Disease Prediction from Medical Data using Machine Learning

Internship: CodeAlpha Machine Learning Internship

Submitted by: SWAYAM KUSANAL

Date: September 2025

Abstract

This project focuses on predicting diseases such as diabetes, heart disease, and breast cancer using structured medical datasets. Various machine learning classification algorithms were applied, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and XGBoost. The models were trained and evaluated on publicly available datasets from the UCI Machine Learning Repository and Kaggle. Results show that ensemble methods like Random Forest and XGBoost perform best, achieving higher accuracy compared to traditional algorithms. This work demonstrates the potential of ML in early disease detection and decision support systems.

Introduction

Importance of early disease prediction in healthcare. Role of machine learning in improving diagnostic accuracy. Objective: Build models that predict disease likelihood based on patient data.

Objectives

- Use structured patient datasets (symptoms, lab tests, demographics).
- Apply multiple classification algorithms for prediction.
- Compare performance across models.
- Identify the best-performing model for each dataset.

Literature Review

Machine Learning has been widely applied in healthcare for diagnosis. Logistic Regression is widely used for binary classification problems. SVM is effective in handling high-dimensional medical data. Random Forest and XGBoost are ensemble methods with higher predictive power.

Methodology

Datasets: Heart Disease (UCI), Diabetes (PIMA), Breast Cancer (Scikit-learn). Steps include preprocessing (scaling, train-test split), model training, and evaluation using Accuracy, Precision, Recall, and F1-score.

Implementation

Language: Python
Libraries: Pandas, NumPy, Scikit-learn, XGBoost
Steps: Data loading → Preprocessing → Model training → Evaluation

Results & Discussion

Breast Cancer Dataset Example:
- Logistic Regression: ~96%
- SVM: ~97%
- Random Forest: ~95%
- XGBoost: ~96%

Observation: Ensemble models performed consistently well. SVM gave strong performance on high-dimensional data.

Conclusion

ML models can effectively predict disease likelihood. SVM, Random Forest, and XGBoost achieved higher accuracy. Future scope includes Deep Learning approaches and deployment in real-time systems.

References

- UCI Machine Learning Repository
- Kaggle Datasets
- Scikit-learn Documentation