**Katherine Garcia Pombal**

DATA MINING **| DR.LEE**

Final project

**📌 Dataset Description and Relevance:**

For this project, I selected the “Diabetes Prediction Dataset” from Kaggle. The dataset contains detailed health-related information from over 1,000 individuals, including variables such as age, gender, BMI, hypertension status, smoking history, HbA1c levels, and blood glucose levels. This dataset is highly relevant for healthcare analytics, as diabetes is a major chronic illness worldwide that places significant strain on healthcare systems.

This project aligns with my career goals in data analytics and healthcare, combining data science skills with real-world impact. The dataset’s variety of health indicators allows for diverse modeling approaches such as classification, regression, and clustering.

**🧠 Problem Statement and Business Value:**

The main objective of this project is to develop machine learning models that support early detection and management of diabetes. Specifically, I will:

1. **Predict whether a patient is likely to develop diabetes (classification)** – enabling early intervention.
2. **Estimate a patient’s HbA1c level (regression)** – helping monitor glucose control and risk.
3. **Segment patients into health risk groups (clustering)** – aiding in personalized treatment strategies.

Solving these problems has strong business value: healthcare providers can prioritize high-risk individuals, allocate resources more efficiently, and implement preventive care programs. Additionally, this data-driven approach supports the transition toward value-based healthcare, where improving long-term outcomes takes precedence over short-term treatments.

**Predictive Modeling for Diabetes Risk Using Machine Learning**

📌 **Introduction**

The **diabetes\_prediction\_dataset.csv file** contains medical and demographic data of patients along with their diabetes status, whether positive or negative. It consists of various features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. The Dataset can be utilized to construct machine learning models that can predict the likelihood of diabetes in patients based on their medical history and demographic details.

The dataset contains 100,000 patient records with demographic and clinical information, including age, gender, BMI, blood glucose level, HbA1c level, smoking history, hypertension, and heart disease status. The primary goal of this project was to build predictive and descriptive models that aid in diabetes detection, management, and patient segmentation.

With diabetes being a major global health concern, early diagnosis and risk segmentation are critical for both public health outcomes and operational efficiency in healthcare institutions. Through classification, regression, and clustering techniques, this project provides tools to support data-driven interventions and targeted care.

**🔍 Methods**

**1. Data Cleaning & Preprocessing**

Initial cleaning involved handling missing values and correcting inconsistencies in categorical labels (e.g., replacing “No Info” in smoking history with “Unknown”). Outliers in BMI were also removed (e.g., values > 60). Categorical variables were encoded using one-hot encoding, and numerical features were standardized using z-score normalization. Feature selection and preprocessing were automated using ColumnTransformer pipelines.

**2. Exploratory Data Analysis**

EDA revealed key patterns:

* ~8.5% of patients had diabetes.
* HbA1c\_level and blood\_glucose\_level had the highest positive correlation with diabetes.
* Diabetic patients tended to be older and have higher BMI and glucose-related metrics.

Visualizations included count plots, histograms, correlation heatmaps, and pairplots to analyze feature relationships.

**3. Classification (Predicting Diabetes)**

A logistic regression model was implemented using a pipeline. Due to class imbalance, class weights were adjusted. Evaluation on the test set gave the following:

* **Accuracy**: 89%
* **Recall (diabetes class)**: 89%
* **Precision (diabetes class)**: 43%
* **F1-score (diabetes class)**: 0.58

The model performed well in capturing actual diabetic cases (high recall), making it suitable for screening applications.

**4. Regression (Predicting HbA1c Level)**

A linear regression model was used to predict HbA1c\_level based on patient features. Performance was measured using RMSE and R²:

* **RMSE**: 1.047
* **R²**: 0.041

The model was moderately successful in predicting HbA1c levels, which can be useful in estimating long-term glucose exposure for patients with incomplete records.

**5. Clustering (Patient Segmentation)**

K-Means clustering was applied on standardized age, BMI, glucose, and HbA1c data. The elbow method suggested k=4. Cluster analysis showed:

* One group had significantly higher HbA1c and glucose levels and a higher diabetes rate.
* Another cluster contained mostly younger, healthy individuals.

These insights can help healthcare providers segment patients into risk categories for proactive care and monitoring.

**📈 Results Summary**

| **Task** | **Method** | **Key Metric(s)** | **Result** |
| --- | --- | --- | --- |
| Classification | Logistic Regression | Recall (diabetes class) | 0.89 |
| Regression | Linear Regression | R², RMSE | [0.041 R² and 1.047 RMSE] |
| Clustering | K-Means | Cluster Profiles & Visualization | 4 clusters (risk groups) |

**✅ Conclusion**

This project successfully demonstrated how machine learning techniques can be used to support diabetes prediction and patient risk profiling. The classification model showed strong performance in identifying diabetic patients, while the clustering model provided valuable patient segmentation. Though the regression model had moderate predictive power, it still contributes to assessing glycemic trends.

For future work, incorporating time-series lab data or lifestyle variables could improve model accuracy. Techniques like SMOTE could enhance class balance for better precision.