ISyE 6740 – Summer 2023
Project Final Report

Team Member Names & GT Id: Ankit Agarwal (aagarwal432), Chetna C Kewalramani (ckewalramani3), Jiangqin Ma (jma416)

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

Table 1 Diabetes Data Set

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 employ a series of feature selection techniques:

Correlation Analysis: to identify multicollinearity and find/remove highly correlated features to improve the model performance.

	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	age_group	smoking_group	gender_group	diabetes
hypertension	1.000000	0.121262	0.147666	0.080939	0.084429	0.249825	0.015438	0.014203	0.197823
heart_disease	0.121262	1.000000	0.061198	0.067589	0.070066	0.229630	0.068298	0.077696	0.171727
bmi	0.147666	0.061198	1.000000	0.082997	0.091261	0.337632	0.029798	-0.022994	0.214357
HbA1c_level	0.080939	0.067589	0.082997	1.000000	0.166733	0.101987	0.014829	0.019957	0.400660
blood_glucose_level	0.084429	0.070066	0.091261	0.166733	1.000000	0.110571	0.016084	0.017199	0.419558
age_group	0.249825	0.229630	0.337632	0.101987	0.110571	1.000000	0.089520	-0.028755	0.258168
smoking_group	0.015438	0.068298	0.029798	0.014829	0.016084	0.089520	1.000000	0.085723	0.038803
gender_group	0.014203	0.077696	-0.022994	0.019957	0.017199	-0.028755	0.085723	1.000000	0.037411
diabetes	0.197823	0.171727	0.214357	0.400660	0.419558	0.258168	0.038803	0.037411	1.000000

Table 2 Correlation between variables

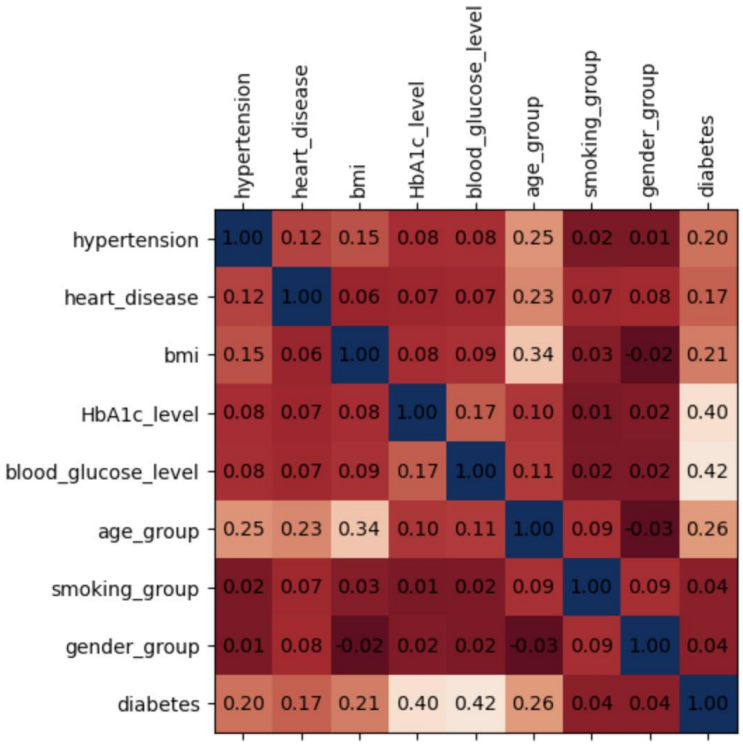

Figure 1 Correlation Matrix

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.

From the above correlation table and matrix, we can see that HbA1c_level has the strongest correlation with the target variable (diabetes status), and we can also find the multicollinearity between features is very weak, so we do not need to perform PCA.

Mutual Information (MI): to assess the information gained by each feature (or principal component) with respect to the target variable.

Feature	MI Score
HbA1c_level	0.13
blood_glucose_level	0.1121
age_group	0.0425
bmi	0.0201
heart_disease	0.0087
smoking_group	0.0086
hypertension	0.0067
gender_group	0.0

Table 3 MI Scores by Feature

MI scores represent the importance of each feature in relation to the target variable (diabetes status). Higher MI scores indicate the feature has a stronger relationship with the target variable. HbA1c_level has the highest MI score, which means that this is the most informative feature for predicting diabetes status. And gender_group has an MI score of 0, which means that it has no mutual information with diabetes status. So, gender might not provide any useful information for predicting diabetes status, we will not select this feature for training our machine learning models.

Anomaly detection

We focus on comparing the effectiveness of Local Outlier Factor (LOF) and one-class SVM techniques for outlier detection in our dataset. We assess the performance of both techniques and determine which one yields better results in terms of outlier detection.

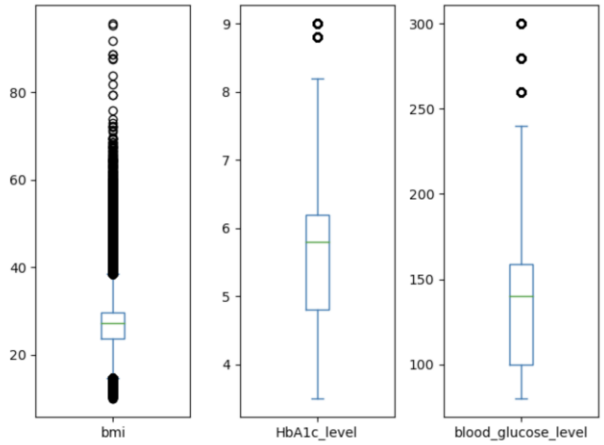

Figure 2 Boxplots

From the boxplots, we can identify that our dataset has some outliers.

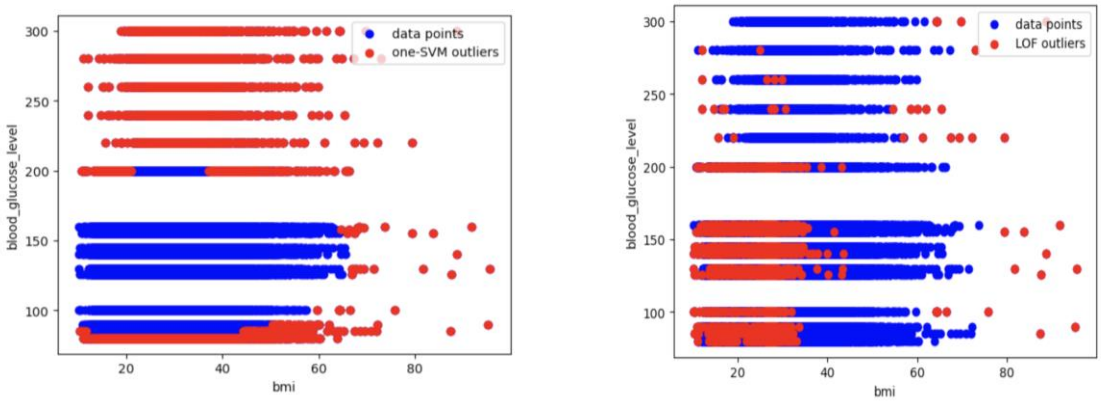

Figure 3 Outlier Distribution for One-SVM (left) and LOF (right) techniques

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 was 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 were 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 evaluated below metrics:

- **Inertia**: This is used to check the compactness of different clusters. Inertia is measured by calculating the sum of squared distances between each data point and the centroid of its assigned cluster thus a lower value suggests better clustering. As per the below values, the inertia is lowest when number of clusters are selected as 4.
- **Silhouette Score**: This metric is used to check the quality of clustering by considering both the compactness of data points within the clusters and the separation between the clusters. It also checks how well each datapoint fits into the assigned cluster. In the table below, the Silhouette score is highest when the number of clusters are selected as 3.
- **Rand Index**: This value is used to measure the similarity between the true cluster labels and the predicted cluster assignments. In the below table, the Rand Index is highest when number of clusters are selected as 4.

No. of Clusters	2	3	4	5	6	7
Inertia	73358312.48	27052900.88	17324266.42	9412360.64	7792406.486	6606542.425
Silhouette Score	0.5380880942	0.6840760276	0.6754599156	0.6282596847	0.596758675	0.5413640635
Rand Index	-0.01116735889	0.05679154362	0.06795424892	0.04520883963	0.04296102074	0.02872242044

Table 4 K-Means Model Analysis

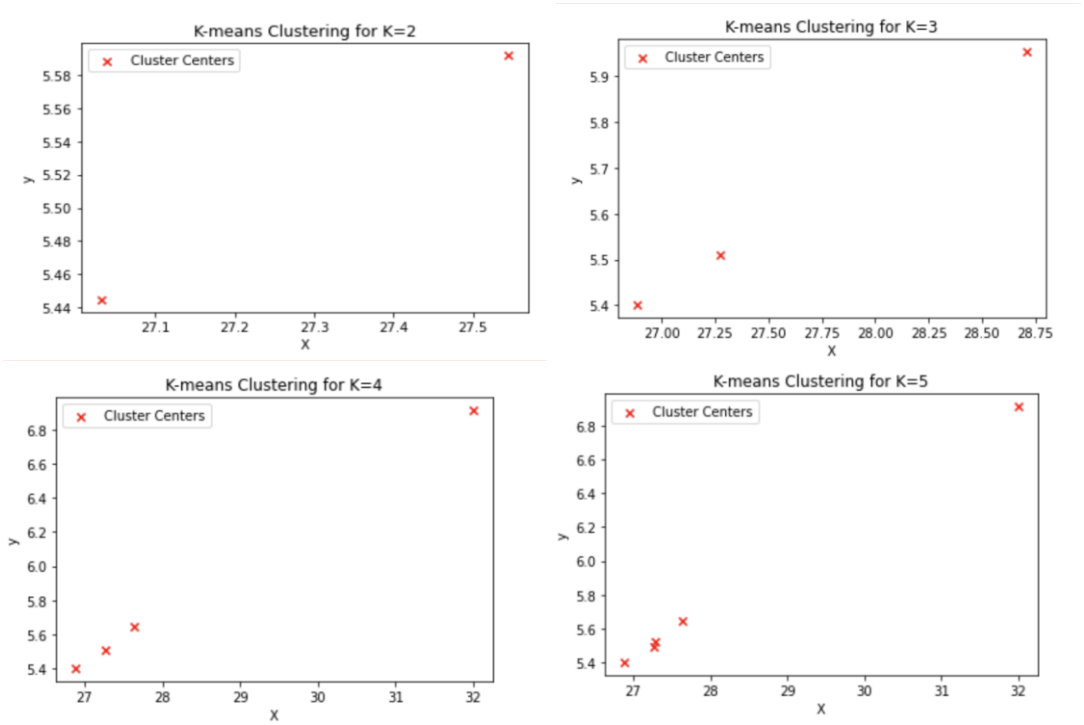

Figure 4 K-Means Clustering Model Analysis

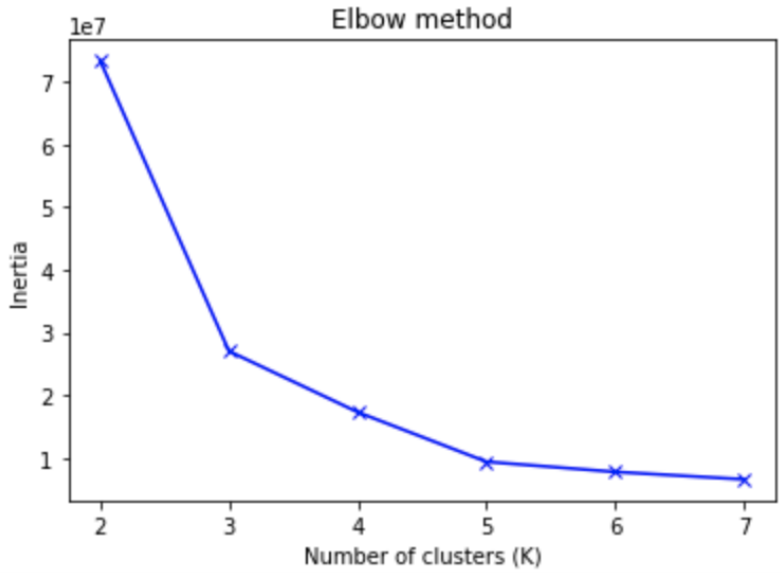

Figure 5 Elbow Plot

For GMM, we evaluate below metrics:

- **Log-Likelihood**: This can 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. In the table below, the log-likelihood value increases as the number of clusters increases. However, this may lead to overfitting and for this dataset this value may not be used to correctly measure the model performance.
- **AIC and BIC**: The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to select a particular model. Lower AIC and BIC values indicate a better trade-off between model fit and complexity. As per the results below, the AIC and BIC both decrease when the number of clusters increases. This may lead to overfitting and may not be a good measure for the model's performance.
- **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. In the table below, the Silhouette score is highest when the number of clusters are selected as 3.

No. of Clusters	2	3	4	5	6	7
Log Likelihood	-9.72688308	-9.556644359	-9.250465242	-8.484855704	-7.857405253	-7.697375002
Silhouette Score	0.0259365143	0.6840760276	0.5986975776	0.6226661691	0.5466601755	0.535968882
Mismatch Rate	0.48664	0.67932	0.90432	0.69245	0.71894	0.92328
AIC	1945368.588	1911335.808	1850164.833	1697062.325	1571594.892	1539610.442
BIC	1945549.333	1911611.682	1850535.837	1697528.458	1572156.155	1540266.834

Table 5 GMM Model Analysis

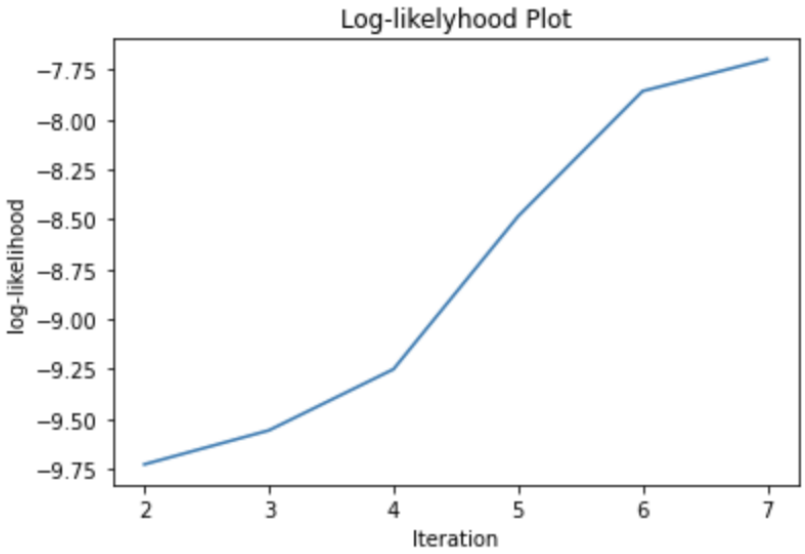

Figure 6 Log-likelihood Plot

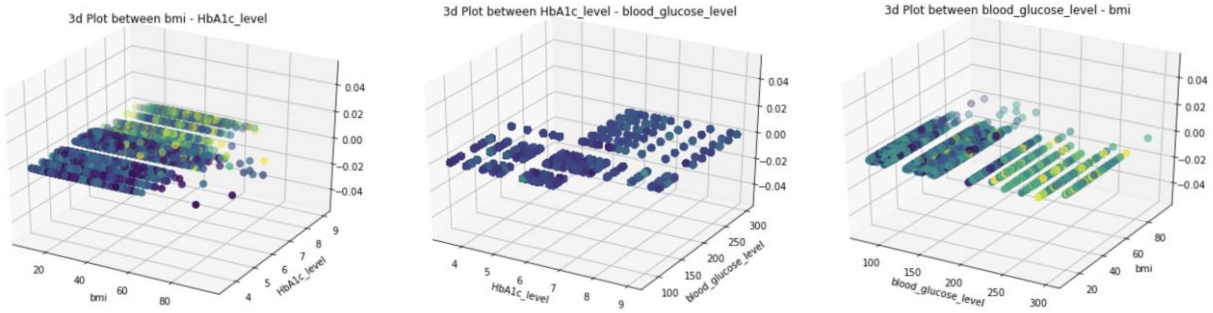

Figure 7 3D plots

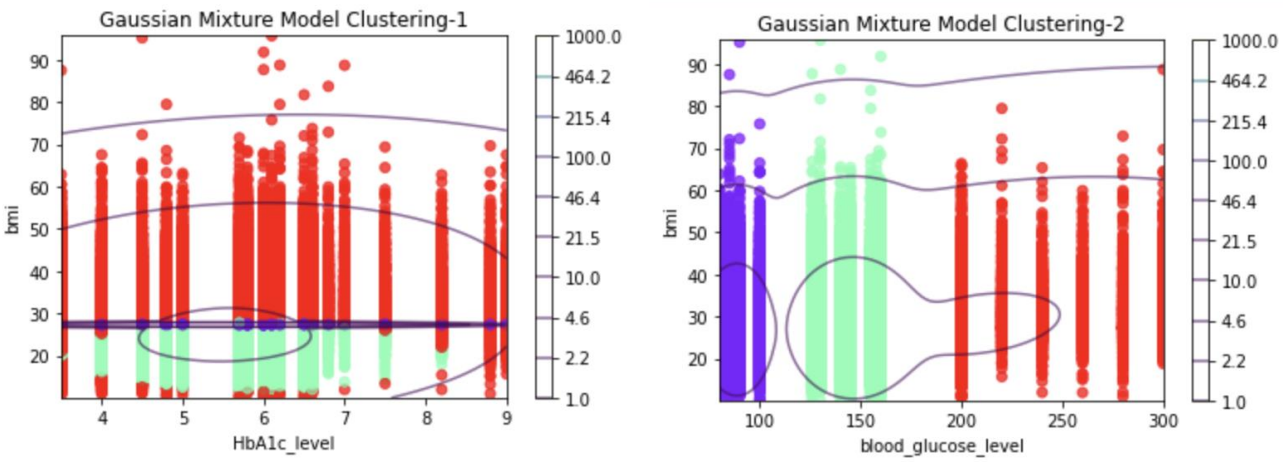

Figure 8 GMM Clustering

K-Nearest Neighbors (KNN)

KNN was 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 was 90% accuracy.

The best number of neighbors (k) was found using the grid search method from sklearn. The k used for this project was 9. Three KNN models were constructed, one with no outliers removed, one with outliers removed using One Class SVM, and one with outlier removed using Local Outlier Factor. The test accuracy of each of the models is summarized in the table below.

	No Outliers Removed	**One Class SVM Outliers Removed**	**Local Outlier Factor Outliers Removed**
Test Accuracy	95.84%	95.63%	96.34%

Table 6 KNN Model Summary

As seen in the table, the model with outliers removed using the Local Outlier Factor method had the highest test accuracy. The models overall did not have much difference in their accuracy scores and they all met our threshold. Even so, the model with the outliers removed using Local Outlier Factor is the best performing one.

Logistic Regression

Various assessment metrics, including confusion matrix, the area under the ROC curve (AUC), precision, recall, accuracy and F1 score, can be used to evaluate the model's performance and its capacity for generalization.

Using the default setting, the accuracy is 95.98%. By tuning its hyperparameters, such as regularization strength (C), penalty type (L1 or L2), and solver (liblinear or saga), we can improve the performance of the model, raising the accuracy to 96.21%.

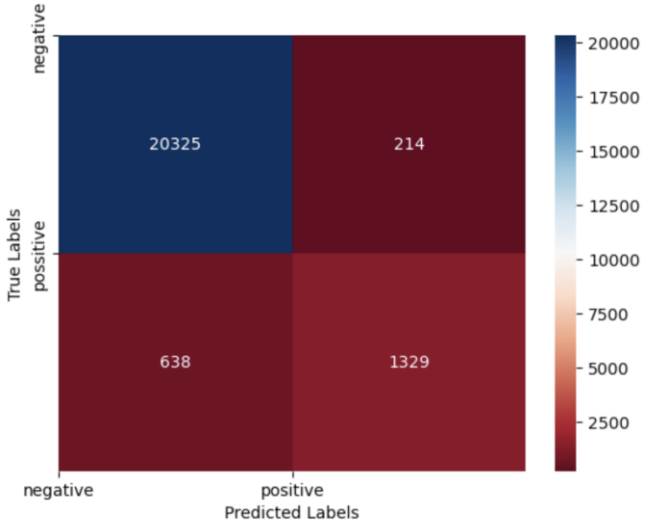

Fig 9 Confusion Matrix

	precision	recall	f1-score	support
0	0.97	0.99	0.98	20539
1	0.86	0.68	0.76	1967
accuracy			0.96	22506
macro avg	0.92	0.83	0.87	22506
weighted avg	0.96	0.96	0.96	22506

Fig 10 Classification Report

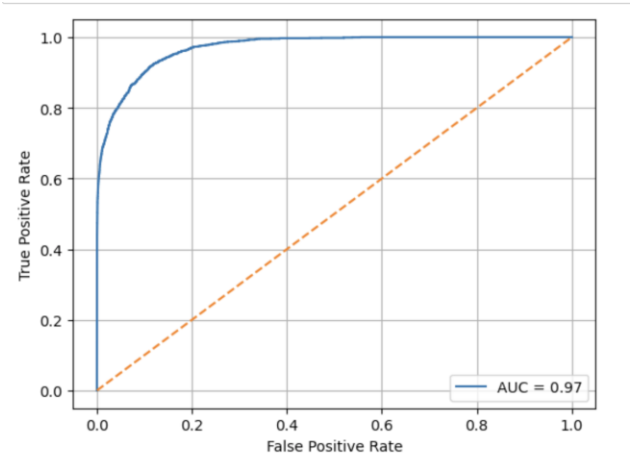

Fig 11 ROC Curve

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

Feature	Likelihood of having Diabetes
HbA1c_level	2.2137039963225504
heart_disease	1.012409569529042
hypertension	0.9930506333630824
age_group	0.5494338178072599
bmi	0.09974983792050005
smoking_group	0.0665618194024914
blood_glucose_level	0.03531426007974127

Table 7 Diabetes Likelihood by Feature

Positive coefficients indicate a positive association with the likelihood of being diabetic, while negative coefficients indicate a negative association. We can see that all the features have positive associations with the likelihood of being diabetic, which means that higher levels of the features are associated with a higher likelihood of being diabetic. HbA1c_level has the strongest relationship.

Model Comparison & Results

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

The Silhouette score is highest and similar for both the models when the number of clusters are selected as 3. Which means that both K-means and GMM suggests the number of clusters should be 3 and both are producing 3 equally well-defined clusters. For other number of clusters Silhouette score of K-means is higher than the GMM, so K-means performs better than GMM in those instances.

Also, the Log-likelihood, AIC and BIC values do not suggest a good fitting for the datapoints thus we can say that K-means performs better than the GMM model for this dataset. K-means and GMM clustering achieve the best performance when using 3 clusters, even though the true labels might have a different number of clusters.

Evaluate the logistic regression model's performance compared to other models used in the project, such as KNN classification and some tree-based models. Since they are all computational effectiveness, we only compare their classification accuracy.

Model	**Accuracy**
Logistic regression	96.21%
KNN	95.63%
Decision Tree model	96.04%
Random Forest model	96.04%
Extra Trees model	96.04%
AdaBoost model	96.04%

Table 8 Classification Models Accuracy

KNN has the best performance overall, since it computes efficiently, achieves the highest accuracy, and it is easily interpreted. The tree base models, such as decision tree, random forest, extra trees, AdaBoost have the same accuracy, which suggest they might be performing similarly with default parameter settings. We can continue to experiment with hyperparameter tuning in our future research to potentially improve their performance.

Contribution

Feature selection, logistic regression models, tree-based models, arranging meetings and coordinating group work, working on report — Jiangqin Ma
KNN models, outlier detection, working on report – Chetna Kewalramani
Data cleaning, K-Means models, GMM models, working on report – Ankit Agarwal

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