**<Heart Disease Prediction>**

**Submitted for**

**Statistical Machine Learning CSET211**

Submitted by:

**(<E23CSEU0295>)** **E.Sai Praneeth Reddy**

**(<E23CSEU0290>)** **Vegesna Hari Ram Varma**

Submitted to

**DR. ASHIMA YADAV**

**July-Dec 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

Description automatically generated

**INDEX**

|  |  |  |
| --- | --- | --- |
| Sr.No | Content | Page No |
| 1 | Abstract,Introduction,Related Work,Methodology | 3 |
| 2  3 | Hardware/Software Required,Experimental Results  Conclusions,Future Scope,Github link | 4  5 |

**1. Abstract**

This project aims to develop a machine learning-based heart disease prediction system that analyses a set of health parameters to assess the likelihood of heart disease. The system provides a binary prediction (0 for no risk, 1 for risk) to indicate the presence of heart disease risk. Additionally, the model includes a unique interpretability feature that highlights the factors most contributing to each individual’s risk—such as high cholesterol or elevated blood pressure. This feature enhances the model’s utility by offering healthcare professionals actionable insights into specific health indicators, allowing for more tailored care and preventive strategies.

**2. Introduction**

Heart disease is a leading cause of mortality worldwide, making early detection and risk assessment essential for effective prevention and treatment. This project leverages machine learning to predict an individual's heart disease risk based on basic health metrics, including age, cholesterol levels, and other cardiovascular indicators. By incorporating a feature that identifies contributing risk factors for each prediction, the system offers greater transparency. This enables patients and healthcare providers to understand potential health risks better, facilitating informed decision-making and personalized intervention plans.

**3. Related Work**

https://www.kaggle.com/code/desalegngeb/heart-disease-predictions/notebook

**4. Methodology**

* Data Collection: The dataset includes age, sex, chest pain type, blood pressure, cholesterol, and ECG readings.
* Data Preprocessing: Categorical features are encoded, missing values addressed, and numeric features standardized. The dataset is split into training and test sets.
* Model Selection: To select the most effective model, we tested a range of classifiers, including:
  + Logistic Regression
  + Linear Discriminant Analysis (LDA)
  + Quadratic Discriminant Analysis (QDA)
  + Random Forest
  + Decision Tree
  + AdaBoost
  + Gradient Boosting
  + Naive Bayes
  + Nu-SVC
  + Neural Network
  + Support Vector Machine (SVM)
  + Nearest Neighbours

These models were evaluated based on accuracy, precision, recall, and interpretability. Feature importance and Shapley values were then used to identify key risk factors for each prediction.

Evaluation: Evaluation metrics included accuracy, confusion matrix, and ROC-AUC, alongside interpretability assessments to validate the model’s predictions and explainability.

**5. Hardware/Software Required**

Hardware: A standard laptop or desktop with at least 4GB of RAM suffices, though a GPU would enhance performance.

Software:

* Programming Language: Python 3.12.6
* Libraries: pandas, numpy, scikit-learn, CatBoost, XGBoost, LightGBM, and SHAP.
* Development Environment: Jupyter Notebook or any Python IDE.

**6. Experimental Results**

Model Performance Metrics:

* Logistic Regression: Accuracy = 86.49%, Precision = 0.92, Recall = 0.91, F1 Score = 0.82, ROC-AUC = 0.86
* Linear DA: Accuracy = 85.14%, Precision = 0.92, Recall = 0.89, F1 Score = 0.82, ROC-AUC = 0.85
* Quadratic DA: Accuracy = 85.14%, Precision = 0.90, Recall = 0.83, F1 Score = 0.85, ROC-AUC = 0.84
* Additional models (Random Forest, AdaBoost, etc.) displayed varying metrics, all detailed to determine the most suitable model for high accuracy and interpretability.

Factor Identification Results: In cases flagged "at risk," high cholesterol and blood pressure were primary factors. SHAP values quantified each factor's impact, with visualizations such as SHAP force plots explaining how features contribute to each risk score.

Visualization: ROC-AUC, feature importance, confusion matrix, and SHAP summary plots provided a comprehensive view of the model’s performance and the features influencing heart disease risk.

**7. Conclusions**

The heart disease prediction model effectively classifies individuals at risk, offering transparency by identifying specific risk factors. This interpretability feature equips healthcare providers with targeted information for better patient care.

**8. Future Scope**

Future work could involve expanding the dataset to include more diverse demographic data, enhancing model robustness across population groups. Additionally, exploring deep learning architectures may reveal complex interactions among features, potentially increasing prediction accuracy. Deploying the model as a web or mobile application would increase accessibility for users, while real-time factor identification could provide even more immediate insights for health monitoring. Finally, extending the factor-identification feature to suggest lifestyle adjustments or treatment recommendations based on identified risks could further benefit users.

**9. GitHub Link**

https://github.com/harivegesna73/HeartdiseasePredictor