UNIVERSIDAD NACIONAL DE CAAGUAZÚ FACULTAD DE CIENCIAS
Y TECNOLOGÍAS CARRERA DE INGENIERÍA EN INFORMÁTICA


# Modelo de alerta temprana para la deserción estudiantil en la Facultad de Ciencias y Tecnologías basado en la estimación de factores académicos.

**Nathalia Romina González Duarte**
Tutor: Ing. Juan Vicente Bogado Machuca

CORONEL OVIEDO, ABRIL DE 2024


Usted es libre de:

- **Compartir — copiar y redistribuir el material en cualquier medio o formato**
- **Adaptar — remezclar, transformar y construir a partir del material**

Bajo los siguientes términos:

- Atribución — Usted debe dar crédito de manera adecuada, brindar un enlace a la licencia, e indicar si se han realizado cambios. Puede hacerlo en cualquier forma razonable, pero no de forma tal que sugiera que usted o su uso tienen el apoyo de la licenciante.
- No Comercial — Usted no puede hacer uso del material con propósitos comerciales.

# DERECHO DE AUTOR

Quien suscriben, Nathalia Gonzalez autora del trabajo de investigación titulado «**Modelo de alerta temprana para la deserción estudiantil en la Facultad de Ciencias y Tecnologías basado en la estimación de factores académicos**», declara/n que voluntariamente cede/n a título gratuito en forma pura y simple ilimitada e irrevocablemente a favor de la Facultad de Ciencias y Tecnologías – UNCA, el derecho de autor de contenido patrimonial, que le corresponde sobre el trabajo de referencia. Conforme a lo anteriormente expresado, esta sesión le otorga a la FCyT la Facultad de comunicar la obra divulgarla, publicarla y reproducirla en soportes analógicos o digitales en la oportunidad que así lo estime conveniente. La FCyT deberá indicar qué autoría o creación del trabajo corresponde a mi persona y hará referencia al autor y a las personas que hayan colaborado en la realización del presente trabajo de investigación. En la ciudad de Coronel Oviedo a los 26, del mes abril del 2024.

.....
Firma

# Página de aprobación

Trabajo de fin de grado para la obtención del Título de Ingeniero en Electricidad aprobado en representación de la Facultad Ciencias y Tecnologías de la Universidad Nacional de Caaguazú, por el Tribunal Examinador constituido por los siguientes profesores.

_____

Prof. Ing.

_____

Prof. Ing.

_____

Prof. Ing.

Acta Nro.: _____

Fecha: _____

Calificación: _____

# Dedicatoria

Este proyecto se lo dedico a mis padres Marina y Ever, quienes me han brindado su amor incondicional y un apoyo constante en cada paso que he dado.

A mi hermanita Ivana por la paciencia de compartir conmigo noches de desvelos.

A Buttowski y Peluchin, por quererme y acompañarme sin limitaciones en los momentos más difíciles.

Y a Dios, por fortalecer mi espíritu y brindarme la salud necesaria para alcanzar mis metas y objetivos.

# Agradecimientos

Mi más sincero agradecimiento al Ingeniero Juan Vicente Bogado, mi tutor, por su dedicación, paciencia y orientación invaluable a lo largo de todo el desarrollo de este proyecto.

A la Facultad de Ciencias y Tecnologías, por proporcionarme los recursos y el ambiente propicio para llevar a cabo este proyecto.

A mis compañeros y pofesores cuya colaboración y apoyo han sido fundamentales en este camino de aprendizaje y crecimiento personal.

# Resumen

Este estudio aborda el desafío de la deserción estudiantil en la Facultad de Ciencias y Tecnologías mediante la construcción de un modelo de alerta temprana basado en factores académicos. Utilizando una metodología de ciencia de datos, se caracterizaron y analizaron registros académicos, aplicando técnicas como el análisis de clusters y el método del codo para optimizar la segmentación de los estudiantes. Se desarrollaron varios modelos predictivos de machine learning, incluyendo regresión logística, árboles de decisión, y k-nearest neighbors, los cuales fueron evaluados mediante métricas de precisión, recall, y F1 Score para determinar su eficacia en la clasificacíon de estados academicos.

Los resultados indicaron una variabilidad en la efectividad de los modelos en función de la carrera y los datos disponibles, para Informática, el mejor modelo resultó ser K-Nearest Neighbors (KNN). Para Electricidad y Civil, el modelo de Árbol de Decisión (DT) fue el más eficaz. Y para Electrónica, la Regresión Logística (RL) y K-Nearest Neighbors (KNN) demostraron un mejor rendimiento, resaltando la importancia de adaptar las intervenciones a las características específicas de los estudiantes y sus entornos académicos..

Las conclusiones subrayan el desempeño de distintos modelo en la identificación temprana de estudiantes en riesgo, proponiendo además la integración de factores socioeconómicos y psicológicos para futuras investigaciones en el área.

**Palabras claves:** Ciencia de datos, Deserción estudiantil, Modelos predictivos, Machine learning, Análisis de clusters.

# Abstract

This study addresses the challenge of student dropout in the Faculty of Sciences and Technologies by constructing an early warning model based on academic factors. Employing a data science methodology, academic records were characterized and analyzed, using techniques such as cluster analysis and the elbow method to optimize student segmentation. Several predictive machine learning models were developed, including logistic regression, decision trees, and k-nearest neighbors, which were evaluated using precision, recall, and F1 Score metrics to determine their effectiveness in classifying academic statuses.

The results indicated variability in model effectiveness depending on the major and available data, for Computer Science, the best model turned out to be K-Nearest Neighbors(KNN), for Electrical and Civil, the Decision Tree (DT) model was the most effective and for Electronics, Logistic Regression (RL) and K-Nearest Neighbors (KNN) demonstrated better performance highlighting the importance of tailoring interventions to students' specific characteristics and academic environments.

The conclusions highlight the performance of different models in early identification of at-risk students, further proposing the integration of socioeconomic and psychological factors for future research in the field.

**Key words:** Data science, Student dropout, Predictive models, Machine learning, Cluster analysis.

# Índice

# Índice de tablas

# Índice de figuras

# 1. Modelo de alerta temprana para la deserción Estudiantil en la Facultad de Ciencias y Tecnologías basado en la estimacón de factores académicos

La deserción universitaria representa un desafío significativo para el sistema educativo, ya que repercute negativamente en diversos aspectos del desarrollo nacional[11]. El presente proyecto, se desarrolló para abordar y comprender mejor los patrones que subyacen a este fenómeno, mediante tecnologias de Machine Learning. La problemática de alta deserción universitaria en Paraguay es un tema de gran preocupación que refleja la dificultad que enfrentan los estudiantes para completar sus estudios en el país. Actualmente, solo alrededor del 10 % de los jóvenes paraguayos logran terminar sus carreras universitarias, lo que deja una abrumadora mayoría del 90 % que en algún momento toma la difícil decisión de abandonar sus estudios [6]. Un estudio realizado en la Universidad Católica Nuestra señora de la Asunción por Norma Coppari, Laura Bagnoli y Paula Maidana sobre la deserción universitaria en donde participaron 119 estudiantes paraguayos, 82 mujeres y 37 varones, solteros, empleados en mayoría, entre 17 y 68 años de edad, de carreras diversas, 31 inscriptos en universidades públicas y 88 en privadas. Solo 45 de la muestra (38 %) cursa regularmente, 48(40 %) lo hace irregularmente con riesgo de abandono, y 26 (22 %) ha dejado la carrera [18]. En la Facultad de Ciencias y Tecnologías, la deserción estudiantil es un problema comúnmente observado y representa una debilidad para la institución educativa, la misma menciona en el plan de desarrollo de una de las carreras, que se requiere consolidar el mecanismo de análisis de retención, deserción, transferencia y promoción para la evaluación de la eficiencia interna de la carrera [5]. La Predicción de la deserción estudiantil utilizando Machine Learning, surgió como una alternativa para afrontar la problemática, se buscó analizar datos históricos con el fin identificar indicadores tempranos que señalasen que un alumno este en riesgo de abandonar sus estudios. Esta identificación temprana podrá ayudar a intervenir de manera oportuna y brindar apoyo personalizado, con el fin de prevenir la deserción. La justificación del proyecto radicó en la creencia de que el modelo predictivo desarrollado con Machine Learning proporcionaria información detallada sobre cada estudiante, lo que facilitaría la creación de estrategias de intervención adaptadas a las necesidades individuales, como tutorías académicas, asesoramiento y programas de mentoría, para mejorar el rendimiento académico y el bienestar de los estudiantes, así como realizar ajustes en caso de ser necesario al mecanismo de orientación académica de la FCYT. El modelo predictivo, desarrollado a lo largo de este proyecto beneficiará a la institución educativa y a los estudiantes, ya que con el desarollo del proyecto se ofrece una mejor visión de la trayectoria de un estudiante. Un enfoque innovador y similar es explorado en el estudio titulado ”Minería de datos: predicción de

la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos", realizado en la Universidad Tecnológica de Izúcar de Matamoros de México, aplican técnicas de minería de datos para predecir la deserción escolar. Utilizando el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos, el estudio demuestra cómo la integración de datos socioeconómicos y académicos puede ser utilizada para identificar tempranamente a estudiantes en riesgo, facilitando intervenciones proactivas. Los autores, Sergio Valero Orea, Alejandro Salvador Vargas, y Marcela García Alonso, detallan el proceso desde la recopilación de datos hasta la creación de modelos predictivos, demostrando cómo estas herramientas pueden identificar a estudiantes en riesgo de abandono escolar.[10]

# 2. Marco Teórico

Para llevar a cabo este proyecto fue escencial establecer los aspectos clave a explorar y cada enfoque necesario que respaldó dicho proyecto. Este proyecto se apoyo en un conjunto de Técnica de ciencia de datos aplicada a la base de datos académicas de estudiantes para caracterizarlo. La ciencia de datos es una disciplina que emplea herramientas, métodos y tecnología con el fin de extraer información significativa a partir de datos[1] En este contexto, se aplicaron diversas técnicas, la limpieza y preparación de datos, esto implica la identificación y corrección de errores en los datos, la gestión de valores faltantes y la estandarización de los formatos[1], también se realizó un analisis exploratorio de datos para comprender la estructura del mismo.

## 2.1. Caracterización de datos

### 2.1.1. Método del codo

Además para caracterizar los datos se determino el número optimo de grupos (o çlusters") en el algoritmo de K-Means, con el **método del codo**, que identifica el punto en el que se produce un cambio. Para aplicarlo, se ejecuta el algoritmo K-means con diferentes cantidades de grupos (k) y se calcula la suma de las distancias al cuadrado de cada punto con respecto a su centroide. Luego, se grafican estos resultados con el número de grupos en el eje x y la suma de distancias al cuadrado en el eje y. Se busca en la gráfica el punto donde la disminución de la suma de distancias al cuadrado se vuelve abrupta, formando una especie de codo como se ve en la figura 1. Este punto corresponde al número óptimo de grupos, conocido como el **codo** notable en la reducción de la variabilidad dentro de

cada grupo[4].

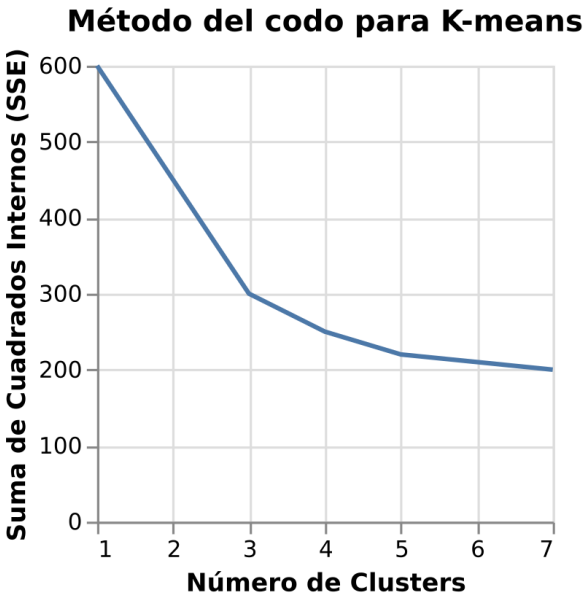


Figura 1: Método del Codo.

### 2.1.2. K-Means

K-Means es un algoritmo de aprendizaje no supervisado simple que se utiliza para agrupar un conjunto de datos en un número predefinido de grupos, conocido como $k$. El proceso comienza estableciendo $k$ centroides inicialmente, ubicándolos estratégicamente lo más lejos posible entre sí para obtener mejores resultados. Luego, cada punto del conjunto de datos se asigna al centroide más cercano, formando grupos preliminares. Después de esta asignación inicial, los centroides se recalculan como el centro de sus respectivos grupos y el proceso se repite, reasignando puntos y recalculando centroides hasta que los centroides ya no cambian su posición, así como se explica en la figura2. El objetivo del algoritmo es minimizar una función de error cuadrático, lo que efectivamente trata de hacer los grupos tan compactos y separados como sea posible[16]. Una vez identificado el número de agrupaciones se caracterizaron los datos.

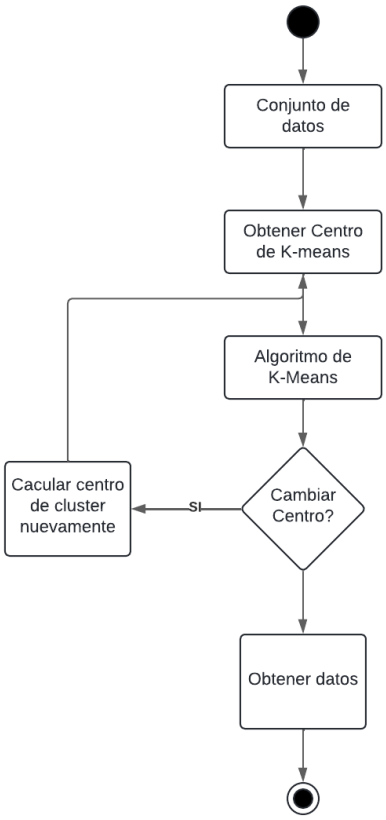


Figura 2: Diagrama de funcionamiento de K-Means

## 2.2. Modelos de Clasificación

Luego se ajustaron modelos predictivos de Machine Learnig, Machine Learning es el estudio científico de algoritmos y modelos estadísticos que los sistemas informáticos utilizan para realizar una tarea específica sin ser programados explícitamente[7]. Para la realización de este proyecto se utilizaron cinco diferentes modelos:

### 2.2.1. Regresión Lógistica(RL)

Se usa para predecir la probabilidad de que un evento ocurra o no. Funciona encontrando una ecuación que mejor estime la probabilidad de pertenencia a una categoría basada en los valores de entrada dados. En términos simples, intenta dibujar una línea que separe de la mejor manera posible las dos categorías utilizando los datos de entrada. Usa una técnica matemática para ajustar la mejor línea que divide las dos categorías y calcula la probabilidad de que una entrada pertenezca a una categoría basada en esta línea[8].

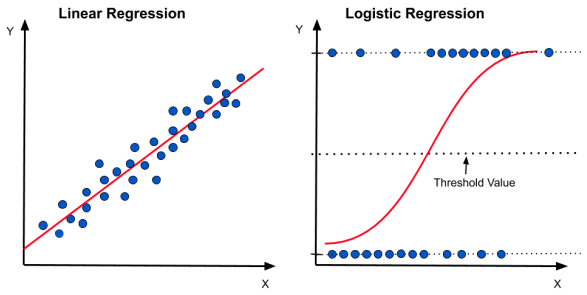


Figura 3: Regresión Lógistica[15]

### 2.2.2. Árbol de Desiciones(DT)

Un árbol de decisión es un método de predicción que se basa en aprender de ejemplos y razonamiento lógico. Se asemeja a los sistemas de predicción basados en reglas, donde se establecen condiciones para resolver un problema. Es uno de los modelos de clasificación más comunes y populares. Durante el proceso de aprendizaje, se construye un árbol que representa el conocimiento adquirido. Gráficamente, se visualiza como un conjunto de nodos, ramas y hojas. Asi como muestra en la figura 4 el nodo principal (raíz) inicia el proceso de clasificación y los nodos internos plantean preguntas sobre atributos específicos. Cada respuesta se representa como un nodo hijo y las ramas muestran los posibles valores de los atributos. Los nodos hoja representan decisiones finales que coinciden con las clases del problema a resolver[13].

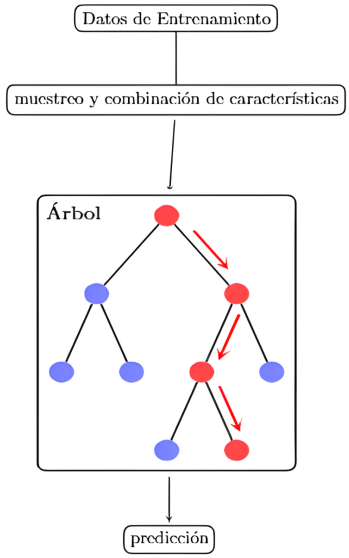


Figura 4: Árbol de Decisiones

### 2.2.3. Random Forest(RF)

Como se ve en la figura 5 Es un tipo de predictor formado por una colección de árboles de regresión base aleatorizados. Estos árboles son generados utilizando una variable aleatoria, independiente de los datos de entrada y del conjunto de entrenamiento, evitando así cualquier técnica de remuestreo.Los árboles se combinan usando un promedio calculado mediante el método de Monte Carlo, basado en la ley de los grandes números, para producir una estimación de regresión agregada. Cada árbol divide el espacio de datos en celdas rectangulares hasta alcanzar un número específico de divisiones, determinado por el usuario. Este enfoque permite que el bosque aleatorio maneje cortes de datos de manera efectiva sin depender directamente de los datos de entrenamiento o de estrategias de división optimizadas basadas en los mismos[8].

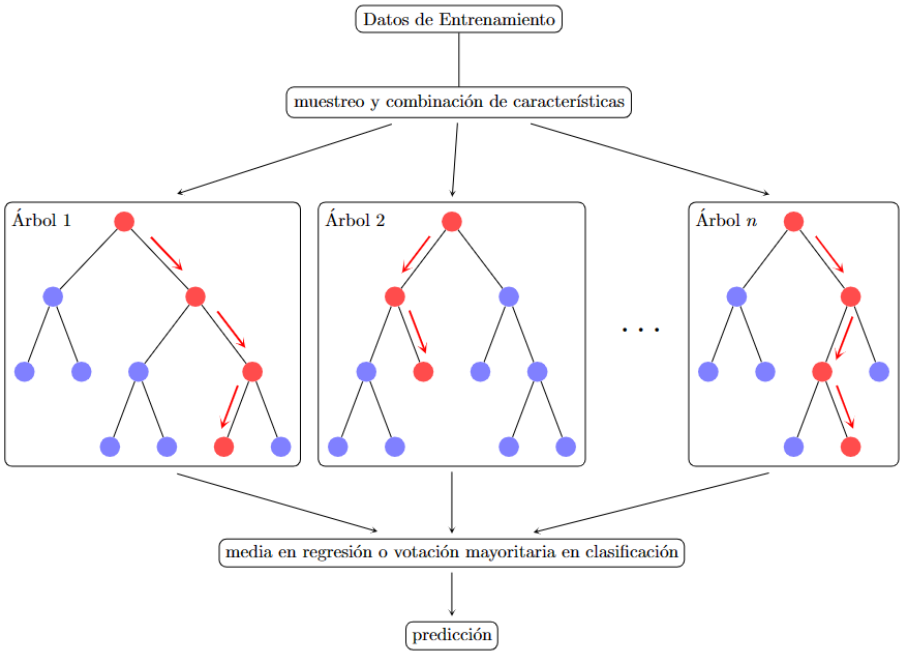


Figura 5: Random Forest[12]

### 2.2.4. Máquina de Soporte Vectorial o Support Vector Machine(SVM)

Es un tipo de modelo de aprendizaje automático que no depende de la cantidad de parámetros predefinidos y que se enseña a sí mismo a partir de ejemplos supervisados (es decir, datos que ya están etiquetados). Utiliza una técnica conocida como ”kernel”, que básicamente permite manejar datos en formas más complejas transformando los datos de entrada a un espacio con muchas dimensiones. Este modelo crea un límite, llamado hiperplano, que funciona bien incluso para datos que no se separan fácilmente en su forma original.

Luego, para hacer predicciones, transforma estos datos de vuelta a su espacio original de una manera que mantiene la efectividad de las decisiones tomadas en el espacio de alta dimensión[8]. En la figura 6 muestra como funciona lo mencionado.

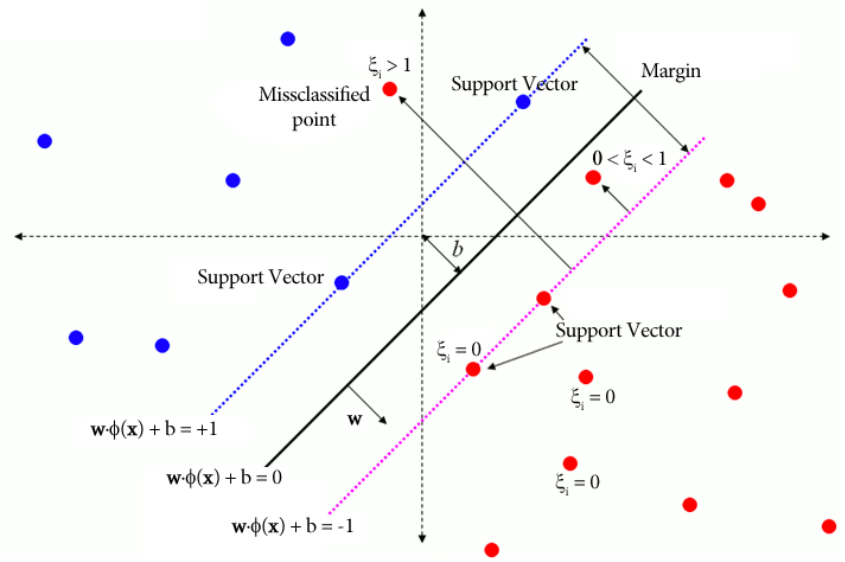


Figura 6: Máquina de Soporte Vectorial[2]

### 2.2.5. K Vecinos más Cercanos o en inglés K-Nearest Neighbors(KNN)

Predice la clase de un elemento basándose en las clases de los 'k' elementos más cercanos a él como se muestra en la figura 7 . Es como preguntar a tus 'k' vecinos más cercanos para decidir sobre un tema y luego ir con la mayoría de sus opiniones. No asume ningún tipo de distribución estadística para los datos, lo que lo hace útil en situaciones donde la relación entre los datos no es conocida. La clase de un nuevo punto se determina por la clase más frecuente entre sus k vecinos más cercanos[8].

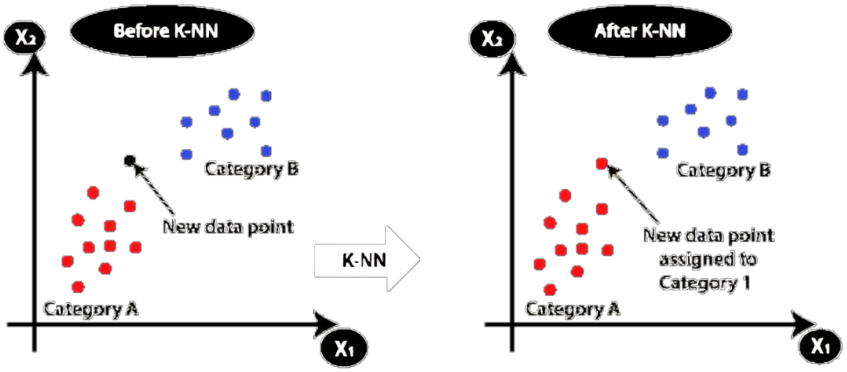


Figura 7: K Vecinos más Cercanos[3]

## 2.3. Métricas de evaluación

Para cada modelo descrito se utilizó cuatro métricas que son:

### 2.3.1. Matriz de Confusión

Por último para analizar el desempeño de los modelos se utilizó la matriz de confusión que es una representación de la calidad de un clasificador, para una clasificación ideal, los valores más altos de la matriz deben estar posicionados en la diagonal de ésta[17]. La matriz de confusión se construye a partir de una imagen de satélite con N celdillas clasificadas en M clases. Sobre las columnas se ordenan las clases reales (verdad-terreno), y sobre las filas las unidades cartográ ficas (unidades -o clases del mapa). Los elementos que aparecen en la diagonal como muestra en la figura 8 nos indican el número de clasificaciones realizadas correctamente, y aquellos que aparecen fuera suponen migraciones o fugas[9].

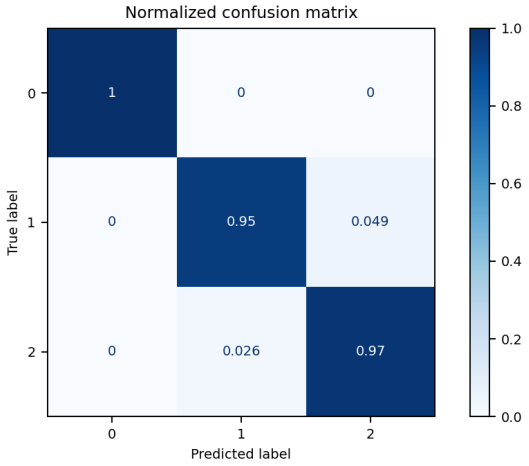


Figura 8: Matriz de Confusión[14]

### 2.3.2. Exactitud o Accuracy

Esta métrica se define como la cantidad de veces que acierta una afirmación, sobre el total de datos de entrada. Esta metrica puede arrojar valores que al parecer son alto cuando en verdad la parte relevante no lo es tanto, y es causado por un desbalance en la cantidad de muestras verdaderas y positivas[17].

$$\text{Exactitud} = \frac{\text{Verdaderos Positivos} + \text{Verdaderos Negativos}}{\text{Total de muestras}}$$

### 2.3.3. Presición

Esta métrica se define como la cantidad de casos verdaderos positivos sobre la cantidad total de todo lo que que era positivo.En otras palabras, de todo lo que el algoritmo predijo como positivo, se evalúa cuánto de eso era cierto[17]. Como se muesta en la siguiente formula:

$$\text{Precisión} = \frac{\text{Verdadero Positivo}}{\text{Verdadero Positivo} + \text{Falso Positivo}}$$

### 2.3.4. Recall

Es una métrica que mide la capacidad de un modelo para identificar correctamente todos los casos relevantes dentro de un conjunto de datos. Se calcula como la proporción de verdaderos positivos entre la suma de los verdaderos positivos y los falsos negativos. Esto implica que el recall evalúa cuántos de los elementos realmente positivos fueron correctamente clasificados como tales por el modelo[17].

$$\text{Recall} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}}$$

### 2.3.5. F1 Score

Sería el doble del producto de precision y recall sobre la suma de estos dos. Una manera visual de entender este concepto es imaginar un espacio tridimensional, donde los ejes X y Y representan la precisión y la recall, respectivamente, mientras que el eje Z representa el f1-score. Cuando los valores de precisión y exhaustividad están cerca de uno entre sí, el valor de f1-score tiende a ser más alto[17].

$$\text{F1-score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

# 3. Objetivos

## 3.1. Objetivo General

Construir un modelo de alerta temprana para la deserción estudiantil en la Facultad de Ciencias y Tecnologías basado en la estimación de factores académicos.

## 3.2. Objetivo específico

- Caracterizar la base de datos académica de estudiantes de la Facultad de Ciencias y Tecnologías aplicando ciencia de datos.
- Ajustar modelos predictivos de machine learning para predecir estado académico de alumnos de la Facultad de Ciencias y Tecnologías.
- Utilizar métricas para obetener el modelo predictivo con mejor rendimiento.

# 4. Metodología

## 4.1. Recolección y preprocesamiento de datos

Los datos utilizados en este estudio fueron obtenidos mediante una solicitud formal al Decano de la Facultad de Ciencias y Tecnologías de la Universidad Nacional de Caaguazú (UNCA). Tras la aprobación de la solicitud, se proporcionó un dataset de todos los registros de alumnos desde el año 2010 hasta el 2023. Este dataset incluye 75.937 registros de 1.422 alumnos distribuidos en las siguientes carreras:

| **Carrera** | **Cantidad de alumnos** |
|---|---|
| Ingeniería en Informática | 248 alumnos |
| Ingeniería Civil | 483 alumnos |
| Ingeniería en Electricidad | 502 alumnos |
| Ingeniería en Electrónica | 189 alumnos |

Tabla 1: Distribución de alumnos por cada carrera.

De estos, 85 alumnos cambiaron de carrera internamente durante su trayectoria académica. El dataset contiene los siguientes campos:

| **Campo** | **Descripción** | **Tipo de dato** |
|---|---|---|
| AñoIngreso | Fecha y hora del ingreso del alumno | Datetime |
| ID-Ingreso | Identificador único del ingreso | Numérico |
| IDAlumno | Número de identificación del alumno | Numérico |
| Sexo | Género del alumno | Numérico |
| Carrera | Carrera del alumno | String |
| id-materia | Identificador de la materia | Numérico |
| Materia | Nombre de la materia | String |
| Calificación | Calificación obtenida | Numérico |
| Periodo_Evaluacion | Periodo de evaluación | String |
| AñoAcademico | Año académico correspondiente | Numérico |
| PeriodoAcadémico | Periodo académico | Numérico |
| FechaExamen | Fecha del examen final | Numérico |

Tabla 2: Nombre, descripción y tipo de dato de cada variable de la base de datos.

Se realizó una limpieza de datos para mejorar la calidad del dataset. Los pasos incluyeron la eliminación de registros duplicados y la corrección de incoherencias entre el año de ingreso y la fecha del primer examen. Durante el proceso se identificaron a los estudiantes cursantes, los mismos fueron excluidos como muestra el algoritmo 1 ya que su situación aún no es definitiva y no aportan a la predicción de los resultados finales.

**Algorithm 1** Proceso de Filtrado de Estudiantes Cursantes

```
procedure FiltrarCursantes
    Data: Base de datos de alumnos
    Result: Base de datos modificada sin alumnos cursantes
    Cargar la base de datos en la variable data
    fechaExamen ← Extraer la columna 'fecha de examen' de data
    anoDeExamen ← Aplicar función para extraer solo el año de fechaExamen
    filtro ← Filtrar anoDeExamen donde el año sea 2022 o 2023
    Crear columna 'esCursante' en data y asignar 1 a todas las filas filtradas
    Eliminar las filas de data que no están en el filtro
    return data
```

Se verificaron las correlatividades de las materias para aquellos alumnos que cambiaron de carrera, asegurando así la consistencia interna de los datos. Además, se separaron los registros por carrera debido a la especificidad de las materias en cada una.

Se obtuvieron métricas adicionales como el tiempo de estudio en años, el promedio de calificaciones, la cantidad de ausencias, aplazos y la cantidad de calificación cinco felicitado(5F). También, cada materia fue transformada en un campo individual dentro del dataset, con la calificación final del alumno en esa materia como valor de dicho campo. Teniendo así un total de 97 columnas en las carrera Civil, Electrica y Elecrónica y 93 columnas en Infórmatica. Así como excluyeron los cursantes, el dataset cuenta para cada carrera correpondientemente 151 filas para Informática, 87 filas para Electrónica, 268 filas para Electricidad y 159 filas para Civil.

## 4.2. Análisis de Clústers

Se realizó un análisis de clústers para determinar la existencia de grupos bien diferenciados en la base datos. Este análisis ayudó a identificar varios estados en los alumnos, basados en su actividad académica y estado actual en las carreras. Se utilizó el algoritmo de K-means, aplicando el método del codo para determinar el número de grupos que se formaron y luego se procedió a analizar cada clúster con técnicas estadísticas para caracterizar los mismos.

## 4.3. Predicción de estado académico con machine learning

Una vez obtenidos la cantidad de clústers existentes en la base de datos, su caracterización, se procedió a implementar técnicas de aprendizaje automático para predecir los patrones de deserción y éxito académico. Se experimentó con diferentes configuraciones de entre-

namiento y validación, lo que permitió mejorar la precisión de los modelos predictivos. La fase de entrenamiento inicial utilizó todos los datos de alumnos de interés, mientras que las fases subsiguientes ajustaron el modelo para predecir con mayor exactitud basándose en los primeros años de los estudiantes. Para ello se utilizó diferentes modelos que incluyen regresión logística, árboles de decisión, random forest, SVM y k-nearest neighbors. se compararon los resultados utilizando como métrica la exactitud, recall, la precisión y F1.

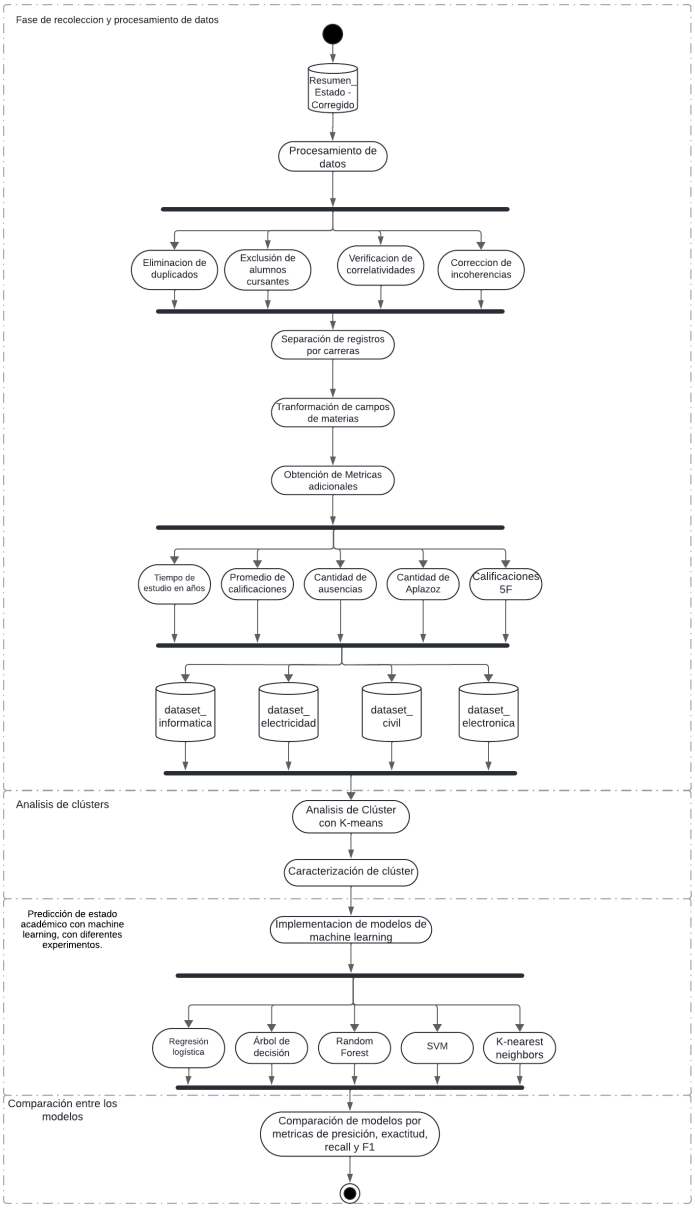


Figura 9: Flujo de trabajo resumido que indica la metodología utilizada. Recolección de datos, análisis de clusters, comparación de modelos.

# 5. Resultados y análisis

En esta sección se presentan los resultados del análisis de la base de datos utilizada en este trabajo, así como el análisis de clusters y los estados acádemicos obtenidos, finalmente se presentarán los experimentos realizados y un tabla comparativa de los mismos.

## 5.1. Análisis de datos

Con la obtención de los datos se procedio a hacer un análisis de los mismos, con un registro de 75.937 de 1.422 alumnos distribuidos en las carreras como se muestra en la tabla 1, este primer análisis se centró en las características académicas clave de los alumnos. A continuación, se presentan las varibles adicionales de cada carrera:

| **Variable** | **Informática** |
|---|---|
| Cantidad de Alumnos (Sexo 1) Masculino | 172 |
| Cantidad de Alumnos (Sexo 2) Femenino | 76 |
| Promedio del Tiempo de Estudio (años) | 3.46 |
| Promedio de Ausencias en Finales | 6.90 |
| Promedio de Aplazos | 4.46 |
| Promedio de las Calificaciones | 3.24 |

Tabla 3: Datos académicos para la Carrera de Informática

| **Variable** | **Civil** |
|---|---|
| Cantidad de Alumnos (Sexo 1) Masculino | 284 |
| Cantidad de Alumnos (Sexo 2) Femenino | 199 |
| Promedio del Tiempo de Estudio (años) | 3.70 |
| Promedio de Ausencias en Finales | 8.00 |
| Promedio de Aplazos | 5.13 |
| Promedio de las Calificaciones | 3.20 |

Tabla 4: Datos académicos para la Carrera de Civil

| **Variable** | **Electricidad** |
|---|---|
| Cantidad de Alumnos (Sexo 1) Masculino | 421 |
| Cantidad de Alumnos (Sexo 2) Femenino | 81 |
| Promedio del Tiempo de Estudio (años) | 3.73 |
| Promedio de Ausencias en Finales | 8.61 |
| Promedio de Aplazos | 5.74 |
| Promedio de las Calificaciones | 3.13 |

Tabla 5: Datos académicos para la Carrera de Electricidad

| **Variable** | **Electrónica** |
|---|---|
| Cantidad de Alumnos (Sexo 1) Maculino | 146 |
| Cantidad de Alumnos (Sexo 2) Femenino | 43 |
| Promedio del Tiempo de Estudio (años) | 3.62 |
| Promedio de Ausencias en Finales | 8.42 |
| Promedio de Aplazos | 6.46 |
| Promedio de las Calificaciones | 3.15 |

Tabla 6: Datos académicos para la Carrera de Electrónica

Tras el análisis de los datos presentados, se observaron distintas características académicas entre los alumnos de las diversas carreras analizadas. Este análisis preliminar ha revelado no solo variaciones en indicadores como el promedio de tiempo de estudio, ausencias en finales, aplazos y calificaciones, sino también diferencias marcadas en la distribución de género entre las carreras. Estas diferencias subrayan la necesidad de entender mejor la agrupación y las similitudes entre los estudiantes, lo que lleva al siguiente paso crucial de la investigación.

## 5.2. Analisis de clústers

Para ver si hay presencia de agrupaciones dentro del dataset y determinar caracteristicas academicas similares entre los alumnos, se empleó un analisis de clustering K-Means. Se utilizo el metodo del codo para identificar el número óptimo de clusters.

El método del codo, representado en las Figuras 30, 31, 32 y 33, respectivamente a las carreras Informática, Electricidad, Electrónica y Civil, muestra la relación entre el número de clusters (k) y la distorsión, que es la suma de las distancias al cuadrado de cada punto

a su centro de cluster más cercano. El punto de codo, donde la reducción de la distorsión comienza a disminuir a un ritmo más lento, se identificó en k=4, en las distintas carreras.

En las Figuras 10, 11, 12 y 13, respectivamente a las diferentes carreras, se presenta una visualización de los clusters resultantes utilizando un análisis de componentes principales (PCA) para reducir la dimensionalidad a dos componentes principales. Esta proyección permite observar la dispersión de los clusters en un espacio bidimensional. Se pueden identificar distintas agrupaciones, con cierta superposición entre ellas, lo que indica variaciones en los perfiles estudiantiles pero también sugiere que existen características comunes que pueden estar presentes en múltiples clusters.

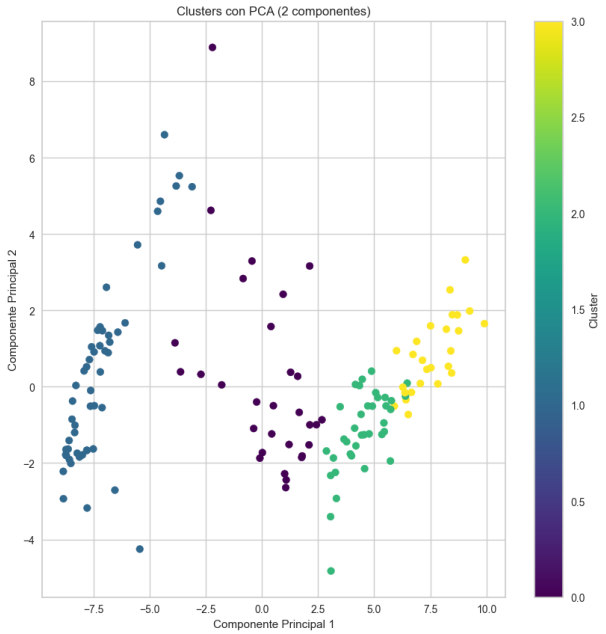


Figura 10: Visualización de clusters resultantes. Carrera Informática

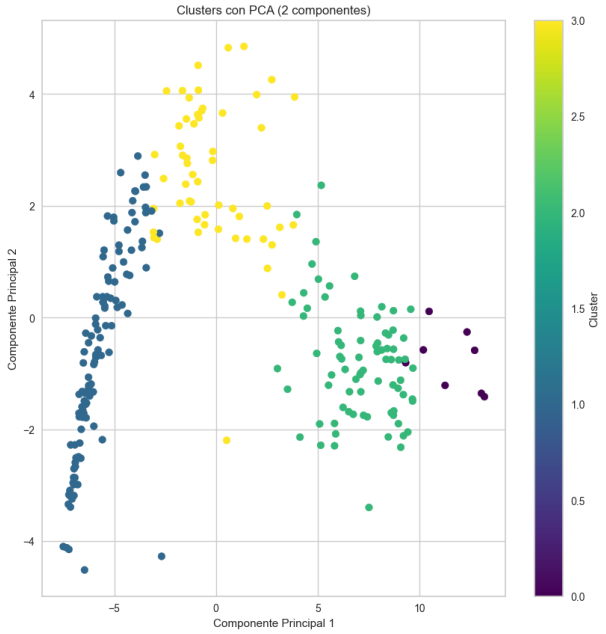


Figura 11: Visualización de clusters resultantes. Carrera Electricidad

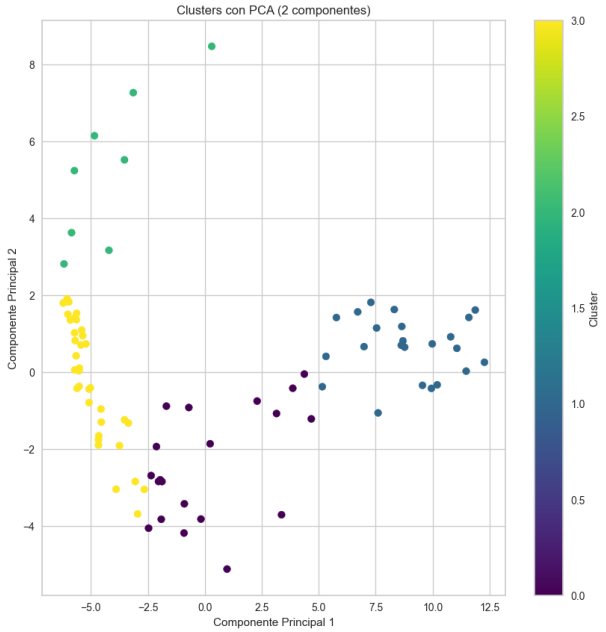


Figura 12: Visualización de clusters resultantes. Carrera Electrónica

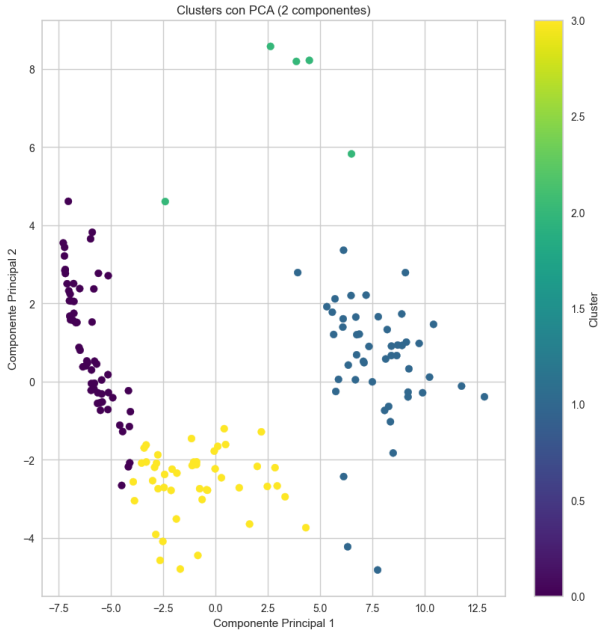


Figura 13: Visualización de clusters resultantes. Carrera Civil

Como el análisis de clustering reveló la presencia de cuatro grupos distintos dentro del conjunto de datos, nos permitió definir cuatro estados académicos específicos de interés para la investigación. A continuación, el algoritmo 2 respresenta el procedimiento que se realizó para la obtención de los estados:

**Algorithm 2** Calcular Estado de Alumno

**Require:** IDAlumno, Materia, AñoIngreso, AñoAcademico , AñoExamen, TiempoEstudio
**Ensure:** estado_alumno

```
Inicializar estado_alumno a "Desconocido"
for cada idAlumno en la base de datos do
    if materia = PG then
        estado_alumno ← "graduado"
    else if ActividadAcademica < 2018 and TiempoEstudio < 5 then
        estado_alumno ← "Desertor temprano"
    else if (carrera = "Informática" and numMaterias=83) or ((carrera = "Civil" or carrera = "Eléctrica" or carrera = "Electrónica") and numMaterias=87) then
        estado_alumno ← "Falta PFG"
    else if ActividadAcademica < 2018 and TiempoEstudio > 5 then
        estado_alumno ← "Desertor Tardio"
```

Estos estados se derivan de combinaciones únicas de características académicas, tales como la completitud de la malla curricular y la actividad académica reciente. A continuación, se presenta una descripción de cada estado:

| Estado Académico | Descripción |
|---|---|
| 2 | Alumnos Graduados: Malla y proyecto final completados. |
| 3 | Desertores Tempranos: Sin actividad desde 2018, estudios de menos de 5 años. |
| 4 | Desertores Tardíos: Sin actividad desde 2018, estudios de más de 5 años. |
| 5 | Proyecto Final Pendiente: Malla completa, proyecto final no aprobado. |

Tabla 7: Descripción de los estados académicos definidos.

### 5.2.1. Desbalance en el Conjunto de Datos

Durante el análisis, se identificó que existe un desequilibrio en la distribución de los estados académicos, con una mayor representación de algunos estados en comparación con otros. Idealmente, un conjunto de datos balanceado permitiría a los modelos aprender con igual representación de cada estado, posibilitando así una evaluación más equitativa de la capacidad predictiva del modelo. No obstante, dada la naturaleza de los datos disponibles y la importancia de retener toda la información posible para una evaluación exhaustiva, no se procedió a equilibrar el conjunto de datos. La decisión de mantener el conjunto de datos en su forma original se sustenta en el número limitado de registros disponibles, lo que confiere un valor significativo a cada instancia individual. Eliminar datos podría resultar en la pérdida de información crítica y potencialmente valiosa, mientras que la generación de datos sintéticos para las clases minoritarias podría introducir sesgos y variaciones que no reflejan la realidad del entorno académico estudiado.

## 5.3. Modelos predictivos

### 5.3.1. Diseño de experimentos

Se utilizaron los siguientes modelos SVM, Random Forest, Regresión Logística, KNN, Decisión tree.

Se comenzó con la construcción del modelo de predicción utilizando el conjunto de datos completo que incluía todas las variables de rendimiento académico ya mencionadas. Este experimento inicial demostró altas métricas de rendimiento cuando fue evaluado utilizando la misma distribución de datos con la que fue entrenado. En la Tabla 8, se observa los resultados del primer experiemento donde ACC es la exactitud, PRESC es la presición, RECALL es la recuperación y F1 el error del tipo F1 que es la varianza entre exactitud y presición.

| Carrera | Métrica | RL | DT | RF | SVM | KNN |
|---|---|---|---|---|---|---|
| Informática | ACC | 0.870 | 0.774 | **0.903** | 0.870 | 0.870 |
| | PRESC | **0.907** | 0.854 | **0.907** | **0.907** | 0.885 |
| | RECALL | 0.870 | 0.774 | **0.903** | 0.870 | 0.870 |
| | F1 | 0.884 | 0.795 | **0.902** | 0.884 | 0.871 |
| Electricidad | ACC | **0.981** | 0.962 | 0.944 | 0.962 | 0.925 |
| | PRESC | **0.981** | 0.927 | 0.951 | 0.927 | 0.940 |
| | RECALL | **0.981** | 0.962 | 0.944 | 0.962 | 0.925 |
| | F1 | **0.978** | 0.944 | 0.940 | 0.944 | 0.919 |
| Electrónica | ACC | **0.944** | 0.888 | **0.944** | **0.944** | **0.944** |
| | PRESC | **0.948** | 0.792 | **0.948** | **0.948** | **0.948** |
| | RECALL | **0.944** | 0.888 | **0.944** | **0.944** | **0.944** |
| | F1 | **0.936** | 0.837 | **0.936** | **0.936** | **0.936** |
| Civil | ACC | **0.846** | 0.807 | 0.807 | **0.846** | 0.807 |
| | PRESC | **0.858** | 0.801 | 0.732 | **0.858** | 0.667 |
| | RECALL | **0.846** | 0.807 | 0.807 | **0.846** | 0.807 |
| | F1 | 0.842 | 0.801 | 0.765 | **0.846** | 0.727 |

Tabla 8: Resumen de Modelos Predictivos por Carrera, entrenamiento y prueba utilizando el mismo conjunto de datos.

En la carrera de Informática el modelo Random Forest (RF) se destaca significativamente, obteniendo las mejores puntuaciones en todas las métricas: exactitud (0.903), precisión (0.907), recall(0.903) y F1 (0.902).

En el caso de la carrera de Electricidad, la Regresión Logística (RL) exhibe un rendimiento sobresaliente, con los valores más altos en exactitud, precisión y recall (0.981), así como en la métrica F1 (0.978). Aunque los demás modelos también muestran un buen desempeño.

Para la carrera de Electrónica, varios modelos, incluyendo RL, RF, SVM, y KNN, comparten el liderazgo en rendimiento. Todos estos modelos alcanzan la máxima exactitud y recall (0.944) y presentan valores muy similares en las métricas de precisión y F1 (0.948 y 0.936, respectivamente).

En la carrera de Civil, el modelo SVM se destacan, alcanzando la mayor exactitud, recall y F1 (0.846) . La RL también lidera en precisión (0.858) al igual que SVM.

Sin embargo, se observó una discrepancia significativa en el rendimiento cuando el modelo fue aplicado en un contexto realista, donde solo estaban disponibles los datos hasta el tercer año. Este fenómeno puso de relieve la existencia de un desajuste de distribución, donde los patrones aprendidos por el modelo no eran totalmente aplicables al escenario de predicción deseado.

Para abordar este problema, se reajusto los datos, para entrenar el modelo únicamente

con datos disponibles del movimiento academico hasta el tercer año del alumnado.

| Carrera | Métrica | RL | DT | RF | SVM | KNN |
|---|---|---|---|---|---|---|
| Informática | ACC | 0.689 | 0.655 | 0.689 | 0.689 | **0.724** |
| | PRESC | 0.720 | 0.692 | 0.727 | 0.742 | **0.822** |
| | RECALL | 0.689 | 0.655 | 0.689 | 0.689 | **0.724** |
| | F1 | 0.692 | 0.665 | 0.701 | 0.693 | **0.720** |
| Electricidad | ACC | 0.745 | **0.803** | **0.803** | 0.705 | 0.705 |
| | PRESC | 0.735 | **0.812** | 0.758 | 0.668 | 0.668 |
| | RECALL | 0.745 | **0.803** | **0.803** | 0.705 | 0.705 |
| | F1 | 0.732 | **0.803** | 0.771 | 0.685 | 0.727 |
| Electrónica | ACC | **0.833** | 0.722 | 0.777 | 0.777 | 0.777 |
| | PRESC | **0.888** | 0.861 | 0.814 | 0.731 | 0.877 |
| | RECALL | **0.833** | 0.722 | 0.777 | 0.777 | 0.777 |
| | F1 | **0.837** | 0.744 | 0.785 | 0.738 | 0.776 |
| Civil | ACC | 0.781 | **0.843** | **0.843** | 0.812 | 0.75 |
| | PRESC | 0.753 | **0.904** | 0.871 | 0.759 | 0.768 |
| | RECALL | 0.781 | **0.843** | **0.843** | 0.812 | 0.75 |
| | F1 | 0.763 | **0.843** | 0.837 | 0.777 | 0.750 |

Tabla 9: Resumen de Modelos Predictivos por Carrera, entrenamiento y prueba con datos de Tercer año

Para la carrera de Informática, el modelo KNN destaca al alcanzar las métricas más altas en todas las categorías: exactitud (0.724), precisión (0.822), recall (0.724) y F1 (0.720).

En Electricidad, el Árbol de Decisión (DT) y el Random Forest (RF) comparten el liderazgo en las métricas de exactitud y recall (0.803), con el DT mostrando la mayor precisión (0.812).

Para Electrónica, la Regresión Logística (RL) presenta el mejor rendimiento con la más alta exactitud (0.833) y recall (0.833), así como la mejor precisión (0.888) y F1 (0.837).

En la carrera de Civil, el Árbol de Decisión nuevamente sobresale, ofreciendo las puntuaciones más altas en exactitud y recall (0.843) y la mayor precisión (0.904).

La aplicación de este experimento reveló diferencias sustanciales en la capacidad de predicción entre los modelos, y se observó que ciertas categorías de estado académico eran confundidas con frecuencia. Destacablemente, las categorías correspondientes a estudiantes egresados (estado 2) y aquellos que habían completado la malla curricular pero no el Proyecto Final de Grado (estado 5) presentaban una alta tasa de confusión. Estas dos categorías comparten la característica común del éxito académico hasta el momento, lo que potencialmente explicaba la dificultad del modelo para distinguirlas.

Tras un análisis detallado, se tomó la decisión metodológica de combinar estos dos Estados

en una sola categoría que representara el éxito académico hasta ese punto. Esta decisión se basó en la lógica de que ambos estados reflejan una transición exitosa a través de la malla curricular y una proximidad a la culminación de sus estudios.

| **Carrera** | **Métrica** | **RL** | **DT** | **RF** | **SVM** | **KNN** |
|---|---|---|---|---|---|---|
| Informática | ACC | 0.827 | 0.793 | 0.862 | 0.862 | **0.896** |
| | PRESC | 0.863 | 0.853 | 0.876 | 0.876 | **0.896** |
| | RECALL | 0.827 | 0.793 | 0.862 | 0.862 | **0.896** |
| | F1 | 0.839 | 0.814 | 0.865 | 0.865 | **0.895** |
| Electricidad | ACC | 0.901 | **0.980** | 0.921 | 0.862 | 0.823 |
| | PRESC | 0.877 | **0.981** | 0.911 | 0.809 | 0.805 |
| | RECALL | 0.901 | **0.980** | 0.921 | 0.862 | 0.823 |
| | F1 | 0.882 | **0.979** | 0.912 | 0.835 | 0.814 |
| Electrónica | ACC | **0.888** | 0.777 | 0.833 | 0.833 | **0.888** |
| | PRESC | **0.898** | 0.809 | 0.740 | 0.740 | **0.898** |
| | RECALL | **0.888** | 0.777 | 0.833 | 0.833 | **0.888** |
| | F1 | **0.882** | 0.790 | 0.783 | 0.783 | **0.882** |
| Civil | ACC | 0.875 | **0.968** | 0.906 | 0.906 | 0.843 |
| | PRESC | 0.889 | **0.976** | 0.920 | 0.909 | 0.850 |
| | RECALL | 0.875 | **0.968** | 0.906 | 0.906 | 0.843 |
| | F1 | 0.878 | **0.970** | 0.911 | 0.904 | 0.843 |

Tabla 10: Resumen de Modelos Predictivos por Carrera, entrenamiento y prueba con datos hasta tercer año fusionando estado 2 y 5

En la carrera de Electricidad, el modelo de árboles de decisión (DT) sobresale con resultados muy cercanos a la perfección, pero hay una disminución en las métricas para KNN en comparación con los resultados previos a la combinación de los estado.

En la carrera de Informática, aunque todos los modelos mejoraron, el KNN destaca, pero sigue habiendo espacio para optimización en los otros modelos.

Para Electrónica, todos los modelos experimentaron una mejora, pero algunos como RF y SVM no alcanzaron los niveles de rendimiento del modelo KNN o RL.

La carrera de Civil muestra mejoras en todas las métricas para todos los modelos, pero la mejora es menos pronunciada para KNN. En esta se destaca el modelo de Árbol de decisiones sobre las demás.

Después de observar la mejora en el rendimiento de los modelos predictivos al combinar las categorías de estudiantes que han terminado sus cursos pero aún no han completado su proyecto final de grado (estado 5) con aquellos que se han graduado (estado 2), se realizó un siguiente experimento. Este consistió en entrenar los modelos utilizando datos correspondientes a los estudiantes en su cuarto curso, manteniendo esta vez separadas

las clases 2 y 5. La premisa de este experimento fue evaluar si, al tener acceso a más información acumulada a lo largo de la carrera estudiantil, el modelo podría diferenciar con mayor efectividad entre estos dos estados académicos distintos.

| **Carrera** | **Métrica** | **RL** | **DT** | **RF** | **SVM** | **KNN** |
|---|---|---|---|---|---|---|
| Informática | ACC | **0.758** | 0.689 | **0.758** | 0.724 | 0.724 |
| | PRESC | **0.844** | 0.706 | 0.790 | 0.822 | 0.770 |
| | RECALL | **0.758** | 0.689 | **0.758** | 0.724 | 0.724 |
| | F1 | 0.725 | 0.695 | **0.762** | 0.720 | 0.732 |
| Electricidad | ACC | 0.784 | **0.882** | 0.803 | 0.725 | 0.725 |
| | PRESC | 0.785 | **0.888** | 0.803 | 0.684 | 0.684 |
| | RECALL | 0.784 | **0.882** | 0.803 | 0.725 | 0.725 |
| | F1 | 0.781 | **0.882** | 0.799 | 0.704 | 0.754 |
| Electrónica | ACC | **0.944** | **0.944** | 0.888 | 0.833 | 0.833 |
| | PRESC | **0.962** | 0.958 | 0.912 | 0.757 | 0.851 |
| | RECALL | **0.944** | **0.944** | 0.888 | 0.833 | 0.833 |
| | F1 | **0.944** | 0.939 | 0.880 | 0.782 | 0.833 |
| Civil | ACC | 0.781 | **0.875** | 0.843 | 0.843 | **0.875** |
| | PRESC | 0.753 | **0.908** | 0.778 | 0.772 | 0.810 |
| | RECALL | 0.781 | **0.875** | 0.843 | 0.843 | **0.875** |
| | F1 | 0.763 | **0.887** | 0.805 | 0.805 | 0.833 |

Tabla 11: Resumen de Modelos Predictivos por Carrera, entrenamiento y prueba con datos hasta cuarto año.

Para la carrera de Informática, la Regresión Logística (RL) y el Random Forest (RF) son destacados, ambos alcanzando la mayor exactitud (0.758). La Regresión Logística también muestra la mejor precisión (0.844) .

En Electricidad, el Árbol de Decisión (DT) exhibe un rendimiento superior, con la mejor exactitud (0.882) y la mejor precisión (0.888).

Para la carrera de Electrónica, tanto la Regresión Logística (RL) como el Árbol de Decisión (DT) comparten la mayor exactitud (0.944), con la Regresión Logística obteniendo además la mayor precisión (0.962) y la mejor puntuación F1 (0.944).

Finalmente, en la carrera de Civil, el Árbol de Decisión (DT) y el K-Nearest Neighbors (KNN) logran la más alta exactitud (0.875). El DT, sin embargo, sobresale con la mejor precisión (0.908) y el mejor F1 (0.887) .

Los resultados de este experimento demuestran que, si bien la inclusión de datos de cuarto año mejora la capacidad predictiva de los modelos en algunas áreas, las confusiones entre las clases de éxito académico (2 y 5) siguen siendo un punto de atención. Incrementar la cantidad de datos hasta el cuarto año académico refuerza significativamente la precisión

de los modelos predictivos, a pesar de esta mejora, se reconoce la existencia de desafíos en la clasificación exacta, particularmente en la diferenciación de los estados mencionados.

Para complementar y visualizar efectivamente los resultados obtenidos en los experimentos, se han generado gráficos detallados para cada carrera, que ilustran el rendimiento de los distintos modelos predictivos utilizados. Estos gráficos reflejan la exactitud, precisión, recall y la puntuación F1 para los modelos SVM, Random Forest, Regresión Logística, KNN y Árbol de Decisión en diversas configuraciones de datos.

### 5.3.2. Comparación de modelos, con las diferentes configuraciones de datos

**Primer Experimento, entrenamiento y prueba utilizando el mismo conjunto de datos.** En el gráfico 14 correspondiente a la carrera de Informática, se observa que el modelo RF presenta las métricas más altas en comparación con los demás modelos, lo cual es consistente con la tabla de resumen 8 que señala al RF como el modelo ganador para esta carrera debido a su alto rendimiento y simplicidad.

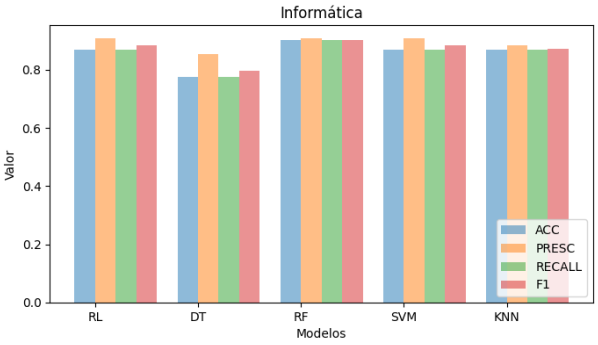


Figura 14: Gráfico primer experimento, entrenamiento y prueba utilizando el mismo conjunto de datos. Carrera Informática.

Para la carrera de Electricidad, el gráfico 15 destaca al RL con valores de rendimiento superiores, especialmente en términos de exactitud y precisión.

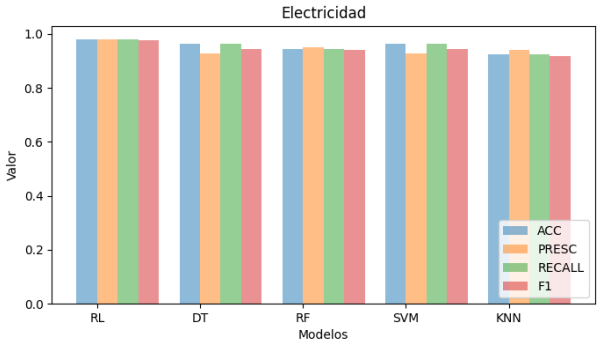


Figura 15: Gráfico primer experimento, entrenamiento y prueba utilizando el mismo conjunto de datos. Carrera Electricidad.

En lo que respecta a Electrónica, el gráfico 16 muestra un rendimiento relativamente equilibrado entre los modelo.

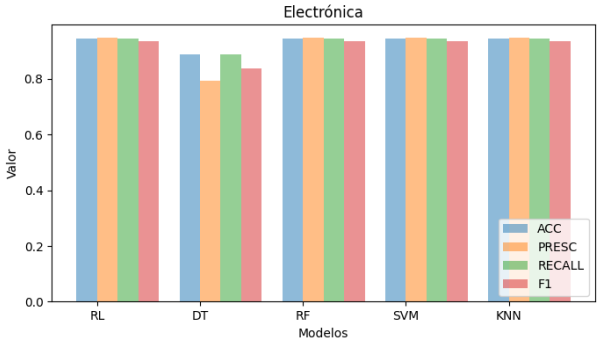


Figura 16: Gráfico primer experimento, entrenamiento y prueba utilizando el mismo conjunto de datos. Carrera Electrónica.

Finalmente, el gráfico 17 de la carrera de Civil resalta a SVM como el modelo con las métricas más altas, similar al RL solo que hay disminución en la metrica de F1.

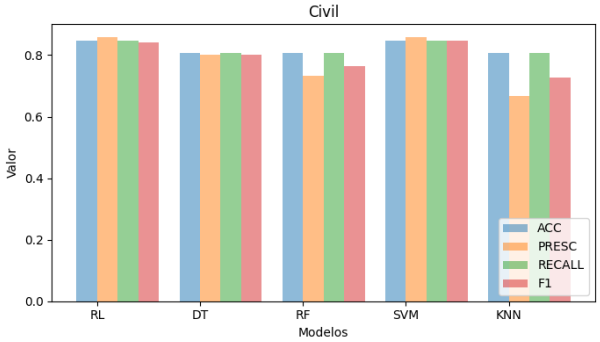


Figura 17: Gráfico primer experimento, entrenamiento y prueba utilizando el mismo conjunto de datos. Carrera Civil.

**Segundo Experimento, entrenamiento y prueba con datos de hasta Tercer curso.** El gráfico 18 para la carrera de Informática destaca al modelo KNN. A pesar de la reducción del conjunto de datos al tercer año, el KNN tuvo un buen rendimiento.

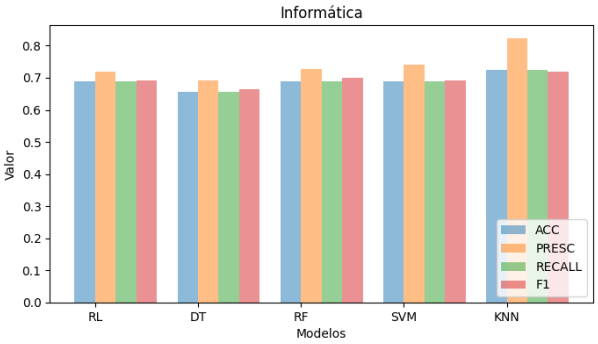


Figura 18: Gráfico segundo experimento, entrenamiento y prueba con datos hasta tercer curso. Carrera Informática.

Para la carrera de Electricidad,en el gráfico 19 se observa que el modelo Árbol de Decisión(DT) muestra un mejor rendimiento.

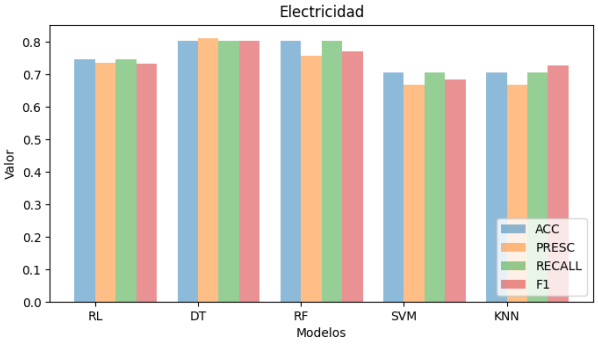


Figura 19: Gráfico segundo experimento, entrenamiento y prueba con datos hasta tercer curso. Carrera Electricidad.

En el caso de Electrónica, el modelo RL tuvo un rendimiento más uniforme a través de las métricas como se ve en el grafico 20. La Regresión Logística mostrando fortalezas en términos de precisión, reafirmando su eficacia incluso con un conjunto de datos más limitado.

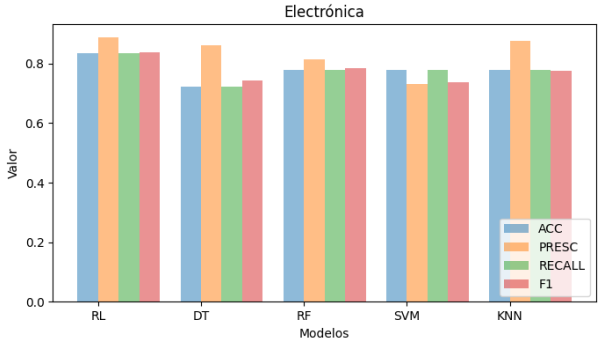


Figura 20: Gráfico segundo experimento, entrenamiento y prueba con datos hasta tercer curso. Carrera Electrónica.

El gráfico 21 de la carrera de Civil muestra al Árbol de Desición(DT) con mejor rendimiento.

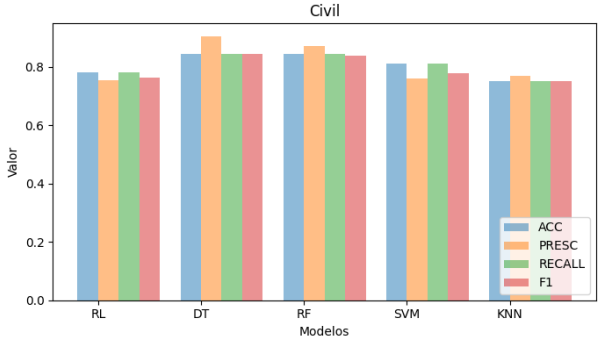


Figura 21: Gráfico segundo experimento, entrenamiento y prueba con datos hasta tercer curso. Carrera Civil.

**Tercer Experimento, entrenamiento y prueba con datos de hasta Tercer curso, fusionando estado 2 y 5.** En la carrera Informática, el grafico 22 muestra nuevamente a modelo KNN con mejor rendimiento para este experimento.

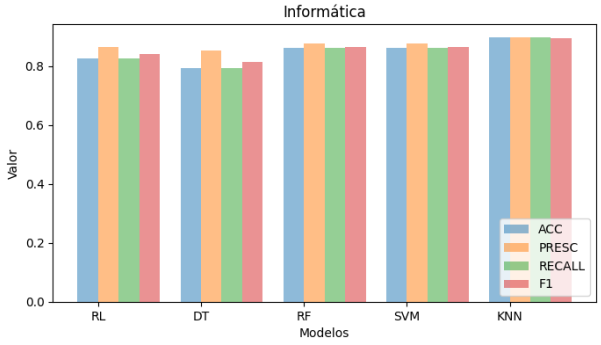


Figura 22: Gráfico tercer experimento, entrenamiento y prueba con datos hasta tercer curso fusionando estado 2 y 5. Carrera Informática.

Para la carrera de Electricidad la grafica 23 muestra al igual que el experimento anterior muestra un mejor rendiminento con el modelo Árbol de Decisión(DT).

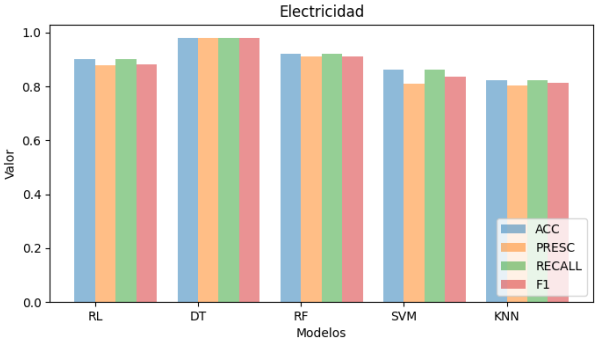


Figura 23: Gráfico tercer experimento, entrenamiento y prueba con datos hasta tercer curso fusionando estado 2 y 5. Carrera Electricidad.

En la carrera de Electrónica, en el grafico 24 se observa que los modelos RL y KNN comparten un mejor rendimiento para este experimento.

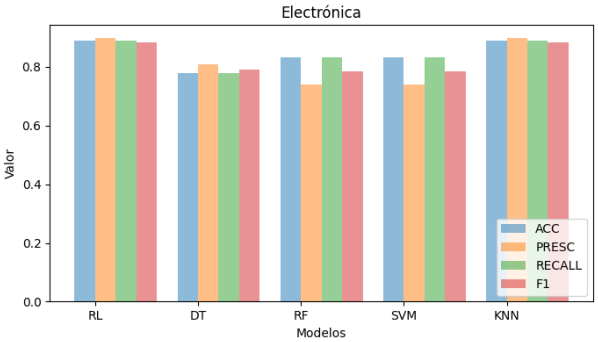


Figura 24: Gráfico tercer experimento, entrenamiento y prueba con datos hasta tercer curso fusionando estado 2 y 5. Carrera Electrónica.

En el grafico 25 correspondiente a la carrera civil, se observa al modelo de Árbol de Decisiones(DT) tener un mejor rendimiento.

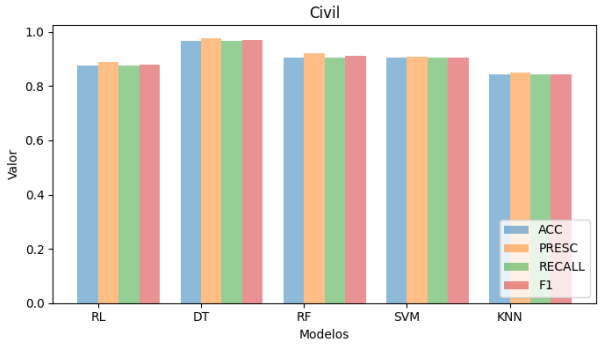


Figura 25: Gráfico tercer experimento, entrenamiento y prueba con datos hasta tercer curso fusionando estado 2 y 5. Carrera Civil.

**Cuarto Experimento, entrenamiento y prueba con datos de hasta Cuarto curso, separando estado 2 y 5.** Para la carrera informatica, en el grafico 26 se observa que hay cierta similitu entre los modelos RL Y RF, donde RL muestra una mejor presicion comparado con los demas modelos.

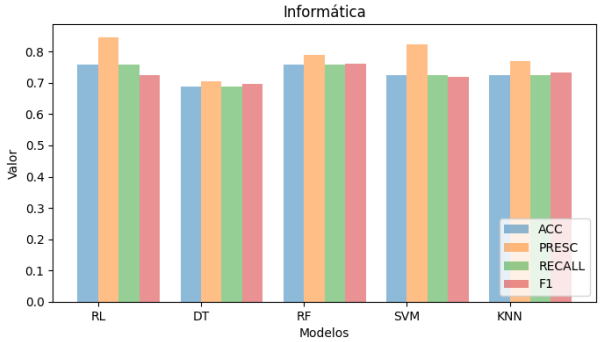


Figura 26: Gráfico Cuarto experimento, entrenamiento y prueba con datos hasta Cuarto curso. Carrera Informática.

El grafico 27 de la carrera Electricidad muestra nuevamente al Árbol de Decisiones(DT) con un mejor rendiminto.

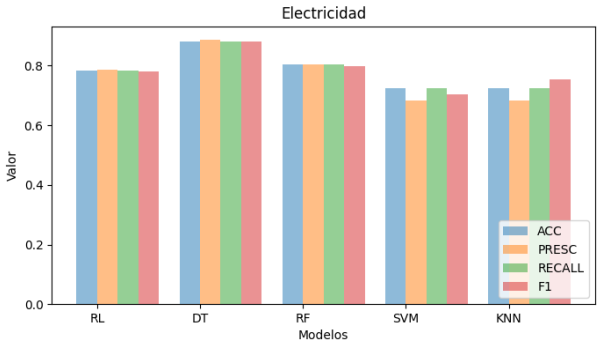


Figura 27: Gráfico Cuarto experimento, entrenamiento y prueba con datos hasta Cuarto curso. Carrera Electricidad.

En la carrera Electrónica, en el grafico 28 se observa también al igual que el anterior experimento al RL con mejor rendimiento.

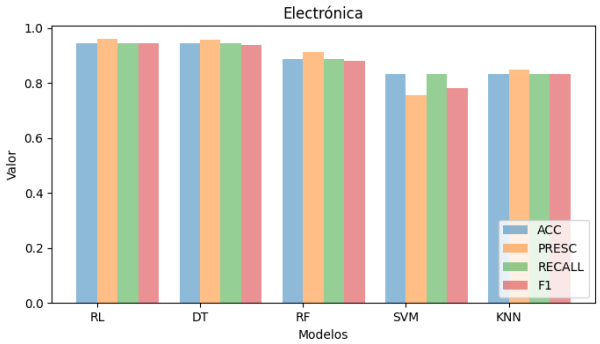


Figura 28: Gráfico Cuarto experimento, entrenamiento y prueba con datos hasta Cuarto curso. Carrera Electrónica.

Por último el gráfico 28 de la carrera civil, muestra nuevamente al igual que el anterior experimento con un buen rendimiento al modelo Árbol de Decisiones(DT)

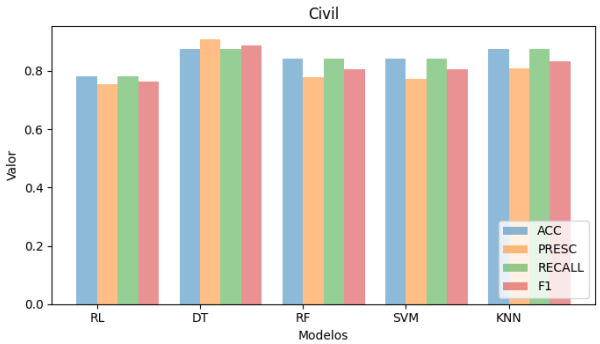


Figura 29: Gráfico Cuarto experimento, entrenamiento y prueba con datos hasta Cuarto curso. Carrera Civil.

En el transcurso de estos experimentos, se implementaron un total de cinco modelos predictivos: SVM, Random Forest, Regresión Logística, KNN y Árbol de Decisión. Cada uno de estos modelos se evaluó en varias configuraciones de datos, que incluyeron conjuntos de datos completos, conjunto de datos solo hasta el tercer año, conjunto de datos solo hasta el tercer año mezclando 2 estados y otro conjunto de datos solo hasta el cuarto año de estudios. Al final de los experimentos, el enfoque seleccionado fue el uso de datos hasta el tercer año con la combinación de las categorías de estados académicos 2 y 5 en una sola. Esta configuración se eligió porque mostró un mejor equilibrio en el rendimiento de predicción entre las diferentes carreras y modelos. Los experimentos muestran que para

| **Carrera** | **Modelo Ganador** | **Métricas Clave** | **Razón de Elección** |
|---|---|---|---|
| Informática | KNN | Exactitud: 0.896, Precisión: 0.896 | Alto rendimiento y simplicidad. |
| Electricidad | DT | Exactitud: 0.980, Precisión: 0.981 | Excelente en precisión, fácil interpretación de resultados. |
| Electrónica | RL | Exactitud: 0.888, Precisión: 0.898 | Alta precisión, proporciona insights claros de las variables. |
| Civil | DT | Exactitud: 0.968, Precisión: 0.970 | Consistentemente alto rendimiento y buena comprensibilidad. |

Tabla 12: Resumen del Mejor Modelo por Carrera, con el enfoque seleccionado.

la carrera de Informática, el modelo KNN con datos hasta el tercer año con la combinación de las categorías de estados académicos 2 y 5 en una sola proporcionó las mejores métricas de rendimiento. En Electricidad, el Árbol de Decisión se destacó especialmente , mostrando una alta precisión y exactitud. Para Electrónica, la Regresión Logística fue el más efectivos. En la carrera de Civil, el Árbol de Decisión mostró una consistente superioridad. Los modelos y configuraciones de datos finales reflejan un enfoque dirigido a optimizar la precisión y la aplicabilidad en escenarios realistas donde no todos los datos de los estudiantes están disponibles de inmediato. La decisión de combinar los estados 2 y 5 resultó ser crítica para mejorar la claridad de la predicción y reducir las tasas de confusión, lo que sugiere que la agrupación de categorías similares puede ser una estrategia efectiva en contextos donde las diferencias entre estados son mínimas pero críticamente importantes. Con esta configuración final, los modelos mejoraron significativamente su capacidad para predecir con precisión los resultados académicos en las cuatro carreras estudiadas. Esta mejora fue especialmente notable en el contexto de predicciones más realistas y prácticas, adecuadas para la toma de decisiones en entornos educativos donde las intervenciones tempranas pueden ser necesarias. Este enfoque subraya la importancia de adaptar los modelos de aprendizaje automático a las peculiaridades de los datos y las necesidades específicas del entorno de aplicación para lograr resultados óptimos.

# 6. Conclusiones y recomendaciones

- Se construyó un modelo de alerta temprana para la deserción estudiantil en la Facultad de Ciencias y Tecnologías, utilizando para ello estimaciones basadas en factores académicos relevantes extraídos de la base de datos académica de la facultad.

- En el estudio realizado, se logró una caracterización efectiva de la base de datos académica mediante técnicas avanzadas de ciencia de datos. Inicialmente, se utilizó el método del codo para determinar el número óptimo de agrupaciones, identificando cuatro grupos distintos. Se segmentó la población estudiantil en: graduados, desertores tempranos (estudiantes que abandonaron antes de los cinco años de carrera), desertores tardíos (aquellos que dejaron la carrera después de cinco años) y estudiantes que completaron la malla curricular pero aún deben presentar su Proyecto de Fin de Grado. Este análisis detallado permitió comprender mejor la distribución académica.

- Los modelos predictivos de machine learning se ajustaron y evaluaron mediante cuatro configuraciones distintas de datos en una serie de experimentos de entrenamiento y predicción. El método más efectivo fue el tercer experimento, que combinaba datos de estudiantes en los estados 2 y 5 hasta el tercer año. Esta combinación creó un conjunto de datos más homogéneo y representativo del éxito académico, permitiendo a los modelos identificar con mayor precisión los patrones y factores clave que predeterminan los resultados académicos exitosos.

- Con la seleccion del tercer experimento, para las distintas carreras, los modelos óptimos variaron: para Informática, el mejor modelo resultó ser K-Nearest Neighbors (KNN) con una extactitud, precisión y recall de 0.896 y F1 de 0.895 en contraste el que tuvo menor desempeño fue Árbol de Decisión (DT) con una exactitud y recall de 0.793, una precisión de 0.853 y F1 de 0.814 ; para Electricidad y Civil, el modelo de Árbol de Decisión (DT) fue el más eficaz, en electricidad con una exactitud y recall de 0.980, una precisión de 0.981, y F1 de 0.979, en la carrera civil con una exactitud y recall de 0.968, una presición de 0.976 y F1 de 0.970 en cambio el que tuvo menor desempeño en las dos carreras Fue KNN, respectivamente para la carrera de Electricidad con una exactitud y recall de 0.823, una precisión de 0.805 y F1 de 0.814 y en la carrera Civil con exatitud, recall y F1 de 0.843 y una precisión de 0.850 ; y para Electrónica, la Regresión Logística (RL) y K-Nearest Neighbors (KNN) demostraron un mejor rendimiento con exactitud y recall de 0.888, con una precisión de 0.898 y F1 de 0.882 a diferencia el Modelo Árbol de Decisión (DT) demostro menor exactitud y racall de 0.777 mejorando un poco en la presicion con

0.809 comparado con los modelos RF y SVM que demostraron una menor precisión de 0.740.

## 6.1. Recomendaciones

Al conlcuir este trabajo se llegaron a las siguientes recomendaciones:

- Promover la implementación de estos modelos de alerta temprana en el sistema académico de la facultad para realizar intervenciones proactivas y reducir la tasa de deserción.
- Considerar la inclusión de variables socioeconómicas y psicológicas en futuras versiones del modelo para mejorar la comprensión y predicción de la deserción estudiantil, utilizando una visión más holística del estudiante.

# Referencias

[1] AWS Amazon, "What is data science?" 2023, [Online]. Available: https://aws.amazon.com/es/what-is/data-science/.

[2] F. Bellas Aláez, "Métodos numéricos de aprendizaje automático supervisado aplicados a la predicción de floraciones masivas de pseudo- nitzschia spp. en las rías baixas gallegas." Ph.D. disertación, 05 2021.

[3] J. A. Camacho, "K-nearest neighbors," 2021, [Online]. Available: https:https://www.jacobsoft.com.mx/en/k-nearest-neighbors/.

[4] Daniel Rodríguez, "Método del codo (elbow method) para seleccionar el número óptimo de clústeres en k-means," 2023, [Online].Available: https://www.analyticslane.com/2023/06/09/metodo-del-codo.

[5] C. D. FCyT, "Plan de desarrollo ingeniería en electricidad," http://www.fctunca.edu.py/application/files/6814/9092/2147/Plan_de_Desarrollo_Ingenieria_en_Electricidad_1.pdf, 2016, [Online; accessed 26-October-2016].

[6] Latitud25, "Alta deserción universitaria: hablemos de lo complejo que es estudiar en paraguay," https://enlatitud25.com/news/alta-desercion-universitaria-hablemos-de-lo-complejo-que-es-estudiar-en-paraguay/, 2023, [Online; accessed 2-February-2023].

[7] B. Mahesh, *Machine Learning Algorithms -A Review*, 2019, [Online]. Available: https:https://www.researchgate.net/publication/344717762_Machine_Learning_Algorithms_-A_Review.

[8] Mohammadmehdi Saberioon. Petr Císař, Laurent Labbé, Pavel Souček, Pablo Pelissier y Thierry Kerneis, "Análisis comparativo del rendimiento de la máquina de vectores de soporte, el bosque aleatorio, la regresión logística y los k-vecinos más cercanos en la clasificación de la trucha arco iris (oncorhynchus mykiss) utilizando características basadas en imágenes," 2018, [Online]. Available: https://www.mdpi.com/1424-8220/18/4/1027.

[9] J. M. S. Muñoz, "Análisis de calidad cartográfica mediante el estudio de la matriz de confusión," *Pensamiento matemático*, vol. 6, no. 2, pp. 9–26, 2016.

[10] S. V. Orea, A. S. Vargas, y M. G. Alonso, "Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos," *Ene*, vol. 779, no. 73, p. 33, 2005.

[11] Pablo Díaz, Alexis Tejedor De Leó, "El cadesun. un nuevo instrumento para analizar la deserción estudiantil universitaria," 2017, [Online]. Available: https://www.redalyc.org/journal/340/34056722012/html/.

[12] J. Riebesell, "Random forest," 2024, [Online]. Available: https:https://tikz.net/random-forest/.

[13] Rocío Barrientos Martínez, Nicandro Ramírez, Héctor, Acosta Mesa, Ivonne Rabatte Suárez, María del Carmen Gogeascoechea Trejo, Patricia Pavón León, Sobeida L. Blázquez Morales, "Árboles de decisión como herramienta en el diagnóstico médico," 2009, [Online]. Available: http://www.soporte.uv.mx/rm/num_anteriores/revmedica_vol9_num2/articulos/arboles.pdf.

[14] R. Srivatsan, P. Indi, S. Agrahari, S. Menon, y S. Ashok, "Machine learning based prognostic model and mobile application software platform for predicting infection susceptibility of covid-19 using healthcare data," *Research on Biomedical Engineering*, vol. 38, pp. 1–12, 11 2020.

[15] S. Swataroy, "Decoding logistic regression: Your key to binary classification in machine learning," 2024, [Online]. Available: https://medium.com/@saraswataroy21/decoding-logistic-regression-your-key-to-binary-classification-in-machine-learning-0b1947ddb3c8.

[16] Trupti Kodinariya, Dr. Prashant Makwana, "Review on determining number of cluster in k-means clustering," 2013, [Online]. Available: https://www.researchgate.net/profile/Trupti-Kodinariya/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering/links/5789fda408ae59aa667931d2/Review-on-Determining-of-Cluster-in-K-means-Clustering.pdf.

[17] Waster, "Explicación alternativa para accuracy, precision, recall y f1-score," 2019, [Online]. Available: https://steemit.com/spanish/@waster/explicacion-alternativa-para-accuracy-precision-recall-y-f1-score.

[18] L. B. y. P. M. Norma Coppari, "Eureka," https://psicoeureka.com.py/sites/default/files/articulos/eureka-16-1-12_0.pdf, 2019, [Online; accessed 10-May-2019].

# Apéndices

## A. Análisis de clusters

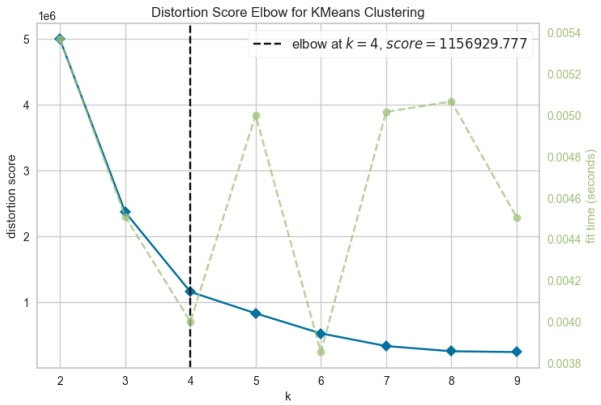


Figura 30: Análisis de clustering, método del codo. Carrera Informatica

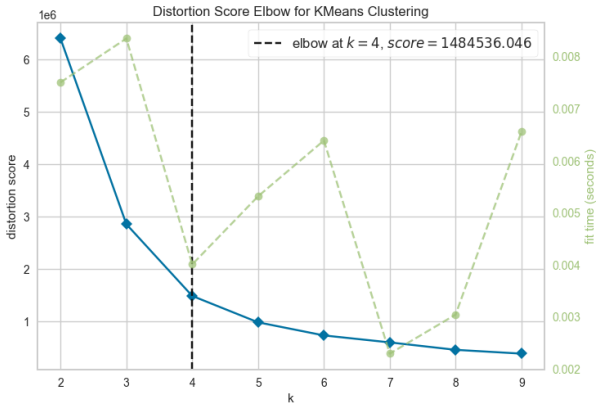


Figura 31: Análisis de clustering, método del codo. Carrera Electricidad

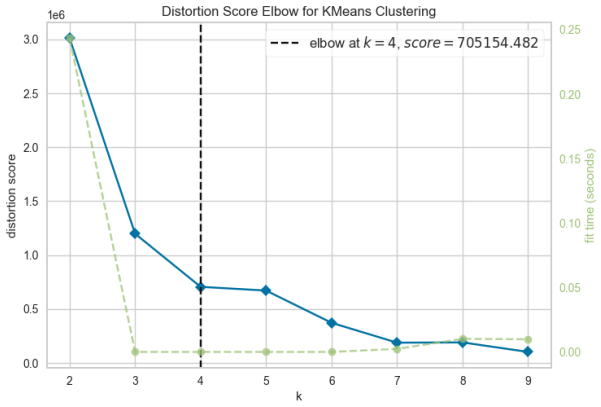


Figura 32: Análisis de clustering, método del codo. Carrera Eletrónica


Figura 33: Análisis de clustering, método del codo. Carrera Informática

# B. Matriz de Confusión

## B.1. Primer Experimento, Entrenamiento y Prueba con todos los datos

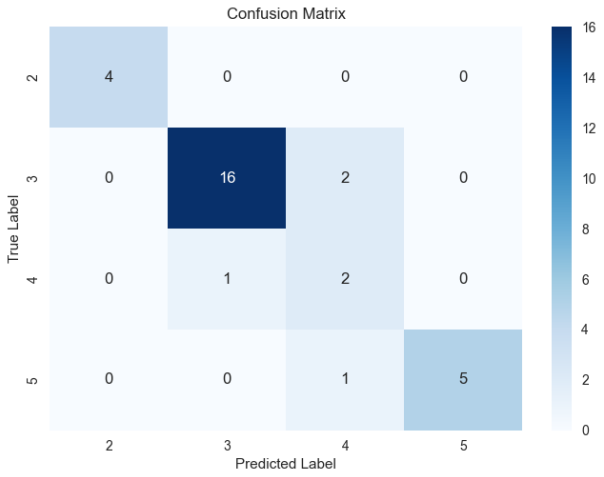


Figura 34: Matriz de Confusión, Modelo RL, carrera Informática, Experimento 1

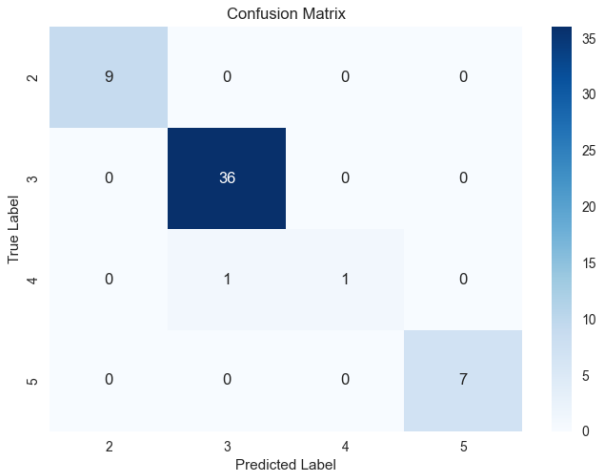


Figura 35: Matriz de Confusión, Modelo RL, carrera Electricidad, Experimento 1

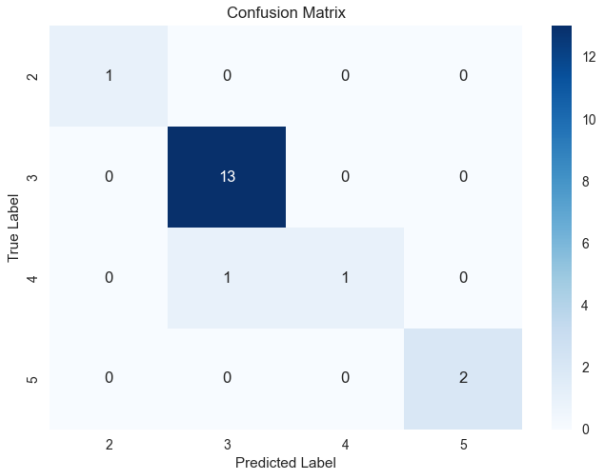


Figura 36: Matriz de Confusión, Modelo RL, carrera Electrónica, Experimento 1

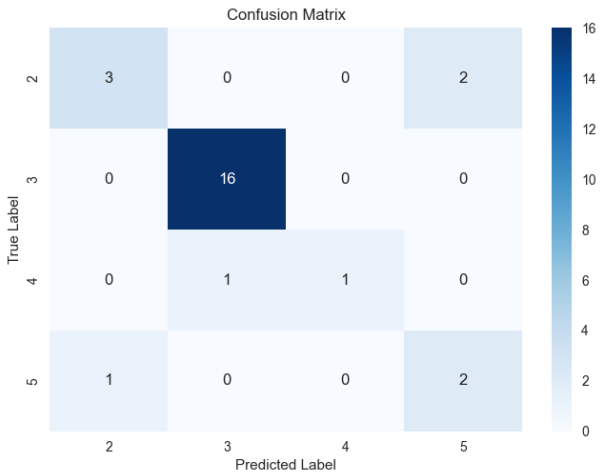


Figura 37: Matriz de Confusión, Modelo RL, carrera Civil, Experimento 1

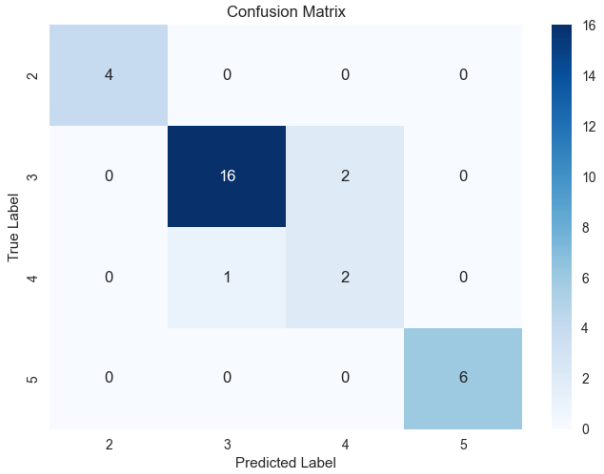


Figura 38: Matriz de Confusión, Modelo DT, carrera Informática, Experimento 1

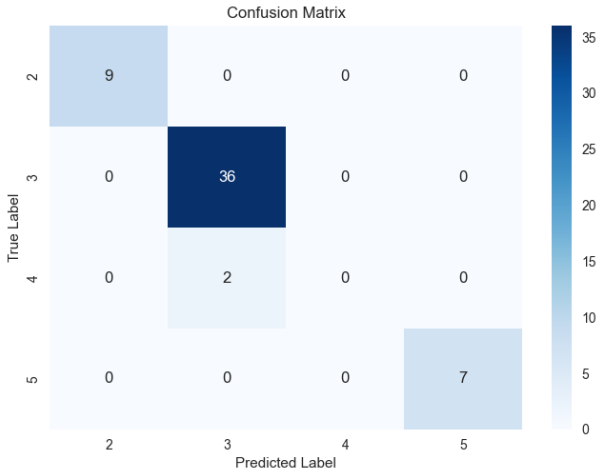


Figura 39: Matriz de Confusión, Modelo DT , carrera Electricidad, Experimento 1

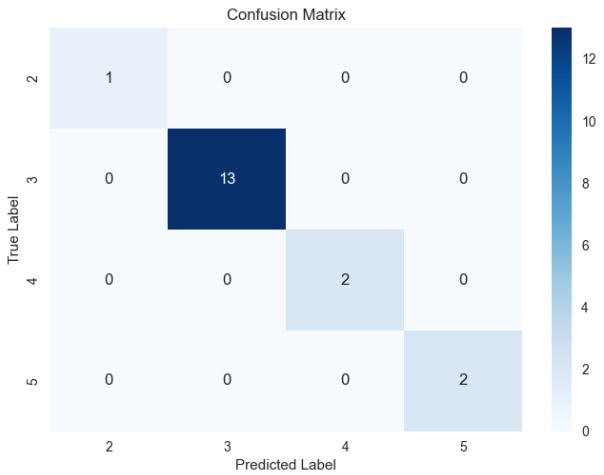


Figura 40: Matriz de Confusión, Modelo DT, carrera Electrónica, Experimento 1

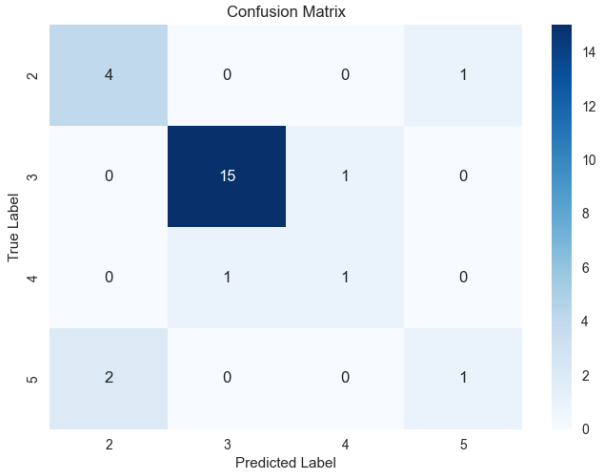


Figura 41: Matriz de Confusión, Modelo DT, carrera Civil, Experimento 1

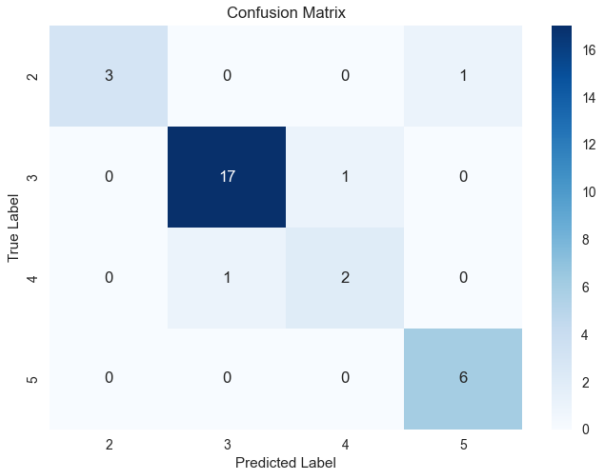


Figura 42: Matriz de Confusión, Modelo RF, carrera Informática, Experimento 1

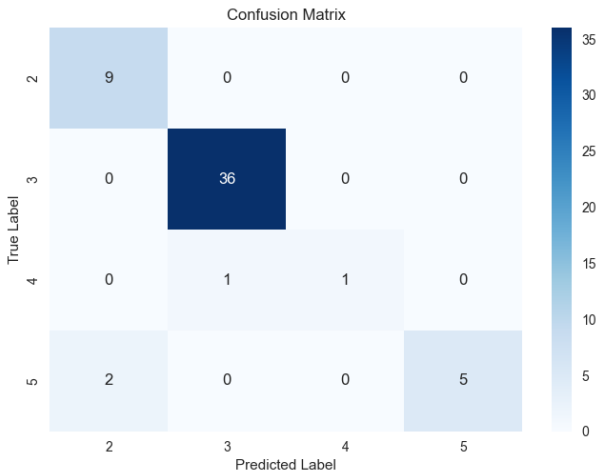


Figura 43: Matriz de Confusión, Modelo RF , carrera Electricidad, Experimento 1

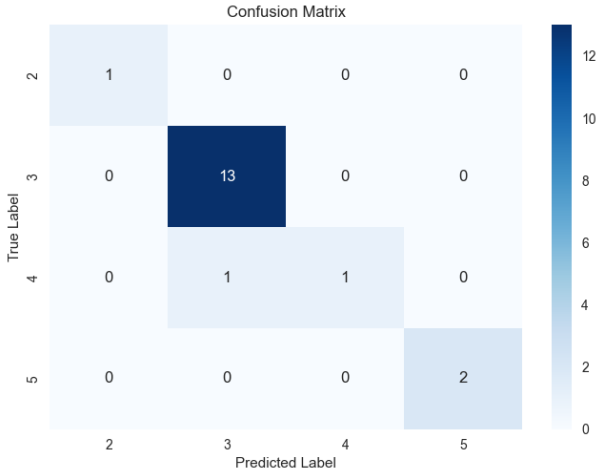


Figura 44: Matriz de Confusión, Modelo RF, carrera Electrónica, Experimento 1

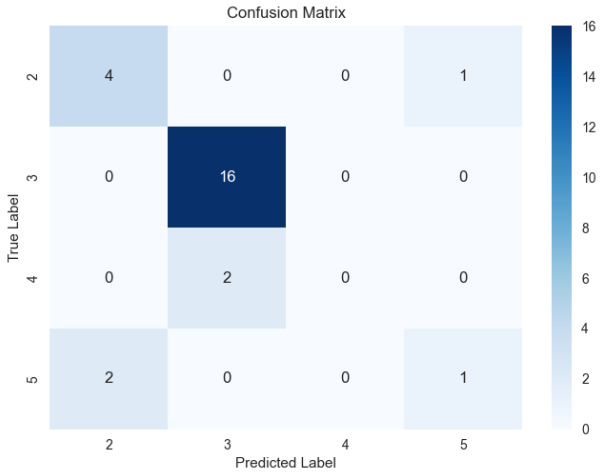


Figura 45: Matriz de Confusión, Modelo RF, carrera Civil, Experimento 1

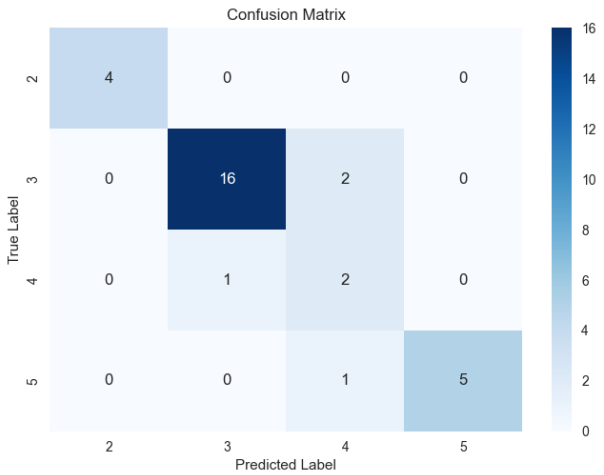


Figura 46: Matriz de Confusión, Modelo SVM, carrera Informática, Experimento 1

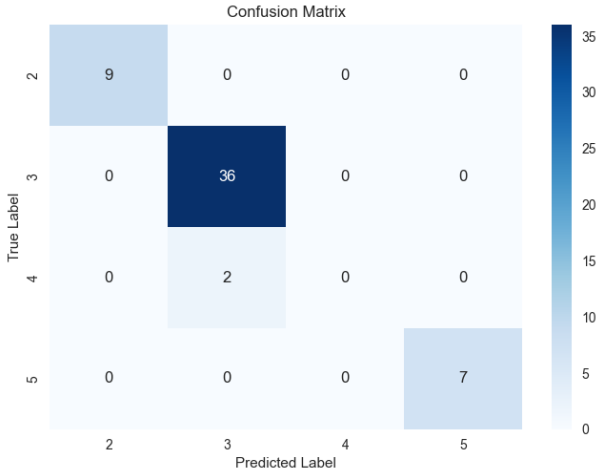


Figura 47: Matriz de Confusión, Modelo SVM, carrera Electricidad, Experimento 1

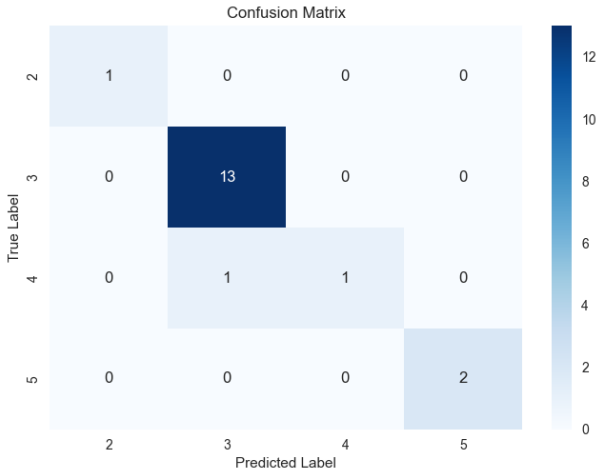


Figura 48: Matriz de Confusión, Modelo SVM, carrera Electrónica, Experimento 1

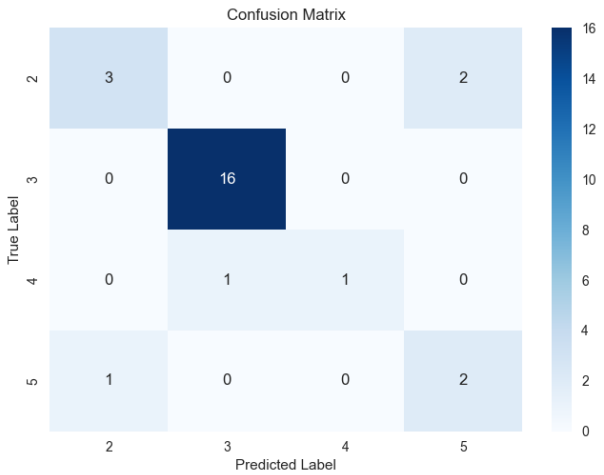


Figura 49: Matriz de Confusión, Modelo SVM, carrera Civil, Experimento 1

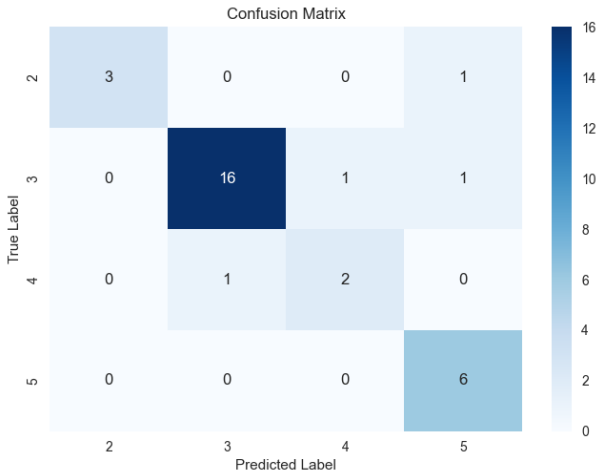


Figura 50: Matriz de Confusión, Modelo KNN, carrera Informática, Experimento 1

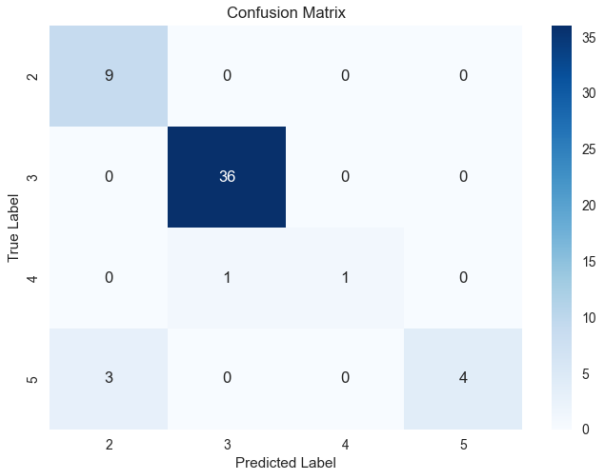


Figura 51: Matriz de Confusión, Modelo KNN, carrera Electricidad, Experimento 1

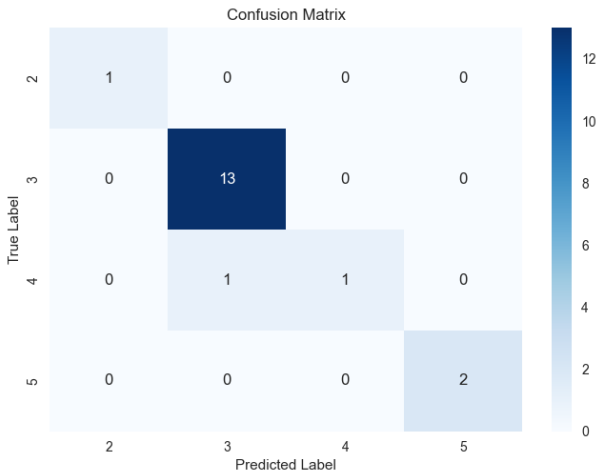


Figura 52: Matriz de Confusión, Modelo KNN, carrera Electrónica, Experimento 1

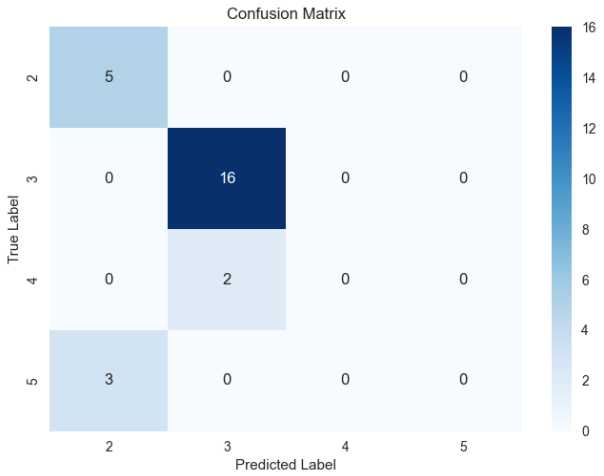


Figura 53: Matriz de Confusión, Modelo KNN, carrera Civil, Experimento 1

## B.2. Segundo Experimento, Entrenamiento y Prueba con datos de hasta Tercer curso

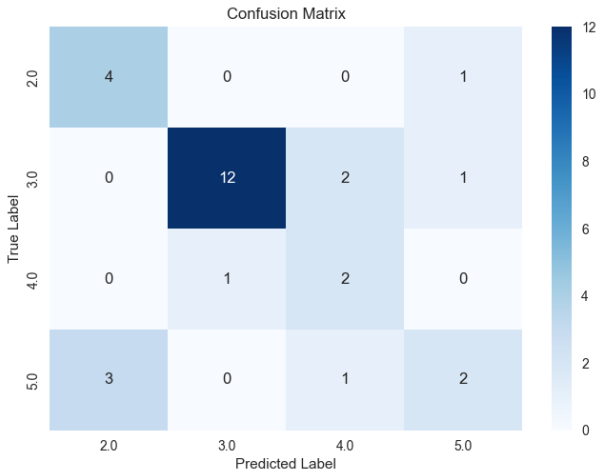


Figura 54: Matriz de Confusión, Modelo RL, carrera Informática, Experimento 2

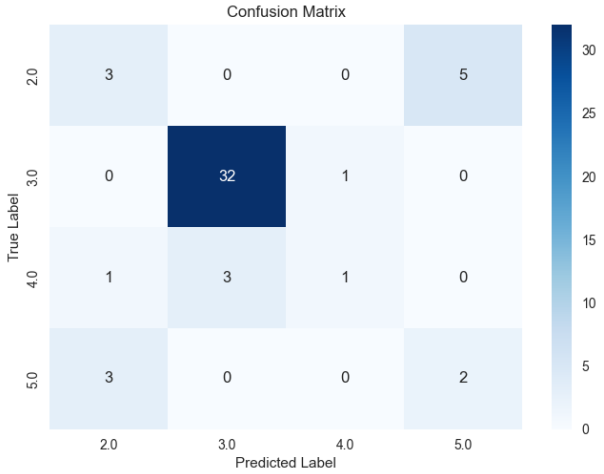


Figura 55: Matriz de Confusión, Modelo RL, carrera Electricidad, Experimento 2

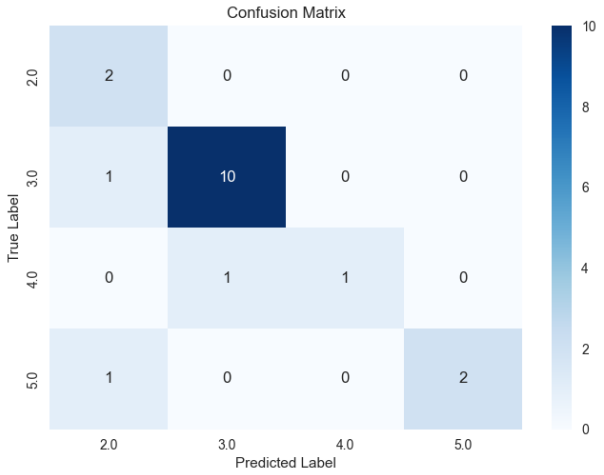


Figura 56: Matriz de Confusión, Modelo RL, carrera Electrónica, Experimento 2

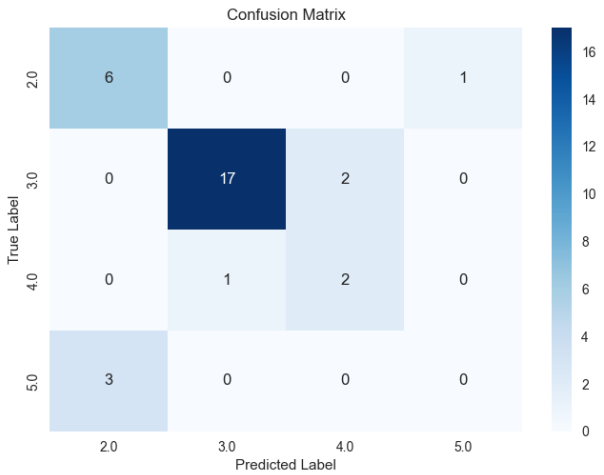


Figura 57: Matriz de Confusión, Modelo RL, carrera Civil, Experimento 2

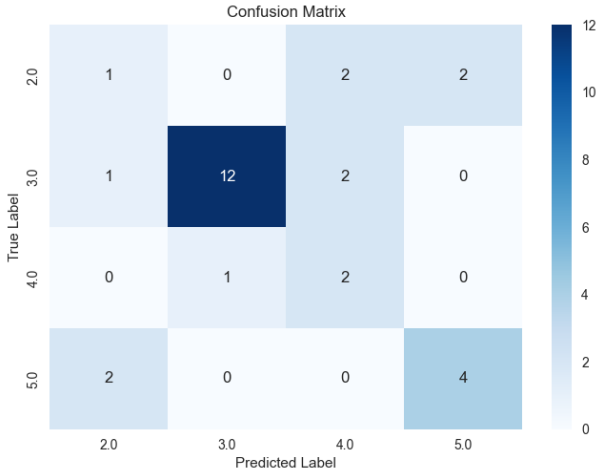


Figura 58: Matriz de Confusión, Modelo DT, carrera Informática, Experimento 2

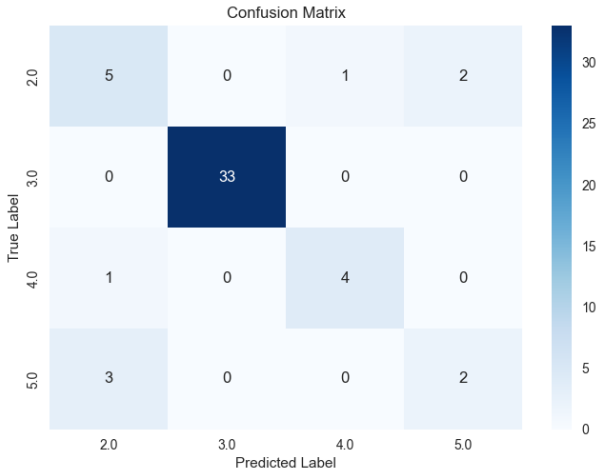


Figura 59: Matriz de Confusión, Modelo DT , carrera Electricidad, Experimento 2

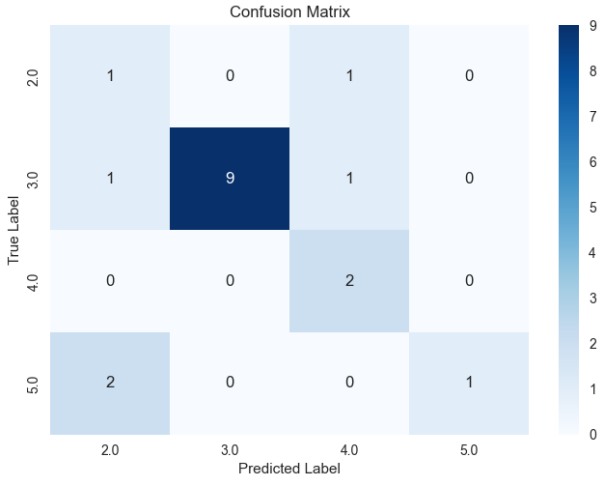


Figura 60: Matriz de Confusión, Modelo DT, carrera Electrónica, Experimento 2

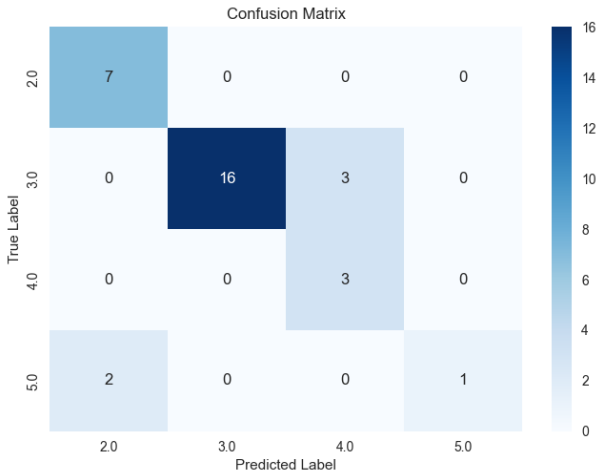


Figura 61: Matriz de Confusión, Modelo DT, carrera Civil, Experimento 2

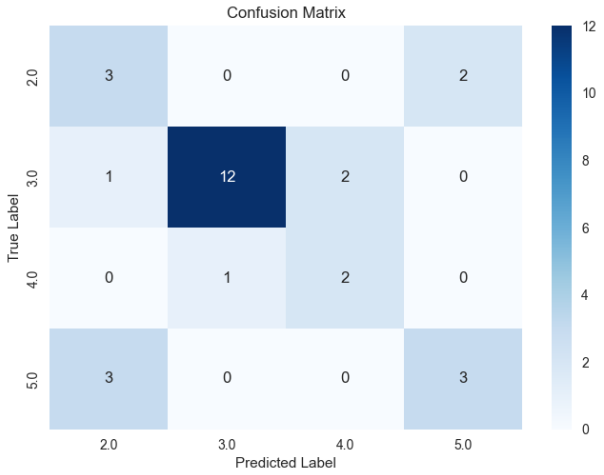


Figura 62: Matriz de Confusión, Modelo RF, carrera Informática, Experimento 2

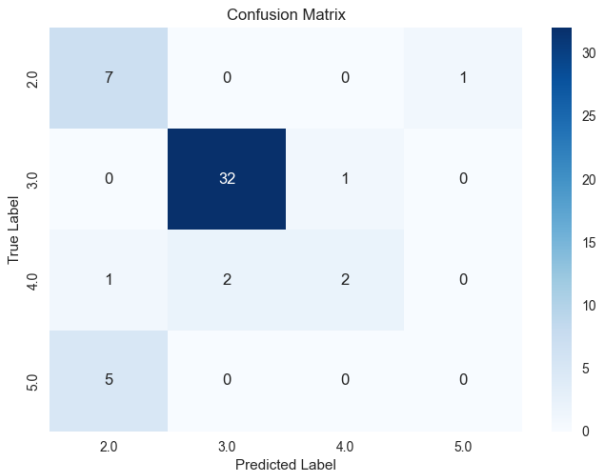


Figura 63: Matriz de Confusión, Modelo RF , carrera Electricidad, Experimento 2

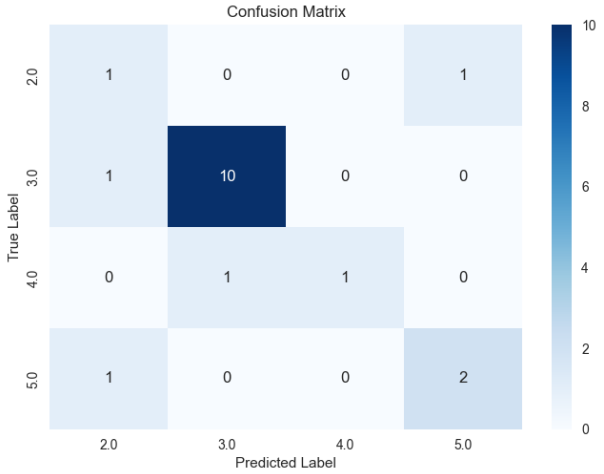


Figura 64: Matriz de Confusión, Modelo RF, carrera Electrónica, Experimento 2

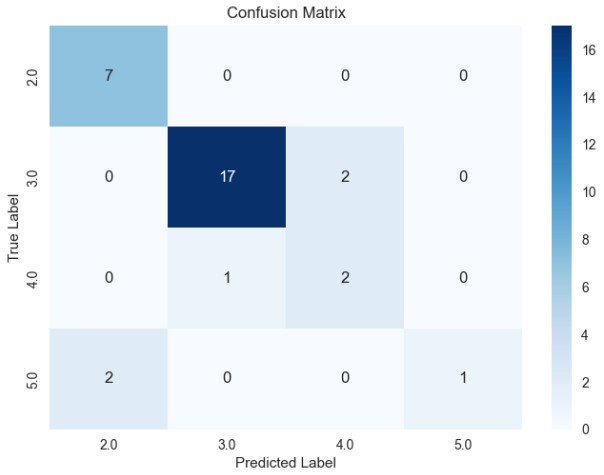


Figura 65: Matriz de Confusión, Modelo RF, carrera Civil, Experimento 2

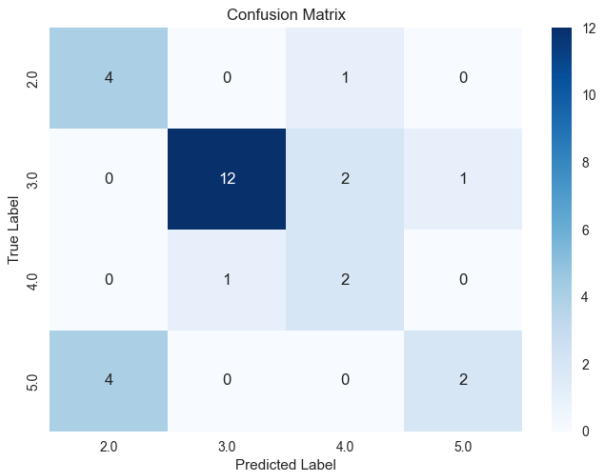


Figura 66: Matriz de Confusión, Modelo SVM, carrera Informática, Experimento 2

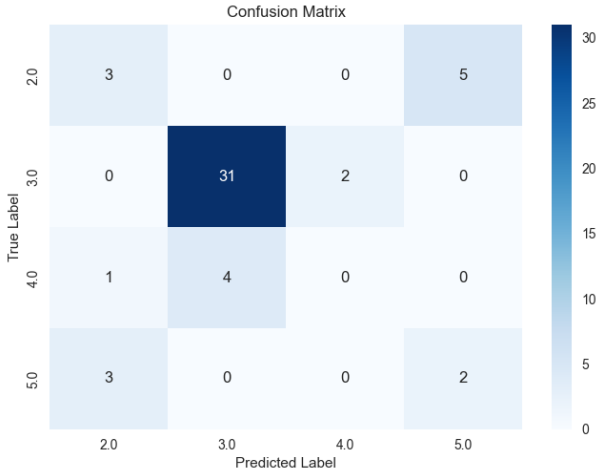


Figura 67: Matriz de Confusión, Modelo SVM, carrera Electricidad, Experimento 2

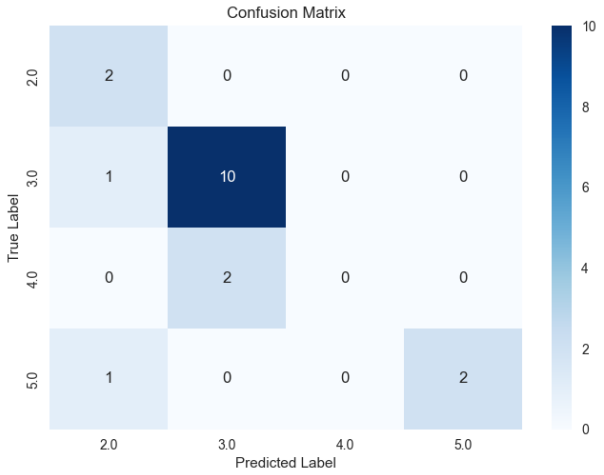


Figura 68: Matriz de Confusión, Modelo SVM, carrera Electrónica, Experimento 2

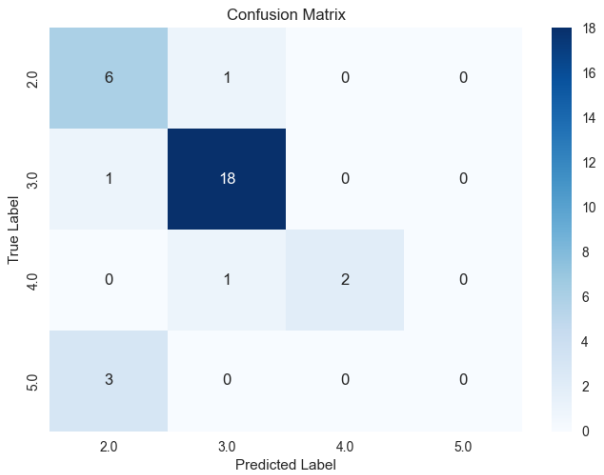


Figura 69: Matriz de Confusión, Modelo SVM, carrera Civil, Experimento 2

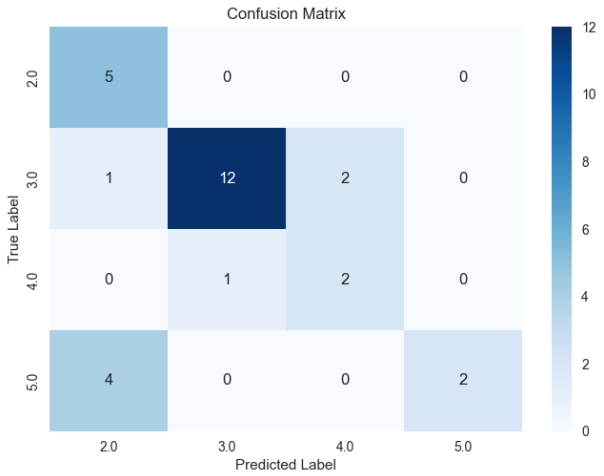


Figura 70: Matriz de Confusión, Modelo KNN, carrera Informática, Experimento 2

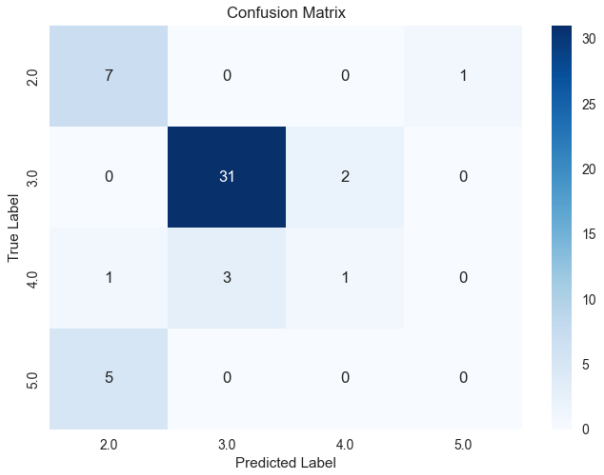


Figura 71: Matriz de Confusión, Modelo KNN, carrera Electricidad, Experimento 2

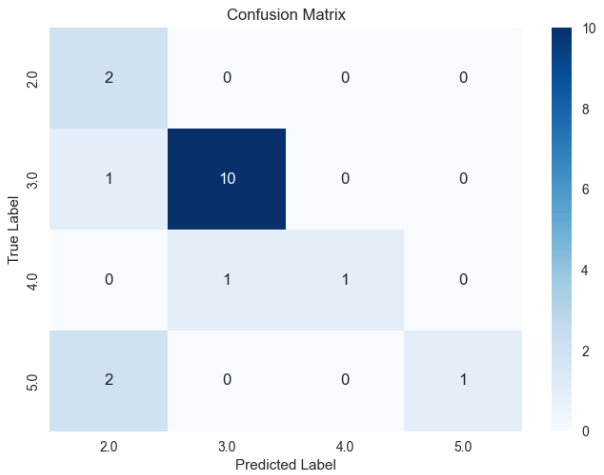


Figura 72: Matriz de Confusión, Modelo KNN, carrera Electrónica, Experimento 2

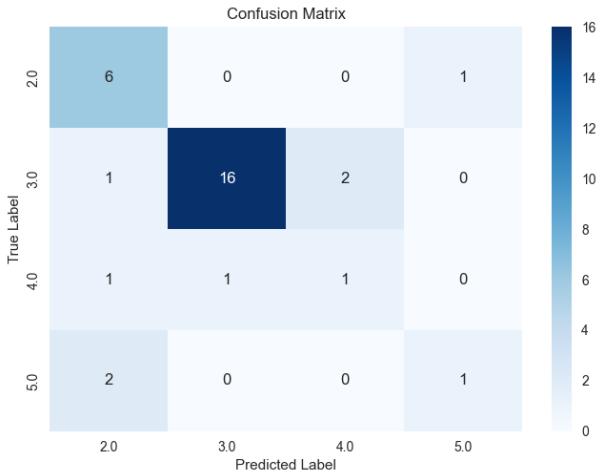


Figura 73: Matriz de Confusión, Modelo KNN, carrera Civil, Experimento 2

## B.3. Tercer Experimento, Entrenamiento y Prueba con datos de hasta Tercer curso, mezclando estado 2 y 5

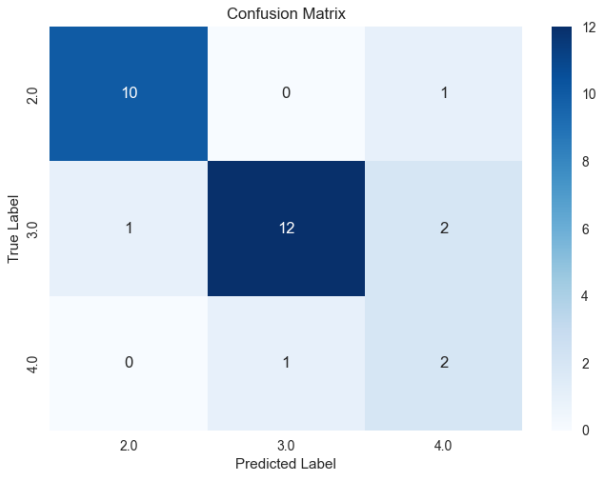


Figura 74: Matriz de Confusión, Modelo RL, carrera Informática, Experimento 3

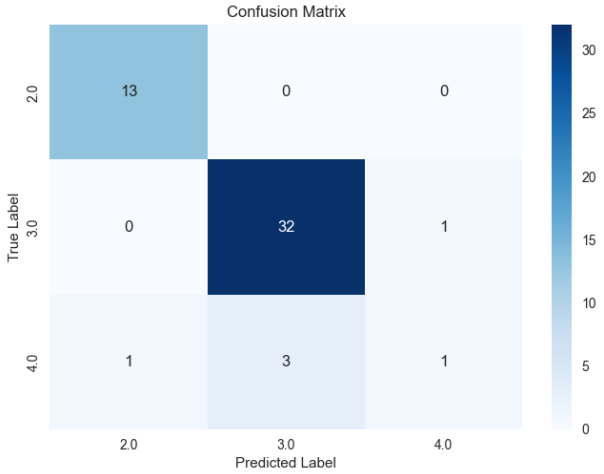


Figura 75: Matriz de Confusión, Modelo RL, carrera Electricidad, Experimento 3

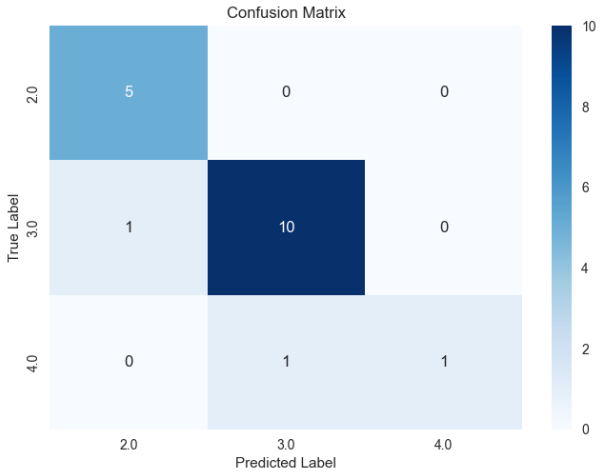


Figura 76: Matriz de Confusión, Modelo RL, carrera Electrónica, Experimento 3

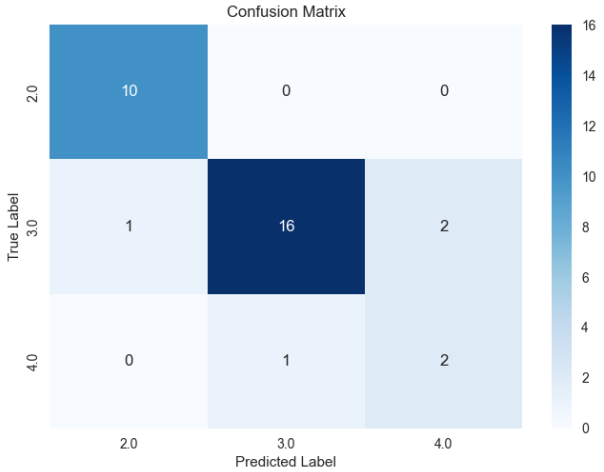


Figura 77: Matriz de Confusión, Modelo RL, carrera Civil, Experimento 3

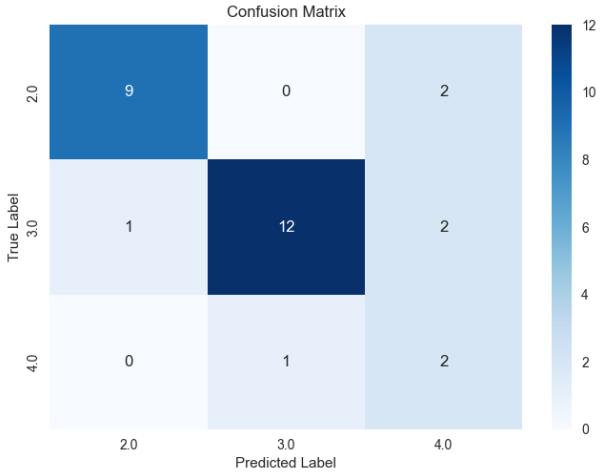


Figura 78: Matriz de Confusión, Modelo DT, carrera Informática, Experimento 3

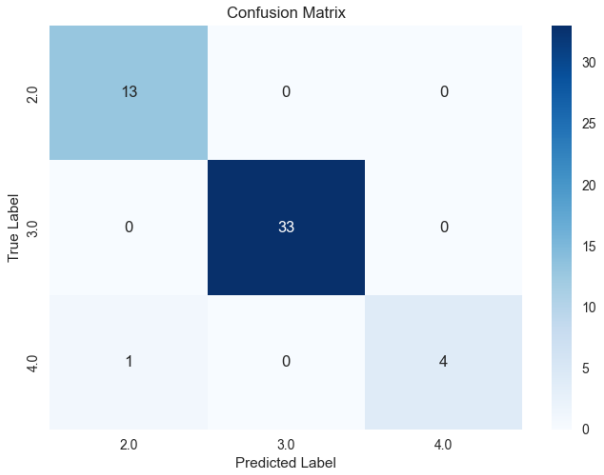


Figura 79: Matriz de Confusión, Modelo DT , carrera Electricidad, Experimento 3

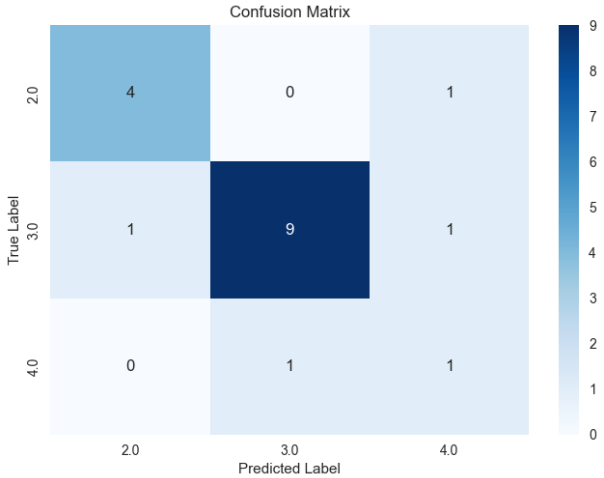


Figura 80: Matriz de Confusión, Modelo DT, carrera Electrónica, Experimento 3

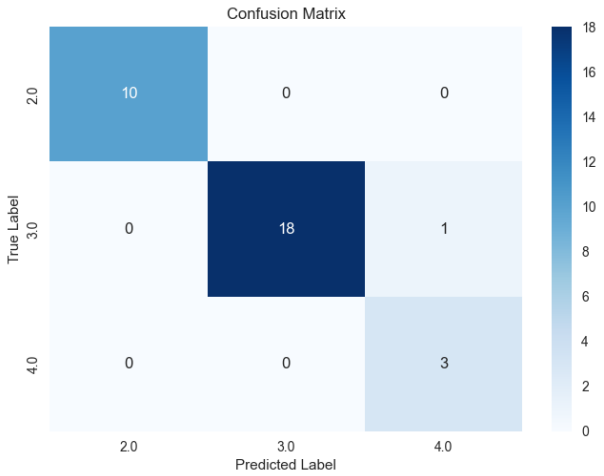


Figura 81: Matriz de Confusión, Modelo DT, carrera Civil, Experimento 3

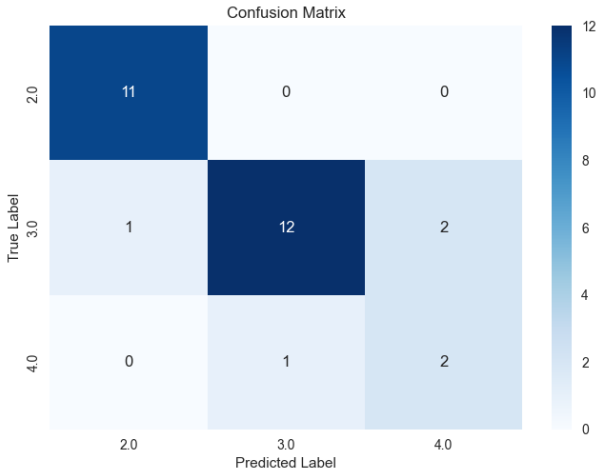


Figura 82: Matriz de Confusión, Modelo RF, carrera Informática, Experimento 3

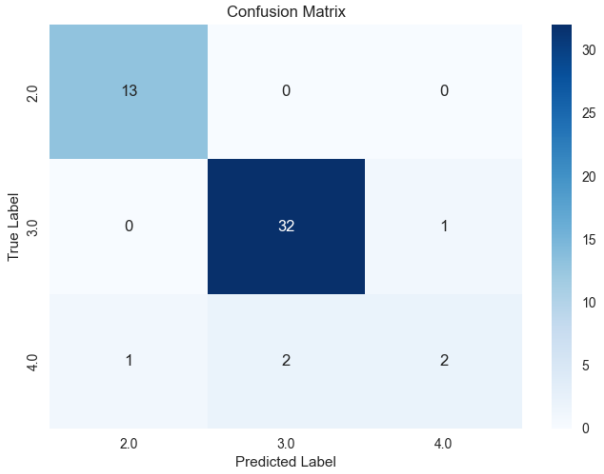


Figura 83: Matriz de Confusión, Modelo RF , carrera Electricidad, Experimento 3

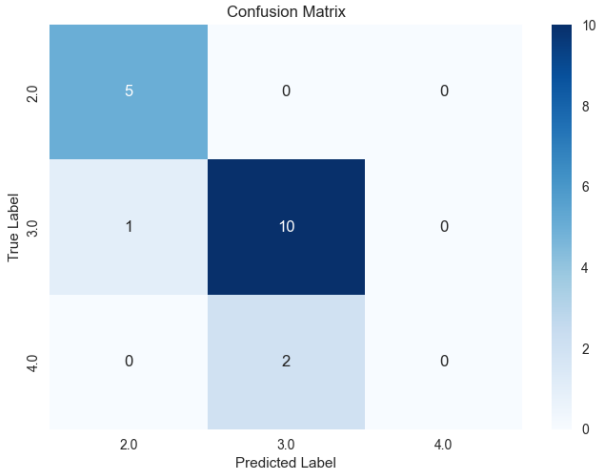


Figura 84: Matriz de Confusión, Modelo RF, carrera Electrónica, Experimento 3

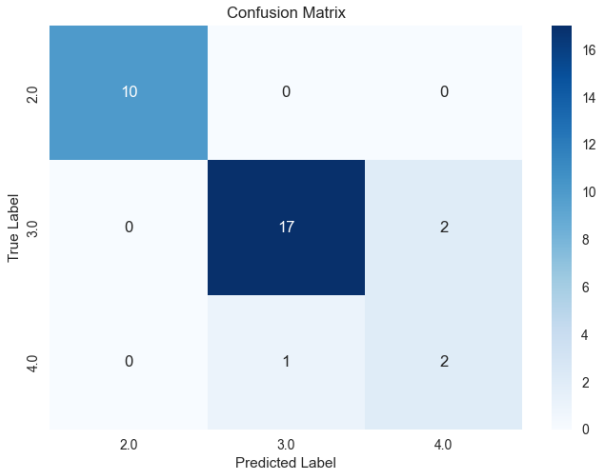


Figura 85: Matriz de Confusión, Modelo RF, carrera Civil, Experimento 3

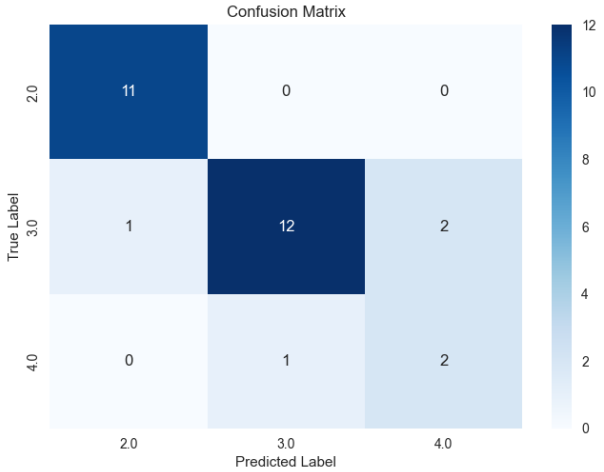


Figura 86: Matriz de Confusión, Modelo SVM, carrera Informática, Experimento 3

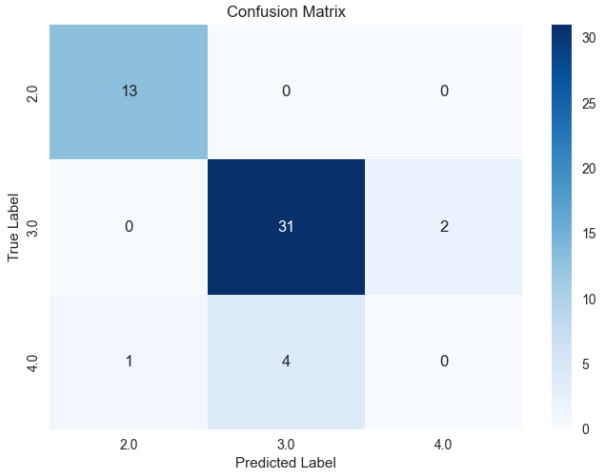


Figura 87: Matriz de Confusión, Modelo SVM, carrera Electricidad, Experimento 3

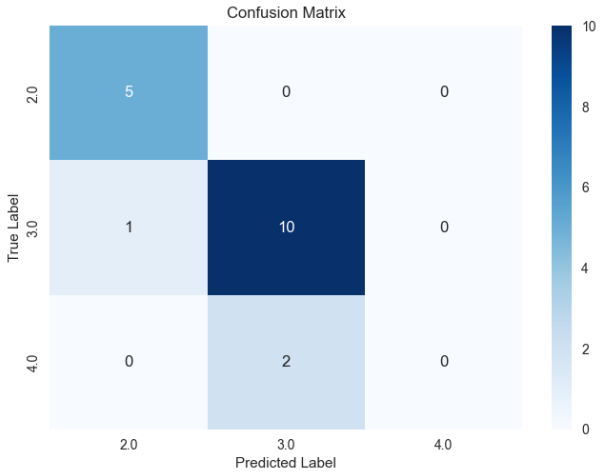


Figura 88: Matriz de Confusión, Modelo SVM, carrera Electrónica, Experimento 3

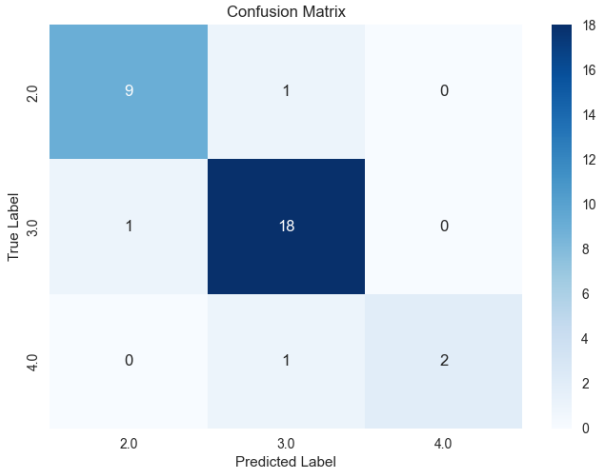


Figura 89: Matriz de Confusión, Modelo SVM, carrera Civil, Experimento 3

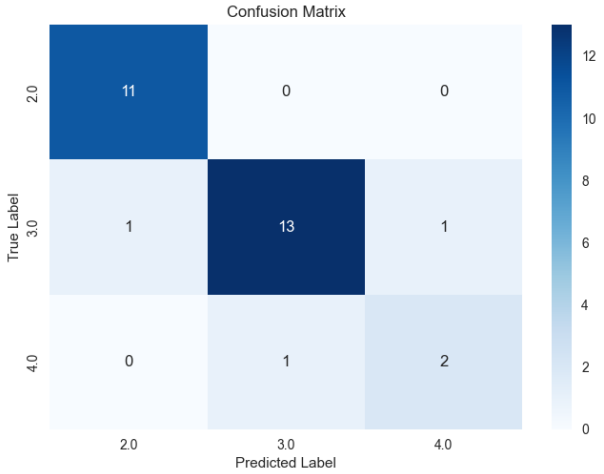


Figura 90: Matriz de Confusión, Modelo KNN, carrera Informática, Experimento 3

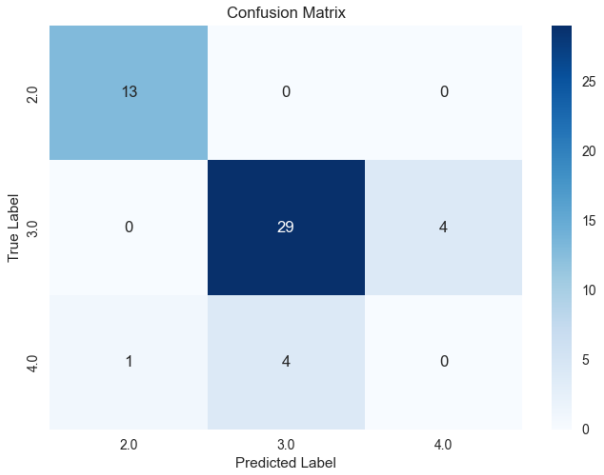


Figura 91: Matriz de Confusión, Modelo KNN, carrera Electricidad, Experimento 3

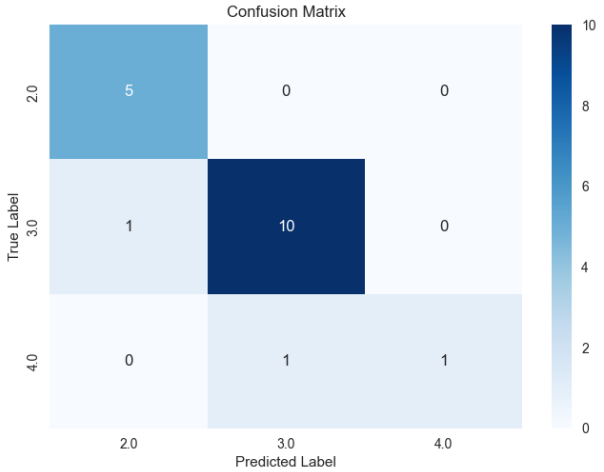


Figura 92: Matriz de Confusión, Modelo KNN, carrera Electrónica, Experimento 3

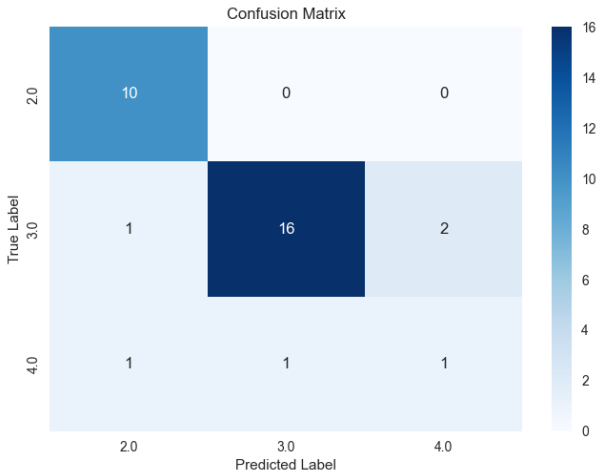


Figura 93: Matriz de Confusión, Modelo KNN, carrera Civil, Experimento 3

## B.4. Cuarto Experimento, Entrenamiento y Prueba con datos de hasta Cuarto curso, separando estado 2 y 5

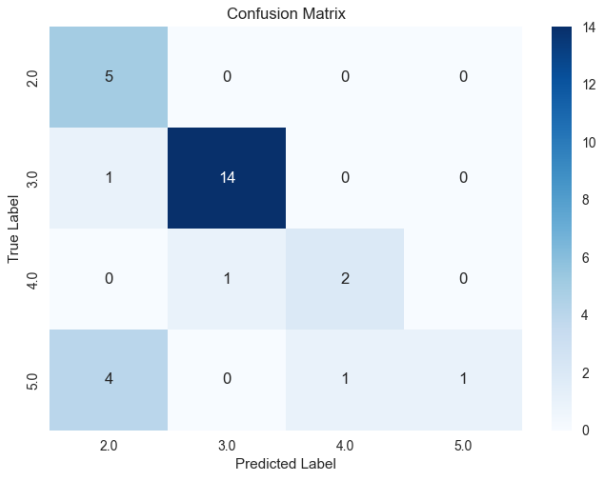


Figura 94: Matriz de Confusión, Modelo RL, carrera Informática, Experimento 4

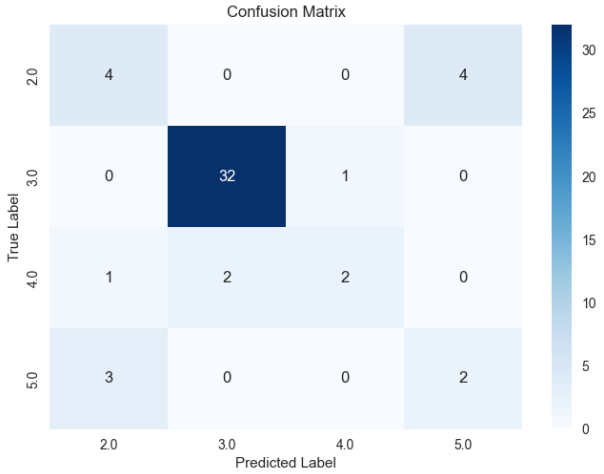


Figura 95: Matriz de Confusión, Modelo RL, carrera Electricidad, Experimento 4

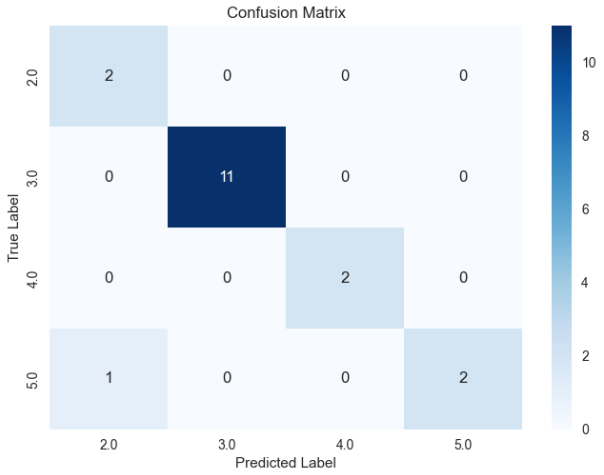


Figura 96: Matriz de Confusión, Modelo RL, carrera Electrónica, Experimento 4

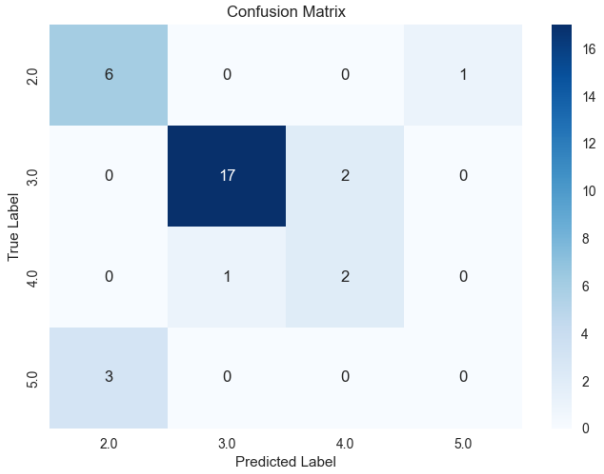


Figura 97: Matriz de Confusión, Modelo RL, carrera Civil, Experimento 4

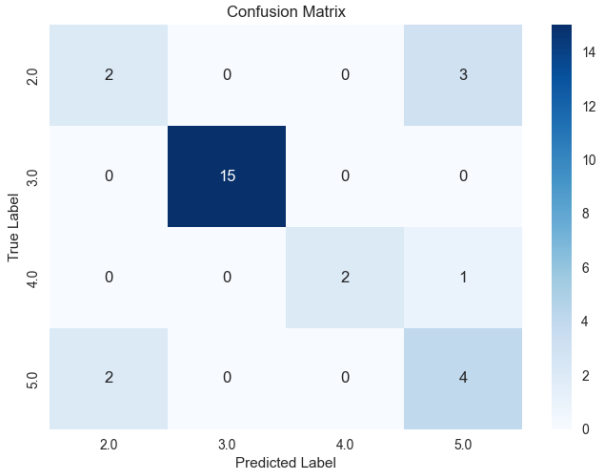


Figura 98: Matriz de Confusión, Modelo DT, carrera Informática, Experimento 4

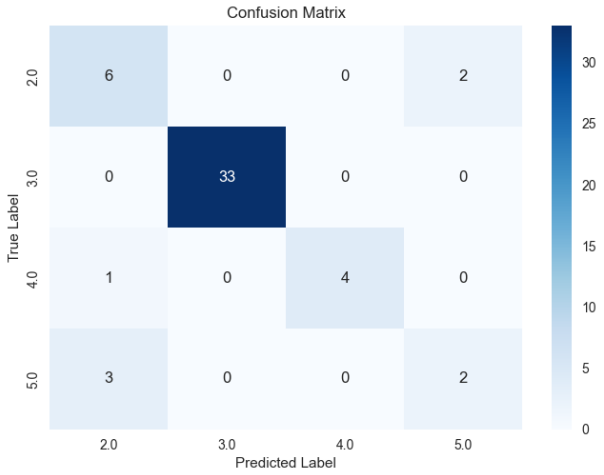


Figura 99: Matriz de Confusión, Modelo DT , carrera Electricidad, Experimento 4

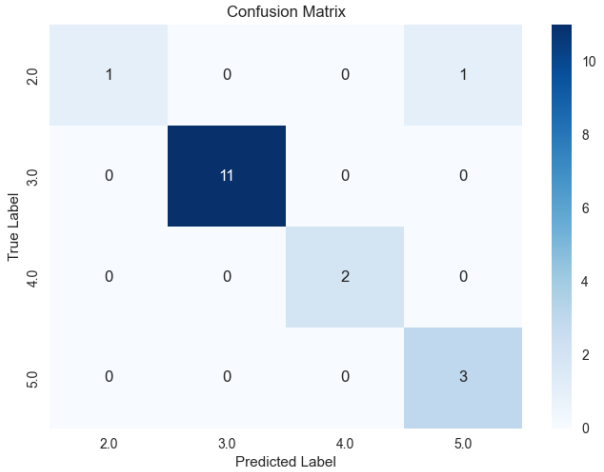


Figura 100: Matriz de Confusión, Modelo DT, carrera Electrónica, Experimento 4

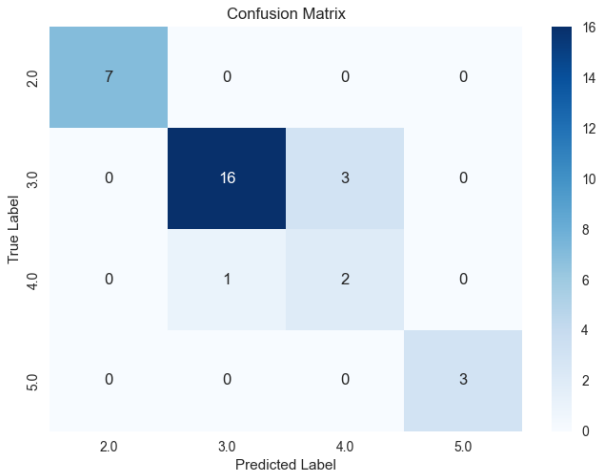


Figura 101: Matriz de Confusión, Modelo DT, carrera Civil, Experimento 4

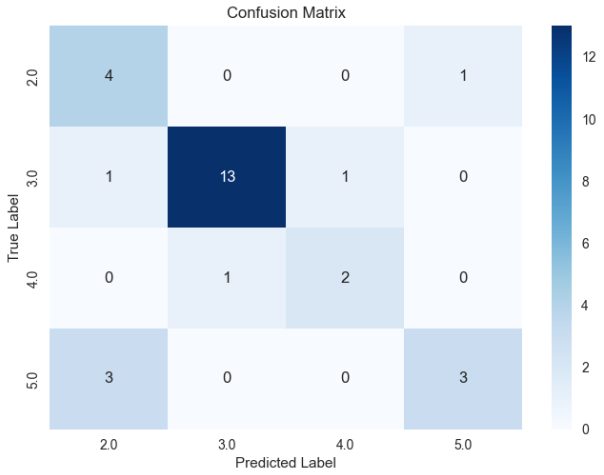


Figura 102: Matriz de Confusión, Modelo RF, carrera Informática, Experimento 4

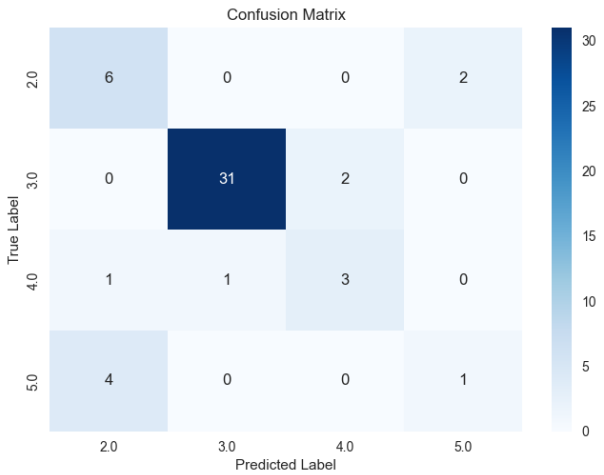


Figura 103: Matriz de Confusión, Modelo RF , carrera Electricidad, Experimento 4

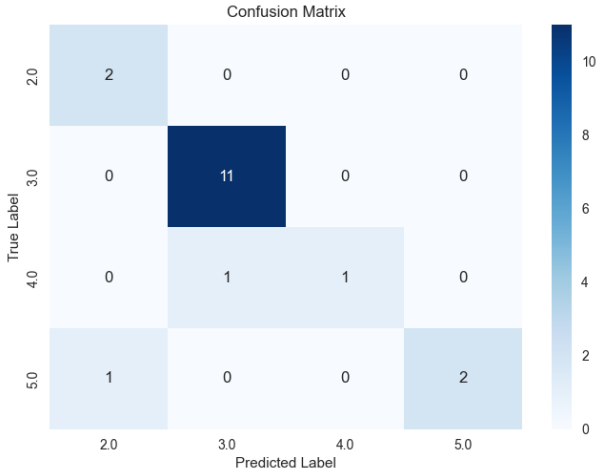


Figura 104: Matriz de Confusión, Modelo RF, carrera Electrónica, Experimento 4

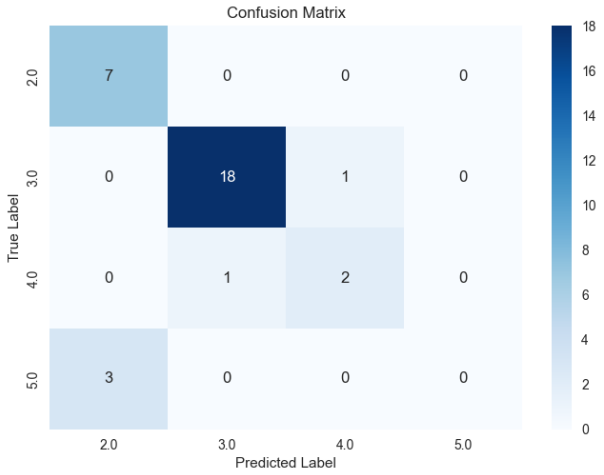


Figura 105: Matriz de Confusión, Modelo RF, carrera Civil, Experimento 4

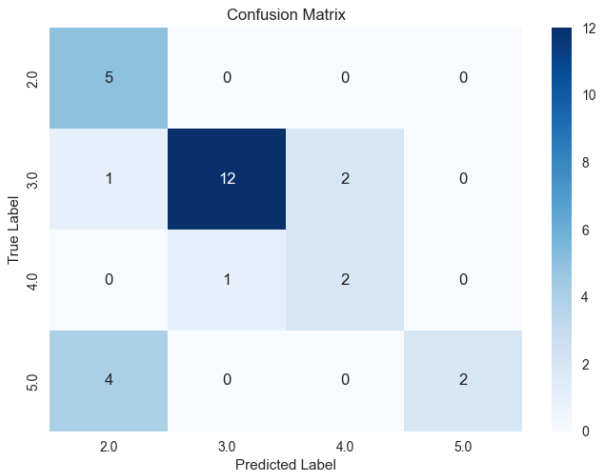


Figura 106: Matriz de Confusión, Modelo SVM, carrera Informática, Experimento 4

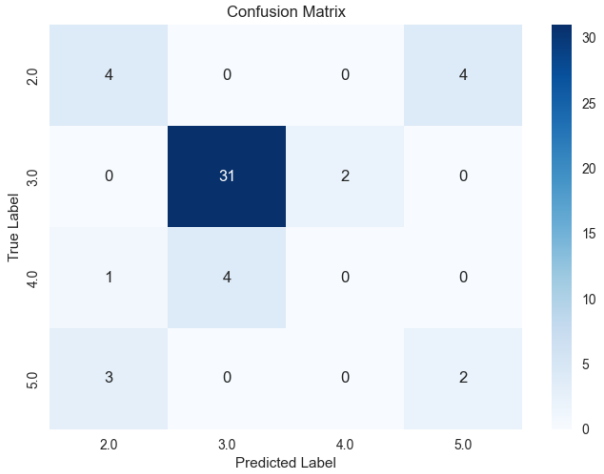


Figura 107: Matriz de Confusión, Modelo SVM, carrera Electricidad, Experimento 4

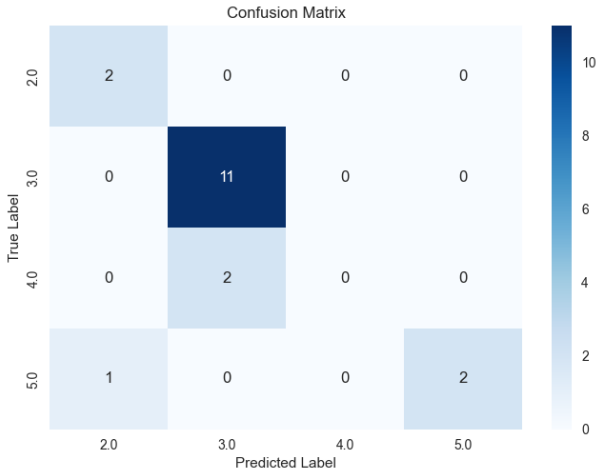


Figura 108: Matriz de Confusión, Modelo SVM, carrera Electrónica, Experimento 4

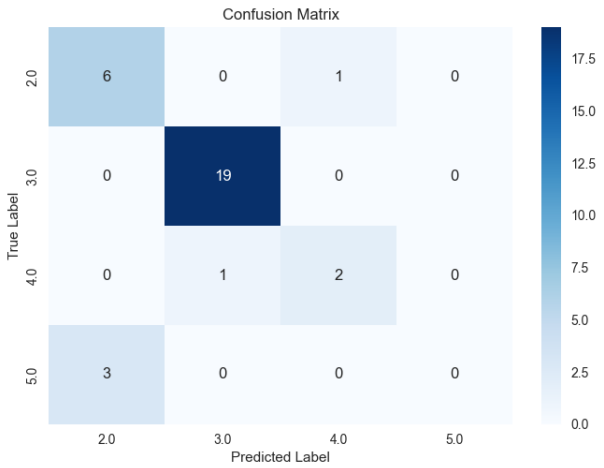


Figura 109: Matriz de Confusión, Modelo SVM, carrera Civil, Experimento 4

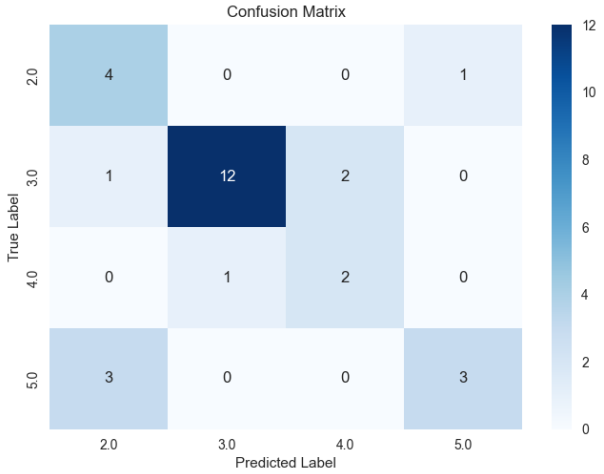


Figura 110: Matriz de Confusión, Modelo KNN, carrera Informática, Experimento 4

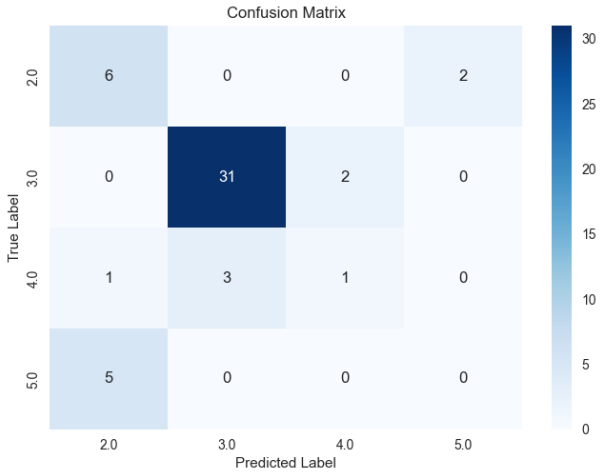


Figura 111: Matriz de Confusión, Modelo KNN, carrera Electricidad, Experimento 4

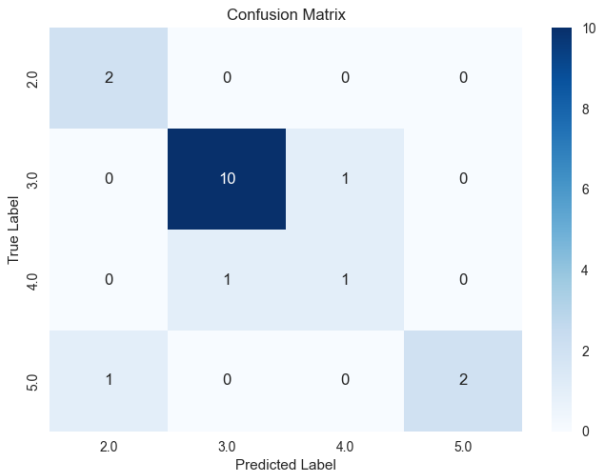


Figura 112: Matriz de Confusión, Modelo KNN, carrera Electrónica, Experimento 4

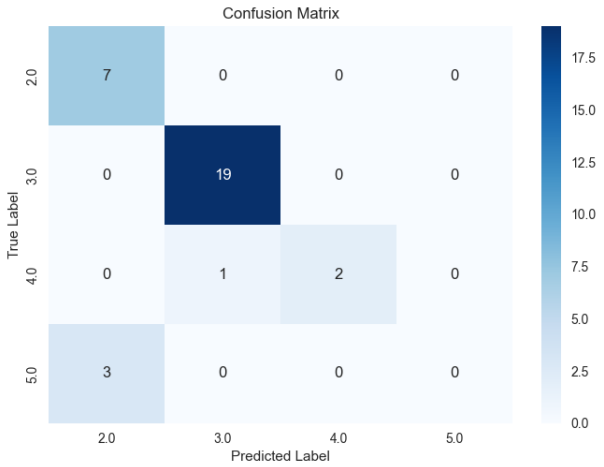


Figura 113: Matriz de Confusión, Modelo KNN, carrera Civil, Experimento 4