ISyE 6740 – Summer 2023
Project Proposal

Team Member Names: Ankit Agarwal, Chetna C Kewalramani, Jiangqin Ma

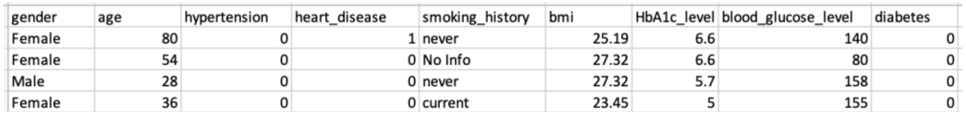
Project Title: Diabetes Data Analysis

Problem Statement

Millions of individuals throughout the world suffer from the common chronic condition known as diabetes. Identifying those who may be at risk of developing the disease can be considerably aided by early detection and precise diabetes prediction. We can examine a patient's medical history and demographic information to predict diabetes by utilizing machine learning algorithms. Additionally, we can investigate the relationships between various demographic and medical traits and the likelihood of developing diabetes.

Data Source

The diabetes prediction dataset, sourced from Kaggle, comprises 100,000 rows and 9 columns, with each row representing a patient. The dataset encompasses variables such as age, gender, hypertension, heart disease, smoking history, body mass index (BMI), HbA1c level, blood glucose levels, and diabetes status (1 represents positive diagnosis and 0 represents negative diagnosis) for each patient.



gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Female	80	0	1	never	25.19	6.6	140	0
Female	54	0	0	No Info	27.32	6.6	80	0
Male	28	0	0	never	27.32	5.7	158	0
Female	36	0	0	current	23.45	5	155	0

Methodology

Data Preprocessing

Preprocess the diabetes dataset by handling missing values, normalizing numerical features, and encoding categorical variables. This step will also include grouping the datapoints based upon the age group and smoking history.

Feature Selection

To enhance the performance of our predictive models and simplify the dimensionality of the dataset, we will employ a series of feature selection techniques:

- **Correlation Analysis:** to identify multicollinearity and find/remove highly correlated features to improve the model performance
- **Principal Component Analysis (PCA):** to transform the original features into a new set of uncorrelated principal components and to capture the most essential information while reducing the dimensionality of the dataset
- **Mutual Information (MI):** to assess the information gained by each feature (or principal component) with respect to the target variable

Anomaly detection

We will focus on comparing the effectiveness of robust covariance and one-class SVM techniques for outlier detection in our dataset. We will assess the performance of both techniques and determine which one yields better results in terms of outlier detection.

Machine Learning Models and Algorithms

K-means & Gaussian Mixture Model

K-means algorithm will be used to find out and analyze various clusters in the dataset using various parameters. The elbow method can be used to find out the correct number of clusters which uses within-cluster sum of squares (WCSS) value. The average silhouette method will be used to determine the quality of clusters.

K-means clustering has some disadvantages like manually choosing the value of K and dependency on the initial centroid locations to name a few. To overcome these, we'll also analyze the diabetes dataset using Gaussian Mixture Models. This will help identify the clusters using density estimation and relative probabilities.

K-Nearest Neighbors (KNN)

KNN is a classification algorithm that assesses the likelihood of a patient developing diabetes based on other patients with similar demographics and health data. This model evaluates the degree of similarity between a patient and nearby data points using demographic and medical information. Our objective is to appropriately categorize people and learn more about their risk of developing diabetes.

Logistic Regression

Logistic regression can be used to predict the probability of a patient to be diabetic. The goal is to find the best-fitting relationship between the various features in the dataset and probability (between 0 and 1) of having diabetes by using the logistic function.

The model will be trained using Maximum Likelihood Estimation (MLE) or gradient descent optimization algorithm. The objective is to find the values of the coefficients that maximizes the likelihood of the observed data and to find the decision boundary for the classification.

Evaluation and Final Results

K-means & GMM

Both K-means and GMM will be used to find the relevant clusters, to test the accuracy of the data and to check the speed of these two models.

For K-means, we'll evaluate below metrics:

- **Inertia**: This will be used to check the compactness of different clusters. A lower value suggests better clustering.
- **Silhouette Score**: This metric will be used to check the quality of clustering by considering both the compactness of data points within the clusters and the separation between the clusters.
- **Rand Index**: This will be used to measure the similarity between the true cluster labels and the predicted cluster assignments.

For GMM, we'll evaluate below metrics:

- **Log-Likelihood**: This will be used to measure how well the GMM model fits the data. It quantifies the probability of observing the given data points under the learned GMM parameters. Higher log-likelihood values indicate a better fit.
- **AIC and BIC**: The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used to select a particular model. Lower AIC and BIC values indicate a better trade-off between model fit and complexity.
- **Silhouette Score**: Similar to its usage in K-Means, the silhouette score can also be applied to evaluate GMM. It measures the compactness and separation of clusters, with higher values indicating better-defined clusters.

K-Nearest Neighbors (KNN)

KNN will be evaluated using the accuracy of the model on the testing dataset. Diabetes is a serious condition that can be managed with lifestyle changes rather than invasive procedures, so the threshold for the test dataset will be 90% accuracy.

We will experiment with various settings for the number of neighbors (k) and choose the one that produces the best accuracy in order to optimize the performance of this model. We can determine the balance between overfitting (low k) and underfitting (high k) by experimenting with different neighbor values. For each neighbor value on the testing dataset, the model with the highest accuracy will be selected.

Logistic Regression

Various assessment metrics, including mis-classification rate, confusion matrix, area under the ROC curve (AUC), precision, recall, and F1 score, can be used to evaluate the model's performance and its capacity for generalization. By tuning its hyperparameters, such as regularization strength (C) or penalty type (L1 or L2), we can improve the performance of the model.

By analyzing the model’s coefficients for various features, we can determine the strength between the features and likelihood of having diabetes.

Model Comparison

The dataset will be divided into training and testing dataset. The training dataset will be used to train the models and testing dataset will be used to evaluate the performance of each model.

Evaluate the logistic regression model's performance compared to other models used in the project, such as K-means, GMM clustering, and KNN classification. We will also check various factors like precision, readability, and computational effectiveness while comparing different models.

Project Timeline

Key dates we hope to achieve certain milestones:
Proposal finished – July 5th
Data cleaning, exploratory analysis, visualization – July 10th
Model training and analysis – July 15th
Model comparison – July 24th
Final Report finished – July 30th