Comp 6235: Coursework3

Predicting Heart Disease Using Machine Learning Algorithms

Group Members	
Kanit Jompuk	31847447
Sharang Gupta	32196946
Abdulaziz Alkhalefah	31432298
Sarah Alharbi	31711413

17.9 million

deaths every year

(World health organization, 2020)

TABLE OF CONTENTS

1

INTRODUCTION

Methodologies, methods and tools that help a data scientist or decision maker

Cardiovascular disease

Group of disorders of the heart and blood vessels

(Cardiovascular diseases (CVDs), 2020)

Confusion Matrix

Actual Values	Predicted Values: Positive	Predicted Values: Negative
Positive	TP Actual : Heart Disease Predict : Heart disease	FN Actual : Heart Disease Predict : No Heart Disease
Negative	FP Actual : No Heart Disease Predict: Heart disease	TN Actual : No Heart Disease Predict : No Heart Disease

2

Case Studies

Case Studies

Predictive analytics to prevent and control chronic diseases (2016)

A novel optimal feature selection technique for medical data classification using ANOVA based whale optimization (2020)

Effective Heart disease prediction using Hybrid Machine learning Techniques (2019)

Dataset summary

UCL Dataset with 13 features
303 patients

Effective Heart disease prediction using Hybrid Machine learning Techniques (2019)

1. Age
2. Sex
3. Cp (Chest pain categorized)
4. Trest bps (Level of blood pressure)
5. Chol (Serum Cholesterol)
6. FBS (blood sugar level on fasting)
7. Resting (Result of electriocardiogram)
8. Thali (The accomplishment of the maximum rate of heart)
9. Exang (Angina induced by exercise)
10. Oldpeak (Exercise – induced ST depression in comparision with the state of rest)
11. Slope (ST segment measured in terms of the slope during peak exercise)
12. Ca (Fluoroscopy colored major vessels)
13. Thal (Status of heart illustrated through three distinctly)

1,190

Patients

12

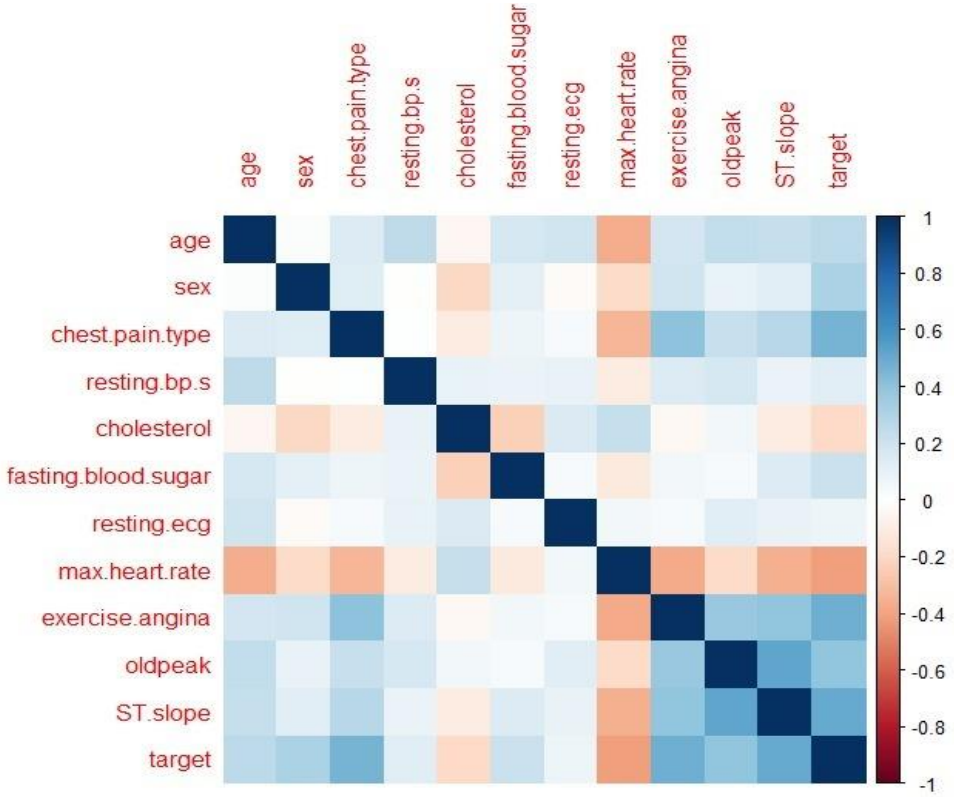
Attributes

All fields numeric

3

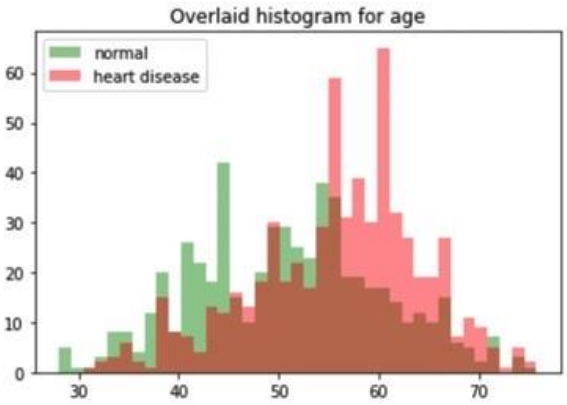
ANALYSIS

The Correlation of the Independent variables and Dependent variable



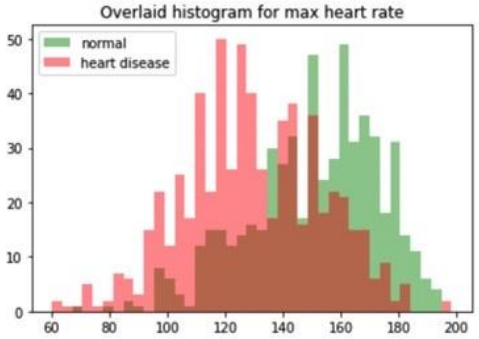
The correlation of the Independent variables and Target
Age = 0.26
Sex = 0.31
Chest.pain.type =0.46
Resting.bp.s = 0.12
Cholesterol = -0.20
Fasting.blood.sugar = 0.22
Resting.ecg = 0.07
Max.heart.rate = -0.41
Exercise.angina = 0.48
Oldpeak = 0.40
ST.slope = 0.51

AGE



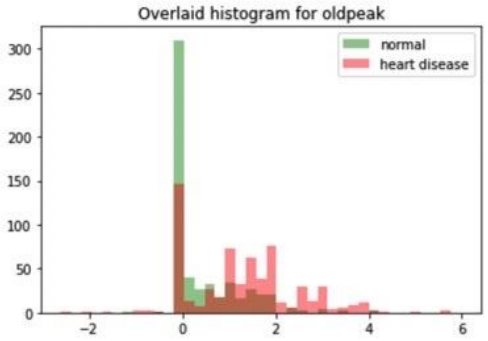
As evident, older people seem to be more susceptible to heart disease

Maximum Heart Rate



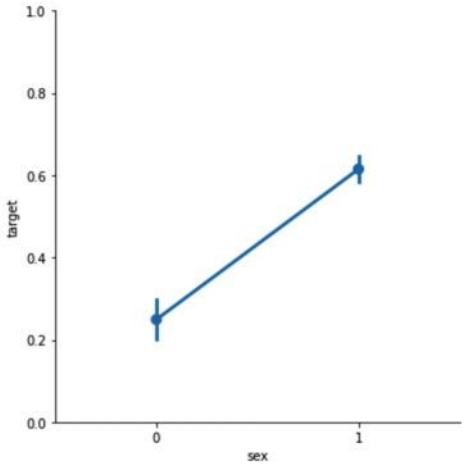
As evident, lower maximum heart rate indicates the presence of heart disease

Exercise ST Depression



A higher exercise induced ST depression seems to indicate the presence of heart disease

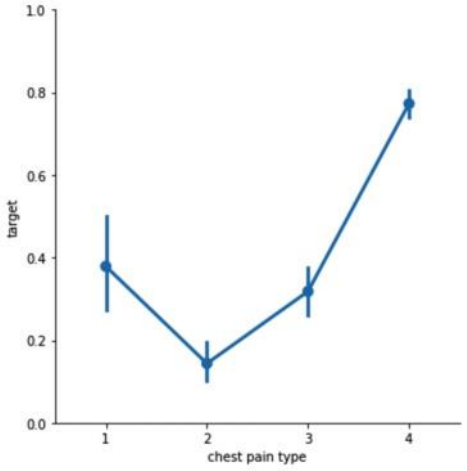
Sex



0 : female
1 : male

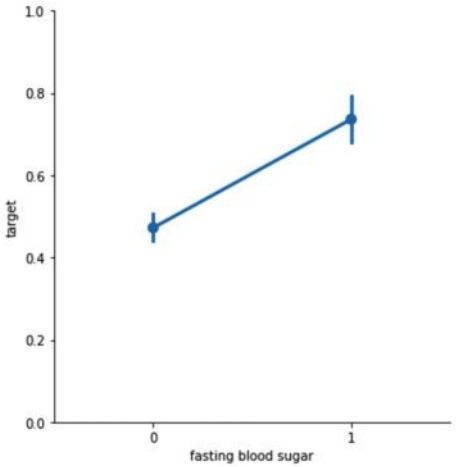
It appears males are more susceptible to heart diseases than females

Chest Pain Type



1 : typical angina
2 : atypical angina
3 : non-anginal pain
4 : asymptomatic

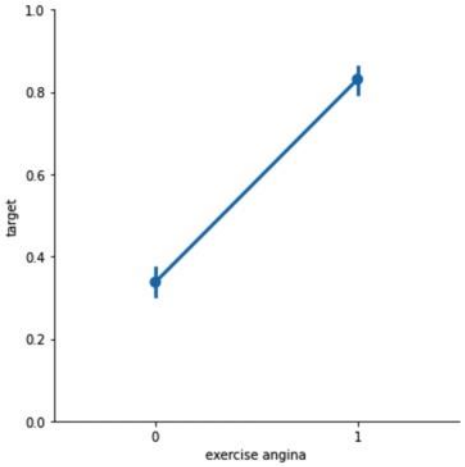
Blood Sugar



0 : sugar levels < 120 mg/dl
1 : sugar levels > 120 mg/dl

A higher blood sugar level seems to indicate heart disease

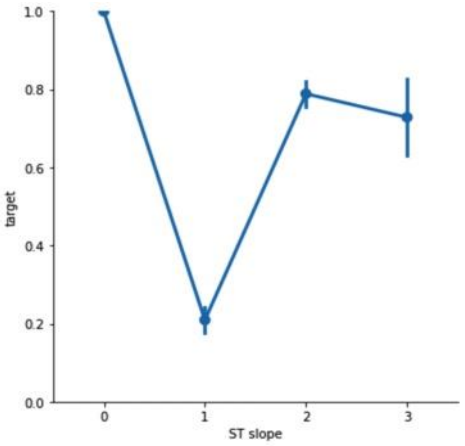
Exercise Induced Angina



0 : No exercise induced angina
1 : Had exercise induced angina

Presence of exercise induced angina is a strong indicator of heart disease!

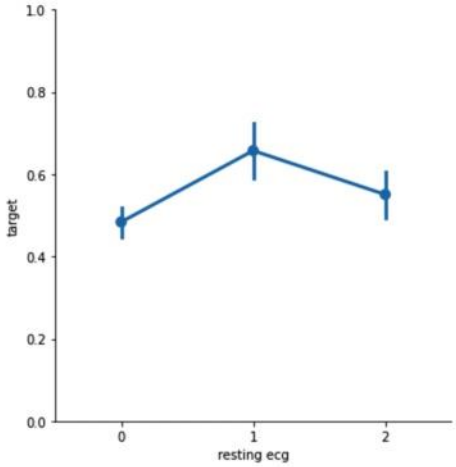
ST slope



1 : upsloping
2 : flat
3 : downsloping

An upsloping slope of the peak exercise ST seems to indicate lower chances of heart disease!

Resting ECG

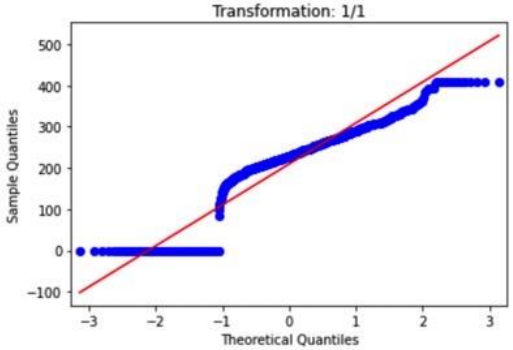


0 : normal
1: ST-T wave abnormality
2 : left ventricular hypertrophy

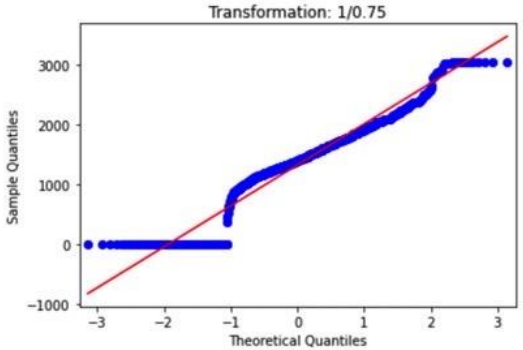
ST-T wave abnormality is the strongest indicator of heart disease amongst the above readings

Data Cleaning

1. Transformation of Skewed Features
2. Detect and Clean Outliers (~99 percentile)
3. $x^{1/n}$ transformations



Cholesterol before transformation



Cholesterol after transformation

Hyper Parameter Tuning

- We used GridSearchCV across 5 algorithms, in order to get the best hyperparameter settings for our algorithm.

- In our analysis, we trained both the raw, uncleaned data as well as the transformed, truncated and cleaned data in order to compare the results. From the results, we noticed a significant improvement in the performance of the model(time to train, time to predict, less hidden layers etc.) when we use the truncated dataset, with a negligible difference in accuracy.

- The random forest classifier yielded the best results, with accuracy of 95% using the following hyperparameters : **(max_depth=16, n_estimators=250)**

- Our analysis was done by 3 different team members, 2 of us using python, and one using R, in order to eliminate mistakes and bias, and our results were replicable across the platforms.

- The trade-off with the random forest classifier was the training and prediction time, where we observed a 10x difference in latency as compared to other models.

Generate 5 models

all independent variables (Python)

Algorithm	Function	The accuracy	The recall *	The precision
Logistic Regression (LR)	sklearn LogisticRegression	0.83	0.8	0.84
K-nearest neighbour (K-NN)	Sklearn KNeighborsClassifier	0.87	0.87	0.87
Decision Tree (DT)	DecisionTreeClassifier	0.90	0.90	0.90
Random Forest (RF)	RandomForestClassifier	0.95	**0.95**	0.95
Support Vector Machine (SVM)	Sklearn svm	0.82	0.92	0.92

Generate 5 models

Except Resting.bp.s and Resting.ecg (python)

Algorithm	Function	The accuracy	The recall *	The precision	Note
Logistic Regression (LR)	sklearn LogisticRegression	0.84	0.85	0.84	
K-nearest neighbour (K-NN)	Sklearn KNeighborsClassifier	0.89	0.90	0.89	
Decision Tree (DT)	DecisionTreeClassifier	0.90	0.90	0.90	
Random Forest (RF)	RandomForestClassifier	0.95	**0.95**	0.95	
Support Vector Machine (SVM)	Sklearn.svm	0.90	0.90	0.90	

Generate 5 models

all independent variables (R)

Algorithm	Function	The accuracy	The recall *	The precision
Logistic Regression (LR)	glm	0.8193	0.8175	0.8374
K-nearest neighbour (K-NN)	knn	0.8824	0.8968	0.8828
Decision Tree (DT)	rpart	0.8109	0.8889	0.7832
Random Forest (RF)	randomForest	0.9076	**0.9365**	0.8940
Support Vector Machine (SVM)	svm	0.8571	0.9127	0.8333

Generate 5 models

Except Resting.bp.s and Resting.ecg (R)

Algorithm	Function	The accuracy	The recall *	The precision
Logistic Regression (LR)	glm	0.8277	0.8333	0.8400
K-nearest neighbour (K-NN)	knn	0.8968	0.8968	0.9113
Decision Tree (DT)	rpart	0.8109	0.8889	0.7832
Random Forest (RF)	randomForest	0.8908	**0.9048**	0.8906
Support Vector Machine (SVM)	svm	0.8529	0.8809	0.8473

The Result

	precision	recall	f1-score	support
0	0.95	0.95	0.95	111
1	0.95	0.95	0.95	127
accuracy			0.95	238
macro avg	0.95	0.95	0.95	238
weighted avg	0.95	0.95	0.95	238

```
[[105   6]
 [  6 121]]
```

```
0.9495798319327731
```

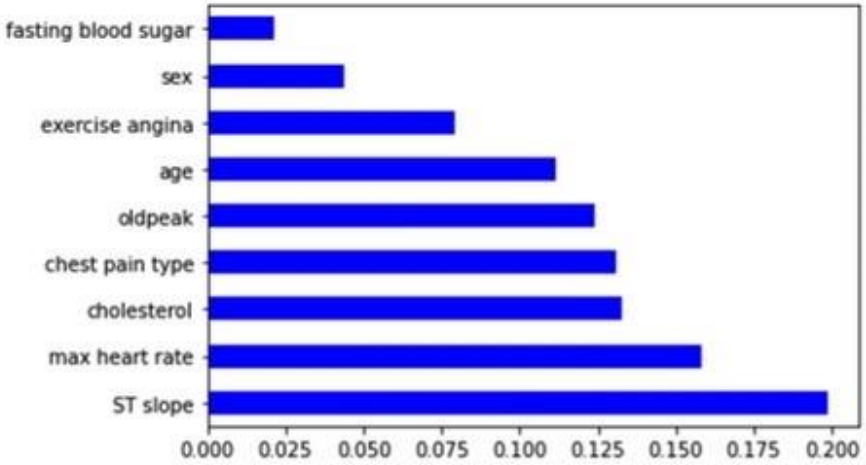
105 are the True Positives in our test data.

There are 6 type 1 error (False Positives)- predicted positive and it's false.

There are 6 type 2 error (False Negatives)- predicted negative and it's false.

121 are the True Negatives in our test data.

The Result



The top 4 significant features in the random forest model are

St slope, max heart rate, chest pain type, cholestrol

4

CONCLUSION

Conclusion

- Cardiovascular disease prediction is challenging and important in the medical field and it will help saving human lives. There is a huge number of machine learning algorithms in predicting cardiovascular disease and most of them have performed well in most cases. (Mohan, Thirumalai and Srivastava, 2019)
- The system was tested on HEART DISEASE DATASET (Comprehensive) with 5 different algorithms: LR, K-NN, DT, FR and SVM using R and Python.
- By analysing the results, it is clear that Random Forest has the highest accuracy rate with 0.90 in R and 0.95 in Python.

References

Deepika, K. and *Seema, S., 2016. Predictive analytics to prevent and control chronic diseases. 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATTccT), pp.381-386. Available at: <https://ieeexplore.ieee.org/abstract/document/7912028>*

Evans, J. D. (1996). Straightforward statistics for the behavioral sciences. Thomson Brooks/Cole Publishing Co.

Moorthy, U. and Gandhi, U., 2020. A novel optimal feature selection technique for medical data classification using ANOVA based whale optimization. *Journal of Ambient Intelligence and Humanized Computing (2020)*, [online] pp.1-12. Available at: <https://link.springer.com/article/10.1007/s12652-020-02592-w#Tab7> [Accessed 7 December 2020].

Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access (Volume: 7)*, [online] 7, pp.81542-81554. Available at: <https://ieeexplore.ieee.org/abstract/document/8740989/authors> [Accessed 7 December 2020].

Siddhartha, M., 2020. Heart Disease Dataset (Comprehensive). [online] IEEE DataPort. Available at: <https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive> [Accessed 25 December 2020].

Who.int. 2020. Cardiovascular Diseases. [online] Available at: <https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1> [Accessed 20 December 2020].

Who.int. 2020. Cardiovascular Diseases (Cvds). [online] Available at: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)> [Accessed 20 December 2020].

Mark split form

Agreed Mark Split

Name	Email	Distribution of work
Kanit Jompuk	kj1a20@soton.ac.uk	25%
	Date	12/30/2020
	Signature	DocuSigned by: Kanit Jompuk A0B859AD89B44D8...
Sharang Gupta	sdg1n20@southampton.ac.uk	25%
	Da Highlights and Notes	12/30/2020
	Signature	DocuSigned by: ED35E29E719543D...
Abdulaziz Alkhalefah	asaa1g19@soton.ac.uk	25%
	Date	12/30/2020
	Signature	DocuSigned by: Abdulaziz 8D654CFC07B6452...
Sarah Alharbi	sama2d19@soton.ac.uk	25%
	Date	12/30/2020
	Signature	DocuSigned by: Sarah Alharbi D050EF968DD9489...

THANK YOU!

1. Kanit Jompuk	31847447
2. Sharang Gupta (deputy Leader)	32196946
3. Abdulaziz Alkhalefah (leader)	31432298
4. Sarah Alharbi	31711413

Do you have any questions?