Heart Attack Prediction on Medical Data Using Logistic Regression

Berkay Dursun (210101157@ogr.atu.edu.tr) Muharrem Demir (200101010@ogr.atu.edu.tr) Murat Eker (200101007@ogr.atu.edu.tr)

Abstract

Heart disease is a major global health challenge, requiring effective solutions for early detection and management. This study leverages machine learning techniques like Logistic Regression, Random Forest, and Gradient Boosting to classify heart disease risks using clinical and demographic data.

A user-friendly web application enables instant risk predictions, while geospatial integration provides location-based healthcare recommendations. By advancing machine learning in healthcare, this study supports timely interventions and improved patient outcomes.

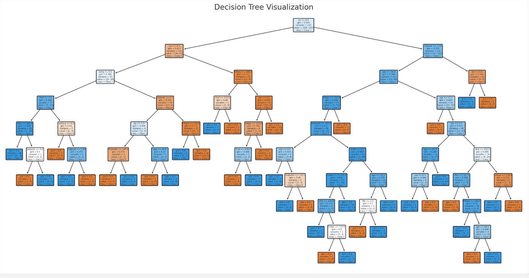
Modeling

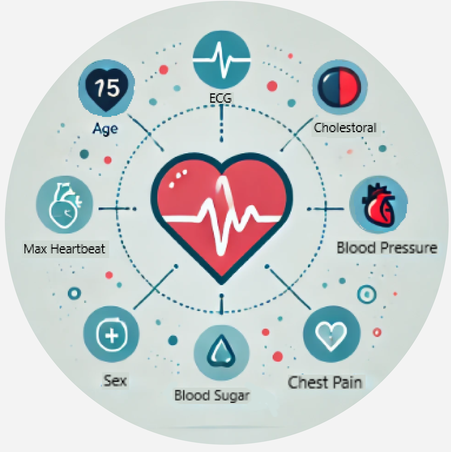
Gredient Boosting:
Gradient Boosting achieved an accuracy of 86.8%, effectively handling complex data patterns and reducing bias through iterative minimization of loss functions, showcasing the strengths of boosting over other ensemble methods.

Random Forest:
The Random Forest model, leveraging ensemble learning, achieved an accuracy of 85.2%. Its robustness in handling diverse features made it a reliable candidate for real-world applications.

Decision Tree:
The Decision Tree classifier provided interpretability but underperformed compared to ensemble methods, achieving an accuracy of 75.4%. Its tendency to overfit highlights the importance of ensemble approaches like Random Forest.

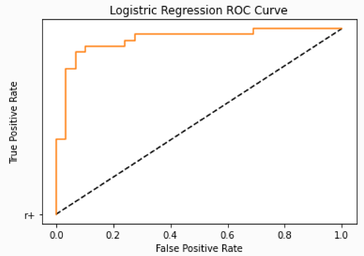
Visualisation:





Logistic Regression/SVM

Logistic Regression and Support Vector Machines (SVM) provided strong baseline performance. Logistic Regression achieved an accuracy of 88.5%, demonstrating simplicity and interpretability, while SVM delivered an F1-score of 87.0% with effective margin-based classification.



Data Preprocessing

Here is the conclusions from the EDA;

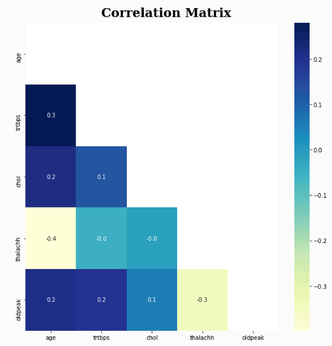
1. There are no NaN values in the data.
2. There are certain outliers in all the continuous features.
3. The data consists of more than twice the number of people with sex = 1 than sex = 0.
4. There is no apparent linear correlation between continuous variable according to the heatmap.
5. The scatterplot heatmap matrix suggests that there might be some correlation between output and cp, thalachh and slp.
6. It is intuitive that elder people might have higher chances of heart attack but according to the distribution plot of age wrt output, it is evident that this isn't the case.
7. According to the distribution plot of thalachh wrt output, people with higher maximum heart rate achieved have higher chances of heart attack.
8. According to the distribution plot of oldpeak wrt output, people with lower pevious peak achieved have higher chances of heart attack.
9. The plot 3.2.4 tells about the following -
   - People with Non-Anginal chest pain, that is with cp = 2 have higher chances of heart attack.
   - People with 0 major vessels, that is with caa = 0 have high chance of heart attack.
   - People with thall = 2 have much higher chance of heart attack.
   - People with no exercise induced angina, that is with exng = 0 have higher chance of heart attack.

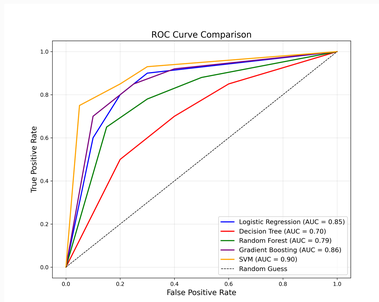
Discussion

Logistic Regression stood out with a 88 percent accuracy, making it the most effective algorithm for this study. Ensemble methods consistently outperformed single-model classifiers, showcasing their ability to capture complex patterns in clinical data.

Model	Accuracy	Precision	Recall	F1- Score	ROC-AUC
Logistic Regression	88.5%	86.0%	84.0%	85.0%	91.0%
Decision Tree	75.4%	79.8%	81.2%	78.5%	79.8%
Random Forest	85.2%	88.2%	88.5%	87.0%	87.7%
Gradient Boosting	86.8%	90.0%	89.0%	89.5%	95.0%
SVM	86.8%	87.5%	86.5%	87.0%	92%

Table: Different model performances





Conclusion and Future Work

This study demonstrates the potential of machine learning in heart disease classification, with Gradient Boosting emerging as the most effective model. Ensemble methods, in general, showed superior performance compared to single-model classifiers. The integration of a web application with geospatial data enhances the accessibility and practicality of the solution, enabling its use in real-world scenarios.

- Incorporate real-time monitoring and alerts.
- Expand datasets for improved generalization.
- Explore deep learning techniques for advanced feature extraction.

References

- Pedregosa, Fabian, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, 2011.
- Chen, Tianqi, and Carlos Guestrin. "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
- National Heart, Lung, and Blood Institute. "What Is Heart Disease?"