Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Report on Mini Project

Machine Learning -I (DJS23DSL402)

AY: 2024-25

DIABETES PREDICTION

OMKAR PRAMOD BANDIKATTE: 60009230009

Guided By

Dr. Kriti Srivastava

CHAPTER 1: INTRODUCTION

Diabetes is a widespread chronic disease characterized by high blood glucose levels, leading to severe complications if untreated. Early detection and preventive care are crucial to reduce the risk of severe health issues. This project aims to develop a machine learning model to predict the likelihood of an individual having diabetes based on medical and lifestyle factors. The dataset used is sourced from Kaggle and includes features like age, gender, BMI, blood glucose level, HbA1c level, hypertension, smoking history, and heart disease status. The objectives of the project include:

• Understanding the dataset through exploratory data analysis.

• Preprocessing the data to handle missing values and outliers.

• Applying and comparing different machine learning models.

• Identifying the most accurate model for diabetes prediction.

This project demonstrates how data-driven methods can assist healthcare professionals in early diagnosis, leading to timely intervention and better patient outcomes.

In addition, the project highlights the importance of using machine learning not just for prediction, but also for gaining deeper insights into the underlying factors contributing to diabetes. By analyzing the relationships between various features, healthcare researchers can better understand risk patterns and recommend more personalized preventive measures for individuals at higher risk.

CHAPTER 2: DATA DESCRIPTION

The dataset utilized in this project, titled **"Diabetes Prediction Dataset,"** is sourced from Kaggle. It contains a comprehensive collection of health records necessary for the prediction task.

The dataset consists of approximately **100,000 records** and **19 columns**, representing a mix of numeric and categorical attributes.

Key Features:

- **age**: Continuous variable representing the age of individuals.
- **gender**: Categorical variable indicating Male, Female, or Other.
- **hypertension**: Binary variable showing the presence (1) or absence (0) of hypertension.
- **heart_disease**: Binary variable representing heart disease status.
- **smoking_history**: Categorical variable indicating the individual's smoking habits.
- **bmi**: Numerical value representing Body Mass Index.
- **HbA1c_level**: Average blood sugar over the past three months.
- **blood_glucose_level**: Current blood glucose level.
- **diabetes**: Target variable indicating whether the individual has diabetes (1) or not (0).

CHAPTER 3: DATA ANALYSIS

Exploratory data analysis was conducted to understand feature distributions and relationships. Key observations include:

- blood_glucose_level and HbA1c_level show strong correlations with diabetes. • Higher BMI values are commonly associated with diabetic individuals.
- Smoking history and presence of hypertension or heart disease also affect diabetes risk.

Outliers in numeric features were identified using box plots and treated where necessary. Categorical variables like gender and smoking_history were analyzed for their distribution. Correlation heatmaps helped identify important features influencing diabetes prediction. The insights gained during analysis guided the feature selection and modeling phases of the project.

Additionally, feature scaling was applied to the numeric variables to ensure uniformity, as large variations in scale could impact model performance. Transformations like standardization helped prepare the data for machine learning algorithms, improving both convergence speed and model accuracy during training.

CHAPTER 4: DATA MODELLING

After preprocessing, several machine learning models were trained and evaluated:

• Logistic Regression

• Random Forest Classifier • Decision Tree Classifier

The data was split into 80% training and 20% testing sets. Categorical variables were encoded, and numerical variables were standardized for better model performance.

Among all models, the Random Forest Classifier performed best, achieving an accuracy of approximately 85%. Feature importance analysis showed that blood_glucose_level, HbA1c_level, and bmi were the most significant predictors. Hyperparameter tuning was done to further improve the model's performance.

Cross-validation techniques were also employed to ensure the model's robustness and prevent overfitting. By validating the model across different subsets of the data, we achieved more reliable evaluation metrics and improved the generalization capability of the final model.

CHAPTER 4: CONCLUSION

This project successfully built a machine learning model to predict diabetes using patient health information.

Random Forest emerged as the best model, offering high accuracy and good interpretability. Important predictors included blood_glucose_level, HbA1c_level, and bmi, aligning with known medical facts.

The model shows potential for practical applications in healthcare, helping doctors in early diagnosis and preventive care.Future work could focus on improving the dataset, addressing class imbalance if present, and exploring deployment as a web-based application to make predictions easily accessible.

Moreover, integrating the model with electronic health record (EHR) systems could automate risk assessments during regular check-ups, enabling healthcare providers to offer timely advice and interventions without additional diagnostic tests.