UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS ECONÓMICO ADMINISTRATIVAS

MASTER IN DATA SCIENCE

Programación II
Reto 3 – Aplicación de ML a Proyecto de Investigación

José de Jesús Nicolás Zúñiga González
224807959
jose.zuniga0795@alumnos.udg.mx

CUCEA
El mejor lugar para el talento

MCD | MAESTRÍA EN CIENCIA DE LOS DATOS

1. Introducción

En el contexto empresarial actual, la asignación de proyectos a proveedores adecuados es un desafío crítico que puede impactar directamente en los resultados de una organización. La incorporación de técnicas de inteligencia artificial (IA) y aprendizaje automático (AA) ofrece nuevas oportunidades para mejorar esta tarea, permitiendo tomar decisiones más informadas, eficientes y precisas. Las empresas de manufactura electrónica (EMS – *Electronic Manufacturer Services*) no son ajenas a este tipo de problemáticas.

La selección de proveedores implica múltiples variables, tales como:

- Costos y presupuestos disponibles.
- Calidad de los servicios o productos proporcionados.
- Plazos de entrega.
- Experiencia previa con proyectos similares.
- Reputación y confiabilidad del proveedor.

Balancear estos factores manualmente puede ser una tarea compleja y propensa a errores, especialmente en proyectos de gran escala o con múltiples opciones de proveedores.

2. Planteamiento del Problema

El departamento de compras de una de las EMS con mayor presencia global está buscando desarrollar un sistema que le permita asignar los proyectos de automatización a proveedores de tal forma que la probabilidad de éxito del proyecto sea mayor si la asignación se hace de manera manual.

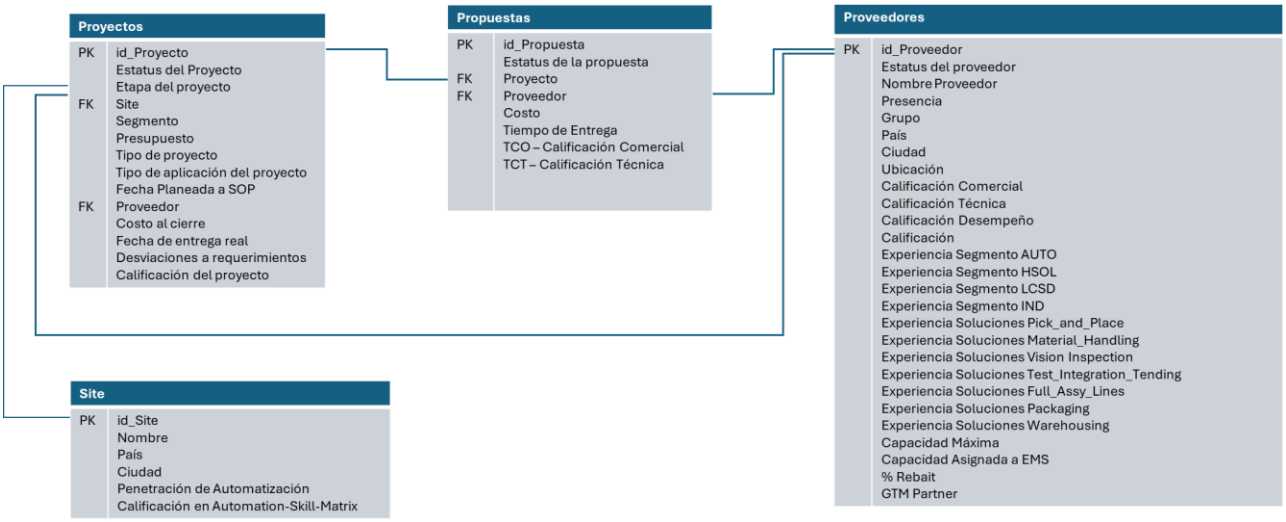
La inteligencia artificial (IA) y el aprendizaje automático (AA) permiten analizar grandes cantidades de datos de manera sistemática y en tiempo real, ofreciendo varias ventajas clave:

- Automatización del proceso: Los algoritmos pueden analizar las características de los proyectos y proveedores, asignándolos de manera eficiente sin intervención humana.
- Modelos predictivos: El aprendizaje automático puede prever el desempeño futuro de un proveedor basándose en datos históricos.

- Personalización: Los sistemas pueden ajustar recomendaciones en función de las necesidades específicas de cada proyecto.
- Optimización: Algoritmos avanzados pueden identificar la combinación más eficiente y rentable de proveedores para un conjunto de proyectos.

3. Propuesta de Solución

3.1. Estructura de la base de datos



Para este proyecto se estará trabajando con 4 tablas diferentes. La primera de ellas **Proyectos** contendrá el listado de todos los proyectos de automatización que están planeados para el periodo fiscal. Su estructura es:

Id_Proyecto	Identificador único para el proyecto de automatización.
Estatus del Proyecto	Variable categórica para determinar el estatus del proyecto: 0 – No ha comenzado 1 – Activo 2 – Terminado
Etapa del Proyecto	Variable categórica para identificar la fase del proyecto: 0 – Proyecto en planeación 1 – Proyecto en licitación 2 – Proyecto en desarrollo 3 – FAT (Factory Acceptance Test) 4 – SAT (Site Acceptance Test) 5 – Proyecto en Implementación 6 – SOP 7 – Proyecto cerrado
Site	Planta en la que se instalará el proyecto
Segmento	Segmento al que pertenece el cliente del proyecto: 0 – Automotriz (AUTO) 1 – Médico (HSOL) 2 – LifeStyle (LSCD)

	3 – Industrial (IND) 4 – Comunicaciones (CEC)
Presupuesto	Capital estimado para invertir en el proyecto.
Tipo de Proyecto	0 – Estándar 1 – Personalizado
Tipo de Aplicación del Proyecto	0 – Manejo de material 1 – Traslado de material 2 – Visión 3 – Pruebas 4 – Línea de ensamble 5 – Empaque 6 – Almacén
Fecha Planeada a SOP	Fecha de cuando tiene que estar la máquina liberada a producción
Proveedor	Id del proveedor seleccionado tras la licitación
Costo al Cierre	Monto final del proyecto
Fecha de entrega real	Fecha en la que se liberó realmente el proyecto
Desviaciones a requerimientos	Cantidad de desviaciones hechas con respecto los requerimientos definidos en el SOW
Calificación del Proyecto	Valor numérico obtenido de la encuesta de cierre del proyecto.

La segunda tabla es **Site**. Esta tabla contiene el listado de todas las plantas que tiene la compañía globalmente.

Id_Site	Identificador de la planta
Nombre	Nombre de la planta
País	País en el que está la planta
Ciudad	Ciudad en la que se ubica la planta
Penetración de Automatización	Índice que indica que tanta experiencia tiene el site con respecto a proyectos de automatización.
Calificación en Automation Skill Matrix	Calificación que representa que tan preparados está la plantilla de trabajadores en temas de automatización.

La tercera tabla, **Propuestas**, enlistará las diferentes propuestas hechas para los proyectos licitados.

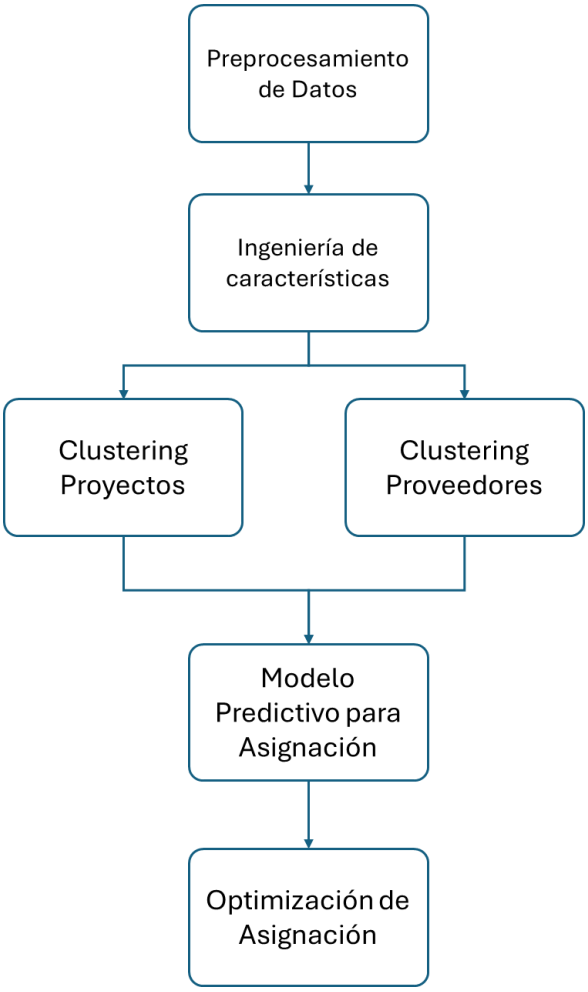
Id_Propuesta	Identificador único por propuesta recibida por proveedor
Estatus de la propuesta	0 – Sin iniciar 1 – En proceso 2 – Entregada 4 – Descartada 5 – Seleccionada
Proyecto	Id del proyecto
Proveedor	Id del proveedor que entregó la propuesta
Costo	Monto por el cual se cotizó el proyecto
Tiempo de Entrega	Cantidad de días para entrega del proyecto
TCO	Calificación comercial de la propuesta.
TCT	Calificación técnica de la propuesta

La última tabla es la que contiene el listado de **Proveedores**.

Id_Proveedor	Identificador único del proveedor
Estatus del proveedor	0 – Proveedor bloqueado 1 – Proveedor aprobado
Nombre del proveedor	Nombre del proveedor

Presencia	0 – Local 1 – Global
Grupo	Si el proveedor pertenece a un grupo comercial
País	País en el que se encuentra el proveedor
Ciudad	Ciudad en la que está el proveedor
Ubicación	Ubicación del proveedor (dirección)
Calificación comercial	Valor obtenido de la evaluación comercial
Calificación técnica	Valor obtenido por la evaluación técnica
Calificación Desempeño	Valor obtenido por la experiencia de los stakeholders
Calificación	Promedio obtenido de las tres calificaciones
Experiencia en segmento AUTO	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Experiencia en segmento HSOL	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Experiencia en segmento LCSD	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Experiencia en segmento IND	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Experiencia en segmento CEC	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Experiencia en Soluciones Pick_and_Place	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Experiencia en Soluciones Material Handling	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Experiencia en Soluciones visión	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Experiencia en Soluciones Pruebas	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Experiencia en Soluciones líneas Ensamble	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Experiencia en Soluciones empaque	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Experiencia en Soluciones de almacén	0 – Poca experiencia 3 – Experiencia intermedia 5 – Alta experiencia
Capacidad máxima	Número de proyectos que puede trabajar al mismo tiempo el proveedor
Capacidad asignada	Procentaje de la capacidad máxima que tiene destinada a la empresa
% Rebait	% de incentivo tras alcanzar un nivel de compra
GTM	0 – No pertenece a la clase 1 – Pertenece a la clase

3.2. Pipeline



3.2.1. Algoritmos de preprocesamiento de datos

El objetivo de este paso es seguir una secuencia de operaciones sistemáticas para limpiar, transformar y preparar datos crudos antes de usarlos en modelos de análisis o aprendizaje automático. El primer paso será la limpieza de los datos. Se eliminarán los registros duplicados y para los valores faltantes en caso de haberlos se hará la imputación de valores. En este caso se utilizará la **imputación con KNN (K-Nearest Neighbors).** Esta consiste en consiste en rellenar valores faltantes de una observación utilizando la información de las observaciones más similares a ella (sus “vecinos más cercanos”) según las demás variables. Este método es más preciso que la media/moda porque considera el contexto multivariable. Sin embargo, puede ser

computacionalmente costoso si el dataset es grande y en caso de que los rangos entre variables sean muy grandes se requiere escalar los datos. Para cuando se trate de valores faltantes en variables categóricas, se utilizará la moda para imputar dichos valores.

3.2.2. Algoritmos de Ingeniería de Características

La ingeniería de características (feature engineering) es el proceso de crear, transformar, seleccionar o eliminar variables (características) para mejorar el rendimiento de modelos de aprendizaje automático. Este paso puede tener un impacto mayor que la elección del algoritmo en sí.

En el caso de las variables numéricas se hará:

- Escalación y normalización de datos mediante el algoritmo de **StandardScaler**. Ésta es una técnica de escalado que transforma las características numéricas para que tengan media 0 y desviación estándar 1, es decir, las convierte en una distribución normal estándar (z-score).
- De ser necesario generación de nuevas características

Para las variables categóricas:

- Se aplicará **Label Encoding** las variables categóricas ordinales. Este método consiste en la codificación de variables categóricas donde cada categoría se reemplaza por un número entero único.
- En el caso de las variables independientes se utilizará el método **One-Hot Enconding,** la cual es es una técnica para convertir variables categóricas en una representación numérica binaria. Consiste en crear una nueva columna para cada categoría y asignar un 1 en la columna correspondiente y 0 en las demás.

3.2.3. Algoritmos de Clusterización

Los algoritmos de clusterización son métodos de aprendizaje no supervisado que agrupan observaciones en conjuntos basados en su similitud. Entre los más comunes se encuentran el algoritmo K-Means, que divide los datos en un

número predefinido de grupos optimizando la distancia intra-clúster, y el algoritmo DBSCAN, que forma clústeres basados en densidades de datos cercanos, ideal para conjuntos con formas no lineales. Estos métodos son útiles para descubrir patrones ocultos y segmentar datos en análisis exploratorios.

Tanto para los proveedores como para los proyectos se puede utilizar un algoritmo de K-Means. Éste es un algoritmo de clustering no supervisado que agrupa datos en K grupos (clústeres) basándose en la distancia entre puntos y centroides. Es simple, rápido y uno de los métodos más utilizados para agrupar datos sin etiquetas.

El objetivo de aplicar estos métodos es para poder agrupar los proyectos y los proveedores en clústeres no tan obvios como por ejemplo dividir los proyectos por tipo de aplicación.

3.2.4. Algoritmos de Predicción Categórica

En el caso para la asignación del proyecto a un proveedor, se están considerando tres posibles algoritmos: **Extreme Gradient Boosting (XGBoost), Regresión Logística** y **K-Nearest Neighbors (KNN).**

XGBoost (Extreme Gradient Boosting) es un algoritmo de aprendizaje automático que se ha destacado en diversas competiciones y aplicaciones debido a su eficiencia y precisión. Es una implementación avanzada del algoritmo de Gradient Boosting, diseñada para optimizar el rendimiento y reducir el tiempo de ejecución mediante mejoras como la regularización L1 y L2, paralelización y manejo de valores faltantes. Este modelo es especialmente potente para tareas de predicción categórica y regresión, ya que puede manejar grandes cantidades de datos y variables complejas con facilidad. Al aplicar XGBoost, es posible explorar configuraciones como el ajuste de hiperparámetros, profundidades de árboles y tasas de aprendizaje, buscando maximizar la precisión mientras se minimiza el sobreajuste.

Por otor lado, la regresión logística es un método que permite modelar relaciones entre una variable dependiente categórica y una o más variables independientes. Este algoritmo es ampliamente empleado en problemas de clasificación, como determinar si un proyecto debe ser asignado a un

proveedor específico o no. Ésta utiliza la función sigmoide para transformar los valores predichos en probabilidades dentro del rango de 0 a 1.

El modelo estima la probabilidad de que ocurra un evento y asigna clasificaciones basadas en un umbral predeterminado. Por ejemplo, un resultado por encima de 0.5 puede asignarse a una clase positiva, y por debajo, a una clase negativa. Además, es posible ajustar el valor del umbral según las necesidades del problema, optimizando así los resultados.

Entre sus ventajas, la regresión logística destaca por su simplicidad y facilidad de interpretación, lo que la convierte en una herramienta ideal en escenarios con conjuntos de datos pequeños y variables independientes linealmente relacionadas. Sin embargo, requiere un análisis cuidadoso de los datos, ya que su desempeño puede verse afectado por problemas como multicolinealidad o valores atípicos.

Con la implementación de regresión logística en el contexto de asignación de proyectos, se abre la posibilidad de explorar cómo las características específicas de los proveedores y proyectos determinan patrones óptimos de asignación, lo que contribuye a decisiones más informadas y efectivas.

Por último, KNN es un algoritmo de aprendizaje supervisado utilizado para tareas de clasificación y regresión. Funciona identificando los "K" puntos más cercanos en el espacio de características a la observación que se desea clasificar o predecir y utiliza las etiquetas o valores de esos vecinos para tomar una decisión. Es sencillo y efectivo, pero su desempeño puede verse afectado por la elección de "K" y la métrica de distancia utilizada. KNN es particularmente útil cuando los datos tienen patrones claros y bien definidos.

En la siguiente tabla podemos observar una pequeña comparativa entre los tres algoritmos:

Característica	KNN	XGBoost	Regresión Logística
Tipo de modelo	No paramétrico	Ensamble (Gradient Boosting)	Lineal paramétrico
Entrenamiento	No entrena (lazy learning)	Sí, entrena múltiples árboles	Sí, ajusta pesos lineales
Precisión	Media	Alta	Media (buena si hay relación lineal)
Velocidad de predicción	Lenta (depende del tamaño del set)	Rápida	Muy rápida
Manejo de variables categóricas	Necesita codificación	Necesita codificación (o usar CatBoost)	Necesita codificación
Escalado de variables	Necesario	No necesario	Necesario
Robustez ante ruido	Baja	Alta	Media
Interpretabilidad	Baja	Media (puedes ver importancia de variables)	Alta
Ideal para...	Datos simples y bien distribuidos	Grandes datos, relaciones complejas	Modelos explicables con relaciones lineales
Riesgo de overfitting	Alto (si K es muy bajo)	Bajo-medio (usa regularización)	Medio

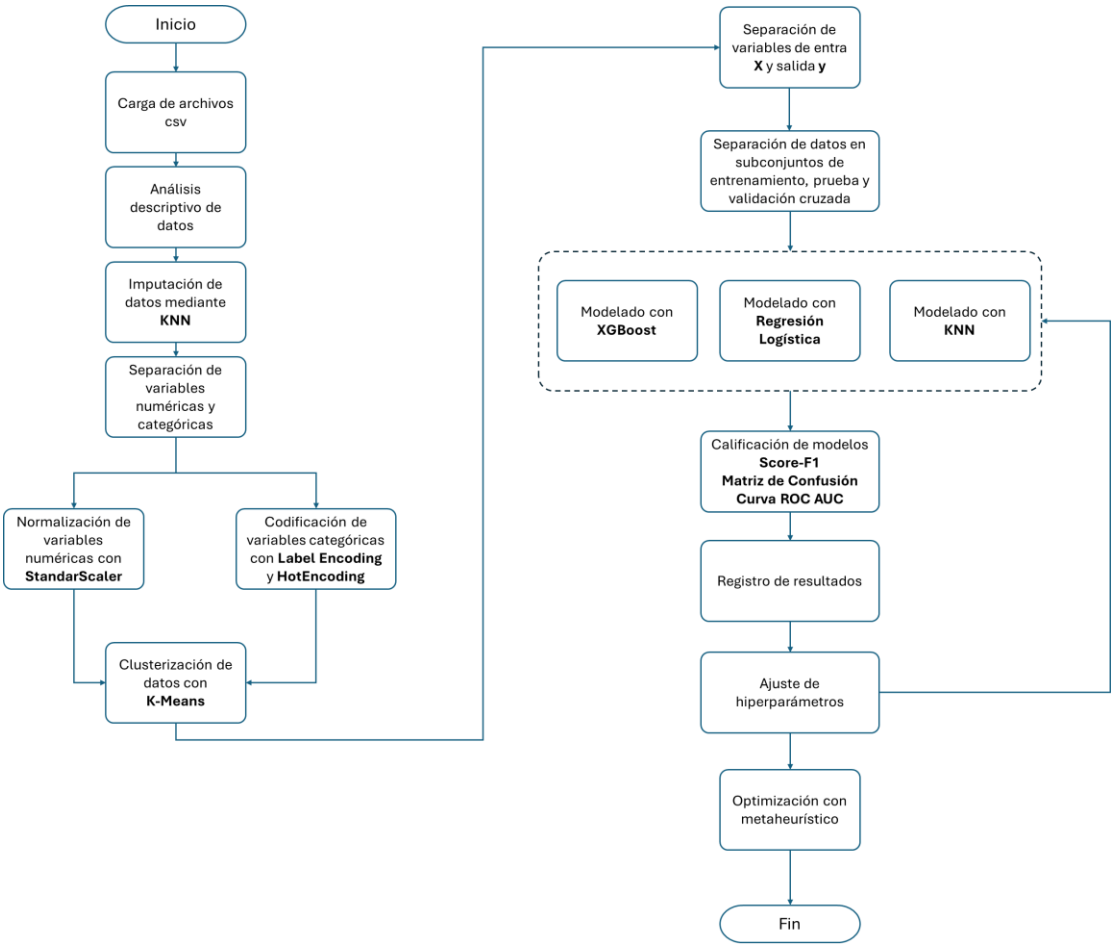
En este punto se trabajaría con los tres modelos para ver cuál ofrece mejores resultados. Para poder llevar un registro del desempeño de los modelos cambiando sus configuraciones se estaría trabajando también con las librerías de **MLOps**.

Nos apoyaremos en el **Score-F1**, **matriz de confusión** y las curvas **ROC AUC** para medir el desempeño de los modelos.

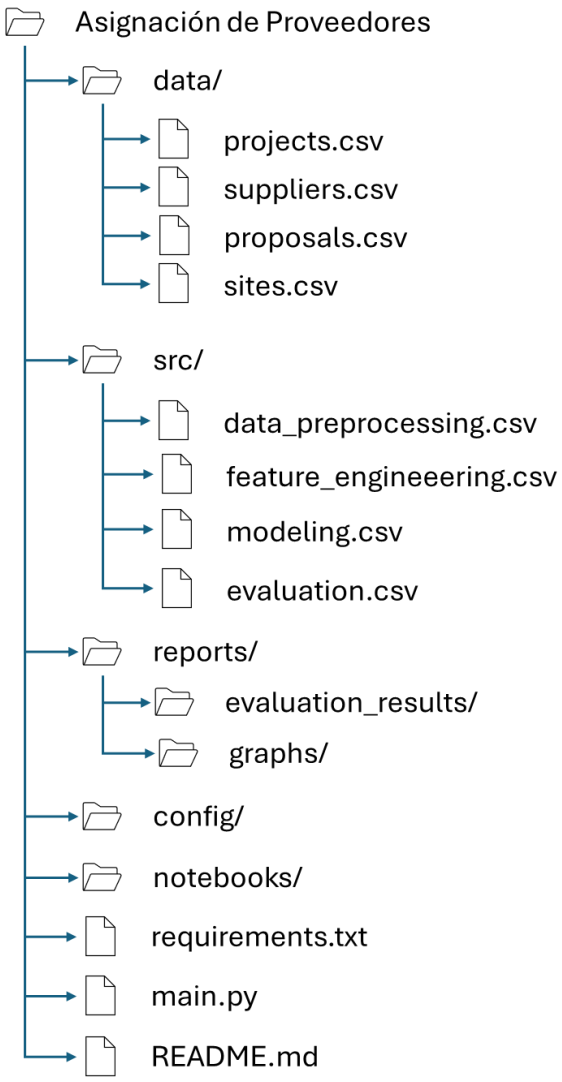
3.2.5. Algoritmos de Optimización

Como parte opcional, se está revisando utilizar algún algoritmo metaheurístico para optimizar la decisión. Se están considerando este tipo de algoritmos porque tiene esa flexibilidad para problemas donde no es fácil modelar restricciones como fórmulas.

La manera en que se está pensando estructurar el procesamiento de los datos se ve reflejada en el siguiente diagrama de flujo.


Inicio
Carga de archivos csv
Análisis descriptivo de datos
Imputación de datos mediante **KNN**
Separación de variables numéricas y categóricas
Normalización de variables numéricas con **StandarScaler**
Codificación de variables categóricas con **Label Encoding** y **HotEncoding**
Clusterización de datos con **K-Means**
Separación de variables de entra **X** y salida **y**
Separación de datos en subconjuntos de entrenamiento, prueba y validación cruzada
Modelado con **XGBoost**
Modelado con **Regresión Logística**
Modelado con **KNN**
Calificación de modelos **Score-F1** **Matriz de Confusión** **Curva ROC AUC**
Registro de resultados
Ajuste de hiperparámetros
Optimización con metaheurístico
Fin

3.3. Estructura del proyecto



4. Conclusión

La inteligencia artificial y el aprendizaje automático representan herramientas poderosas para mejorar la asignación de proyectos a proveedores. Estos enfoques no solo optimizan la toma de decisiones, sino que también permiten a las empresas adaptarse a entornos cambiantes y mantenerse competitivas. Con una implementación adecuada, estas tecnologías tienen el potencial de revolucionar la manera en que se gestionan proyectos y proveedores, marcando un nuevo estándar de eficiencia y precisión en el mundo empresarial.

5. Bibliografía

[1]. Zhou, Z. H. (2016). *Machine learning* (1st ed.). Springer.

[2]. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection techniques. Computers & Electrical Engineering, 40(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[3]. Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

[4]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer.

[5]. Günther, M., & Fritsch, S. (2010). Neural networks: A comprehensive foundation (2nd ed.). Pearson Education.