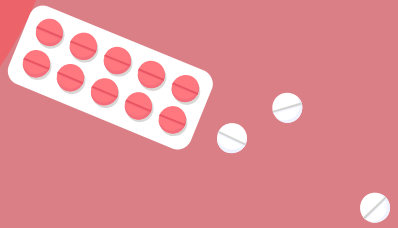
Predicting Heart Disease

Kailey Carbone, Michaela Johnson, Tahseen Shaik, Aleid van der Zel

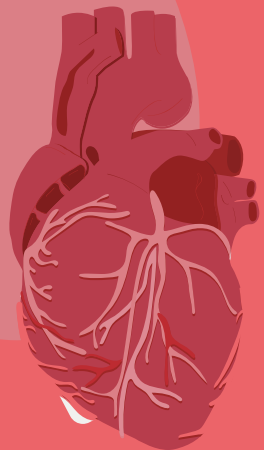


Summary

The purpose of this project was to analyze a dataset using various models to predict heart disease. The dataset consisted of 270 individuals classified as having heart disease based on cardiac catheterizations using 13 variables.

The models that were tested were: Random Forest, Neural Network, Logistic Regression, and SVM.

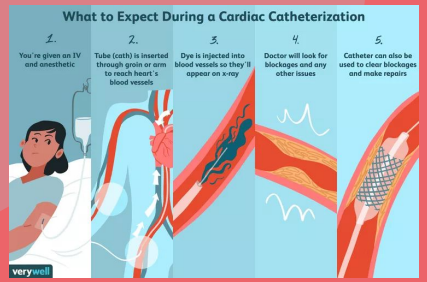
It was determined that Logistic Regression was the best predictive model with an accuracy score of 0.92



Data Set

“This dataset contains 270 case studies of individuals classified as either having or not having heart disease based on results from cardiac catheterizations - the gold standard in heart health assessment. Each patient is identified by 13 independent predictive variables revealing their age, sex, chest pain type, blood pressure measurements, cholesterol levels, electrocardiogram results, exercise-induced angina symptoms, and the number of vessels seen on fluoroscopy showing narrowing of their coronary arteries.”

Source of Data: Predicting Heart Disease Using Clinical Variables (kaggle.com)
Attributes: https://archive.ics.uci.edu/dataset/45/heart+disease
(Credit: Robert Hoyt MD)



Data Set

Exploring patient demographics using SparkSQL

Age:

min_value	max_value	mean_value	median_value	std_deviation	count
29	77	54.4	55.0	9.1	270

Sex:

Sex	Frequency
0	87
1	183

Heart Disease:

Sex	Heart_Disease	Frequency
0	0	67
0	1	20
1	0	83
1	1	100

Key:

sex: 0 = female, 1 = male

heart disease: 0 = false, 1 = true

Data Set

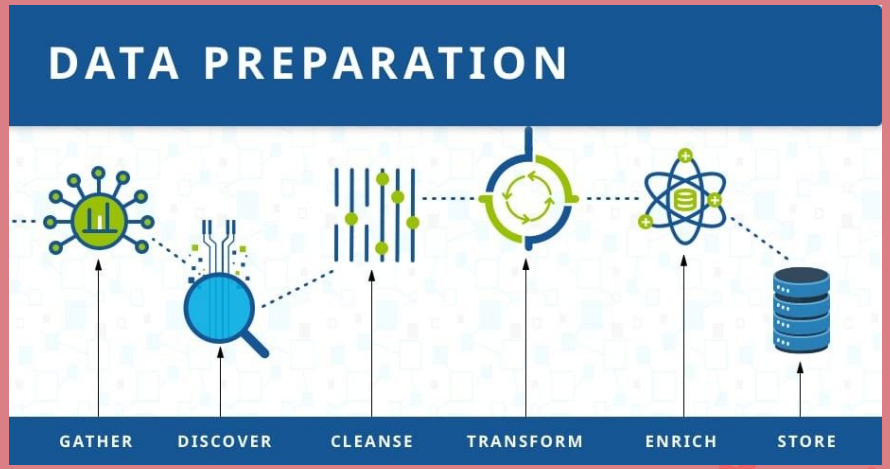
Understanding the data using Tableau LINK

Trends:

- **A large number of individual with heart disease were asymptomatic for chest pains compared to those without heart disease**
- **A higher proportion of patients with heart disease had left ventricular hypertrophy compared to those without heart disease**
- Avg resting BP on admission slightly higher with heart disease
- Avg cholesterol slightly higher with heart disease but big min and max range
- Max heart rate lower for individuals with heart disease
- For males with heart disease the thallium test showed a large number with reversible defects
- There appear to be more fluoresing blood vessels for males with heart disease
- **No obvious relationship between fasting blood sugar value and heart disease**

Preprocessing Data

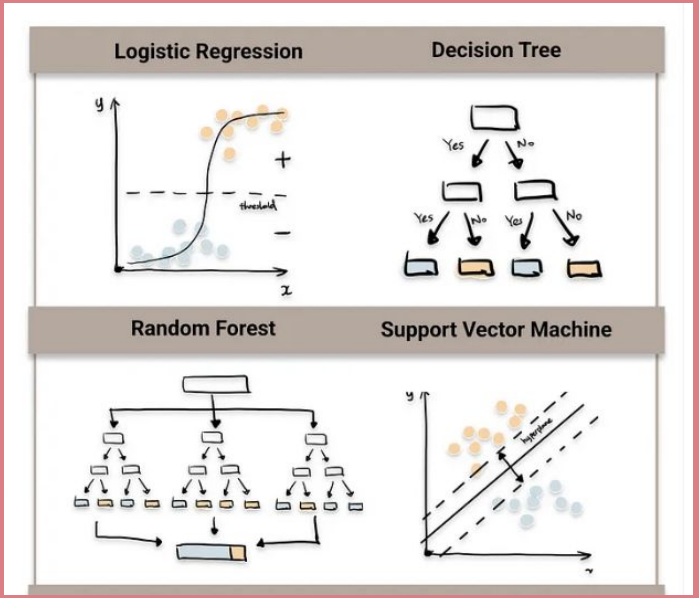
- Changed categorical values to numeric values
- Changed all column types to integers
- Scaled using `get_dummies` on necessary columns
- Converted dataframe to Pandas



Models

Binary Models for Classification

- Neural Network
- Random Forest
- Logistic Regression
- Support Vector Model



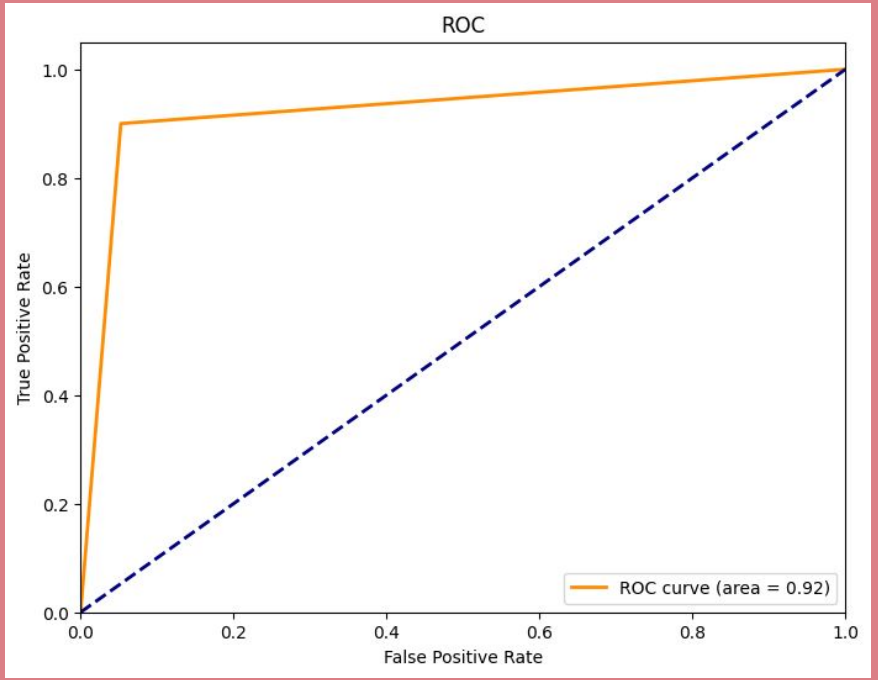
Comparing Model Metrics

The models resulted in the following metrics:

Model	Accuracy	Precision	Recall
Neural Network	0.88	N/A	N/A
Random Forest	0.89	0.87	0.90
Logistic Regression	**0.92**	**0.93**	**0.90**
Support Vector Model	0.89	0.90	0.92

Best Model

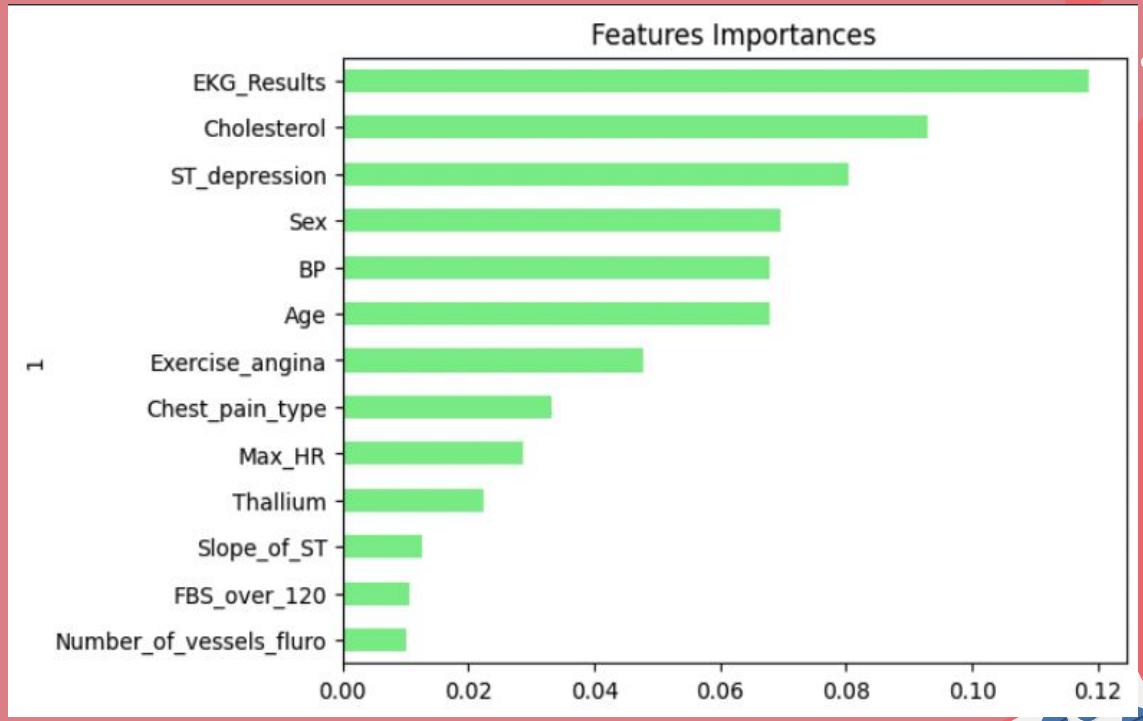
- ROC: trade-off between sensitivity (or TPR) and specificity (1 – FPR)
- Want the ROC to be 1 for the ideal model
- Logistic Regression ROC: .923



Feature Importances

- **EKG_Results** was the highest column for feature importance

- **Number_of_vessels_fluro, FBS_over_120, and Slope_of_ST** were the lowest columns for feature importance



"Slimmer pickin's"

- We re-ran the binary models after trimming the less predictive data.
- We determined the 6 most predictive datasets by using Feature Importance to sort the columns.
- This did NOT improve the Accuracy score, in fact it reduced it significantly.

Model	Accuracy	Precision	Recall
Neural Network	0.75	N/A	N/A
Random Forest	0.71	0.82	0.82
Logistic Regression	0.73	0.78	0.74
Support Vector Model	0.76	0.81	0.76

Conclusion

- Logistic Regression model on original data set with 14 features had the highest accuracy at 0.92
- All of the models had an accuracy of 0.88 or higher
- Decreasing the number of features based on based on RF model using feature_importances did not increase the accuracy on any of the model (actually decreased it)

Clinical Conclusion

- Features such as Sex, BP and Cholesterol which are available before admission will be useful to predict the onset of heart disease
- We need to explore other clinical features for these prediction to be useful for medical professional to raise awareness and potential prevent heart disease
- Select high risk patients for frequency EKG

Potential Next Steps

Possible next steps to optimize model:

- Split sex in dataset based on our data exploration
- Bigger Dataset
- Use different methods for feature selection
  - Handling outliers
  - Feature splitting
  - Backward selection

WITH THAT BEING SAID...

JUST POPPING IN

NO WORRIES


YOU'RE AWESOME
THANK YOU!
imgflip.com