Report on Naive Bayes Classification Model for Diabetes Prediction

Introduction

This analysis was conducted to predict the likelihood of diabetes using a Naive Bayes model on a dataset containing various health-related features. The goal of this analysis is to enhance our understanding of the factors influencing diabetes onset and to provide accurate predictions.

Data Preparation

- **Data Loading**: The dataset was imported from a CSV file.
- **Data Processing**: Features were separated from the target variable ("diabetes"). Categorical variables were converted to numerical values using one-hot encoding.
- **Data Splitting**: The dataset was divided into a training set (80%) and a testing set (20%).

Model Building

The Gaussian Naive Bayes model was trained using the training dataset to identify relationships between features and diabetes status.

Evaluation and Results

- **Overall Accuracy**: 0.869 (86.9%), indicating correct predictions for 86.9% of test cases.
- **Classification Report**:
    - **Class 0 (No Diabetes)**:
        - Precision: 0.98
        - Recall: 0.88
        - F1-score: 0.92
        - Support: 18292 samples
    - **Class 1 (Diabetes)**:
        - Precision: 0.37
        - Recall: 0.79
        - F1-score: 0.51
        - Support: 1708 samples

Conclusion

The model successfully achieved a good accuracy in classifying diabetes cases overall; however, there is a need for improvement in the accuracy of classifying positive diabetes cases.