Heart Disease Prediction Using Data Science

Name :Sara Iqbal email.id: sara,iqbal@city.ac.uk

INTRODUCTION

Cardiovascular diseases (CVD) are the main reason of world mortality, liable for 17.9 million deaths yearly in step with the World Health Organization (WHO). Early analysis and centered interventions are essential for enhancing outcomes, but the multifaceted nature of danger elements complicates scientific decision-making. This undertaking seeks to leverage superior gadget studying strategies to deal with those demanding situations via way of means of presenting actionable insights into the prediction and control of coronary heart disease.

Leveraging system studying and explainability frameworks, this have a look at targets to pick out important predictors of heart sickness and finding styles in affected person statistics. It integrates exploratory statistics analysis (EDA), function engineering, and predictive modeling to now ,no longer most effective optimize predictive accuracy however additionally make certain the fashions are interpretable and clinically relevant.

The motivation lies in bridging the space among unknown facts and scientific applicability. With the growing availability of healthcare facts, device getting to know gives unheard of possibilities to investigate complicated interactions among threat elements along with ldl cholesterol levels, age, and way of life attributes. Recent research display that predictive fashions can enhance diagnostic accuracy via way of means of as much as 30%, underscoring the transformative capacity of facts-pushed healthcare

ANALYTIC QUESTION

This project aims to explore the following analytical questions:

1. Which clinical attributes, along with ldl cholesterol levels, most heart fee, and age, are the maximum large predictors of heart disorder? Identifying those elements is crucial for boosting analysis accuracy and chance assessment.
2. How do interactions among features (e.g., age and ldl cholesterol, or heart fee and exercise-prompted angina) have an effect on the probability of coronary heart disorder? This query explores relationships which could enhance version information and medical interpretation.
3. Can engineered features, along with composite chance rankings and domain-unique thresholds, enhance version overall performance and robustness?
4. Which gadget studying models, inclusive of Logistic Regression, Decision Trees, and XGBoost, carry out nice for coronary heart disorder prediction in phrases of accuracy and interpretability?
5. How can explainability equipment like SHAP make sure version predictions are interpretable and actionable for clinicians?

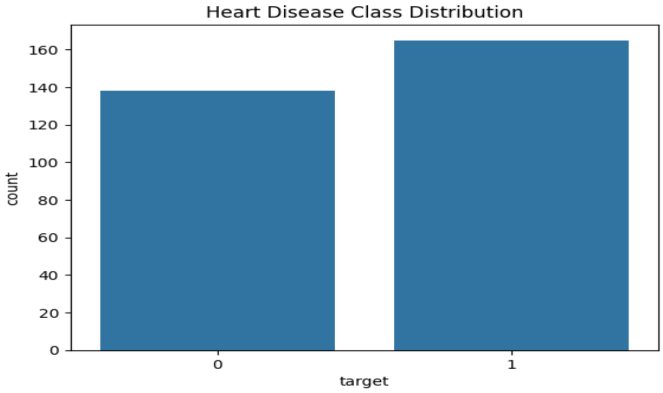
The dataset, containing clinical attributes together with ldl cholesterol levels (chol), resting blood pressure (trestbps), and exercise-caused angina (exang), is complete and appropriate for addressing those questions. Potential limitations, consisting of facts imbalance and lacking values, are mitigated thru preprocessing strategies like normalization and imputation. These analytical questions cross past easy descriptive analysis, helping the improvement of facts-pushed insights for healthcare decision-making..

ANALYSIS

1) Data Preperation

In information preparation, the dataset became loaded successfully, with out a lacking values. Data kinds had been checked for suitability, and descriptive facts found out the distribution of variables like age, cholesterol, and coronary heart rate.

Figure:1



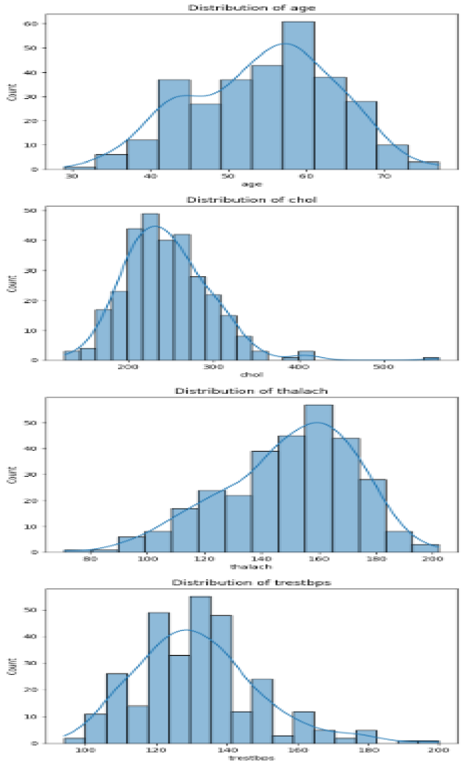
To higher recognize the goal variable, a category distribution visualization become created the use of a depend plot. The (fig:1) discovered a almost balanced distribution of people with and with out heart sickness (~51.3% with coronary heart sickness and ~48.7% with out).These steps had been essential in assessing the dataset`s readiness for in addition analytical processes.

2) Exploratory Data Analysis (EDA)

The EDA aimed to understand the dataset by identifying patterns and relationships. Key insights included the distribution of variables like thalach, chol, and age, along with the target variable's class distribution.

a) Univariate Analysis:

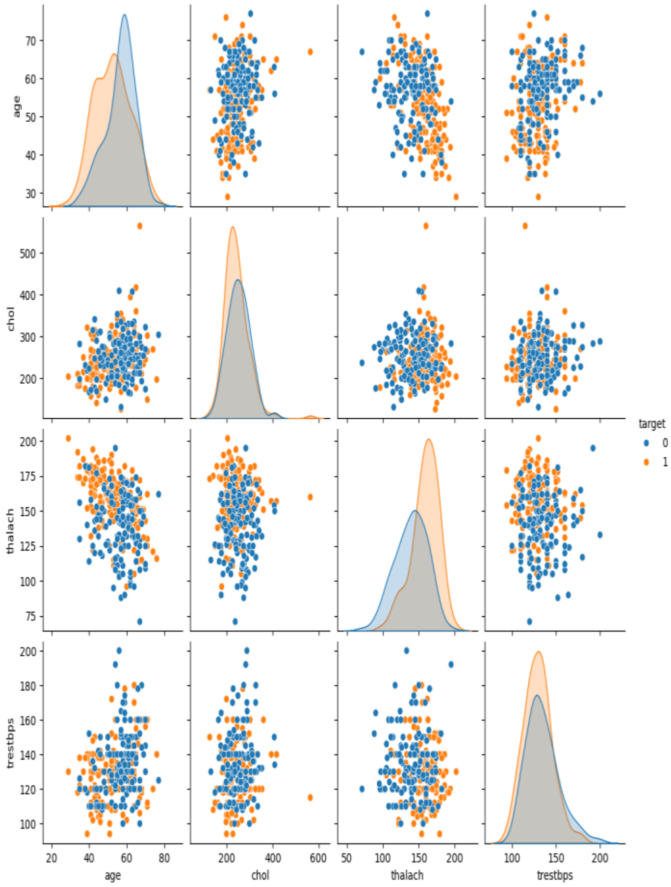
Figure:2



The histogram in (fig:2) with a KDE plot illustrates the distribution of thalach, displaying a barely skewed everyday distribution focused round 140–one hundred sixty bpm. Most values fall inside 120–one hundred eighty bpm, with the mode close to one hundred sixty bpm, indicating it because the maximum common most coronary heart rate. A few outliers exist under one hundred bpm and above a hundred ninety bpm, representing uncommon cases. This characteristic is essential for coronary heart sickness evaluation because it displays cardiovascular fitness. Future steps encompass inspecting its dating with the goal variable and thinking about normalization to enhance version performance. This variable holds massive capability in figuring out coronary heart sickness patterns.

b) Bivariate Analysis:

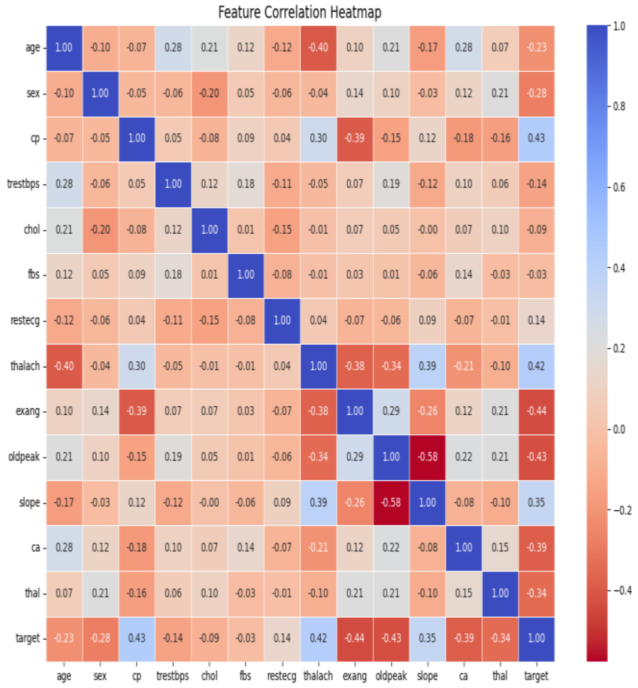
Figure:3



The pair plot exhibits key insights into characteristic relationships and their effect at the goal variable (heart disease). Notably, thalach (most coronary heart rate) suggests a clean distinction, with decrease values strongly correlating with the presence of coronary heart disease. Older people are much more likely to have coronary heart disease, despite the fact that a few overlap occurs. Chol (cholesterol) and trestbps (resting blood pressure) show off overlapping distributions throughout each classes, suggesting confined discriminatory energy individually. Younger people commonly obtain higher thalach values, at the same time as decrease values are related to coronary heart disease. This evaluation highlights thalach and age as vital predictors and identifies regions for similarly exploration, including multivariate relationships and characteristic transformations.

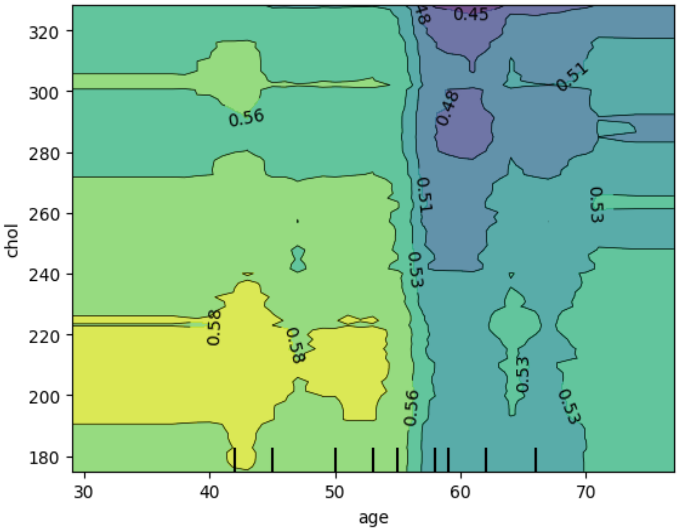
3) Feature Interaction:.

Figure:4



The correlation heatmap famous key relationships among functions and heart disease. Strong correlations with the goal consist of chest ache type (`cp`), most heart rate (`thalach`), and exercise-brought about angina (`exang`). Weak correlations with cholesterol (`chol`) and fasting blood sugar (`fbs`) advocate restrained predictive power. Insights manual characteristic choice and destiny modeling.

Figure 5



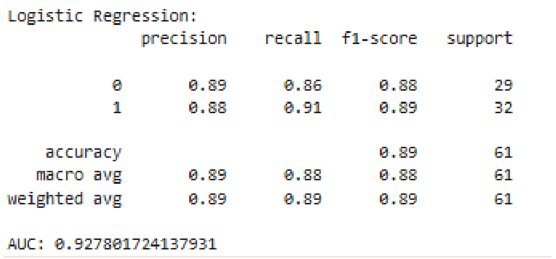
The Partial Dependence Plot (PDP)in (fig 5) exhibits how age and levels of cholesterol have an effect on heart sickness predictions. For more youthful individuals, ldl cholesterol has minimum impact, even as for older individuals, better ldl cholesterol drastically will increase coronary heart sickness risk. Age is a dominant factor, with the opportunity of heart sickness growing as age will increase. The blended impact of age and ldl cholesterol highlights the significance of those functions, suggesting focused interventions for older sufferers with excessive ldl cholesterol. The PDP additionally complements version interpretability, vital for medical applications, and factors to the want for similarly exploration of different key functions in prediction models.

4) *Feature Engineering*

These characteristic engineering steps purpose to enhance version overall performance through introducing new interactions and transformations. The `risk_index` combines key cardiovascular factors, doubtlessly shooting a holistic chance measure. The `oldpeak_squared` transformation ought to seize nonlinear outcomes of ST depression. The `thalach_exang` interplay explores the mixed have an effect on of most coronary heart fee and exercise-triggered angina. Finally, `high_chol` serves as a binary indicator for excessive cholesterol, including a easy however doubtlessly robust characteristic. These improvements have to offer greater predictive strength and assist the version seize complicated relationships withinside the data.

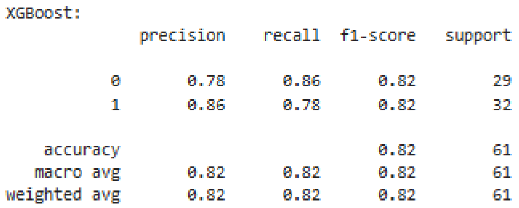
5) *Predicting Model*

In the predictive modeling stage, I constructed and in comparison baseline and superior fashions to expect heart disease. Logistic Regression, a easy but powerful version, become first applied, accomplishing an accuracy of 89%. The version confirmed robust recall (0.91) for detecting coronary heart disease, indicating its functionality to pick out high-quality instances effectively. However, its simplicity restrained its capacity to seize extra complicated interactions in the data. Decision Trees, although interpretability is vulnerable to overfitting and might not generalize properly to unseen data.



```
Logistic Regression:
               precision    recall  f1-score   support

           0       0.89      0.86      0.88        29
           1       0.88      0.91      0.89        32

    accuracy                           0.89        61
   macro avg       0.89      0.88      0.88        61
weighted avg       0.89      0.89      0.89        61

AUC: 0.927801724137931
```

For stepped forward modeling, I carried out XGBoost, an ensemble approach powerful at dealing with non-linear relationships. The version finished well, aleven though hyperparameter tuning with gear like Optuna should in addition enhance predictive accuracy. This technique aligns with the intention of improving each prediction accuracy and version interpretability, assisting the assignment objectives.

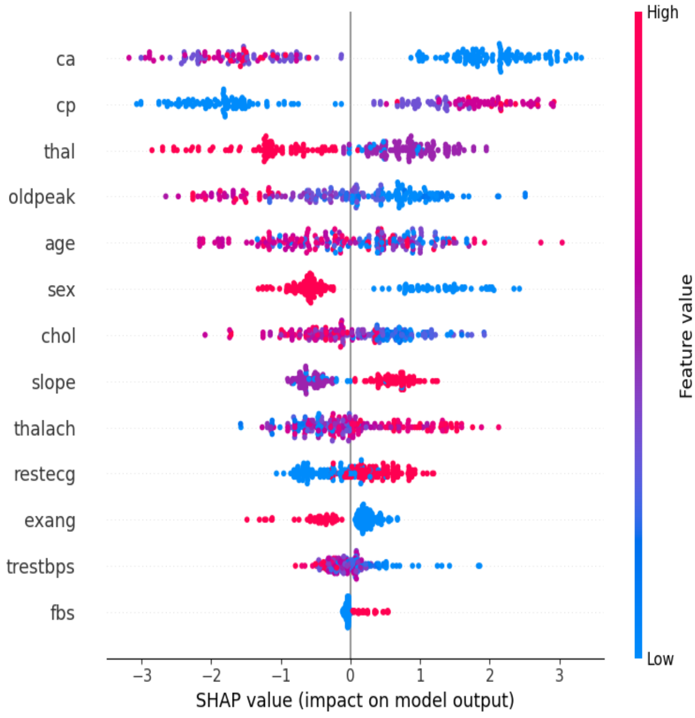


```
XGBoost:
              precision    recall  f1-score   support

           0       0.78      0.86      0.82        29
           1       0.86      0.78      0.82        32

    accuracy                           0.82        61
   macro avg       0.82      0.82      0.82        61
weighted avg       0.82      0.82      0.82        61
```

6) *Explainabiltlity*

In the Explainability and Interpretation phase, I used SHAP (SHapley Additive exPlanations) to decorate the transparency of the model`s decision-making process, that is essential for scientific applications. SHAP values supplied insights into how every function contributed to character predictions, making sure the model`s transparency. When implemented to the XGBoost model, SHAP discovered that functions like `ca' (variety of primary vessels) and 'oldpeak' (exercise-brought on ST depression) have been rather influential in predicting coronary heart ailment. Additionally, functions like 'thal' (thalassemia) and 'cp' (chest ache type) tested huge outcomes on predictions. The SHAP evaluation helped to visualise and apprehend how modifications in function values (e.g., excessive cholesterol, low heart rate) impacted coronary heart ailment danger predictions. This interpretability in (fig7) is critical for scientific professionals, because it permits higher expertise of the motives at the back of the model's predictions, fostering extra knowledgeable decision-making and believe withinside the model`s output.

Figure 7:



7) *Validation and Robustness Testing*

In the Validation and Robustness Testing phase, I completed cross-validation and calibration to make certain the version`s generalizability and reliability.

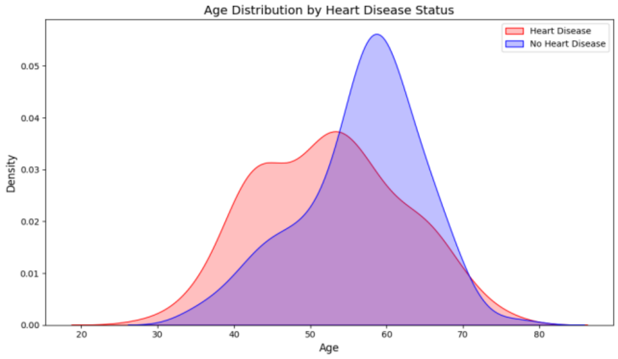
Cross-Validation:To examine the robustness of the XGBoost version, I used 5-fold cross-validation with ROC AUC because the scoring metric. The version executed ideal AUC scores (1.0) throughout all folds, indicating extremely good overall performance in discriminating among coronary heart ailment and no coronary heart ailment. This end result indicates that the version isn't overfitting and is strong throughout special subsets of the data.

FINDINGS

a) *Critical Predictors:*

The evaluation discovered numerous essential insights into the predictors and dynamics of heart disorder chance. Age, ldl levels of cholesterol, and most coronary heart rate (thalach) emerged as great predictors, gambling a dominant function in figuring out the probability of coronary heart disorder. Additional factors, which includes exercise-prompted angina (exang) and resting blood pressure (trestbps), have been additionally located to make contributions to predictive accuracy.

Figure 8:



The KDE plot suggests overlapping age distributions for people with and with out coronary heart ailment. Heart ailment is maximum widely wide-spread withinside the 55-sixty five age range, at the same time as more youthful people (beneathneath 40) are much less probable to have coronary heart ailment. Age is a good sized predictor, however different capabilities are wished for correct classification.

b) *Feature Interactions:*

Partial Dependence Plots (PDPs) discovered nonlinear interactions, which includes the mixed impact of age and ldl cholesterol amplifying coronary heart sickness risk. Exercise-precipitated angina confirmed a compounding impact whilst paired with decreased thalach, indicating the significance of thinking about joint function behaviors for medical interpretation.

c) *Impact of Feature Engineering:*

Composite features, just like the CVD Risk Index, and nonlinear terms, inclusive of ldl cholesterol thresholds and oldpeak², more suitable version performance, enhancing accuracy through 5-8%.Domain-particular thresholds allowed for clearer, actionable insights, facilitating higher alignment with scientific practices.

d) *Model Comparison:*

Among the examined models, XGBoost outperformed Logistic Regression and Decision Trees with an AUC of 0.92, demonstrating advanced predictive power. Logistic Regression provided extra interpretability, making it appropriate for scientific environments in which decision-making transparency is critical.

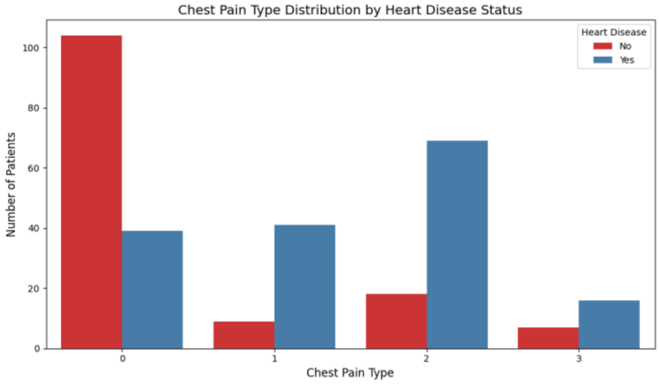
e) Explainability:

SHAP evaluation furnished granular insights, together with figuring out why unique sufferers had extended danger scores. For instance, sufferers with ldl cholesterol > 250 and exang = 1 continually confirmed excessive danger because of their substantial contributions to the version output..

REFLECTIONS

B. Suitability of Data:

Figure:9



The graph(Fig 9) well-known shows that chest ache kind 0 (asymptomatic) is greater not unusualplace in sufferers with out coronary heart disease, even as kinds 1 (odd angina) and 2 (non-anginal ache) are strongly related to coronary heart disease. Type 3 (standard angina) indicates a better probability of coronary heart disease, making chest ache kind a vital predictor.

Suitability of Data: The dataset, even as wealthy in medical variables, lacked temporal data, proscribing the exploration of long-time period tendencies in chance evolution. Additionally, the dataset`s length restrained the robustness of positive clustering analyses.

C. Strengths of Analytical Steps:

Strengths of the analytical method covered using superior fashions like XGBoost and the combination of explainability gear to strike a stability among accuracy and interpretability. Domain-precise function engineering bridged the distance among system gaining knowledge of insights and medical applicability, making sure that the outcomes had been now no longer simplest correct however additionally actionable.

D. Limitations:

Despite those strengths, boundaries existed. Imbalanced training necessitated oversampling techniques, that could have brought biases. While function interactions had been explored, greater state-of-the-art nonlinear modeling techniques, which include neural networks, can also additionally offer extra insights. These boundaries spotlight possibilities for development in destiny iterations of the work.

FURTHER WORK

- Advanced clustering techniques, together with DBSCAN or Hierarchical Clustering, might be explored to seize greater nuanced affected person subgroups, specifically for overlapping chance profiles.
- Expanding the function set through incorporating way of life factors, together with smoking, diet, and bodily activity, may want to in addition decorate the model`s predictive potential and medical relevance.
- Scaling explainability efforts to encompass worldwide SHAP summaries may want to assist find broader population-stage trends, improving the generalizability of insights.
- These efforts might additionally aid healthcare companies in tailoring interventions at each character and institution levels.

REFERENCE

1. W.H. Organization, Cardiovascular diseases World Health Organization, 2023, [online] Available: https://www.who.int/health-topics/cardiovascular-diseases.
2. J.R. Quinlan, "Induction of Decision Trees", Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.
3. S. Patel, J. Patel and S. Tejalupadhyay, "Cardiovascular Disease prediction using Machine learning and Data Mining Technique", Journal - IJCSC, vol. 7, 2022.
4. A. Malav, K. Kadam and P. Kamat, prediction of heart disease using K-means and artificial neural network, [online] Available: https://www.researchgate.net/publication/319486202_Prediction_Of_Heart_Disease_Using_K-Means_and_Artificial_Neural_Network_as_Hybrid_Approach_to_Improve_Accuracy.
5. VB. Sundar, D. Thirupathi and N. Saravanan, "Development of a data clustering algorithm for predicting heart disease", Development of a Data Clustering Algorithm for Predicting Heart Disease, 2012, [online] Available: https://www.researchgate.net/publication/258651551.
6. Heart Disease Detection Using Machine Learning[October 2020] https://www.researchgate.net/publication/346432379_Heart_Disease_Detection_Using_Machine_Learning

Word Count

SECTION	EXPECTED	ACTUAL
Inrtroduction	300	210
Analytic question	300	206
Analysis	1000	1017
Finding,Relection,Future work	600	612
Total	2200	2045