UNIVERSIDAD MARIANO GÁLVEZ DE GUATEMALA

Facultad de ingeniería en sistemas

Curso: Inteligencia Artificial

PROYECTO FINAL: CLASIFICADOR DE SENTIMIENTOS DE RESEÑAS DE AMAZON

Estudiante: Rudy Alexander Lorenzana Silva

Carné: 3490-21-4968

Fecha de entrega: 31 de mayo de 2025

1. Objetivo del Proyecto

El objetivo principal de este proyecto es desarrollar un sistema de clasificación automática de sentimientos mediante el uso de técnicas de minería de texto y algoritmos de aprendizaje automático utilizando la plataforma KNIME. Específicamente, el sistema debe ser capaz de predecir si una reseña escrita por un cliente sobre un producto en Amazon refleja una opinión positiva, negativa o neutral. Esto permitirá a las empresas automatizar el análisis de la satisfacción del cliente y mejorar la toma de decisiones basada en las opiniones expresadas en lenguaje natural.

El proyecto no solo tiene fines académicos, sino también aplicaciones reales en el ámbito empresarial. La capacidad de procesar grandes cantidades de texto y extraer valor a partir de sentimientos expresados por los usuarios representa una herramienta poderosa para empresas que dependen de las reseñas de productos, especialmente en plataformas de comercio electrónico.

2. Descripción del Dataset

El dataset utilizado fue un subconjunto del conjunto de datos "Amazon Fine Food Reviews", ampliamente conocido por contener miles de reseñas de productos alimenticios. Este conjunto de datos incluye información como identificadores de productos, perfiles de usuarios, número de votos útiles, puntuaciones otorgadas y el texto libre de la reseña.

Para este proyecto se seleccionaron exclusivamente las siguientes columnas:

- **reviewText**: campo que contiene el texto de la reseña escrita por el cliente.
- **overall**: puntuación numérica entre 1 y 5 asignada por el usuario.

Se optó por convertir la variable **overall** en una variable categórica denominada **Sentiment**, con el siguiente esquema:

- Puntuaciones de 1 o 2: **Negativo**
- Puntuaciones de 4 o 5: **Positivo**
- Puntuación de 3: **Neutral** (omitida para evitar ambigüedades en el entrenamiento)

Este enfoque de etiquetado binario permite simplificar la tarea de clasificación y centrarse en los casos claramente positivos o negativos.

3. Procesos Aplicados

El flujo de trabajo se construyó de forma modular utilizando KNIME Analytics Platform, permitiendo organizar cada fase de forma clara y secuencial.

3.1 Adquisición de datos:

- Se usó el nodo **CSV Reader** para cargar el archivo con las reseñas.
- Con **Table Manipulator** se seleccionaron solo las columnas necesarias para el análisis.

3.2 Preprocesamiento:

- Con el nodo **Rule Engine** se creó una nueva columna llamada *Sentiment* usando reglas lógicas basadas en la puntuación.
- Se aplicó **Row Filter** para eliminar registros con puntuación 3 (neutral).
- El nodo **Missing Value** eliminó reseñas con campos vacíos.
- **Strings to Document** transformó los textos a tipo Document, lo cual es necesario para el procesamiento de lenguaje natural.

3.3 Procesamiento de texto:

- **Punctuation Erasure** eliminó signos de puntuación innecesarios.
- **Case Converter** transformó todas las letras a minúsculas para uniformizar los datos.
- **Stop Word Filter** eliminó palabras comunes sin valor semántico como "el", "la", "de".
- **Number Filter** removió dígitos y números.
- Se aplicó la combinación de **TF** (Term Frequency) e **IDF** (Inverse Document Frequency) para generar representaciones numéricas del texto.

3.4 Entrenamiento de modelo:

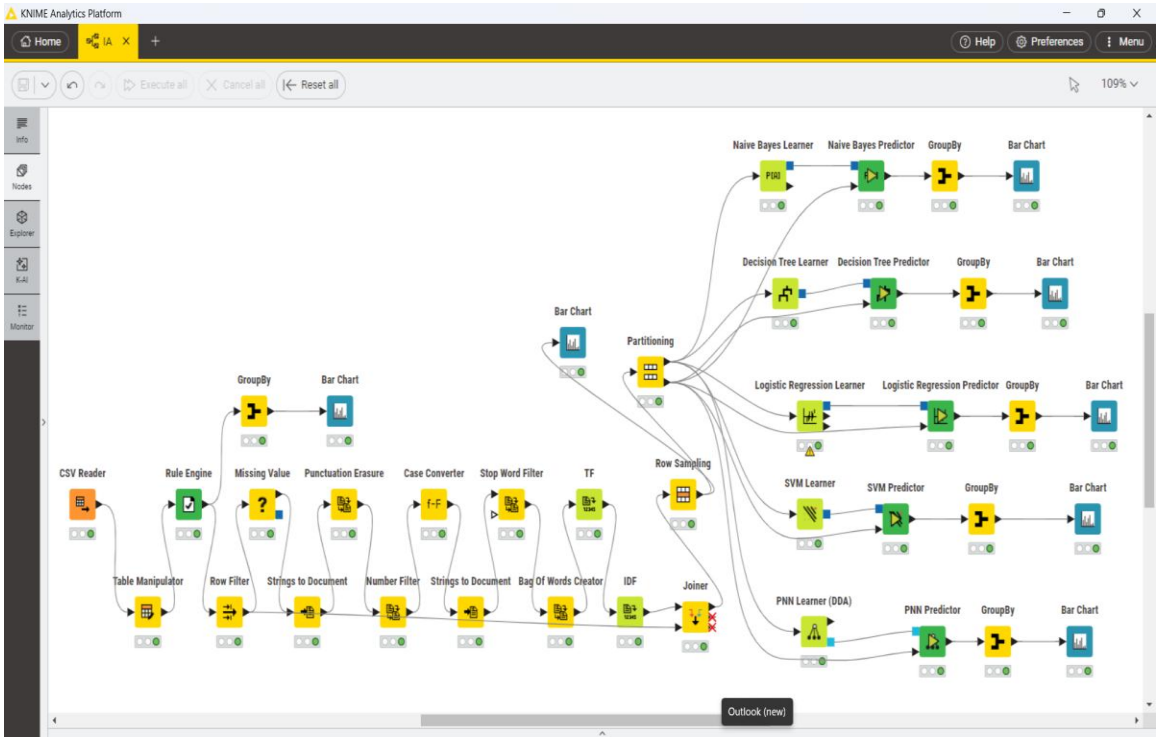
- Se usó el nodo **Partitioning** para separar los datos en 80% para entrenamiento y 20% para prueba.

3.5 Evaluación:

- **Naive Bayes Learner y Predictor**: modelo probabilístico simple pero efectivo.
- **Decision Tree Learner y Predictor**: modelo jerárquico basado en reglas.
- **Logistic Regression Learner y Predictor**: modelo lineal para clasificación.
- **PNN Learner y Predictor**: modelo basado en redes neuronales probabilísticas.
- **SVM Learner y Predictor**: modelo con márgenes de clasificación.

3.6 Visualización:

- Nodos **GroupBy** y **Bar Chart** se utilizaron para visualizar la distribución de sentimientos reales y predichos por cada modelo.
- Un nodo adicional de **Bar Chart** conectado al nodo inicial muestra la distribución de sentimientos reales desde el principio.



4. Modelo y Métrica Principal

La evaluación se basó principalmente en la métrica F1, que combina precisión y recall. Esta métrica es especialmente relevante en casos de desbalance de clases, como ocurre en datasets de reseñas.

Resumen de métricas por modelo:

- **Naive Bayes**: Precisión ~78%, F1 ~78%
- **Árbol de Decisión**: Precisión ~82.85%, F1 ~83%
- **Regresión Logística**: Precisión ~80%, F1 ~80%
- **Redes Neuronales (PNN)**: Precisión ~82%, F1 ~82%
- **SVM**: Precisión ~81%, F1 ~81%

El modelo de mejor desempeño fue el Árbol de Decisión, no solo por su buena precisión sino también por su interpretabilidad y velocidad de ejecución.

5. Conclusión

Este proyecto demuestra que es posible implementar un sistema eficaz de clasificación de sentimientos a partir de texto utilizando KNIME sin necesidad de programación avanzada. La estructura visual de flujos de trabajo permite comprender cada etapa del proceso y facilita futuras modificaciones.

El uso de preprocesamiento riguroso y transformaciones de texto adecuadas fue clave para mejorar el rendimiento de los modelos. Además, se comprobó que modelos como el Árbol de Decisión y PNN ofrecen un equilibrio entre precisión y tiempo de entrenamiento.

Finalmente, este tipo de soluciones puede ser aplicado en escenarios reales, como en empresas que desean automatizar el análisis de comentarios en redes sociales, formularios de opinión o reseñas de productos, apoyando la inteligencia comercial con tecnología basada en IA.