Missouri University of Science and Technology

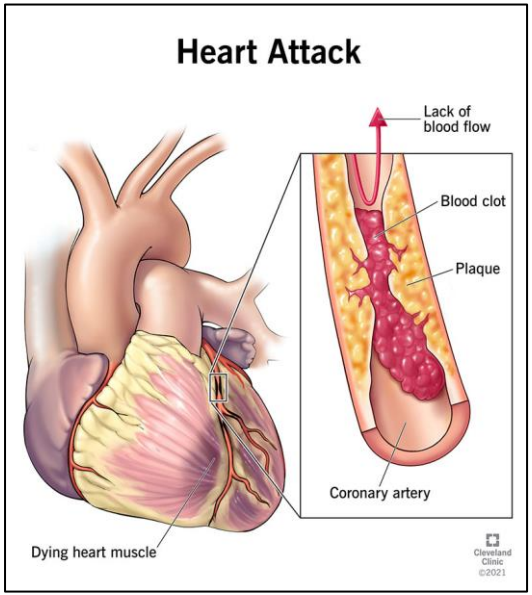
Heart Attack Prediction Project

Saifullah Shoaib

Dec. 11, 2024

Introduction

- What is a heart attack?
  - A heart attack is an extremely dangerous condition that happens because you don't have enough blood flow to some of your heart muscle.
  - Every year, more than 800,000 people in the U.S. have a heart attack. [1]
  - Life threatening emergency and global health concern.
- Early detection improves outcomes.
- Machine learning offers a potential solution for rapid and reliable diagnosis.



[1] Cleveland Clinic (https://my.clevelandclinic.org/health/diseases/16818-heart-attack-myocardial-infarction)

Dataset Selection

- Kaggle “Heart Attack Analysis & Prediction Dataset”:
  - 13 features and 1 target variable.
  - Binary classification: High/low likelihood of heart attack.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

1. EDA (Exploratory Data Analysis)

- Understand the features and dataset shape.
- Univariate analysis.
  - Categorical features.
  - Continuous features.
  - Target.
- Bivariate analysis.
  - Categorical features.
  - Continuous features.
  - Features to features relationships.

2. Machine Learning Models

- Preprocessing:
  - Encoding categorical features.
  - Normalization.
  - Train/test split.
- Model Training:
  - Logistic Regression
  - Random Forest
  - Gradient Boosting
  - XGBoost
  - Support Vector Machine (SVM)
  - Naïve-Bayes
  - K-nearest neighbors

2. Machine Learning Models

- Evaluation Metrics:
  - Accuracy
  - Precision
  - Recall
  - F1-score
  - ROC-AUC (ROC graph)
- Hyperparameter Tuning:
  - Logistic Regression
  - K-nearest neighbors

Results

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.8852	0.9091	0.8824	0.8955	0.9390
Random Forest	0.8361	0.8529	0.8529	0.8529	0.9270
Gradient Boosting	0.8361	0.8750	0.8235	0.8485	0.9216
XGBoost	0.8525	0.9032	0.8235	0.8615	0.9379
SVM	0.8525	0.9310	0.7941	0.8571	0.9336
Naïve Bayes	0.8525	0.8571	0.8824	0.8696	0.8845
K-nearest neighbors	0.8852	0.8857	0.9118	0.8986	0.9265

3. Model Explainability

- Used SHAP for feature importance.
- Applied to Logistic Regression and K-nearest neighbor models.
- Yielded different feature importances.

4. Interactive Visualizations

- Feature correlation heatmap.
- Continuous variable pair plot.
- 2D scatter plot.
- 3D scatter plot.
- Violin plot.

Conclusion

- Key insights:
  - Logistic regression and K-nearest neighbor models performed the best.
  - Machine learning shows promise as a diagnostic aid.
  - Skills for future projects.
- Future work:
  - Expanding dataset.
  - Collaborating with physicians for real-world testing and implementation.

Thank you!

Questions?