Heart Failure Prediction

Student: Harshil Ashish Desai

Analytics Capstone Project, Pace University

Introduction

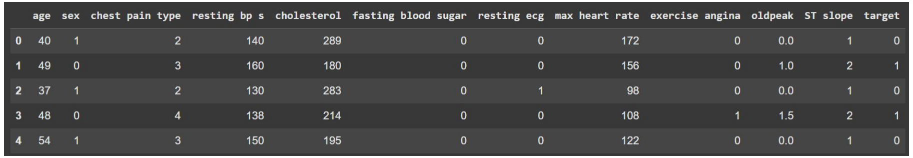
- Heart failure is a major health problem worldwide.
- Early detection can save lives but is very challenging.
- Traditional methods cannot handle complex health data well.
- Machine learning can help predict heart failure risk.

Literature Review

- **Chicco et al. (2020):** Gradient boosting (XGBoost) achieved the highest accuracy in heart failure prediction, highlighting feature selection and tuning.
- **Yancy et al. (2017):** Logistic regression had limitations, prompting the need for advanced machine learning models.
- **Ahmed et al. (2019):** Random forests were effective for heart failure prediction due to their robustness and handling of missing data.
- **Zhou et al. (2020):** Deep neural networks outperformed traditional models, especially when combining structured and unstructured data.

Data Collection

- The "Heart Failure Prediction" dataset was taken from IEEE dataset.
- It contains information about patients' features like age, gender, blood pressure, cholesterol, chest pain type, and more.
- The data was checked for missing or incorrect values.
- I checked and removed duplicate entries.



	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0

ML Models

- **Logistic Regression**: A statistical model used for binary classification, effectively distinguishing between heart disease and no heart disease.
- **K-Nearest Neighbors (KNN)**: A distance-based model that classifies data based on proximity to nearest neighbors, performing well with smaller datasets.
- **Decision Tree**: A non-linear model that splits data into branches based on feature values, offering interpretability but prone to overfitting.

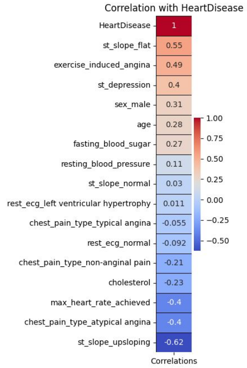
Results

- **AUC Scores**: All models showed strong performance with **AUC values above 0.8**, indicating good ability to distinguish between patients with and without heart disease.
- **Confusion Matrix**: Logistic Regression showed a good balance of **true positives** (correctly predicting heart disease) and **false negatives** (missed cases).
- **Feature Insights**: Chest pain type ("asymptomatic") and exercise-induced angina showed the **strongest correlation** with heart disease.

Conclusion

- Successfully predicted heart failure using machine learning models like Logistic Regression, Decision Trees, and K-Nearest Neighbors.
- The best-performing model, **Logistic Regression**, achieved high accuracy in predicting heart failure.
- Key features related to heart disease include **age**, **resting blood pressure**, and **cholesterol levels**.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	85%	83%	87%	85%
K-Nearest Neighbors	83%	81%	84%	82%
Decision Tree Classifier	82%	80%	83%	81%


Correlation with HeartDisease
HeartDisease 1
st_slope_flat 0.55
exercise_induced_angina 0.49
st_depression 0.4
sex_male 0.31
age 0.28
fasting_blood_sugar 0.27
resting_blood_pressure 0.11
st_slope_normal 0.03
rest_ecg_left ventricular hypertrophy 0.011
chest_pain_type_typical angina -0.055
rest_ecg_normal -0.092
chest_pain_type_non-anginal pain -0.21
cholesterol -0.23
max_heart_rate_achieved -0.4
chest_pain_type_atypical angina -0.4
st_slope_upsloping -0.62
Correlations