**CBIO313: Machine Learning & Data mining**

**Final Project Report**

**Investigating the Role of HbA1c Level in Predicting Diabetes Using Different Machine Learning Models**

**Presented by: Malak Haitham Mohamed**

**ID: 221001396**

**Supervised by: Dr. Muhammed Elsayeh**

**1.Abstract**

This project investigates the predictive power of the HbA1c level in diagnosing diabetes using different machine learning models on a large-scale health dataset of 100,000 individuals. The data includes demographic and clinical variables such as age, gender, BMI, blood glucose, smoking history and hypertension status. A particular focus was placed on evaluating how well HbA1c levels indicate diabetes presence. Multiple classification models were trained and compared including Random Forest, XGBoost, Logistic Regression, and K-Nearest Neighbors and others, using performance metrics like accuracy, precision, recall, and F1 score. The Random Forest model outperformed others, offering the best balance across all evaluation metrics. Results support the strong diagnostic relevance of HbA1c, confirming its clinical utility for early detection when combined with other features.

**2.Introduction**

Diabetes is a chronic metabolic disorder that affects millions of individuals globally and poses significant public health challenges. It is associated with severe long-term complications such as cardiovascular diseases, kidney failure, nerve damage, and vision loss if left undiagnosed or untreated. Therefore, early and accurate diagnosis is critical to managing and mitigating its progression. In clinical practice, Hemoglobin A1c (HbA1c) has emerged as a standard biomarker for the diagnosis and monitoring of diabetes. HbA1c reflects average blood glucose levels over a period of two to three months, offering a more stable measure compared to daily glucose readings.

Despite its wide acceptance, questions remain about the independent predictive power of HbA1c when used in isolation versus when combined with other demographic and clinical features. The availability of large-scale structured health data opens new opportunities to explore these relationships using machine learning techniques, which are capable of uncovering complex nonlinear interactions among variables.

This study aims to assess the diagnostic value of HbA1c using a variety of supervised classification models, ranging from simple linear classifiers to sophisticated ensemble methods. By incorporating additional features such as age, gender, BMI, blood glucose levels, smoking history, hypertension, and heart disease, we aim to build predictive models that can assist healthcare professionals in making data-driven decisions. The key objective is to compare the performance of these models and identify the one that provides the most reliable predictions in diagnosing diabetes. In doing so, this project also evaluates the practical utility of HbA1c as a central feature in predictive analytics frameworks within healthcare applications.

**3.Methodology**

The dataset used in this study comprises health records of 100,000 individuals with features including age, gender, body mass index (BMI), blood glucose levels, HbA1c levels, smoking history, hypertension, and a binary label for diabetes diagnosis. Data preprocessing involved removing rows with missing values in essential columns to ensure model accuracy and integrity. Categorical features such as smoking history were label encoded, and continuous features were scaled where necessary, particularly for models like K-Nearest Neighbors that are sensitive to feature magnitude. The dataset was split into training (80%) and testing (20%) subsets to evaluate model generalization.

Exploratory Data Analysis (EDA) was performed to uncover trends and relationships within the data. Distribution plots showed that HbA1c and glucose levels were significantly higher in diabetic patients. Correlation heatmaps reinforced the importance of HbA1c as a key predictor, and boxplots illustrated the differences in BMI and age between diabetic and non-diabetic groups. These insights guided feature selection and modeling strategies.

Thirteen machine learning models were trained and tested: Logistic Regression, Decision Tree, K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Random Forest, Gradient Boosting, XGBoost, AdaBoost, Bagging, Stacking, Hard Voting, and Soft Voting classifiers. The models were evaluated using accuracy, precision, recall, and F1 score. A summary DataFrame was created with these metrics for comparison, and results were visualized using a barplot.

**4.Results & Discussion**

The Bagging Classifier demonstrated robust performance with an accuracy of 0.97, a precision of 0.94, recall of 0.70, and an F1 score of 0.80 as shown in *fig.1.*It was particularly effective in balancing precision and recall, indicating reliable detection of diabetic patients without producing excessive false positives.

*Fig.1*

The Stacking Classifier also achieved an accuracy of 0.97. Its precision was slightly lower at 0.91, with a recall of 0.67 and an F1 score of 0.77 as shown in Fig.2. These results reflect a strong overall capability in predicting diabetes but with a moderate tendency to miss some positive cases.

*Fig.2*

The Hard Voting Classifier reported an accuracy of 0.9652, precision of 0.93, recall of 0.64, and F1 score of 0.76. as shown in fig.3 While it achieved high precision, its lower recall suggests it may overlook more actual diabetic cases. Similarly, the Soft Voting Classifier scored 0.9649 accuracy, 0.89 precision, 0.67 recall, and 0.77 F1 score as shown in fig.4 , showing slightly more balanced behavior between false positives and false negatives.

*Fig.3 Fig.4*

XGBoost, a powerful gradient-boosted decision tree algorithm, performed very well with 0.97 accuracy, 0.96 precision, 0.69 recall, and an F1 score of 0.81 as shown in Fig.5. It combined high accuracy and a strong balance between precision and recall, making it one of the top-performing models.

*Fig.5*

Gradient Boosting also excelled, reaching 0.97 accuracy and a standout precision of 0.99. However, its recall dropped slightly to 0.68, resulting in an F1 score of 0.81 as shown in Fig.6. This suggests it is very confident in its predictions but may fail to detect some diabetic patients.

Fig.6

AdaBoost followed closely with similar metrics: 0.97 accuracy, 0.97 precision, 0.69 recall, and an F1 score of 0.81 as shown in Fig.7. Its balanced performance and high precision make it another reliable model for deployment.

Fig.7

Naive Bayes showed a noticeable drop in performance with 0.91 accuracy, 0.46 precision, 0.66 recall, and 0.54 F1 score as shown in Fig.8. Its simplicity and assumption of feature independence likely limited its ability to handle complex feature interactions.

Fig.8

The Support Vector Machine (SVM) model performed poorly, with near-zero performance metrics as shown in Fig.9. This may be attributed to improper parameter tuning or the model’s sensitivity to feature scaling and class imbalance.

A blue and white chart with numbers and labels

AI-generated content may be incorrect.

Fig.9

The Decision Tree model delivered a respectable 0.95 accuracy, 0.70 precision, 0.73 recall, and 0.72 F1 score as shown in Fig.10. While not the most accurate, it offered a clear and interpretable structure for identifying diabetes risk.

Fig.10

Logistic Regression achieved solid performance with 0.96 accuracy, 0.87 precision, 0.62 recall, and 0.73 F1 score as shown in Fig.11. Its strength lies in interpretability, but it struggled with capturing complex patterns compared to ensemble methods.

Fig.11

K-Nearest Neighbors (KNN) showed reasonable results with 0.96 accuracy, 0.90 precision, 0.57 recall, and 0.70 F1 score as shown in Fig.12. While it had high precision, the relatively low recall indicates challenges in detecting all diabetic cases

Fig.12

The Random Forest Classifier emerged as the most balanced model overall, with 0.97 accuracy, 0.97 precision, 0.68 recall, and 0.80 F1 score as shown in Fig.13. It consistently delivered high performance across all metrics and handled both feature importance and interaction effectively, making it the best candidate for deployment.

*Fig.13*

To visualize the comparison between the different models, a grouped bar plot was created showing accuracy, precision, recall, and F1 scores for all 13 models as shown in Fig.14. As seen in the graph, most ensemble models such as Random Forest, Gradient Boosting, AdaBoost, and XGBoost cluster toward the top with consistently high metrics. In contrast, models like Naive Bayes and SVM fall significantly behind, especially in terms of precision and F1 scores. This plot reinforces the findings discussed above and clearly illustrates the superior performance of ensemble approaches compared to simpler classification techniques.

*Fig.14*

**4. Deployment**

The trained Random Forest model was saved using the joblib library and deployed through a Flask web application. The application takes user input for features such as age, BMI, glucose, and HbA1c levels and returns a prediction on diabetes status. The Flask app (app.py) loads the model (model.pkl), accepts form input from an HTML page (index.html), and displays the prediction result. This demonstrates a complete machine learning pipeline from data ingestion and model training to real-time deployment. The application was tested locally on the localhost as shown in Fig.15.

*Fig.15*

**5. Conclusion**

This study confirmed that HbA1c is a highly informative feature in machine learning models for predicting diabetes. Ensemble methods, particularly Random Forest, Gradient Boosting, and XGBoost, outperformed traditional classifiers, demonstrating excellent performance in terms of both precision and recall. The project also highlights the importance of model comparison and visualization in selecting the best algorithm. Furthermore, deployment of the final model as a web application shows the practical value of machine learning in healthcare settings, offering a user friendly tool for diabetes risk assessment. Future directions include, external validation on other datasets, and expansion to multi-class classification for prediabetes stages.