COGNITIVE COMPUTING
(UCS420)

Multi-Modal Diabetes Prediction

Submitted by:
Aishwary Srivastava (102317117)
Group: 3Q15

Submitted to:
Mr. Sukhpal Singh

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)
Thapar Institute of Engineering and Technology
Patiala, Punjab

November 2025

Introduction and Problem Statement

India is the diabetes capital of the world! The diabetic population in India is 77 million. Therefore, to better manage and treat the condition, early detection and diagnosis are extremely important. Our proposed model assists in predicting whether a patient is diabetic or not based on some key patient information such as HbA1c level, blood glucose level, BMI, gender, and other considerations. Predicting whether one is at risk for diabetes can save millions of lives and offset a burden on healthcare systems—especially since such simple metrics, which are often found during routine health check-ups, are all that is necessary to assess risk. This project assesses whether diabetes prediction is possible through machine learning using support vector machines, logistic regression, and neural networks for accurate prediction.

We are using Diabetes Prediction Dataset:-
https://huggingface.co/datasets/marianeft/diabetes_prediction_dataset

Data Exploration and Preprocessing Data

Our data set has 9 columns and 100,00 instances out of which 25,000 are labeled as diabetic. The dataset includes the following features:

1. Gender: Male, Female, or Other
2. Age: Age of the patient
3. Hypertension: indicating the presence of hypertension (0 = No, 1 = Yes)
4. Heart Disease: Binary feature indicating the presence of heart disease (0 = No, 1 = Yes)
5. Smoking History: Patient's smoking status (never, former, current, not current, ever and No Info.)
6. BMI: Body Mass Index
7. HbA1c Level
8. Blood Glucose Level
9. Diabetes

Preprocessing

We have checked for null values in the dataset to ensure quality. For example, we mapped the gender column into numerical values for that variable. We also one-hot encoded the Smoking History column since this contained many values and can significantly differentiate the results. Finally, we used a Standard Scaler of Scikit-Learn to scale numerical values including Age, BMI, HbA1c, and Blood Glucose Level. This was critical to ensure accuracy because this allowed each feature to have equal weight during training and prevented features with larger scales from biasing the model training.

Model Implementation and Evaluation

Logistic Regression

Logistic Regression is a statistical method used for binary classification problems. It estimates the probability of an instance belonging to a particular class using a logistic function. Our Logistic Regression model utilises the linear solver (Liblinear), which is suitable for relatively small datasets and provides good classification performance.

Support Vector Machine (SVM)

Support Vector Machine constructs a hyperplane in a high-dimensional space to separate different classes. Our SVM implementation uses the Radial Basis Function (RBF) kernel, which can efficiently handle non-linear relationships between features. The RBF kernel maps input features into an infinite-dimensional space, allowing the model to capture complex interactions between health parameters that may not be apparent in the original feature space.

Neural Networks

Our deep learning framework utilizes two hidden layers where ReLU (Rectified Linear Unit) acts as the activation function and an output layer that utilizes sigmoid as the activation function to give a binary final output (0 for No diabetes and 1 for Diabetes). This architecture allows the model to learn complex patterns and relationships within the health data that simpler models might miss.

Model Training and Evaluation

We trained all models on 75% of the dataset, reserving the remaining 25% for testing. The model performance was evaluated using multiple metrics:

- Precision: Measures the proportion of correct identifications
- Recall: Measures of the proportion of actual positives that were correctly identified
- F1 Score: The harmonic mean of precision and recall
- Accuracy: The overall correctness of predictions
- ROC Curve: Graphical representation of the diagnostic ability of the model
- AUC Score: Area under the ROC curve, measuring the model's ability to distinguish between classes

Visualizations and Key Findings

Through our exploratory data analysis, we used pie charts and bar charts to determine the impact of gender and smoking history on diabetes. This visualization approach helped us encode the values of these attributes in a weighted way to improve model performance.

Key findings from our analysis include:

1. According to the dataset, diabetes is more prevalent in Males than Females. This gender disparity could be attributed to various lifestyle and physiological factors.
2. Smokers in general are more prone to having diabetes than non-smokers. We observed a clear correlation between smoking history and diabetes risk, with current smokers showing the highest risk followed by former smokers.
3. HbA1c Level and Blood Glucose Level were identified as the strongest predictors of diabetes, as expected given their direct relationship with blood sugar regulation.

Challenges and Future Improvements

Challenges

The first challenge was to decide on the encoding strategy that was to be used on the smoking history column since the values were diverse yet impactful on the final output. We decided to go with one hot encoding because it allowed us to keep all the values of the smoking history column and saved us from assigning any inherent numeric weight to the values.

The second challenge was to manually tune the hyperparameters of the models and decide what works best. Multiple changes to the hyperparameters were made and tested to produce the most accurate result.

Future Improvements

Several avenues for future improvement have been identified:

1. **Hyperparameter Tuning**: Implementing automated hyperparameter optimization techniques such as Grid Search, Random Search, or Bayesian Optimization could further enhance model performance.
2. **Feature Selection**: Applying feature selection methods to identify the most predictive subset of features could improve model efficiency and interpretability.
3. **Training on More Data**: Expanding the dataset or incorporating data from diverse populations could improve model generalizability and robustness.
4. **Model Deployment**: Developing a user-friendly interface or web application to make the predictive model accessible to healthcare professionals for clinical use.