Lead University

Administración de datos

Trabajo en grupo 1

Carolina Salas y Kristhel Porras

**1. Introducción**

El conjunto de datos Seeds proporciona un problema de clasificación estructurado y bien definido, lo que lo convierte en un candidato ideal para probar el rendimiento de algoritmos de aprendizaje automático. En este estudio, proponemos el uso del algoritmo eXtreme Gradient Boosting (XGBoost) para clasificar diferentes variedades de semillas de trigo en función de sus propiedades geométricas. XGBoost es ampliamente reconocido por su eficiencia y precisión en el manejo de datos estructurados, lo que lo convierte en una excelente elección para esta tarea.

**2. Objetivo**

El objetivo principal de esta propuesta es desarrollar un modelo de clasificación con XGBoost que distinga con precisión entre las tres variedades de trigo en función de sus atributos numéricos. Se espera obtener un modelo robusto que logre una alta precisión mientras mantiene la eficiencia computacional.

**3. Descripción del Conjunto de Datos**

El conjunto de datos Seeds consta de 210 instancias con 7 atributos numéricos que representan diversas propiedades geométricas de los granos de trigo. Las características incluidas son:

* Área
* Perímetro
* Compacidad
* Longitud del núcleo
* Ancho del núcleo
* Coeficiente de asimetría
* Longitud del surco del núcleo

Cada instancia está clasificada en una de las tres variedades de trigo:

1. Kama
2. Rosa
3. Canadian

**4. Metodología**

El proyecto seguirá los siguientes pasos clave:

**4.1 Desarrollo de un Data Pipeline Completo**

* **Selección y Extracción de Datos**
* **Almacenamiento y Protección de Datos**
* **Limpieza y Transformación de Datos**

**4.2 Desarrollo del Modelo de Inteligencia Artificial**

* Implementar XGBoost utilizando XGBClassifier de la biblioteca XGBoost.
* Optimizar hiperparámetros (tasa de aprendizaje, profundidad máxima, número de estimadores) mediante validación cruzada.
* Evaluar la importancia de las características para comprender qué atributos contribuyen más a la clasificación.
* **Nota:** No se utilizarán regresiones lineales o regresión logística en este estudio.

**4.3 Evaluación del Modelo y Métricas**

* Evaluar el rendimiento del modelo utilizando:
  + **Precisión (Accuracy):** Corrección general de las predicciones.
  + **Matriz de Confusión:** Análisis de errores de clasificación.
  + **Distribución de las Clases:** Comprobar equilibrio en las predicciones.
  + **Importancia de las Variables:** Determinar cuáles tienen mayor influencia en la clasificación.
  + **Impacto de la Normalización o Estandarización:** Comparar modelos con y sin preprocesamiento de escalado.

**4.4 Implementación en una Aplicación Web**

* Desarrollar una aplicación web interactiva utilizando **Streamlit** (o una alternativa equivalente).
* Funcionalidades de la aplicación:
  + **Visualización del Data Pipeline** con los procesos de extracción, limpieza y transformación de datos.
  + **Presentación de los resultados del modelo de IA**, incluyendo predicciones, métricas y gráficos de desempeño.
  + **Interactividad para el usuario**, permitiendo carga de datos y exploración de predicciones en tiempo real.

**5. Bibliotecas de Python Requeridas**

La implementación de este estudio utilizará las siguientes bibliotecas de Python:

* xgboost: Para entrenar y ajustar el modelo XGBoost.
* numpy: Para cálculos numéricos.
* pandas: Para manipulación y gestión del conjunto de datos.
* sklearn: Para preprocesamiento de datos, división en conjuntos de entrenamiento/prueba y métricas de evaluación.
* matplotlib/seaborn: Para visualización de datos y análisis de importancia de características.
* streamlit: Para la implementación de la aplicación web.
* cryptography o pycryptodome: Para encriptación y seguridad de los datos.

**6. Preguntas de Investigación**

A través de este estudio, buscamos responder las siguientes preguntas que pueden resolverse mediante un pipeline de procesamiento y análisis exploratorio de datos (EDA):

1. ¿Cómo afecta la normalización o estandarización de las características al rendimiento del modelo?
2. ¿Cuáles son las características más importantes para predecir la variedad de trigo?
3. ¿Existe un patrón de correlación fuerte entre algunas características del conjunto de datos?
4. ¿Qué impacto tiene la cantidad de árboles (n\_estimators) en XGBoost sobre la precisión del modelo?
5. ¿Cómo se comparan los tiempos de procesamiento y rendimiento del modelo con diferentes configuraciones de hiperparámetros?