ILLINOIS TECH

# CSP 571 DATA PREPARATION AND ANALYSIS

# PROJECT REPORT

## Predictive Analysis of Diabetes Using Machine Learning Techniques

## Team Members:



| | |
|---|---|
| **Motupalli Lakshmi Lahitha** | **A20550762** |
| **Rutuja Jadhav** | **A20539073** |
| **Lekhana Sai Mikkilineni** | **A20545604** |
| **Deepika Regulavalasa** | **A20526663** |

## 1.Abstract

This report explores predictive analytics for diabetes using the Diabetes dataset, which comprises 768 samples and 8 medical features. Our approach integrates data preprocessing, exploratory data analysis, unsupervised clustering, and supervised learning models to derive actionable insights. The report details methodologies, model performances, and potential future enhancements.

## 2.Introduction

### Objective

To analyze the diabetes dataset and predict diabetes onset using machine learning techniques.

### Dataset Overview

The dataset contains 8 medical features, including glucose levels, BMI, and age, along with a binary target variable indicating diabetes presence. This dataset's structured nature makes it well-suited for predictive analytics.

## 3.SCOPE

The scope of this project encompasses comprehensive data preprocessing, including handling missing values, normalizing data for consistency, and selecting key features to enhance model performance. Exploratory Data Analysis (EDA) was conducted to uncover correlations among features and assess class separability, providing insights into the relationships within the data. Clustering techniques, such as K-Means, were employed to group data into diabetic and non-diabetic categories, leveraging feature-based separability. Finally, classification models, including Logistic Regression, Random Forest, and XGBoost, were implemented to predict diabetes outcomes, with an emphasis on evaluating and optimizing model performance.

# 4.Methodology

**1. Data Preprocessing:**The dataset was cleaned and prepared to ensure consistency and reliability-

- **Handling Missing Values:** Filled nulls with median/mean values.
- **Normalization:** Ensured consistent data scaling.
- **Feature Selection:** Reduced noise and improved interpretability by focusing on key features like Glucose and BMI.

**2. Exploratory Data Analysis (EDA):**EDA helped uncover patterns and relationships in the data-

- **Correlations:** A strong positive correlation was identified between Glucose and diabetes outcome, highlighting its predictive value.
- **Visualization:** Techniques like heatmaps and dimensionality reduction (PCA, UMAP, t-SNE) were used to explore patterns and visualize separability between diabetic and non-diabetic groups.

**3. Unsupervised Learning:**Clustering techniques revealed natural groupings in the data-

- **K-Means Clustering:** Grouped data into diabetic and non-diabetic clusters.
- **Visualization:** PCA-based visualization highlighted cluster separability.

**4. Supervised Learning:**Predictive models were built to classify diabetes outcomes-

- **Models Implemented:**
    - Logistic Regression
    - Random Forest
    - XGBoost
- **Hyperparameter Tuning:** Randomized and grid search techniques were employed to optimize model parameters. Key hyperparameters, such as the number of estimators and maximum depth for Random Forest, were fine-tuned. This resulted in a significant accuracy improvement, with Random Forest achieving a maximum accuracy of 81.2%.
- **Evaluation**: Metrics such as accuracy, precision, recall, and F1-score were used to assess model performance, ensuring reliable predictions.

---

## 5.Results and Insights

**Results-**

- **Model Performance:**The supervised models showed varying levels of accuracy, with XGBoost leading slightly-
    - **XGBoost:**Achieved the highest accuracy of 80.5%, excelling in capturing complex relationships in the data
    - **Random Forest:** Performed well with 79.7% accuracy, which increased to 81.2% after hyperparameter tuning. This highlights its flexibility and adaptability when fine-tuned.
    - **Logistic Regression:** Logistic Regression achieved 76.6% accuracy, doing a good job of capturing simple patterns in the data but not quite matching the performance of the more advanced ensemble models.
- **Feature Importance:** Glucose, BMI, and Insulin emerged as the most significant predictors.

**Insights-**

- **Dimensionality Reduction**: Methods like PCA and t-SNE made it easier to spot patterns and separate diabetic from non-diabetic groups, thus enhancing clustering and classification.
- **Hyperparameter Tuning**: Adjusting settings for models like Random Forest significantly improved accuracy and reliability.
- **Feature Engineering**: Focusing on key features like Glucose, BMI, and Insulin simplified the models while keeping accuracy high.

---

## 6.Conclusion

Machine learning analysis on the Diabetes dataset demonstrates enormous potential within healthcare, mainly for diagnosis and early interventions in diabetic conditions. The following project successfully implements a wide range of pipelines consisting of data preprocessing, exploratory data analysis, clustering, and supervised learning that have extracted actionable insight through an integrated approach. XGBoost turned out to be the best among all tried, with an accuracy of 80.5%. However, Random Forest showed 81.2% after hyperparameter tuning, which really gives reason for model optimization. Logistic Regression provided a great baseline that validated the performances of more advanced models. Key features identified included Glucose, BMI, and Insulin, which were the most critical predictors, hence their clinical relevance in assessing diabetes risk. Such methods of dimensionality reduction like PCA, UMAP, and t-SNE facilitate better visualization and enhance class separation.

---

## 7.Future Work

- Improvement in Model Accuracy: Utilizing higher-order ensemble techniques like stacking and bagging, along with the use of deep learning models such as neural networks to better capture complex patterns in data.
- Real-Time Prediction Systems: We can  create a web-based or mobile application integrated with wearable devices to provide real-time predictions of diabetes risk, monitoring via continuous glucose monitors.
- Feature Engineering and Innovation: Including derived features like age-adjusted glucose level or summary health index to capture more predictors of the variation in results, improving model performance.
- Longitudinal Data Analysis: The dataset can be extended by including time-series data which can then be used to track a patient's health over time, enabling predictions of not only the current risk but also future diabetes onset.

---

## 8.GitHub Repository Link

https://github.com/rutujdv/DiabetesDataAnalysis

Link to group 22 github repository.

---