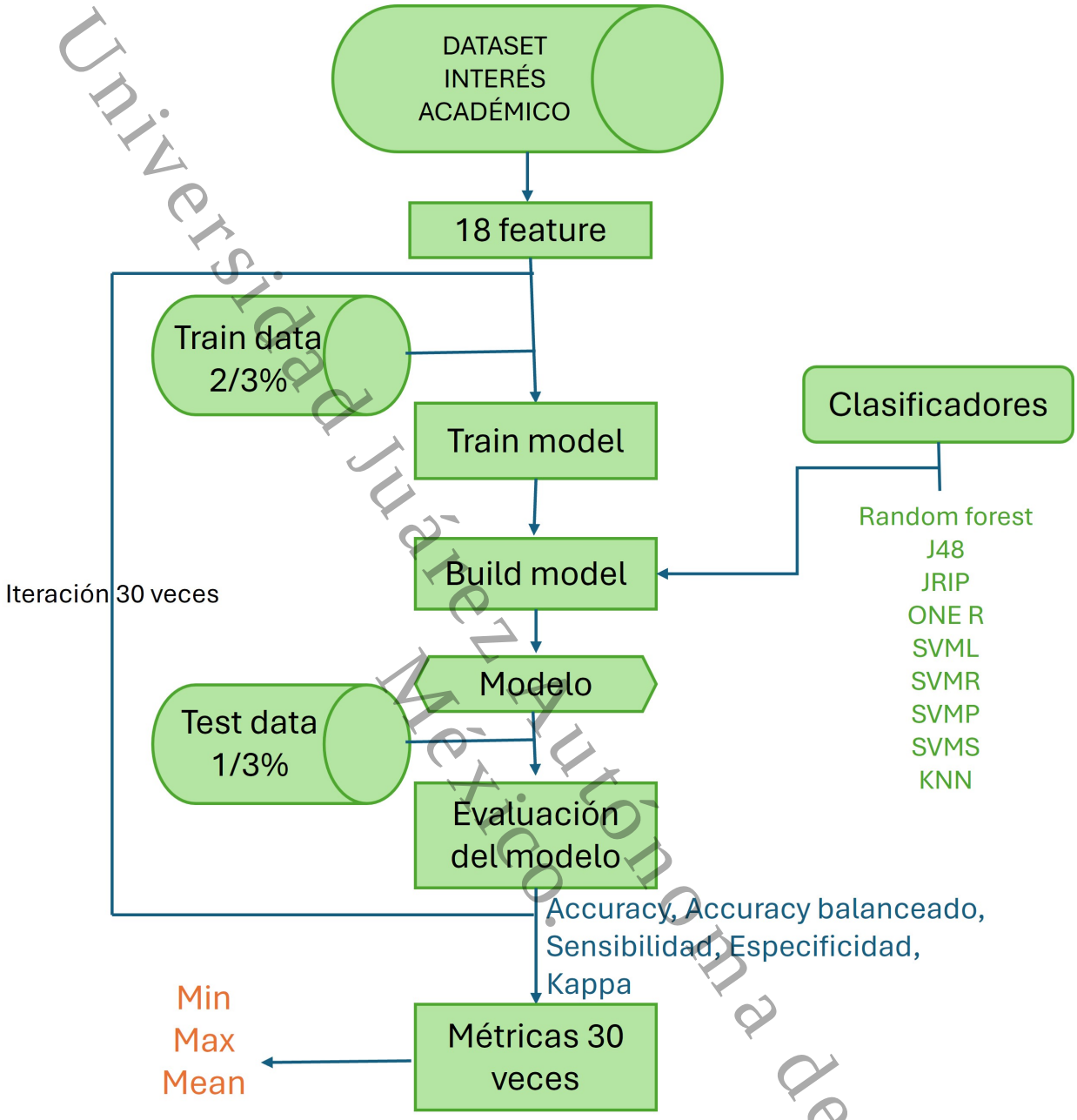

Figura 5.8. Diagrama de flujo del proceso de los datos, utilizando las variables relevantes obtenidas por el método filtro *chi-squared*.

En la Tabla 5.13 se muestra el rendimiento de los modelos utilizando las métricas ya mencionadas. Como se puede observar en todos los clasificadores se obtiene el 100 % en todas las métricas.

Tabla 5.13. Medidas de rendimiento de los clasificadores.

Clasificador	Accuracy	Especificidad	Sensibilidad	Kappa	Accuracy balanceado
Random forest	1	1	1	1	1
J48	1	1	1	1	1
JRIP	1	1	1	1	1
OneR	1	1	1	1	1
SVMKL	1	1	1	1	1
SVMKP	1	1	1	1	1
SVMKR	1	1	1	1	1
SVMKS	1	1	1	1	1
k-NN	1	1	1	1	1

Cada modelo se ejecutó 30 veces con diferentes semillas, para corroborar si los resultados son consistentes a lo largo de varias ejecuciones.

Tabla 5.14. Resultado de las 30 ejecuciones de cada modelo, mostrando los valores máximo, promedio y mínimo de cada uno de los clasificadores. En donde podemos observar que nos da los mismos resultados que en el experimento 4, en donde se utiliza variable relevante encontrada por el filtro CFS.

Nombre del clasificador	Máximo	Mínimo	promedio
Random forest	**1**	**1**	**1**
J48	1	0.95	0.9933
JRIP	1	0.95	0.9966
OneR	**1**	**1**	**1**
SVMKL	**1**	**1**	**1**
SVMKP	**1**	**1**	**1**
SVMKR	**1**	**1**	**1**
SVMKS	**1**	**1**	**1**
k-NN	**1**	**1**	**1**

5.6. Experimento # 6: Utilizando las variables relevantes obtenidas mediante el filtro *Information Gain*

Este experimento tiene el objetivo de crear modelos predictivos utilizando las variables relevantes encontradas por el filtro *Information gain*.

5.6.1. Datos

Los datos utilizados en este experimento son los atributos encontrados mediante el filtro *Information gain*, es decir 18 variables, mismas que se muestran a continuación.

Tabla 5.15. Variables relevantes mediante el filtro *Information gain*.

N°	Nombre del atributo	N°	Nombre del atributo
1	X9CalificacionEntornoSocial	2	X6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo
3	X8QueMetodologiaDeTrabajoEnElAreaEntornoSocialEnTITeInteresaMas	4	X10VecesReprobadasEntornoSocial
5	X2ExperienciaPreviaEntornoSocial	6	X5QueHerramientaDigitalTeInteresaMasPTrabajarEnAreaEntornoSocialEnTI
7	X3TeInteresaInvestigacionEnElAreaEntornoSocial	8	X7QueTeMotivaEntornoSocial
9	X4AprenderMas.4	10	X9VecesReprobadasProgramacion
11	X7QueTeMotivaProgramacion	12	X3TeInteresaInvestigacionProgramacion
13	X8CalificacionProgramacion	14	X2ExperienciaPreviaProgramacion
15	X6ContinuarEstProgramacion	16	X4AprenderMas.1
17	X5LenguajesProgramFav	18	Genero

Como se puede observar el filtro *Information gain* encontró las mismas variables que el filtro *Chi-squared*.

5.6.2. Diseño experimental

En este experimento se crearon modelos con los clasificadores, *Random forest*, J48, JRIP, ONER, *Support vector machine(SVM)* con kernel lineal(SVMKL), SVM kernel polinomial (SVMKP), SVM kernel radial (SVMKR), SVM kernel sigmoidal (SVMKS), k-NN.

En la Figura 5.9 se muestra el proceso que se llevó a cabo para la creación de cada uno de los modelos. Como se mencionó anteriormente, se utilizaron las variables relevantes obtenidas mediante el método filtro *Information Gain*, es decir 18 atributos. Para este experimento se tomó dos terceras partes de los datos para entrenamiento y una tercera parte para prueba, se aplicó cada clasificador de manera independiente y se evaluó el rendimiento de los modelos. Para evaluar los modelos se tomaron en cuenta las métricas de rendimiento: *accuracy*, sensibilidad, especificidad, kappa y *accuracy* balanceado.

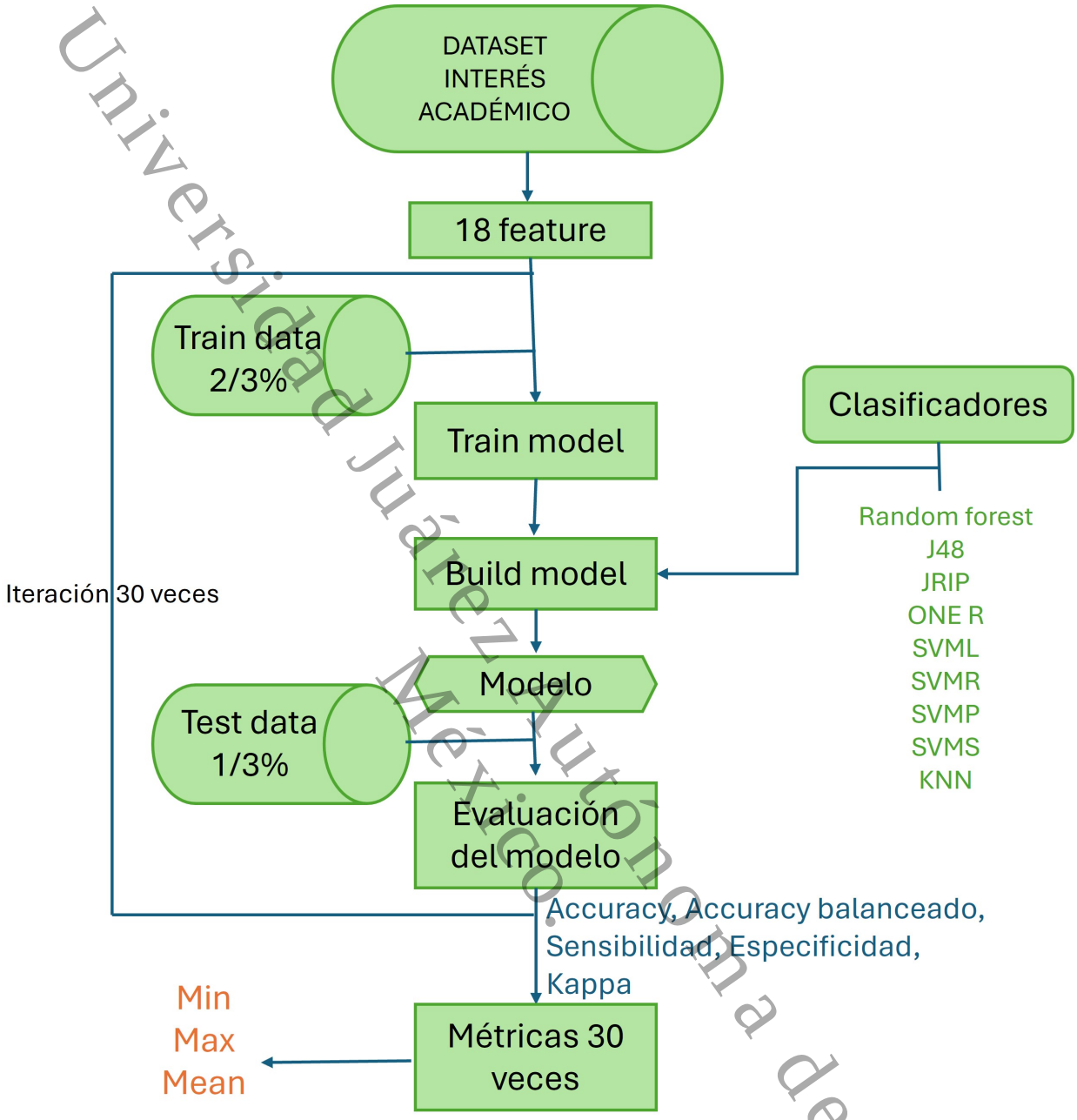

Figura 5.9. Diagrama de flujo del proceso de los datos, utilizando las variables relevantes obtenidas por el método filtro *information gain*.

En la Tabla 5.16 se muestra el rendimiento de los modelos utilizando las métricas ya mencionadas. Como se puede observar en todos los clasificadores se obtiene el 100 % en todas las métricas.

Tabla 5.16. Medidas de rendimiento de los clasificadores, cada modelo se ejecutó 30 veces con diferentes semillas, para corroborar si los resultados son consistentes a lo largo de varias ejecuciones.

Clasificador	Accuracy	Especificidad	Sensibilidad	Kappa	Accuracy balanceado
Random forest	1	1	1	1	1
J48	1	1	1	1	1
JRIP	1	1	1	1	1
OneR	1	1	1	1	1
SVMKL	1	1	1	1	1
SVMKP	1	1	1	1	1
SVMKR	1	1	1	1	1
SVMKS	1	1	1	1	1
k-NN	1	1	1	1	1

Tabla 5.17. Resultado de las 30 ejecuciones de cada modelo, mostrando los resultados en la métrica *accuracy*.

Nombre del clasificador	Máximo	Mínimo	promedio
Random forest	**1**	**1**	**1**
J48	1	0.95	0.9933
JRIP	1	0.95	0.9966
OneR	**1**	**1**	**1**
SVMKL	**1**	**1**	**1**
SVMKP	**1**	**1**	**1**
SVMKR	**1**	**1**	**1**
SVMKS	**1**	**1**	**1**
k-NN	**1**	**1**	**1**

En la Tabla 5.17 se puede observar los resultados de las 30 ejecuciones, mostrando los valores máximo, promedio y mínimo de cada uno de los clasificadores. Podemos observar que obtenemos los mismos resultados que en el experimento 4 y 5, en donde se utiliza variable relevante encontrada por el filtro CFS y *Chi-squared*. En conclusión, de este experimento al ser las mismas variables relevantes encontradas por el filtro *Chi-squared*, los resultados son los mismos.

5.7. Experimento utilizando el conjunto de datos obtenido de la carrera de Ingeniería en Informática (IIA)

5.7.1. Datos

Para este experimento se utilizaron todas las variables del conjunto de datos correspondiente a la carrera de IIA, el cual tiene 74 variables y 32 instancias. En experimentos anteriores se ha

mencionado que las celdas en blanco le colocamos el número 99, para que no afectara a los experimentos y se pudieran trabajar los datos, se le asigno el número 99 como valor centinela, y de esta manera poder identificar esas celdas como no aplicable, evitando así que estas observaciones fueran interpretadas erróneamente durante el análisis. En este conjunto de datos las respuestas en la pregunta del área de interés se dividen de la siguiente manera.

Tabla 5.18. Tabla donde se muestra las 6 clases utilizadas en este experimento.

Área de interés	Número de estudiantes
Redes	6
Programación	7
Ingeniería de software	3
Tratamiento de la información	5
Entorno social	9
Interacción hombre-máquina	2

5.7.2. Diseño experimental

En este experimento primero convertimos todos los tipos de datos a entero. Por ende, la variable de género se dividió en 2 clases: 1 para femenino y 0 para masculino. Posteriormente se realizó una matriz de correlación para analizar la relación entre las variables. En la imagen 5.10 se puede observar la gráfica de correlación entre las variables.

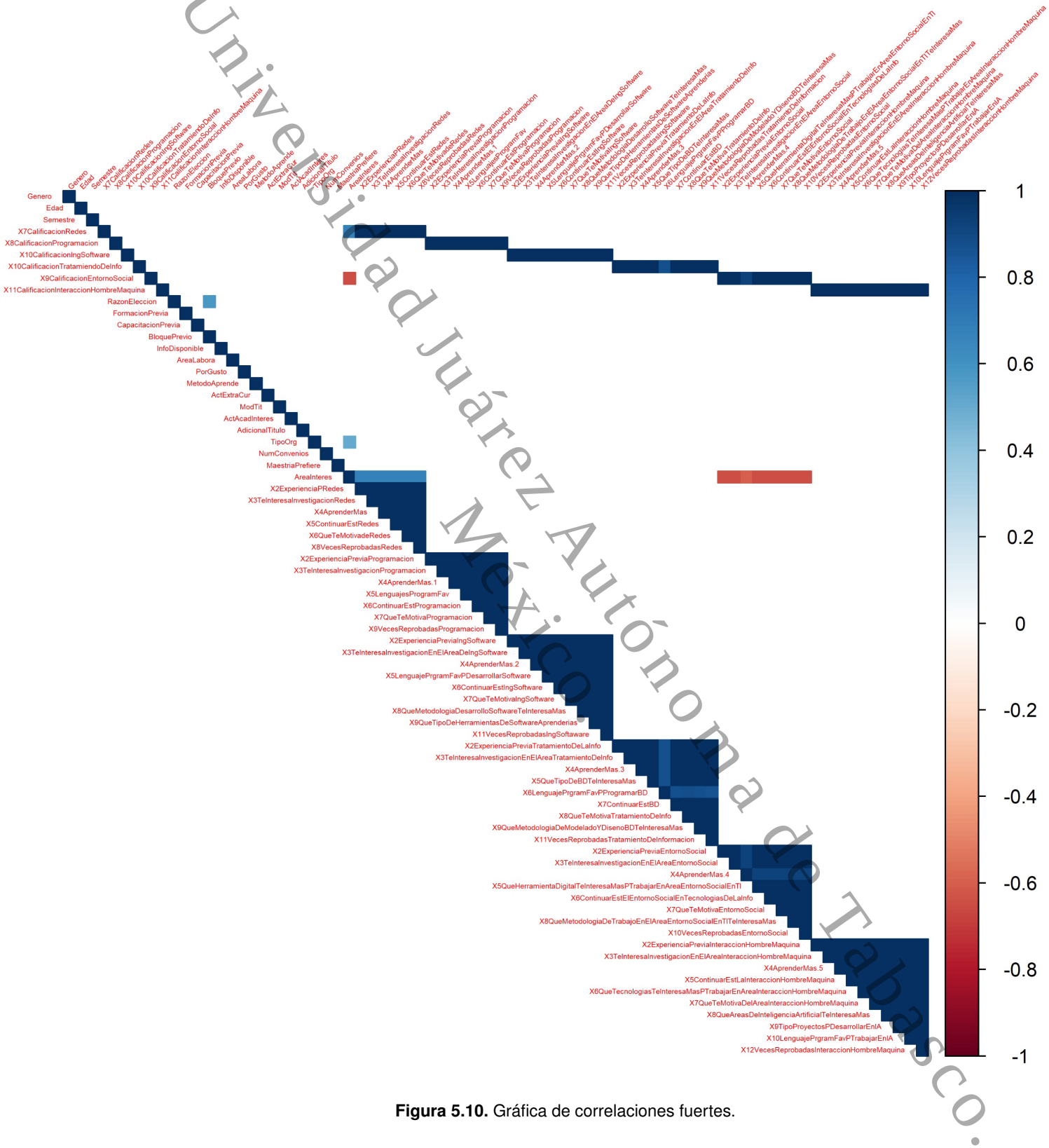

Figura 5.10. Gráfica de correlaciones fuertes.

Como resultados se observa que las variables de calificación de cada área correspondiente

muestran una correlación fuerte con las variables de la misma área, así como la variable del área de interés, muestra una relación media-fuerte con la calificación de redes y una relación fuerte negativa con la variable de calificación de entorno social.

Posteriormente se crearon modelos con los clasificadores, *Random forest*, J48, Jrip,OneR, SVM con kernel lineal, SVM kernel polinomial, SVM kernel radial, SVM kernel sigmoidal. En la Figura 5.11 se observa el proceso que se llevó a cabo para la creación de cada uno de los modelos. Como se mencionó anteriormente, se utilizaron todos los atributos del dataset, es decir 74 atributos y 32 instancias. Para este experimento se tomó el 2/3 de los datos para entrenamiento y 1/3 para prueba, se aplicó cada clasificador de manera independiente y se evaluó el rendimiento de los modelos.

Para evaluar los modelos se tomaron en cuenta las métricas de rendimiento: accuracy, sensibilidad, especificidad, kappa y accuracy balanceado, Estas se analizaron para cada una de las seis clases presentes en el conjunto de datos.

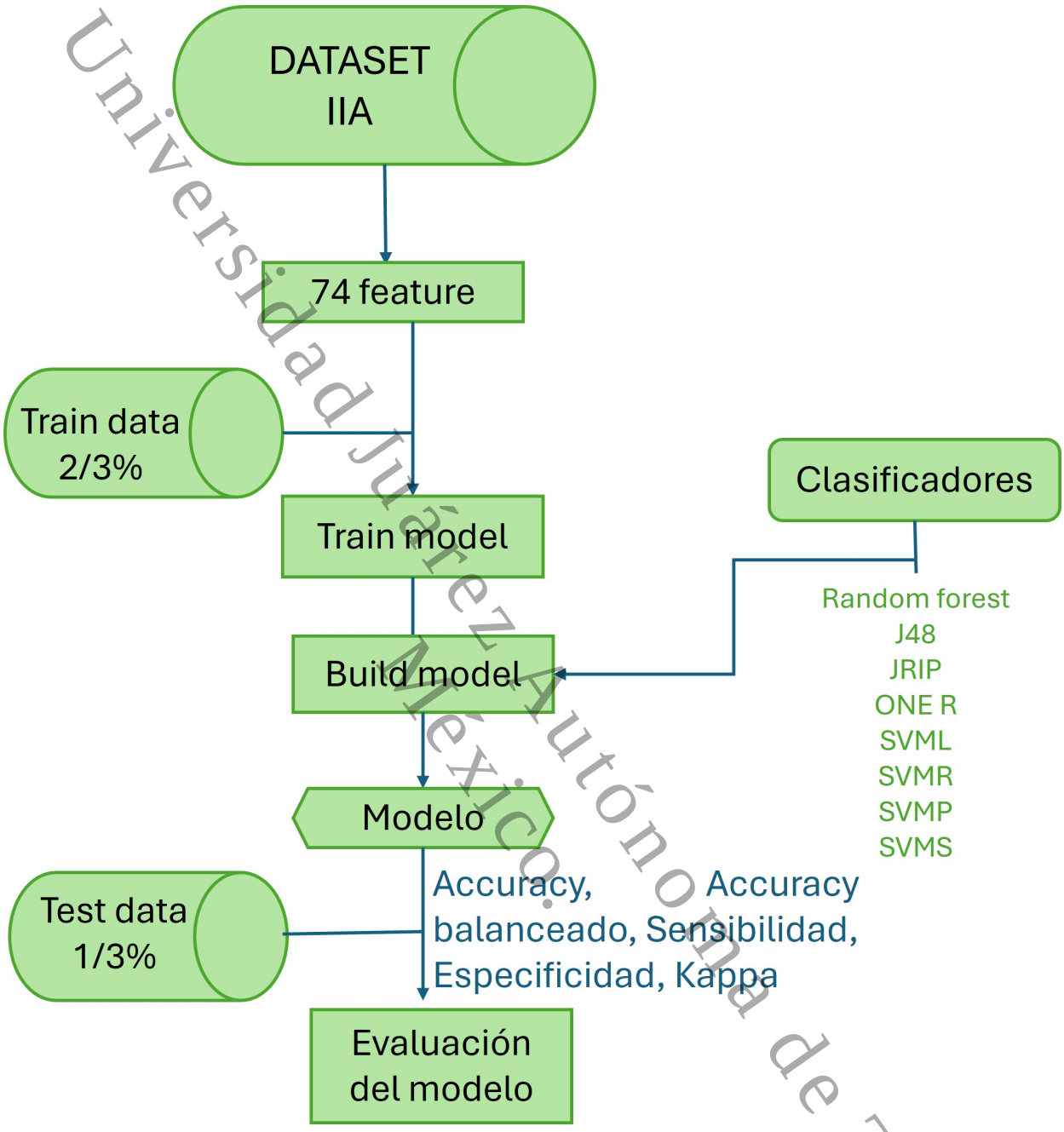

Figura 5.11. Diagrama de flujo del proceso de los datos, utilizando el *dataset* de IIA.

En la Tabla 5.19 se muestra un resumen general del rendimiento de los modelos utilizando las métricas ya mencionadas. El *accuracy* y Kappa corresponden a la clasificación global, mientras que la sensibilidad, especificidad y el *accuracy* balanceado se calculó con el promedio obtenido por clase.

Tabla 5.19. Medidas de rendimiento de los clasificadores.

Clasificador	Accuracy	Kappa	Especificidad	Sensibilidad	Accuracy balanceado
Random forest	**1**	**1**	**1**	**0.83**	**0.83**
J48	**1**	**1**	**1**	**0.83**	**0.83**
JRIP	**1**	**1**	**1**	**0.83**	**0.83**
OneR	0.5	0.35	0.89	0.33	0.53
SVMKL	0.9	0.87	0.98	0.77	0.80
SVMKP	0.8	0.74	0.95	0.69	0.74
SVMKR	0.9	0.87	0.97	0.77	0.79
SVMKS	0.9	0.87	0.98	0.77	0.80

5.7.3. Modelos con mejor desempeño

Random forest, J48 y Jrip presentan un desempeño perfecto en términos de *accuracy* y Kappa, lo que indica una clasificación totalmente correcta en el conjunto de datos utilizado. Sin embargo, su sensibilidad y el *accuracy* balanceado promedio fueron de 0.83, lo que sugiere que, aunque en general clasificaron correctamente, algunas clases fueron menos detectadas. Por otro lado, SVM con los 3 kernels utilizados también mostraron un rendimiento alto y consistente, con un *accuracy* de 0.90 y valores de Kappa de 0.87,mientras que su sensibilidad promedio es de 0.77 y *accuracy* balanceado de 0.80 los posiciona como modelos robustos en contextos de clases desbalanceadas.

5.7.4. Modelo con peor desempeño

OneR tuvo el rendimiento más bajo, con un *accuracy* de 0.50 y un Kappa de 0.35. Además, su sensibilidad fue baja (0.33), reflejando una alta tasa de falsos negativos, y su *accuracy* balanceado fue solo de 0.53, lo que demuestra que no clasificó bien de forma uniforme entre clases.

En la Tabla 5.20 se presentan los valores promedio por clase de las métricas de sensibilidad, especificidad y *accuracy* balanceado, calculadas a partir del desempeño de los distintos modelos evaluados. Este análisis permite identificar cómo se comportaron en conjunto, los modelos frente a cada una de las clases consideradas.

Tabla 5.20. Medidas de rendimiento de los clasificadores.

Clase	Sensibilidad	Especificidad	Accuracy balanceado
1	**0.875**	**1**	**0.937**
2	0.937	0.90	0.92
3	**0.875**	**1**	**0.937**
4	0.875	0.983	0.928
5	0.83	0.981	0.906
6	NA	0.975	NA

Como se observa, las clases 1 a 5 presentan valores altos tanto de sensibilidad como de especificidad, lo cual se traduce en valores de *accuracy* balanceado superiores al 90 %. Esto nos indica un desempeño uniforme y satisfactorio por parte del conjunto de modelos. En particular, las clases 1 y 3 alcanzaron una especificidad perfecta (1), lo que implica que los modelos evitaron totalmente los falsos positivos para esas clases.

Sin embargo, para la clase 6 no se pudo calcular la sensibilidad ni el *accuracy* balanceado. Esto sugiere un desbalance en la distribución de clases y representa una limitación importante que debe ser considerada en futuros trabajos, ya sea mediante técnicas de muestreo.

En general, estos resultados confirman que los modelos evaluados tienen un rendimiento alto y constante en la mayoría de las clases, es decir que logran identificar correctamente el área de interés de los estudiantes de la Ingeniería en Informática Administrativa, cabe destacar que la clase 6 corresponde al área de interés Interacción hombre-máquina y cuenta únicamente con dos casos. Debido a esta baja representación, no fue posible calcular métricas como la sensibilidad y el *accuracy* balanceado para esta categoría.

De igual forma se aplicaron los filtros CFS, *Chi-squared*, *Information gain*, *Random forest*.

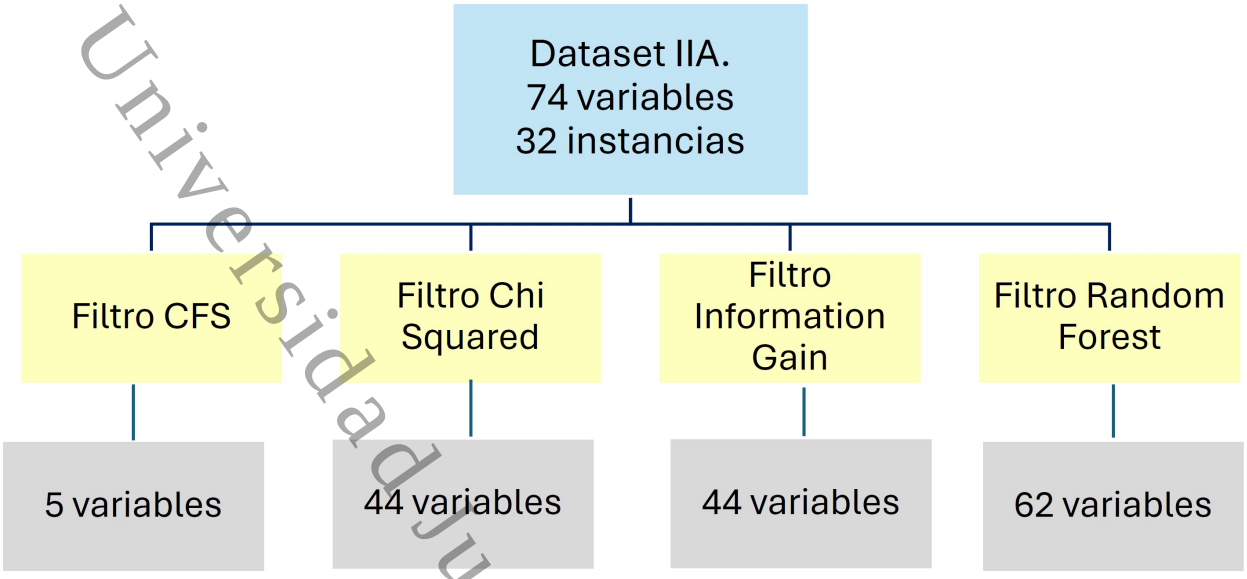

Figura 5.12. Resultados de los métodos filtro utilizando todas las variables del conjunto de datos.

Las variables relevantes encontradas por los métodos filtro son las siguientes:

- **Variables relevantes encontradas por el filtro CFS**
  1. X7CalificacionRedes
  2. X8CalificacionProgramacion
  3. X9CalificacionEntornoSocial
  4. X2ExperienciaPreviaIngSoftware
  5. X2ExperienciaPreviaTratamientoDeLaInfo
- **Variables relevantes encontradas por los filtros *chi-squared* e *information gain***
  1. X7CalificacionRedes
  2. X8CalificacionProgramacion
  3. X10CalificacionIngSoftware
  4. X10CalificacionTratamiendoDeInfo
  5. X9CalificacionEntornoSocial
  6. X2ExperienciaPRedes

7. X3TeInteresaInvestigacionRedes
8. X4AprenderMas
9. X5ContinuarEstRedes
10. X6QueTeMotivadeRedes
11. X8VecesReprobadasRedes
12. X2ExperienciaPreviaProgramacion
13. X3TeInteresaInvestigacionProgramacion
14. X4AprenderMas.1
15. X5LenguajesProgramFav
16. X6ContinuarEstProgramacion
17. X7QueTeMotivaProgramacion
18. X9VecesReprobadasProgramacion
19. X2ExperienciaPreviaIngSoftware
20. X3TeInteresaInvestigacionEnElAreaDeIngSoftware
21. X4AprenderMas.2
22. X5LenguajePrgramFavPDesarrollarSoftware
23. X6ContinuarEstIngSoftware
24. X7QueTeMotivaIngSoftware
25. X8QueMetodologiaDesarrolloSoftwareTeInteresaMas
26. X9QueTipoDeHerramientasDeSoftwareAprenderias
27. X11VecesReprobadasIngSoftaware
28. X2ExperienciaPreviaTratamientoDeLaInfo
29. X3TeInteresaInvestigacionEnElAreaTratamientoDeInfo
30. X4AprenderMas.3
31. X5QueTipoDeBDTeInteresaMas
32. X7ContinuarEstBD

33. X8QueTeMotivaTratamientoDeInfo
34. X9QueMetodologiaDeModeladoYDisenoBDTeInteresaMas
35. X11VecesReprobadasTratamientoDeInformacion
36. X2ExperienciaPreviaEntornoSocial
37. X3TeInteresaInvestigacionEnElAreaEntornoSocial
38. X5QueHerramientaDigitalTeInteresaMasPTrabajarEnAreaEntornoSocialEnTI
39. X6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo
40. X7QueTeMotivaEntornoSocial
41. X8QueMetodologiaDeTrabajoEnElAreaEntornoSocialEnTITeInteresaMas
42. X10VecesReprobadasEntornoSocial
43. X4AprenderMas.4
44. X6LenguajePrgramFavPProgramarBD

- **Variables relevantes encontradas por el filtro *random forest***

1. X9CalificacionEntornoSocial
2. X4AprenderMas
3. X2ExperienciaPreviaEntornoSocial
4. X7CalificacionRedes
5. X10VecesReprobadasEntornoSocial
6. X3TeInteresaInvestigacionProgramacion
7. X5QueHerramientaDigitalTeInteresaMasPTrabajarEnAreaEntornoSocialEnTI
8. X8QueMetodologiaDeTrabajoEnElAreaEntornoSocialEnTITeInteresaMas
9. X9VecesReprobadasProgramacion
10. X2ExperienciaPreviaProgramacion
11. X6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo
12. X4AprenderMas.1

13. X6QueTeMotivadeRedes
14. X3TeInteresaInvestigacionEnElAreaEntornoSocial
15. X3TeInteresaInvestigacionRedes
16. X6ContinuarEstProgramacion
17. X7QueTeMotivaEntornoSocial
18. X8CalificacionProgramacion
19. X5LenguajesProgramFav
20. X5ContinuarEstRedes
21. X7QueTeMotivaProgramacion
22. X2ExperienciaPRedes
23. X8VecesReprobadasRedes
24. X10CalificacionTratamiendoDeInfo
25. X8QueTeMotivaTratamientoDeInfo
26. X2ExperienciaPreviaTratamientoDeLaInfo
27. X5QueTipoDeBDTeInteresaMas
28. X7ContinuarEstBD
29. X11VecesReprobadasTratamientoDeInformacion
30. X3TeInteresaInvestigacionEnElAreaTratamientoDeInfo
31. X4AprenderMas.3
32. X9QueMetodologiaDeModeladoYDisenoBDTeInteresaMas
33. X4AprenderMas.4
34. X3TeInteresaInvestigacionEnElAreaDeIngSoftware
35. X8QueMetodologiaDesarrolloSoftwareTeInteresaMas
36. X7QueTeMotivaIngSoftware
37. X4AprenderMas.2
38. X6ContinuarEstIngSoftware

39. X11VecesReprobadasIngSoftaware
40. X5LenguajePrgramFavPDesarrollarSoftware
41. X2ExperienciaPreviaIngSoftware
42. X10CalificacionIngSoftware
43. X5ContinuarEstLaInteraccionHombreMaquina
44. X9QueTipoDeHerramientasDeSoftwareAprenderias
45. X2ExperienciaPreviaInteraccionHombreMaquina
46. X9TipoProyectosPDesarrollarEnIA
47. X8QueAreasDeInteligenciaArtificialTeInteresaMas
48. X6QueTecnologiasTeInteresaMasPTrabajarEnAreaInteraccionHombreMaquina
49. X10LenguajePrgramFavPTrabajarEnIA
50. X6LenguajePrgramFavPProgramarBD
51. X4AprenderMas.5
52. X11CalificacionInteraccionHombreMaquina
53. X7QueTeMotivaDelAreaInteraccionHombreMaquina
54. TipoOrg
55. X12VecesReprobadasInteraccionHombreMaquina
56. X3TeInteresaInvestigacionEnElAreaInteraccionHombreMaquina
57. ModTit
58. PorGusto
59. AdicionalTitulo
60. MaestriaPrefiere
61. BloquePrevio
62. RazonEleccion

Como podemos observar en la Figura 5.12 el filtro *Chi-squared* y el filtro *Information gain* dan la misma cantidad de variables relevantes, y en el listado de las variables se muestra que estas

son iguales. Mientras que con *Random forest* se obtuvieron 62 variables relevantes en las que se encuentran inmersas las variables relevantes encontradas con los otros filtros. El filtro con menor cantidad de variables es CFS, con 5 variables relevantes.

5.8. Experimento utilizando el conjunto de datos obtenido de la carrera de Ingeniería en Sistemas Computacionales (ISC)

5.8.1. Datos

Para este experimento se utilizó el conjunto de datos recolectado de los estudiantes de ISC, en donde se encuentran todas las respuestas de los estudiantes, omitiendo la pregunta de Razones por la cual eligió esa área de interés y la marca temporal que es la fecha y hora en que respondió el cuestionario el estudiante, por lo tanto, son 74 variables y 103 observaciones.

En experimentos anteriores se ha mencionado que las celdas en blanco le colocamos el número 99, para que no afectara a los experimentos y se pudieran trabajar los datos, se le asigno el número 99 como valor centinela, y de esta manera poder identificar esas celdas como no aplicable, evitando así que estas observaciones fueran interpretadas erróneamente durante el análisis. En este conjunto de datos las respuestas de la pregunta del área de interés se dividen de la siguiente manera.

Tabla 5.21. Tabla de clases que se encuentran en el dataset ISC.

Área de interés	Número de estudiantes
Redes	14
Programación	44
Ingeniería de software	17
Tratamiento de la información	18
Entorno social	6
Interacción hombre-máquina	4

5.8.2. Diseño experimental

En este experimento primero convertimos todos los tipos de datos a entero, por ende, la variable de género se dividió en 2 clases: 1 para femenino y 0 para masculino. Posteriormente se

crearon modelos con los clasificadores *Random forest*, J48, Jrip, OneR, SVM con kernel lineal, SVM kernel polinomial, SVM kernel radial, SVM kernel sigmoidal.

En la Figura 5.13 se observa el proceso que se llevó a cabo para la creación de cada uno de los modelos. Como se mencionó anteriormente, se utilizaron 74 atributos y 103 instancias. Para este experimento se tomaron $\frac{2}{3}$ de los datos para entrenamiento y $\frac{1}{3}$ para pruebas. Se aplicó cada clasificador de manera independiente y se evaluó el rendimiento de los modelos. Para evaluar los modelos se tomaron en cuenta las métricas de rendimiento *accuracy*, sensibilidad, especificidad, kappa y *accuracy* balanceado. Estas se analizaron para cada una de las seis clases presentes en el conjunto de datos.

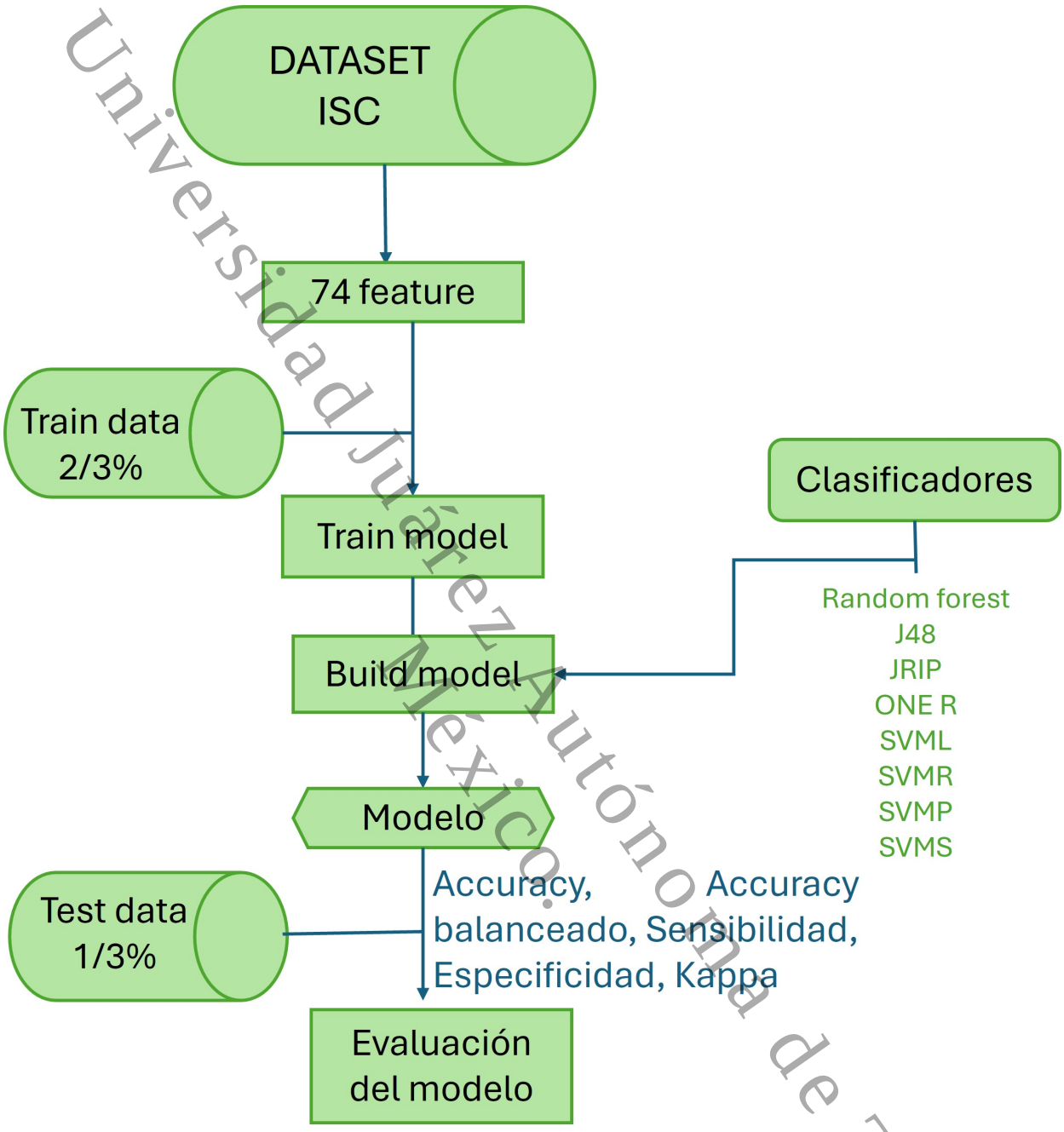

Figura 5.13. Diagrama de flujo del proceso de los datos, utilizando el *datset* de ISC.

En la Tabla 5.22 se muestra un resumen general del rendimiento de los modelos utilizando las métricas ya mencionadas. El *accuracy* y Kappa corresponden a la clasificación global, mientras que la sensibilidad, especificidad y el *accuracy* balanceado se calcularon con el promedio obtenido por clase.

Tabla 5.22. Medidas de rendimiento de los clasificadores.

Clasificador	Accuracy	Kappa	Especificidad	Sensibilidad	Accuracy balanceado
Random forest	**1**	**1**	**1**	**1**	**1**
J48	0.94	0.92	0.98	0.98	0.98
JRIP	0.97	0.96	0.97	0.99	0.98
OneR	0.60	0.44	0.33	0.92	0.62
SVMKL	**1**	**1**	**1**	**1**	**1**
SVMKP	**1**	**1**	**1**	**1**	**1**
SVMKR	**1**	**1**	**1**	**1**	**1**
SVMKS	**1**	**1**	**1**	**1**	**1**

5.8.3. Modelos con mejor desempeño

Random forest, SVM con kernel lineal, polinomial, radial y sigmoidal presentan un desempeño perfecto en todas las métricas, lo que indica una clasificación totalmente correcta en el conjunto de datos utilizado. J48 y Jrip también mostraron un rendimiento alto y consistente, en donde las métricas muestran un resultado por encima del 0.90.

5.8.4. Modelo con peor desempeño

OneR tuvo el rendimiento más bajo, con un *accuracy* de 0.60 y un Kappa de 0.44, además, su sensibilidad fue baja (0.33), reflejando una alta tasa de falsos negativos, y su accuracy balanceado fue solo de 0.62. En general, estos resultados confirman que los modelos evaluados tienen un rendimiento alto y constante en la mayoría de las clases, es decir que logran identificar correctamente el área de interés de los estudiantes de la Ingeniería en Sistemas Computacionales. De igual forma se aplicaron los filtros CFS, *Chi-squared*, *Information gain*, *Random forest*.

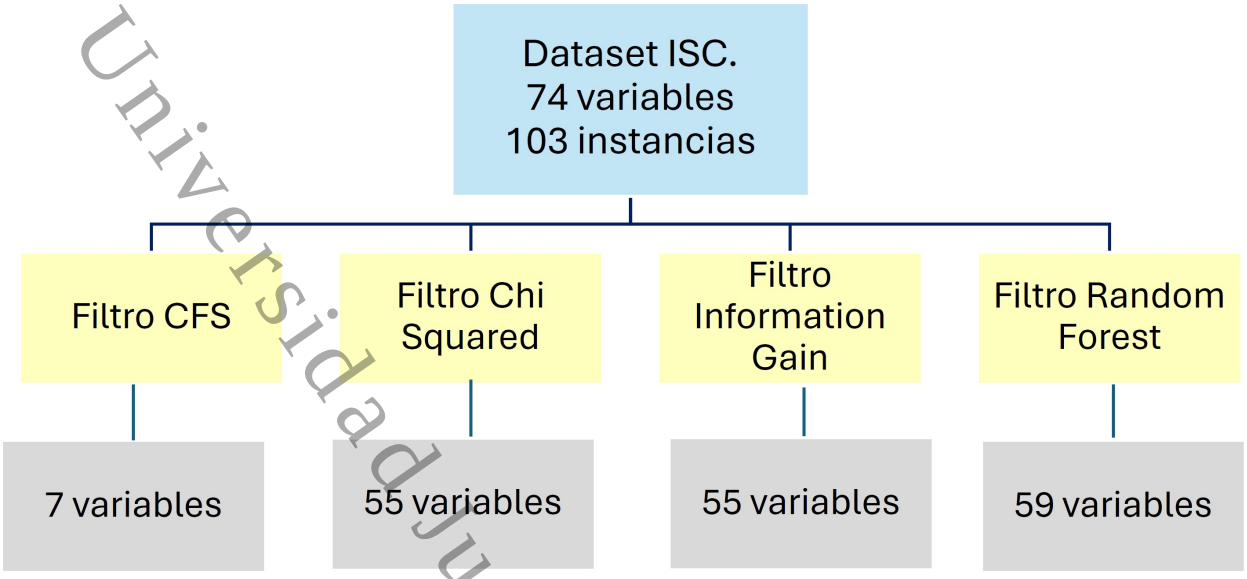

Figura 5.14. Resultados de los métodos filtro utilizando todas las variables del conjunto de datos.

Como podemos observar en la Figura 5.14 el filtro *Chi-squared* y el filtro *Information gain* dan las mismas 55 variables relevantes, mientras que con *Random forest* se obtuvieron 59 variables relevantes en las que se encuentran inmersas las variables relevantes encontradas con los otros filtros. El filtro con menor cantidad de variables es CFS, con 7 variables relevantes.

Las variables relevantes encontradas por los métodos filtro son las siguientes:

- **Variables relevantes encontradas por el filtro CFS**
  1. X7CalificacionRedes
  2. X8CalificacionProgramacion
  3. X10CalificacionIngSoftware
  4. X10CalificacionTratamiendoDeInfo
  5. X9CalificacionEntornoSocial
  6. X11CalificacionInteraccionHombreMaquina
  7. X3TeInteresaInvestigacionProgramacion
- **Variables relevantes encontradas por los filtros *chi-squared* e *information gain***
  1. X7CalificacionRedes

2. X8CalificacionProgramacion
3. X10CalificacionIngSoftware
4. X10CalificacionTratamiendoDeInfo
5. X9CalificacionEntornoSocial
6. X11CalificacionInteraccionHombreMaquina
7. X2ExperienciaPRedes
8. X3TeInteresaInvestigacionRedes
9. X4AprenderMas
10. X5ContinuarEstRedes
11. X6QueTeMotivadeRedes
12. X8VecesReprobadasRedes
13. X3TeInteresaInvestigacionProgramacion
14. X6ContinuarEstProgramacion
15. X7QueTeMotivaProgramacion
16. X9VecesReprobadasProgramacion
17. X2ExperienciaPreviaIngSoftware
18. X3TeInteresaInvestigacionEnElAreaDeIngSoftware
19. X4AprenderMas.2
20. X6ContinuarEstIngSoftware
21. X7QueTeMotivaIngSoftware
22. X8QueMetodologiaDesarrolloSoftwareTeInteresaMas
23. X9QueTipoDeHerramientasDeSoftwareAprenderias
24. X11VecesReprobadasIngSoftaware
25. X2ExperienciaPreviaTratamientoDeLaInfo
26. X3TeInteresaInvestigacionEnElAreaTratamientoDeInfo
27. X4AprenderMas.3

28. X5QueTipoDeBDTeInteresaMas
29. X6LenguajePrgramFavPProgramarBD
30. X7ContinuarEstBD
31. X8QueTeMotivaTratamientoDeInfo
32. X9QueMetodologiaDeModeladoYDisenoBDTeInteresaMas
33. X11VecesReprobadasTratamientoDeInformacion
34. X2ExperienciaPreviaEntornoSocial
35. X3TeInteresaInvestigacionEnElAreaEntornoSocial
36. X4AprenderMas.4
37. X5QueHerramientaDigitalTeInteresaMasPTrabajarEnAreaEntornoSocialEnTI
38. X6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo
39. X7QueTeMotivaEntornoSocial
40. X8QueMetodologiaDeTrabajoEnElAreaEntornoSocialEnTITeInteresaMas
41. X10VecesReprobadasEntornoSocial
42. X2ExperienciaPreviaInteraccionHombreMaquina
43. X3TeInteresaInvestigacionEnElAreaInteraccionHombreMaquina
44. X4AprenderMas.5
45. X5ContinuarEstLaInteraccionHombreMaquina
46. X6QueTecnologiasTeInteresaMasPTrabajarEnAreaInteraccionHombreMaquina
47. X7QueTeMotivaDelAreaInteraccionHombreMaquina
48. X8QueAreasDeInteligenciaArtificialTeInteresaMas
49. X9TipoProyectosPDesarrollarEnIA
50. X10LenguajePrgramFavPTrabajarEnIA
51. X12VecesReprobadasInteraccionHombreMaquina
52. X2ExperienciaPreviaProgramacion
53. X4AprenderMas.1

54. X5LenguajesProgramFav
55. X5LenguajePrgramFavPDesarrollarSoftware

- **Variables relevantes encontradas por el filtro *random forest***

1. ActExtraCur
2. BloquePrevio
3. MetodoAprende
4. ModTit
5. X10CalificacionIngSoftware
6. X10CalificacionTratamiendoDeInfo
7. X10LenguajePrgramFavPTrabajarEnIA
8. X10VecesReprobadasEntornoSocial
9. X11CalificacionInteraccionHombreMaquina
10. X11VecesReprobadasIngSoftaware
11. X11VecesReprobadasTratamientoDeInformacion
12. X12VecesReprobadasInteraccionHombreMaquina
13. X2ExperienciaPRedes
14. X2ExperienciaPreviaEntornoSocial
15. X2ExperienciaPreviaIngSoftware
16. X2ExperienciaPreviaInteraccionHombreMaquina
17. X2ExperienciaPreviaProgramacion
18. X2ExperienciaPreviaTratamientoDeLaInfo
19. X3TeInteresaInvestigacionEnElAreaDeIngSoftware
20. X3TeInteresaInvestigacionEnElAreaEntornoSocial
21. X3TeInteresaInvestigacionEnElAreaInteraccionHombreMaquina
22. X3TeInteresaInvestigacionEnElAreaTratamientoDeInfo

23. X3TeInteresaInvestigacionProgramacion
24. X3TeInteresaInvestigacionRedes
25. X4AprenderMas
26. X4AprenderMas.1
27. X4AprenderMas.2
28. X4AprenderMas.3
29. X4AprenderMas.4
30. X4AprenderMas.5
31. X5ContinuarEstLaInteraccionHombreMaquina
32. X5ContinuarEstRedes
33. X5LenguajePrgramFavPDesarrollarSoftware
34. X5LenguajesProgramFav
35. X5QueHerramientaDigitalTeInteresaMasPTrabajarEnAreaEntornoSocialEnTI
36. X5QueTipoDeBDTeInteresaMas
37. X6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo
38. X6ContinuarEstIngSoftware
39. X6ContinuarEstProgramacion
40. X6LenguajePrgramFavPProgramarBD
41. X6QueTecnologiasTeInteresaMasPTrabajarEnAreaInteraccionHombreMaquina
42. X6QueTeMotivadeRedes
43. X7CalificacionRedes
44. X7ContinuarEstBD
45. X7QueTeMotivaDelAreaInteraccionHombreMaquina
46. X7QueTeMotivaEntornoSocial
47. X7QueTeMotivaIngSoftware
48. X7QueTeMotivaProgramacion

49. X8CalificacionProgramacion
50. X8QueAreasDeInteligenciaArtificialTeInteresaMas
51. X8QueMetodologiaDesarrolloSoftwareTeInteresaMas
52. X8QueMetodologiaDeTrabajoEnElAreaEntornoSocialEnTITeInteresaMas
53. X8QueTeMotivaTratamientoDeInfo
54. X8VecesReprobadasRedes
55. X9CalificacionEntornoSocial
56. X9QueMetodologiaDeModeladoYDisenoBDTeInteresaMas
57. X9QueTipoDeHerramientasDeSoftwareAprenderias
58. X9TipoProyectosPDesarrollarEnIA
59. X9VecesReprobadasProgramacion

5.9. Experimento utilizando el conjunto de datos obtenido de la carrera de Licenciatura en Sistemas Computacionales (LSC)

5.9.1. Datos

Para este experimento se utilizó el conjunto de datos recolectado de los estudiantes de LSC, en donde se encuentran todas las respuestas de los estudiantes, omitiendo la pregunta de Razones por la cual eligió esa área de interés y la marca temporal que es la fecha y hora en que respondió el cuestionario el estudiante. De igual manera se omitieron 3 áreas de interés ya que no hubo alumnos que le interesaran esas áreas; por lo tanto, el dataset cuenta con 47 variables y 11 observaciones.

En experimentos anteriores se ha mencionado que las celdas en blanco le colocamos el número 99, para que no afectara a los experimentos y se pudieran trabajar los datos, se le asigno el número 99 como valor centinela, y de esta manera poder identificar esas celdas como no aplicable, evitando así que estas observaciones fueran interpretadas erróneamente durante el análisis. En este conjunto de datos las respuestas de la pregunta del área de interés se dividen de la siguiente manera.

Tabla 5.23. Tabla de clases del dataset LSC.

Área de interés	Número de estudiantes
Programación	8
Ingeniería de software	2
Tratamiento de la información	1

5.9.2. Diseño experimental

En este experimento primero convertimos todos los tipos de datos a entero, por ende, la variable de género se dividió en 2 clases: 1 para femenino y 0 para masculino. Posteriormente se crearon modelos con los clasificadores *Random forest*, J48, Jrip, OneR, SVM con kernel lineal, SVM kernel polinomial, SVM kernel radial, SVM kernel sigmoidal.

En la Figura 5.15 se observa el proceso que se llevó a cabo para la creación de cada uno de los modelos. Como se mencionó anteriormente, se utilizaron 47 atributos y 11 instancias. Para este experimento se tomó el 2/3 de los datos para entrenamiento y 1/3 para prueba. Se aplicó cada clasificador de manera independiente y se evaluó el rendimiento de los modelos. Para evaluar los modelos se tomaron en cuenta las métricas de rendimiento *accuracy*, sensibilidad, especificidad, kappa y *accuracy* balanceado.

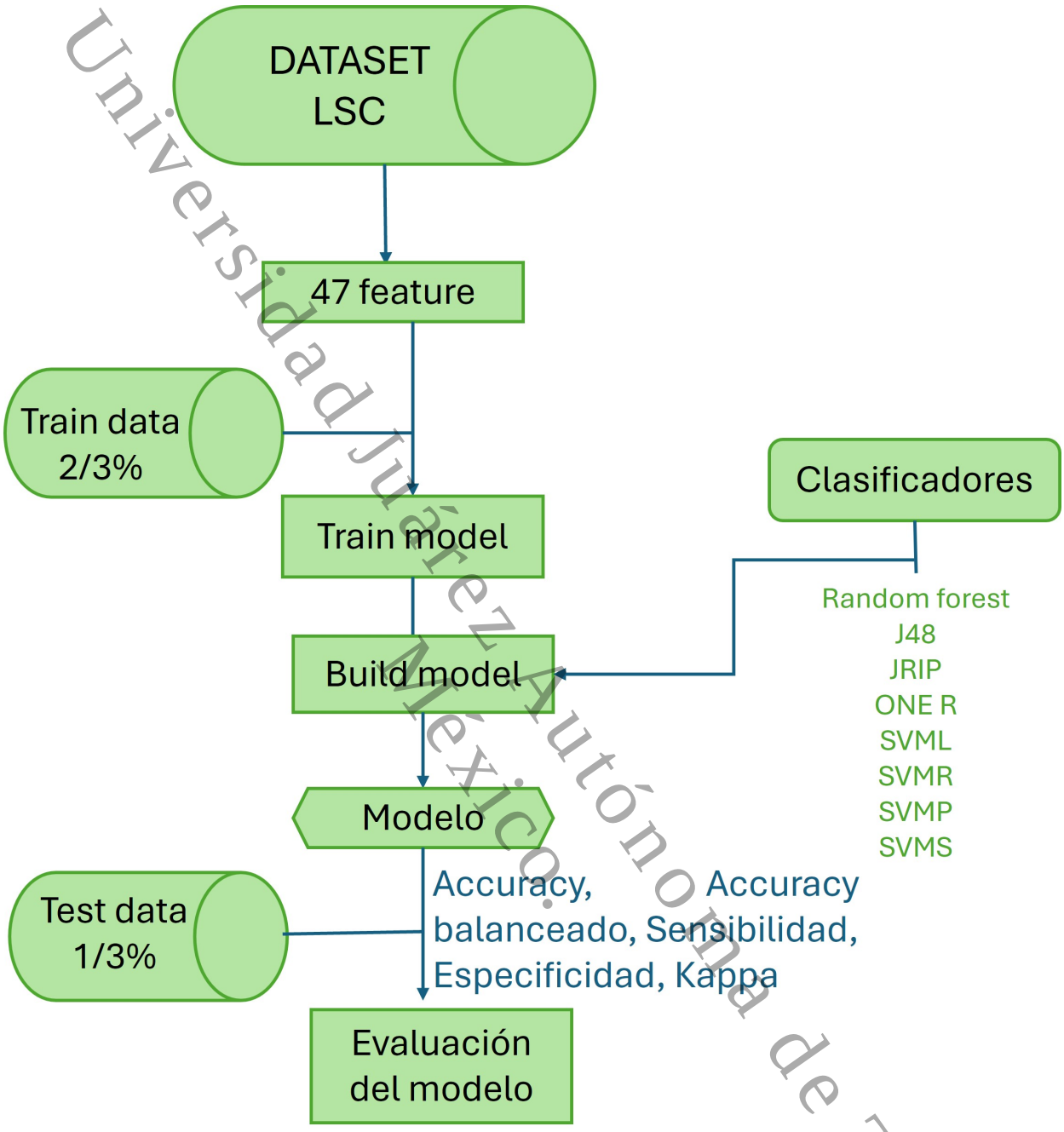

Figura 5.15. Diagrama de flujo del proceso de los datos, utilizando el *dataset* LSC.

En la Tabla 5.24 se muestra un resumen general del rendimiento de los modelos utilizando las métricas ya mencionadas. El *accuracy* y Kappa corresponden a la clasificación global, mientras que la sensibilidad, especificidad y el *accuracy* balanceado está dividido por clase, en este caso 1: Programación, 2:Ingeniería de software, 3: Tratamiento de la información.

Tabla 5.24. Medidas de rendimiento de los clasificadores.

Clasificador	Accuracy	Kappa	Especificidad			Sensibilidad			Accuracy balanceado		
Random forest	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
J48	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
JRIP	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
OneR	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
SVMKL	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
SVMKP	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
SVMKR	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
SVMKS	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA

Como podemos observar en la Tabla 5.24, debido a que en la clase 2 y 3 la cantidad de estudiantes es de uno y dos estudiantes respectivamente, los modelos no logran clasificar de manera correcta, todo esto a pesar de que en el *accuracy* nos da un resultado igual a 1. A pesar de los resultados se aplicaron los filtros CFS, *Chi-squared*, *Information gain* y *Random forest*.

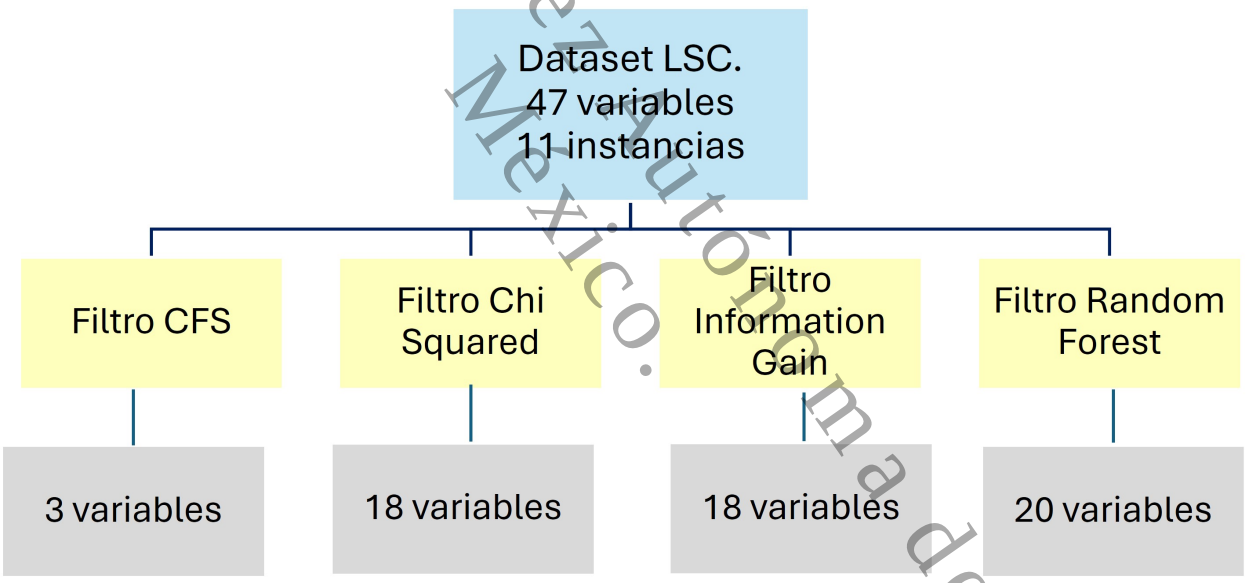

Figura 5.16. Resultados de los métodos filtro utilizando todas las variables del conjunto de datos.

Como podemos observar en la Figura 5.16, el filtro CFS nos da un total de 3 variables relevantes, el filtro *Chi-squared* y el filtro *Information gain* dan los mismos resultados, es decir, 18 variables relevantes, mientras que con *Random forest* se obtuvieron 20 variables relevantes en las que se encuentran inmersas las variables relevantes encontradas con los otros filtros.

A continuación se muestran las variables relevantes encontradas por los métodos filtro.

- **Variables relevantes encontradas por el filtro CFS**

  1. X8CalificacionProgramacion
  2. X2ExperienciaPreviaProgramacion
  3. X10CalificacionIngSoftware

- **Variables relevantes encontradas por los filtros *information gain* y *chi-squared***

  1. X8CalificacionProgramacion
  2. X10CalificacionIngSoftware
  3. X2ExperienciaPreviaProgramacion
  4. X3TeInteresaInvestigacionProgramacion
  5. X4AprenderMas
  6. X5LenguajesProgramFav
  7. X6ContinuarEstProgramacion
  8. X7QueTeMotivaProgramacion
  9. X9VecesReprobadasProgramacion
  10. X2ExperienciaPreviaIngSoftware
  11. X3TeInteresaInvestigacionEnElAreaDeIngSoftware
  12. X4AprenderMas.1
  13. X5LenguajePrgramFavPDesarrollarSoftware
  14. X6ContinuarEstIngSoftware
  15. X7QueTeMotivaIngSoftware
  16. X8QueMetodologiaDesarrolloSoftwareTeInteresaMas
  17. X9QueTipoDeHerramientasDeSoftwareAprenderias
  18. X11VecesReprobadasIngSoftaware

- **Variables relevantes encontradas por el filtro *random forest***

  1. ActAcadInteres

2. MetodoAprende
3. X10CalificacionIngSoftware
4. X11VecesReprobadasIngSoftaware
5. X2ExperienciaPreviaIngSoftware
6. X2ExperienciaPreviaProgramacion
7. X3TeInteresaInvestigacionEnElAreaDeIngSoftware
8. X3TeInteresaInvestigacionProgramacion
9. X4AprenderMas
10. X4AprenderMas.1
11. X5LenguajePrgramFavPDesarrollarSoftware
12. X5LenguajesProgramFav
13. X6ContinuarEstIngSoftware
14. X6ContinuarEstProgramacion
15. X7QueTeMotivaIngSoftware
16. X7QueTeMotivaProgramacion
17. X8CalificacionProgramacion
18. X8QueMetodologiaDesarrolloSoftwareTeInteresaMas
19. X9QueTipoDeHerramientasDeSoftwareAprenderias
20. X9VecesReprobadasProgramacion

5.10. Experimento utilizando el conjunto de datos obtenido de la carrera de Licenciatura en Tecnologías de la Información (LTI)

5.10.1. Datos

Para este experimento se utilizó el conjunto de datos recolectado de los estudiantes de LTI, donde se encuentran todas las respuestas de los estudiantes, omitiendo la pregunta de Razones

por la cuál eligió esa área de interés y la marca temporal que es la fecha y hora en que respondió el cuestionario el estudiante. De igual manera se omitieron 3 áreas de interés ya que en no hubo alumnos que le interesaran esas áreas; por lo tanto, el dataset cuenta con 53 variables y 06 observaciones.

En experimentos anteriores se ha mencionado que las celdas en blanco le colocamos el número 99, para que no afectara a los experimentos y se pudieran trabajar los datos, se le asigno el número 99 como valor centinela, y de esta manera poder identificar esas celdas como no aplicable, evitando así que estas observaciones fueran interpretadas erróneamente durante el análisis. En este conjunto de datos las respuestas de la pregunta del área de interés se dividen de la siguiente manera.

Tabla 5.25. Tabla de clases.

Área de interés	Número de estudiantes
Programación	3
Ingeniería de software	1
Entorno social	2

5.10.2. Diseño experimental

En este experimento primero convertimos todos los tipos de datos a entero, por ende, la variable de género se dividió en 2 clases: 1 para femenino y 0 para masculino. Posteriormente se crearon modelos con los clasificadores *Random forest*, J48, Jrip, OneR, SVM con kernel lineal, SVM kernel polinomial, SVM kernel radial, SVM kernel sigmoidal.

En la Figura 5.17 se observa el proceso que se llevó a cabo para la creación de cada uno de los modelos. Como se mencionó anteriormente, se utilizaron 47 atributos y 11 instancias. Para este experimento se tomó el $\frac{2}{3}$ de los datos para entrenamiento y $\frac{1}{3}$ para prueba. Se aplicó cada clasificador de manera independiente y se evaluó el rendimiento de los modelos. Para evaluar los modelos se tomaron en cuenta las métricas de rendimiento: *accuracy*, sensibilidad, especificidad, kappa y *accuracy* balanceado.

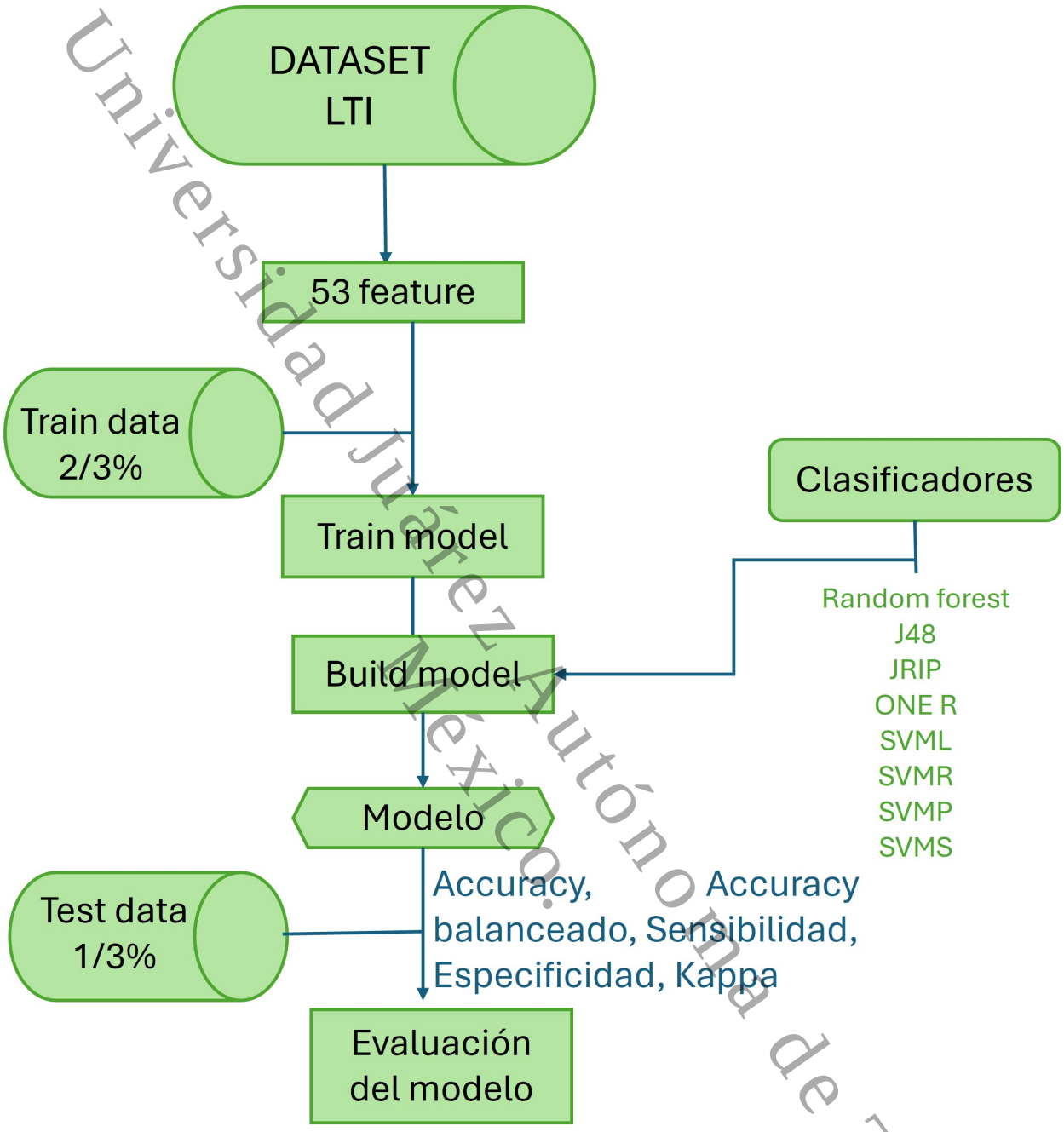

Figura 5.17. Diagrama de flujo del proceso de los datos, utilizando el *dataset* de LTI.

En la Tabla 5.26 se muestra un resumen general del rendimiento de los modelos utilizando las métricas ya mencionadas. El *accuracy* y Kappa corresponden a la clasificación global, mientras que la sensibilidad, especificidad y el *accuracy* balanceado está dividido por clase, en este caso 1: Programación, 2:Ingeniería de software, 3: Entorno social.

Tabla 5.26. Medidas de rendimiento de los clasificadores.

Clasificador	Accuracy	Kappa	Especificidad			Sensibilidad			Accuracy Balanceado		
Random forest	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
J48	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
JRIP	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
OneR	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
SVMKL	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
SVMKP	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
SVMKR	1	NAN	1	NA	NA	NA	1	1	NA	NA	NA
SVMKS	0	0	0	NA	NA	NA	1	0	NA	NA	NA

Como podemos observar en la Tabla 5.26 debido a que en la clase 2 y 3 la cantidad de estudiantes es de uno y dos estudiantes, los modelos no logran clasificar de manera correcta, todo esto a pesar de que en el accuracy nos da un resultado igual 1. A pesar de los resultados se aplicaron los filtros: CFS, *Chi-squared*, *Information gain* y *Random forest*.

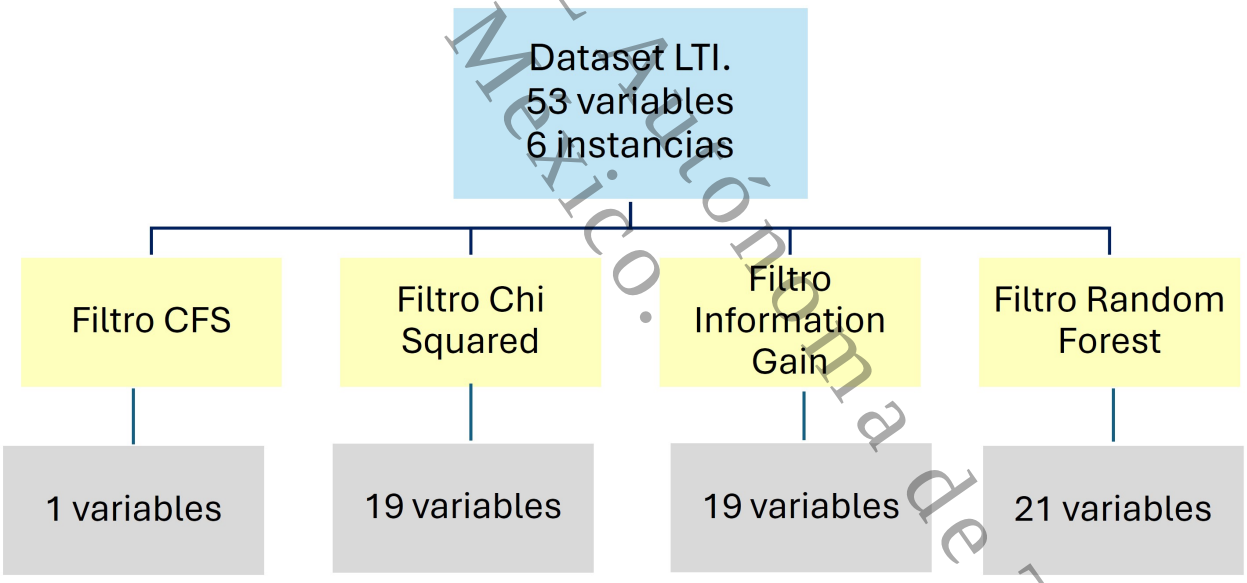

Figura 5.18. Resultados de los métodos filtro utilizando todas las variables del conjunto de datos.

Como podemos observar en la Figura 5.18 el filtro CFS nos da un total de 3 variables relevantes, el filtro *Chi-squared* y el filtro *Information gain* dan los mismos resultados, es decir 18 variables relevantes, mientras que con *Random forest* se obtuvieron 20 variables relevantes en las que se encuentran inmersas las variables relevantes encontradas con los otros filtros.

A continuación se muestran las variables relevantes encontradas por los métodos filtro.

- **Variables relevantes encontradas por el filtro CFS**

  1. Edad

- **Variables relevantes encontradas por los filtros *information gain* y *chi-squared***

  1. Edad
  2. X8CalificacionProgramacion
  3. MetodoAprende
  4. X2ExperienciaPreviaProgramacion
  5. X3TeInteresaInvestigacionProgramacion
  6. X5LenguajesProgramFav
  7. X6ContinuarEstProgramacion
  8. X7QueTeMotivaProgramacion
  9. X9VecesReprobadasProgramacion
  10. X9CalificacionEntornoSocial
  11. AreaLabora
  12. X2ExperienciaPreviaEntornoSocial
  13. X3TeInteresaInvestigacionEnElAreaEntornoSocial
  14. X4AprenderMas.3
  15. X5QueHerramientaDigitalTeInteresaMasPTrabajarEnAreaEntornoSocialEnTI
  16. X6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo
  17. X7QueTeMotivaEntornoSocial
  18. X8QueMetodologiaDeTrabajoEnElAreaEntornoSocialEnTITeInteresaMas
  19. X10VecesReprobadasEntornoSocial

- **Variables relevantes encontradas por el filtro *random forest***

  1. X8CalificacionProgramacion

2. X2ExperienciaPreviaProgramacion
3. X5LenguajesProgramFav
4. X5QueHerramientaDigitalTeInteresaMasPTrabajarEnAreaEntornoSocialEnTI
5. X9VecesReprobadasProgramacion
6. X6ContinuarEstProgramacion
7. MetodoAprende
8. X4AprenderMas.3
9. X3TeInteresaInvestigacionProgramacion
10. X7QueTeMotivaEntornoSocial
11. X7QueTeMotivaProgramacion
12. X10VecesReprobadasEntornoSocial
13. X6ContinuarEstElEntornoSocialEnTecnologiasDeLaInfo
14. X2ExperienciaPreviaEntornoSocial
15. X8QueMetodologiaDeTrabajoEnElAreaEntornoSocialEnTITeInteresaMas
16. X9CalificacionEntornoSocial
17. X3TeInteresaInvestigacionEnElAreaEntornoSocial
18. Edad
19. AreaLabora
20. ModTit
21. BloquePrevio

Capítulo 6

Contribuciones, conclusiones y trabajos futuros

6.1. Conclusiones

El objetivo principal de esta investigavción fue conocer los intereses académicos de la comunidad estudiantil de la DACyTI. En total se realizaron 10 experimentos, utilizando los clasificadores: *Random forest*, J48, JRip, OneR, Máquina de Vectores de Soporte (SVM) con kernel lineal, polinomial, radial y sigmoide; en algunos experimentos se utilizó k-NN. De igual manera se aplicaron técnicas de selección de atributos en Minería de datos, tales como: CFS, *Chi-squared*, *Information gain*, y *Random forest*. Estos se utilizaron para reducir la cantidad de variables en los conjuntos de datos.

Como resultados de los experimentos utilizando los clasificadores, pudimos observar que muestra mejor rendimiento *random forest*, ya que en la mayoria de las métricas da como resultado 1, esto quiere decir que logra predecir correctamente la totalidad de los casos en el conjunto de prueba. Esto significa que el modelo es altamente preciso al clasificar los datos, sin cometer errores en las predicciones realizadas bajo las condiciones del experimento. Este resultado sugiere que el *Random Forest* es una técnica robusta para este tipo de problema, ya que combina múltiples árboles de decisión y, mediante el voto mayoritario, reduce el riesgo de sobreajuste y mejora la capacidad de generalización. Sin embargo, es importante considerar que un rendimiento perfecto

podría estar asociado a características propias del conjunto de datos o al posible sobreajuste, por lo que sería recomendable validar el modelo con nuevos datos para confirmar su efectividad en escenarios reales.

En algunos experimentos realizados con los conjuntos de datos de LSC y LTI (ver los experimentos detalladamente 5.9, 5.10), debido al desbalanceo de clases presente en los datos, los clasificadores no pudieron calcular ciertas métricas evaluativas, resultando en valores NaN (Not a Number). Esta situación se debe a la insuficiente representación de algunas clases, lo que impide el cálculo adecuado de métricas como kappa, sensibilidad, especificidad y accuracy balanceado.

Con las técnicas de selección de características de los métodos filtro, se encontraron diferentes propuestas en la selección de las variables relevantes para la identificación del interés acdémico de los estuadiantes. Los filtros *chi-squared e information gain* obtenian los mismos resultados, mientras que *random forest* obtenia una mayor cantidad de variables en las cuales estaban inmersas las variables obtenidas por los otros filtros. La variables relevantes encontradas por los filtros *CFS, chi-squared, e information gain* se basan en las calificaciones obtenidas en asignaturas del área de interés, la experiencia previa a la elección de la carrera, que los motiva a estudiar, entre otras, esto quiere decir que para identificar los intereses académicos de los estudiantes de licenciatura con estos métodos filtro se toman en cuenta el rendimiento académico de los estudiantes, su formación previa (estudios de bachillerato, vida laboral, investigaciones adicionales) y la motivación actual. Mientras que utilizando el método filtro *random forest* nos muestra que adicional a lo anterior mencionado, también toma en cuenta variables del área general, tales como la modalidad de titulación, si les gustaria seguir estudiando, en que tipo de organización les gustaria laborar, su razón de elección de esa área de interés, su método de aprencdizaje favorito, entre otras. Esto significa que con RF, si toma en cuenta el rendimiento académico de los estudiantes, su formación previa, pero tambien toma en cuenta lo que les gustaria a los estudiantes, si estan interesados en un posgrado, si les gustaria especializarse en un lenguaje de programación especifico, entre otros.

Con estos resultados se puede proponer a la institución educativa que:

- Analice el plan de estudios flexible, que permita integrar asignaturas optativas relacionadas con las áreas de interés predominantes, así como módulos que fortalezcan competencias clave (lenguajes de programación, metodologías de investigación, habilidades blandas).
- Diseñe planes de acompañamiento académico y vocacional personalizados, basados en las variables más relevantes detectadas (rendimiento en asignaturas clave, experiencia previa y motivaciones), con el fin de incrementar la satisfacción y permanencia estudiantil.
- Integre metodologías activas y recursos adaptados al estilo de aprendizaje predominante identificado en los estudiantes, como el aprendizaje práctico, el trabajo colaborativo o el autoaprendizaje guiado.
- Genere programas de orientación para la elección de modalidades de titulación y especialización, considerando las preferencias detectadas, así como la posibilidad de vincular a los estudiantes con organizaciones afines a sus intereses laborales.
- Promueva la continuidad académica mediante la difusión de opciones de posgrado y cursos de especialización en áreas de alta demanda, alineadas con las aspiraciones identificadas.
- Aproveche el modelo predictivo como herramienta de apoyo a la gestión educativa, permitiendo monitorear y actualizar periódicamente las variables clave para tomar decisiones basadas en datos y no únicamente en percepciones.

Respecto al objetivo general de esta investigación, de realizar un análisis preliminar de los intereses académicos de los estudiantes de licenciatura, aplicando minería de datos, éste se ha alcanzado. De igual manera los objetivos específicos se cumplieron adecuadamente.

Contar con un conjunto de datos propio ofrece múltiples beneficios para una investigación, ya que garantiza que la información es real y concisa y se puede utilizar para realizar otras investigaciones.

Como se mencionó en los experimentos, una de las variables que se eliminaron para realizarlos fueron las Razones por las cuales elegían el área de interés correspondientes, sin embargo, a continuación, se muestran una nube de palabras con las respuestas de los estudiantes por área de interés. Cada nube de palabras está construida a partir de los datos recolectados en las en-

cuestas aplicadas. Esta representación visual permite identificar de manera rápida porqué les interesa esa área, siendo las palabras más grandes aquellas con mayor frecuencia de aparición.

Figura 6.1. Nube de palabras generada a partir de los intereses académicos hacia el área de Redes, reportados por los estudiantes en la encuesta aplicada.

Figura 6.2. Nube de palabras generada a partir de los intereses académicos en el área de Programación, reportados por los estudiantes en la encuesta aplicada.

Figura 6.3. Nube de palabras generada a partir de los intereses académicos en el área de Ingeniería de software, reportados por los estudiantes en la encuesta aplicada.

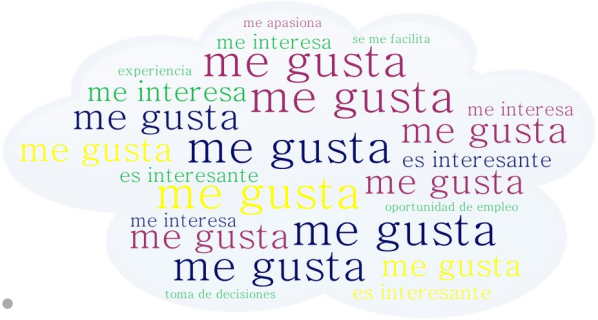

Figura 6.4. Nube de palabras generada a partir de los intereses académicos en el área de Tratamiento de la información, reportados por los estudiantes en la encuesta aplicada.

Figura 6.5. Nube de palabras generada a partir de los intereses académicos en el área de Entorno social, reportados por los estudiantes en la encuesta aplicada.

Figura 6.6. Nube de palabras generada a partir de los intereses académicos en el área de Interacción hombre-máquina, reportados por los estudiantes en la encuesta aplicada.

6.2. Trabajos futuros

Como trabajo futuro, se propone la incorporación de metaheurísticas (como Algoritmos genéticos, Recocido simulado o Enjambre de partículas) para optimizar el proceso de selección de atributos. A diferencia de los filtros utilizados en este estudio (*Chi-squared*, *Information gain*,

CFS y *Random forest importance*), las metaheurísticas permiten explorar de forma más flexible combinaciones de variables que podrían mejorar aún más el rendimiento de los clasificadores.

De igual manera, se recomienda implementar técnicas de ensamble de clasificadores, tales como *Bagging*, *Boosting* o *Stacking*, con el fin de combinar los modelos aplicados en esta investigación (SVM con diferentes kernels, *Random forest*, J48, Jrip y OneR). El uso de ensambles permitiría aprovechar las fortalezas individuales de cada modelo para obtener un sistema de clasificación más preciso y robusto, contribuyendo así a una mejor identificación de los patrones ocultos en los intereses académicos de los estudiantes.

Además se sugiere que se implemente la validación de los instrumentos, dado que en este trabajo ya se aplicó la encuesta a 152 estudiantes, se puede calcular el alfa de Cronbach y se puede realizar un análisis factorial para respaldar la validez. Esto no requiere cambiar las respuestas de la encuesta, solo analizarlas estadísticamente.

Dado que la distribución de las clases de los datasets LSC y LTI usados en este trabajo presentan un desbalanceo, se propone realizar experimentos con técnicas de balanceo de clases para incrementar estadísticamente el rendimiento de los modelos predictivos.

De igual forma con los *datasets* obtenidos se pueden realizar otras investigaciones, tales como: modelos predictivos de deserción escolar, estudio sobre el impacto de la motivación en el rendimiento académico, segmentación de estudiantes según perfiles de aprendizaje y preferencias, análisis del impacto de la formación previa (bachillerato, experiencia laboral) en la adaptación universitaria, entre otros.

Alojamiento de la Tesis en el Repositorio Institucional	
Título de la tesis:	Minado preliminar de datos de los intereses académicos de estudiantes de licenciatura
Autor:	Fátima Guadalupe Montejo Collado
ORCID:	https://orcid.org/0009-0007-2694-0307
Resumen:	Esta investigación se enfoca en identificar los intereses académicos de los estudiantes universitarios de la División Académica de Ciencias y Tecnologías de la Información (DACYTI) de la Universidad Juárez Autónoma de Tabasco (UJAT). Esta División cuenta con cuatro carreras activas, por lo que se diseñó un cuestionario como instrumento de recolección de datos por carrera, basados en el mapa curricular. Los cuestionarios fueron elaborados con la finalidad de recabar información precisa y contextualizada sobre las preferencias, motivaciones y áreas de interés académico de los estudiantes. Los instrumentos fueron enviados y aprobados por el Comité de Ética de la Universidad. La muestra estuvo conformada por 152 estudiantes, quienes debían tener al menos el 40 % de avance curricular. Para el análisis de la información se aplicaron técnicas de minería de datos, y para tareas de clasificación se utilizaron algoritmos como: *Random Forest*, J48, JRip, OneR, *K-Nearest Neighbors* (KNN), Máquinas de Vectores de Soporte (SVM) con diferentes kernels. También se aplicaron filtros de selección de variables relevantes: CFS (*Correlation-based Feature Selection*), *Chi Squared*, *Information Gain* y *Random Forest*. Para evaluar cada modelo se tomaron en cuenta las métricas: *Accuracy*, Kappa, Sensibilidad, Especificidad y *Balanced Accuracy*. Con base en los experimentos realizados, el clasificador con mejores resultados en todos los experimentos fue *Random Forest* y el filtro que menos variables relevantes generó fue CFS. Algunas de las variables más frecuente en los experimentos que mostró mayor relevancia fue la calificación del área de interés y la experiencia previa que los estudiantes tenían antes de ingresar a la carrera. Esto sugiere que tanto el rendimiento académico en asignaturas como la experiencia previa a temas del área influyen significativamente en los intereses académicos de los alumnos. Para futuras investigaciones, se propone emplear metaheurísticas y ensamble de clasificadores, así como también realizar un análisis profundo por carrera balanceando los datos de las clases.
Palabras clave:	Inteligencia Artificial, Minería de datos, Intereses académicos
Referencias citadas:	En la siguiente página se muestran las referencias.

Bibliografía

M. P. Aliende. La gestión de datos de investigación. 2017. https://repositorio.uam.es/bitstream/handle/10486/678601/gestion_perez_us_2017_4.pdf?sequence=.

A. O. Alsayed, M. S. M. Rahim, I. AlBidewi, M. Hussain, S. H. Jabeen, N. Alromema, S. Hussain, & M. L. Jibril. Selection of the right undergraduate major by students using supervised learning techniques. *applied sciences*, 11(22):10639, 2021. https://doi.org/10.3390/app112210639.

A. Ballesteros Román, D. Sánchez-Guzmán, & R. García Salcedo. Minería de datos educativa: Una herramienta para la investigación de patroness de aprendizaje sobre un contexto educativo. *Latin-American Journal of Physics Education*, 7(4), 2013. http://www.lajpe.org/dec13/22-LAJPE_814_bis_Alejandro_Ballesteros.pdf.

F. Berzal. Clustering jerarquico, 2025. https://www.passeidireto.com/es/content/132694081/41-clustering-jerarquico-presentacion-autor-fernando-berzal?

Y. Cardoso García & L. Arza Valdés. Algoritmo oner. su aplicación en ensayos clínicos. *Revista Cubana de Ciencias Informáticas*, 11(2):61–72, 2017. ISSN 2227-1899. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992017000200005&nrm=iso.

E. Cruz, M. González, & J. C. Rangel. Técnicas de machine learning aplicadas a la evaluación del rendimiento ya la predicción de la deserción de estudiantes universitarios, una revisión. *Prisma Tecnológico*, 13(1):77–87, 2022. https://doi.org/10.33412/pri.v13.1.3039.

D. S. G. César Pérez López. *Minería de datos, técnicas y herramientas.* 2007. https://books.google.com.mx/books?id=wz-D_8uPFCEC&pg=PA699&dq=definicion+de+redes+neuronales&hl=es-419&sa=X&ved=2ahUKEwifkI7Q8rr8AhUCIOQIHYm9AfwQuwV6BAgIEAg#v=onepage&q=definicion%20de%20redes%20neuronales&f=false.

M. A. Hall. Correlation-based feature selection for machine learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pages 359–366, 1999. https://www.researchgate.net/publication/2805648_Correlation-Based_Feature_Selection_for_Machine_Learning.

K. R. L. Laura & C. Baluarte. Evaluación de técnicas de minería de datos para la predicción del

rendimiento académico. 2017. https://dialnet.unirioja.es/servlet/articulo?codigo=7352939.

Y. Liu, Y. Wang, & J. Zhang. New machine learning algorithm: Random forest. In *International conference on information computing and applications*, pages 246–252. Springer, 2012. https://link.springer.com/chapter/10.1007/978-3-642-34062-8_32.

J. R. LLevot. Medicina predictiva, aprendizaje automático y anestesia. 2020. https://pesquisa.bvsalud.org/portal/resource/pt/ibc-200721.

G. Martínez. Minería de datos. *Cómo hallar una aguja en un pajar. Ingenierías*, 14(53):53–66, 2001. https://www.cs.buap.mx/~bbeltran/NotasMD.pdf.

J. M. Moine, A. S. Haedo, & S. E. Gordillo. Estudio comparativo de metodologías para minería de datos. In *XIII Workshop de Investigadores en Ciencias de la Computación*, 2011. https://sedici.unlp.edu.ar/handle/10915/20034.

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993. ISBN 978-1-55860-238-0. https://books.google.com.mx/books?hl=es&lr=&id=b3ujBQAAQBAJ&oi=fnd&pg=PP1&dq=%40book%7Bquinlan1993c4.5,+%09title+++++%3D+%7BC4.5:+Programs+for+Machine+Learning%7D,+%09author++++%3D+%7BQuinlan,+J.+Ross%7D,+%09year++++++%3D+%7B1993%7D,+%09publisher+%3D+%7BMorgan+Kaufmann%7D,+%09address+++%3D+%7BSan+Mateo,+CA%7D,+%09isbn++++++%3D+%7B978-1-55860-238-0%7D,&ots=sS6rYSDoz9&sig=aTjM1Gx6DsNR5NYvMODQOZyh51k&redir_esc=y#v=onepage&q&f=false.

A. Rajput, R. P. Aharwal, M. Dubey, S. Saxena, & M. Raghuvanshi. J48 and jrip rules for e-governance data. *International Journal of Computer Science and Security (IJCSS)*, 5(2):201, 2011. https://scholar.google.com/scholar?hl=es&as_sdt=0%2C5&q=J48+and+JRIP+rules+for+e-governance+data&btnG=.

N. E. Rodrıguez-Maya, C. Lara-Álvarez, O. May-Tzuc, & B. A. Suárez-Carranza. Modeling students' dropout in mexican universities. *Research in Computing Science*, 139:163–175, 2017. https://rcs.cic.ipn.mx/2017_139/Modeling%20Students_%20Dropout%20in%20Mexican%20Universities.pdf.

E. M. Rojas. Machine learning: análisis de lenguajes de programación y herramientas para desarrollo. *Iberian Journal of Information Systems and Technologies*, 2020. https://www.proquest.com/docview/2388304894?pq-origsite=gscholar&fromopenview=true.

S. Romero et al. Uso de técnicas de machine learning para predecir el rendimiento académico de los estudiantes de la carrera de ingeniería civil en informática de la universidad del bío-bío, chillán. 2015. http://repobib.ubiobio.cl/jspui/bitstream/123456789/2610/1/Soto%20Romero%2C%20Gaspar.pdf.

D. scientest. Cross-validation : definición e importancia en machine learning, 2025. https://datascientest.com/es/cross-validation-definicion-e-importancia#:~:text=El%20proceso%20de%20validaci%C3%B3n%20consiste,que%20probarlo%20con%20nuevos%20datos.

P.-N. Tan, M. Steinbach, & V. Kumar. Data mining introduction. *People's Posts and Telecommunications Publishing House, Beijing*, 2006. https://books.google.com.mx/books/about/Introduction_to_Data_Mining.html?id=KZQ0jgEACAAJ&redir_esc=y.

Translational Interventional Radiology. Chi-square test, 2023. https://www.sciencedirect.com/topics/medicine-and-dentistry/chi-square-test.

T. Z. Win & N. S. M. Kham. *Information gain measured feature selection to reduce high dimensional data*. PhD thesis, MERAL Portal, 2019. https://meral.edu.mm/record/3413/file_preview/ICCA%202019%20Proceedings%20Book-pages-79-84.pdf.

M. R. Zumárraga, M. I. Castro, J. C. Romero, P. Escobar, M. J. Boada, R. Armas, J. Luzuriaga, & Y. Gonzáles. Medición de intereses profesionales en estudiantes universitarios y un abordaje exploratorio de su relación con el desempeño académico. In *Congresos CLABES*, 2017. https://pure.ups.edu.ec/en/publications/measurement-of-professional-interests-in-university-students-and-.