Internship Report: Disease Prediction from Medical Data

College/Institute Name

Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology

Submitted by

Kusuma Lahari Sunkari
Roll No: 22UECS0662
Department: B.Tech Computer Science and Engineering
Email: lahari090422@gmail.com

Internship Under

Code Alpha

https://www.codealpha.tech

Internship Duration

From: 01-July-2025 To: 31-July-2025

Title of Project

Disease Prediction from Medical Data

Introduction

With the rise in medical data availability, artificial intelligence and machine learning techniques are being applied to improve healthcare diagnosis and prediction. This project focuses on developing a system that predicts the presence of a disease using patient health indicators. It aims to assist doctors by providing early warnings and predictive analysis from structured medical data.

2. Objective

The main goal of this project is to build a machine learning model that can analyze a patient's health data and predict the presence of a disease. This can help healthcare professionals in making early diagnoses and planning preventive treatments.

3. Tools and Technologies Used

- **Google Colab** – Cloud platform for running Python code
- **Python** – Programming language used
- **Pandas** – Data analysis and manipulation
- **Scikit-learn** – For building machine learning models
- **Matplotlib & Seaborn** – For data visualization
- **GitHub** – Version control and code repository

4. Dataset Description

The dataset used in this project includes several features that represent medical data from patients, such as:

- Age
- Gender
- Blood Pressure
- Glucose Levels
- Body Mass Index (BMI)
- Insulin levels
- Diabetes Pedigree Function
- Outcome (0 = No Disease, 1 = Disease)

This dataset was preloaded into Google Colab for processing and training.

5. Methodology

1. **Data Loading & Cleaning:**
   Loaded the dataset and handled any missing or inconsistent values.
2. **Data Exploration:**
   Used visualization tools to understand correlations and patterns

3. **Feature Scaling & Splitting:**
   Standardized the features and split the data into training and test sets.
4. **Model Training:**
   Trained the model using Random Forest Classifier on the training set.
5. **Model Evaluation:**
   Evaluated the model's accuracy using test data, confusion matrix, and classification report.

6. Model Used

Random Forest Classifier

A Random Forest is an ensemble machine learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. It is widely used for classification tasks due to its robustness.

7. Results

- **Accuracy of the model:** 85%
- **Confusion Matrix and Classification Report** showed high precision and recall
- The model was able to predict the presence of disease with good accuracy on unseen data.

8. Conclusion

This project demonstrates how machine learning can be effectively used for disease prediction based on health data. With the help of tools like Google Colab and Scikit-learn, a functional and accurate prediction model was developed. The project highlights the importance of data preprocessing and model evaluation in building efficient ML applications.

9. References

- Scikit-learn Documentation
- Code Alpha Internship Material
- Machine Learning with Python (by Coursera & Kaggle resources)

Declaration

I hereby declare that this internship report and the project work titled **"Disease Prediction from Medical Data"** is my original work, completed under the guidance of **Code Alpha,** and has not been submitted elsewhere for any other certification.