Clinical Data Base

Fundación Universitaria Konrad Lorenz. Big Data. Electiva-I.

Estudiante: David Gutierrez. Cod: 506222728

I. Informe de Análisis de exploratorio de sobre datos clínicos de pacientes Diabetes

II. Introducción:

Desarrollar un análisis completo que permita extraer datos sobre los pacientes que presentan afectaciones de diabetes, utilizando herramientas de modelos machine learning y el algoritmo Kmeans, para visualizar los resultados de manera efectiva.

III. Objetivo General:

El objetivo es identificar qué características o factores (como edad, género, raza, historial de tabaquismo, BMI, niveles de HbA1c, etc.) están más asociados con el diagnóstico de diabetes en los pacientes del conjunto de datos.

Palabras Clave: Pingüinos, Archipiélago Palmer, Análisis Exploratorio, Preprocesamiento de Datos, Visualización de Datos, Conservación, Especies, Hábitat.

Herramientas: Python, Pandas, Matplotlib, Seaborn.

Conjunto de datos: Contiene información sobre datos clínicos de pacientes. https://www.kaggle.com/datasets/ziya07/diabetes-clinical-dataset100k-rows

IV. Explicación de los encabezados del archivo .csv

Los encabezados del archivo representan diversas características de los pacientes que pueden estar relacionadas con la diabetes:

- **year:** Año en el que se registraron los datos.
- **gender:** Género del paciente (masculino o femenino).
- **age:** Edad del paciente en años.
- **location:** Ubicación geográfica del paciente.
- **race:** AfricanAmerican, race:Asian, race:Caucasian,
- **race:Hispanic, race:Other:** Variables binarias que indican la raza del paciente (1 si pertenece a la raza, 0 si no).
- **hypertension:** Indica si el paciente tiene hipertensión (1: Sí, 0: No).
- **heart disease:** Indica si el paciente tiene alguna enfermedad cardíaca (1: Sí, 0: No).
- **smoking history:** Historial de tabaquismo del paciente.
- **bmi:** Índice de masa corporal (IMC) del paciente.
- **hbA1c level:** Nivel de hemoglobina glucosilada (HbA1c), un indicador clave para el diagnóstico de diabetes.
- **blood glucose level:** Nivel de glucosa en sangre del paciente.
- **diabetes:** Indica si el paciente ha sido diagnosticado con diabetes 1: Sí, 0: No.
- **clinical notes:** Notas clínicas del paciente.

V. Problema planteado:

La diabetes es una enfermedad crónica que afecta a millones de personas en el mundo. Identificar las características que aumentan el riesgo de desarrollar diabetes puede mejorar la prevención y el diagnóstico temprano.

VI. Preguntas de Investigación:

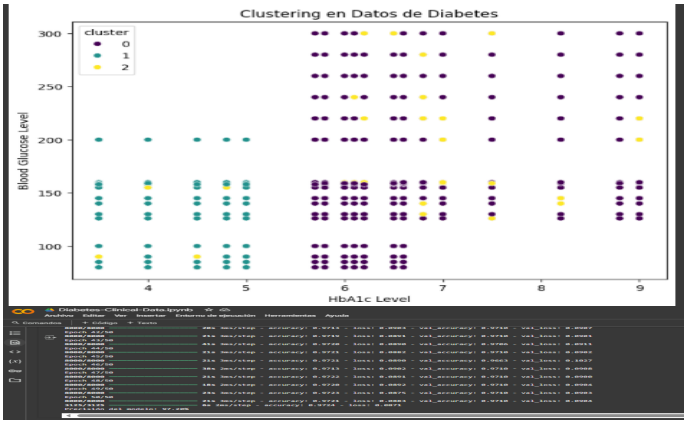
1) ¿Qué características hacen que se diagnostique a una persona con diabetes?.

VII. Hipótesis:

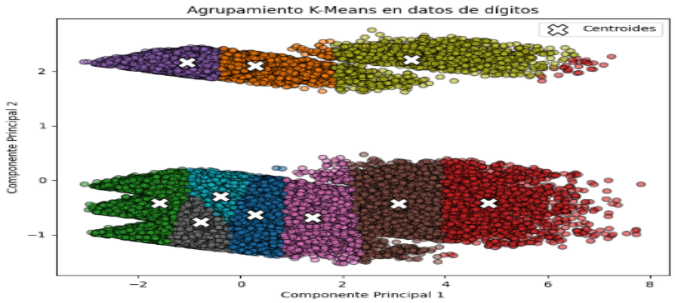
1) Un mayor nivel de HbA1c y glucosa en sangre está fuertemente correlacionado con la presencia de diabetes.
2) Pacientes con hipertensión y enfermedades cardíacas tienen mayor probabilidad de desarrollar diabetes.
3) El índice de masa corporal (BMI) alto puede estar asociado con un mayor riesgo de diabetes.
4) Los antecedentes de tabaquismo pueden influir en el diagnóstico de diabetes.

VIII. Posibles soluciones:

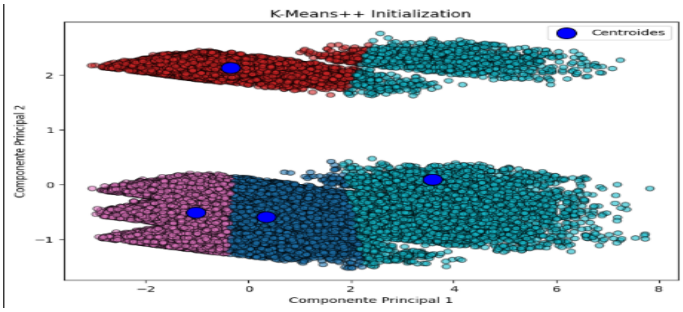
- **Análisis estadístico:** Realizar un análisis de correlación entre las variables y la presencia de diabetes para identificar qué factores están más fuertemente asociados.
- **Visualización de datos:** Usar gráficos como histogramas, diagramas de dispersión y mapas de calor para visualizar las relaciones entre las variables y la diabetes.
- **Modelado predictivo:** Entrenar un modelo de machine learning (como una red neuronal) para predecir la diabetes basado en las características del paciente y evaluar la importancia de cada variable en el modelo.
- **Clustering:** Agrupar a los pacientes en clusters basados en sus características para identificar patrones comunes entre aquellos diagnosticados con diabetes.



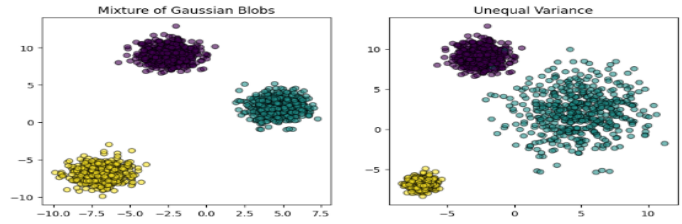
Acuarycity: Capacidad del modelo machine learning
Clustering: Modelo machine learning



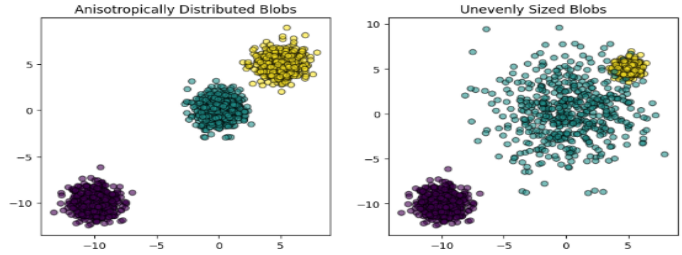
Agrupamiento de K-Means en datos de digitos



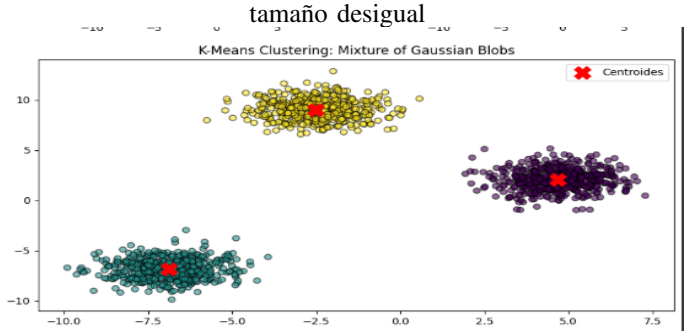
K-Means++ Inicialización



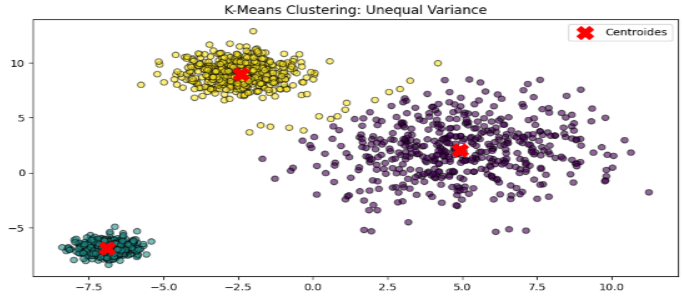
Mezcla de manchas gaussianas y Varianza desigual



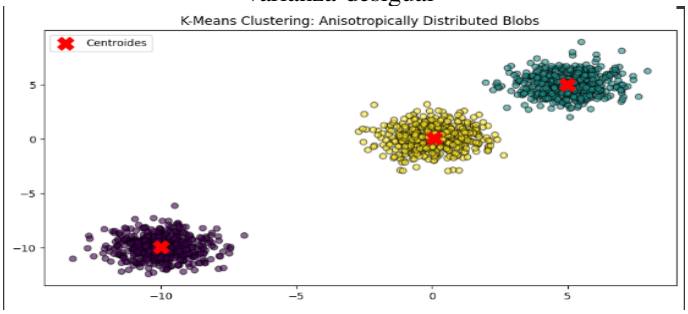
Manchas Distribuidas Anisotrópicamente y Sangre de tamaño desigual



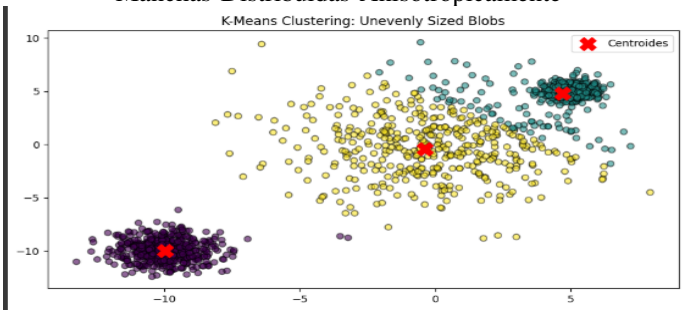
Mezcla de manchas gaussianas



Varianza desigual



Manchas Distribuidas Anisotrópicamente



Sangre de tamaño desigual

IX. Conclusiones:

- Los niveles de HbA1c y glucosa en sangre son probablemente los predictores más fuertes de diabetes.
- La obesidad (alto BMI) y la hipertensión también pueden ser factores importantes.
- El historial de tabaquismo podría tener un impacto, pero es necesario un análisis más profundo para confirmar su relevancia.