Chronic Disease Prediction Using Machine Learning

Abstract

This project aims to predict chronic disease outcomes using supervised machine learning models. A Kaggle healthcare dataset was used to perform data preprocessing, exploratory data analysis, model training, and performance evaluation. Multiple classifiers were compared to identify the most accurate and reliable model.

Tools and Technologies

Python, VS Code, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

Methodology

The workflow included data cleaning, encoding categorical variables, exploratory data analysis using visualizations and heatmaps, and training multiple classification models. The dataset was split into training and testing sets before model evaluation.

Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.95	1.00	0.92	0.96
KNN	0.76	0.88	0.71	0.79
Random Forest	0.98	0.97	1.00	0.98
Decision Tree	0.97	0.97	0.98	0.98
SVM (Linear)	0.95	0.95	0.97	0.96

Results and Conclusion

Among all models, the Random Forest Classifier achieved the highest performance with an accuracy of 98% and perfect recall, making it the most reliable model for chronic disease prediction. Decision Tree and SVM models also performed strongly, while KNN showed comparatively lower accuracy.

Future Scope

Future enhancements include hyperparameter tuning, cross-validation, feature selection, and deploying the model using web frameworks such as Flask or Streamlit.

Author: Arun Santhosh M