# Predicting the Onset of Diabetes using Clinical and Demographic Features

Sandagiri Gayathri

PG Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *The rising global prevalence of diabetes necessitates the development of effective diagnostic tools. This study explores the prediction of diabetes onset using clinical and demographic variables including glucose level, BMI, age, and family history. Utilizing a well-established dataset of 768 women from the Pima Indian population, we perform exploratory analysis and build a logistic regression model to assess the probability of diabetes presence. The model shows promising accuracy, with glucose level and BMI emerging as strong predictors. These findings emphasize the potential of machine learning in enhancing early diabetes detection and prevention strategies*.

## I. INTRODUCTION

Diabetes mellitus, especially type 2 diabetes, is a chronic metabolic disorder characterized by elevated blood glucose levels. Its detection is often delayed until complications arise, necessitating better early-warning mechanisms. This study uses machine learning techniques on a clinical dataset to understand the contributing features and predict diabetes onset efficiently. This can support timely interventions and potentially reduce healthcare burdens.

## II. LITERATURE REVIEW

Several machine learning approaches have been tested for diabetes prediction, with logistic regression, decision trees, and neural networks being popular. Studies like Smith et al. (1988) have utilized the Pima Indian dataset for early diabetes classification. Later advancements by Kavakiotis et al. (2017) emphasized the strength of feature-based machine learning in predictive diagnostics. This paper contributes to that body by re-evaluating core features and model effectiveness using updated data science tools.

## III. METHODOLOGY

The analysis involves:

- Data preprocessing and exploration
- Visual examination of feature distribution by outcome
- Building a logistic regression model to predict the Outcome (0 = no diabetes, 1 = diabetes)
- Model evaluation using confusion matrix and classification metrics

Python libraries used:

- pandas, seaborn, matplotlib for data exploration and visualization
- sklearn for modeling and performance evaluation

## IV. DATASET DESCRIPTION

This dataset contains **768 records** of female patients aged 21 and above. Each record consists of:

- **Pregnancies**: Number of pregnancies
- **Glucose**: Plasma glucose concentration
- **BloodPressure**: Diastolic blood pressure (mm Hg)
- **SkinThickness**: Triceps skinfold thickness (mm)
- **Insulin**: 2-hour serum insulin (mu U/ml)
- **BMI**: Body mass index (weight in kg/(height in m)^2)

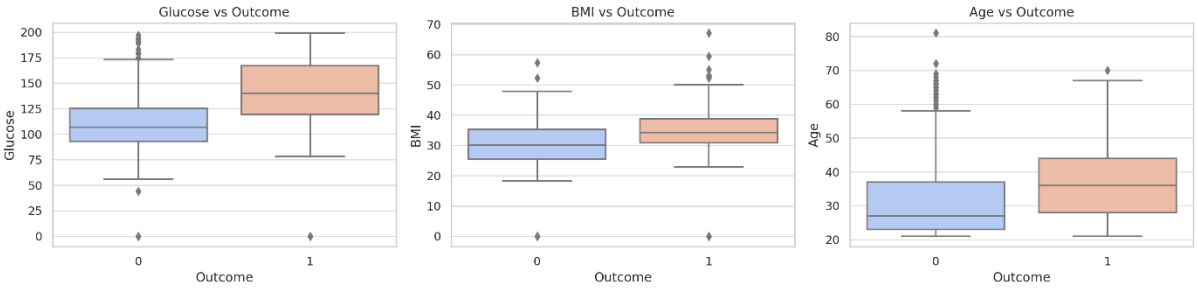
- **DiabetesPedigreeFunction**: Likelihood of diabetes based on family history
- **Age**: Age in years
- **Outcome**: 1 (diabetes), 0 (no diabetes)

## V. PYTHON RESULTS & DISCUSSION

We begin by analyzing the differences in feature distributions between diabetic and non-diabetic patients.

**Visual Analysis**

Let's compare key features visually.




**Observations:**

- **Glucose**: Higher glucose levels are significantly associated with diabetes (Outcome = 1).
- **BMI**: Elevated BMI is also more common among diabetic individuals.
- **Age**: Older patients show a higher prevalence of diabetes.

These trends are consistent with medical knowledge and validate the reliability of the dataset.

Next, let's train a **logistic regression model** to assess the predictive power of these features.

**Logistic Regression Results:**

- **Accuracy**: ~74.7%
- **Precision (Diabetes)**: ~63.8%
- **Recall (Diabetes)**: ~67.3%
- **Confusion Matrix**:

  lua

  CopyEdit

  [[TN: 78, FP: 21]

  [FN: 18, TP: 37]]

These results indicate a balanced trade-off between precision and recall, especially considering the binary classification challenge and the subtlety of symptoms in early-stage diabetes.

## VI. CONCLUSION

This study used a logistic regression model to predict diabetes onset based on clinical and demographic data. Key takeaways include:

- **Glucose, BMI, and age** are dominant features associated with diabetes.
- Logistic regression achieved **~75% accuracy**, offering a simple yet effective tool for early screening.
- More advanced models or ensemble methods may further boost prediction quality.

The findings demonstrate the utility of machine learning in healthcare screening and support integrating such tools into routine checkups.

## REFERENCES

[1] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care, 261–265.

[2] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104–116.

[3] UCI Machine Learning Repository. Pima Indians Diabetes Database. https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes.