Brinda Nathvani Shrey Shah Saumil Trivedi

MIS-637
Completed Project Proposal

1. Introduction

Project Title

- "Data Analysis and Prediction of Diabetes Using Machine Learning"

Project Context

- This college project focuses on analyzing a diabetes dataset to predict the onset of diabetes using various machine learning algorithms. The project aims to explore the effectiveness of these algorithms in a healthcare context, emphasizing predictive precision.

Project Goals

- To conduct in-depth data analysis and exploratory data visualization.
- To preprocess and prepare the dataset for modeling.
- To apply and evaluate multiple machine learning algorithms, with a particular focus on the CatBoost algorithm.
- To address challenges such as imbalanced data to ensure model reliability.

2. Dataset and Preprocessing

Dataset Overview

- The dataset comprises medical records, including health metrics relevant to diabetes.

Preprocessing Activities

- Cleaning data and handling missing values.
- Normalizing and scaling features.
- Feature selection and engineering for optimal model performance.

3. Exploratory Data Analysis (EDA)

Approach

- Using Python libraries (e.g., Seaborn, Pandas) for visualization and statistical analysis.
- Identifying patterns, correlations, and anomalies in the dataset.

Expected Insights

- Comprehensive understanding of the dataset's characteristics and how they may influence model development.

4. Model Development and Evaluation

Machine Learning Algorithms

- Linear Regression: A fundamental approach for understanding relationships between variables.
- Logistic Regression: Ideal for binary classification problems like diabetes prediction.
- K-Nearest Neighbors (KNN): A non-parametric method used for classification and regression.
- Decision Trees and Random Forest: Effective for capturing non-linear relationships in data.
- Gradient Boost and XGBoost: Advanced techniques that combine simple models into a strong learner.
- CatBoost: An algorithm known for its efficiency with categorical data and has shown promising results.
- Artificial Neural Networks: Complex models capable of capturing intricate patterns in data.

Evaluation Metrics

- Precision and Recall: Critical in medical diagnosis to minimize false positives and negatives.
- F1 Score: A harmonic mean of precision and recall.
- Cross-validation: Employed to ensure the robustness and generalizability of the models.

Handling Imbalanced Data

- Application of techniques like oversampling, undersampling, or SMOTE to balance the dataset and enhance model accuracy.