A
Report On

PROJECT - II

For
Machine Learning Course
Fuse.ai Microdegree in AI

Submitted By:
Anish Bhusal
bhusal.anish12@gmail.com

Date of Submission: 3rd April,2020

ABSTRACT

This report consists of analysis of two datasets: Sentiment Analysis and Heart Disease. For these datasets, five classifiers were trained independently and best parameters were chosen for each using grid search. The best models selected were evaluated using the test set. Of all the classifiers, Support Vector Machines(SVM) provided better results with accuracy of 71% on sentiment analysis and 84% on heart disease classification.

1. DATA EXPLORATION

This project report analyzes two datasets:

a. Sentiment Analysis Dataset

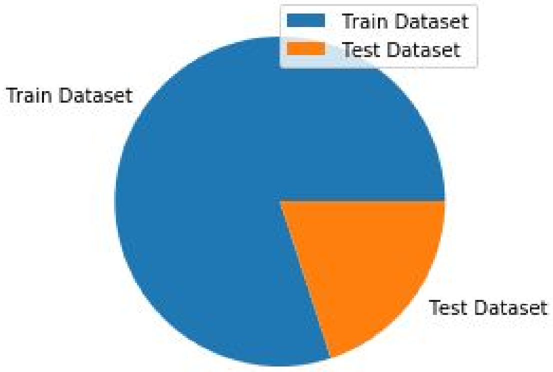
b. UCI ML Heart Disease Dataset

a. Sentiment Analysis Dataset

This dataset consists of Phrases which are categorized into five types of sentiments.

Dataset was provided into two parts: train and test set.

Training Set Examples	14711
Test Set Examples	3678



The table below shows number of phrases by sentiment types in training dataset:

Sentiment Type	0	1	2	3	4
Number of Phrases	988	1165	1876	7033	3649

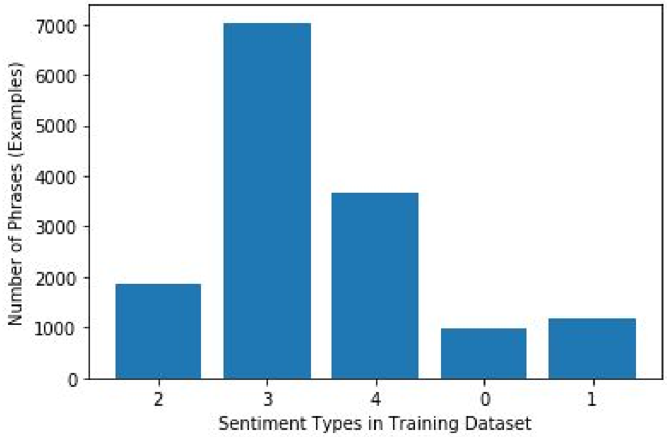


Fig.:- Bar Graph showing counts of Sentiment Types in Training Set

b. UCI ML Heart Disease Dataset

This is a heart disease dataset provided at UCI ML Repository. Originally, it consists of 76 attributes but for experimental purposes 14 attributes are taken. The provided dataset for this project consisted of following number of examples and attributes:

	Training Set	Test Set
Number of examples	242	61
Number of attributes	14	14

The dataset consisted of following attributes:

Numerical Type Attributes	Categorical Attributes
'age', 'sex', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'oldpeak', 'ca', 'target'	'cp', 'exang', 'slope', 'thal'

The figures below show some analysis done over the dataset:

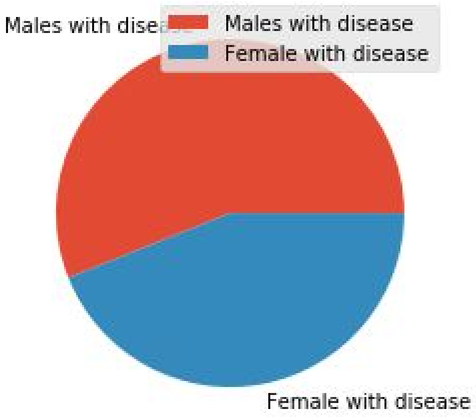


Fig.: Pie Chart showing Percentage of Males and Females having heart disease

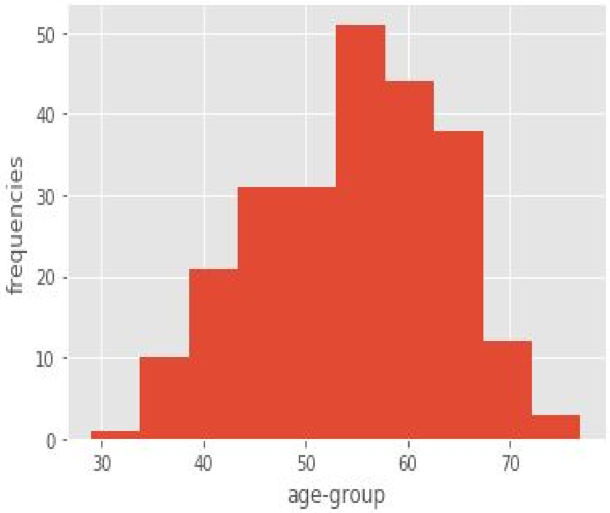
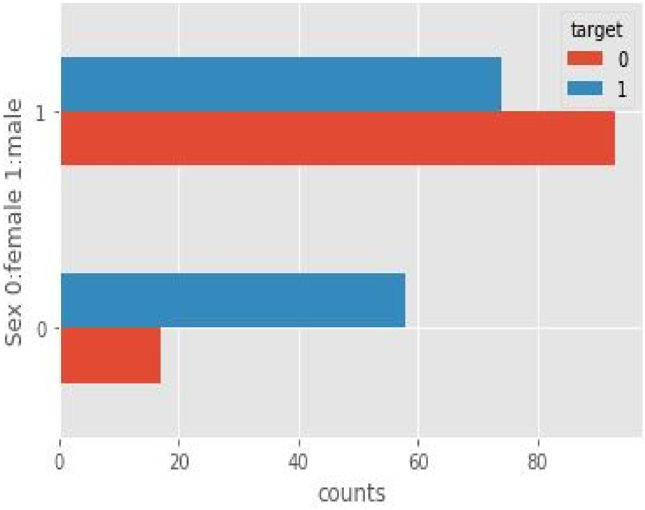


Fig.: Histogram plot shows that this dataset consists of maximum number of data of people from age group 50-60

Fig.: Number of males and females with and without disease. This figure shows that the number of males having heart disease is more than the number of females which means males are more vulnerable to heart disease than females.



Now let’s visualize the correlation between features and target in training dataset:

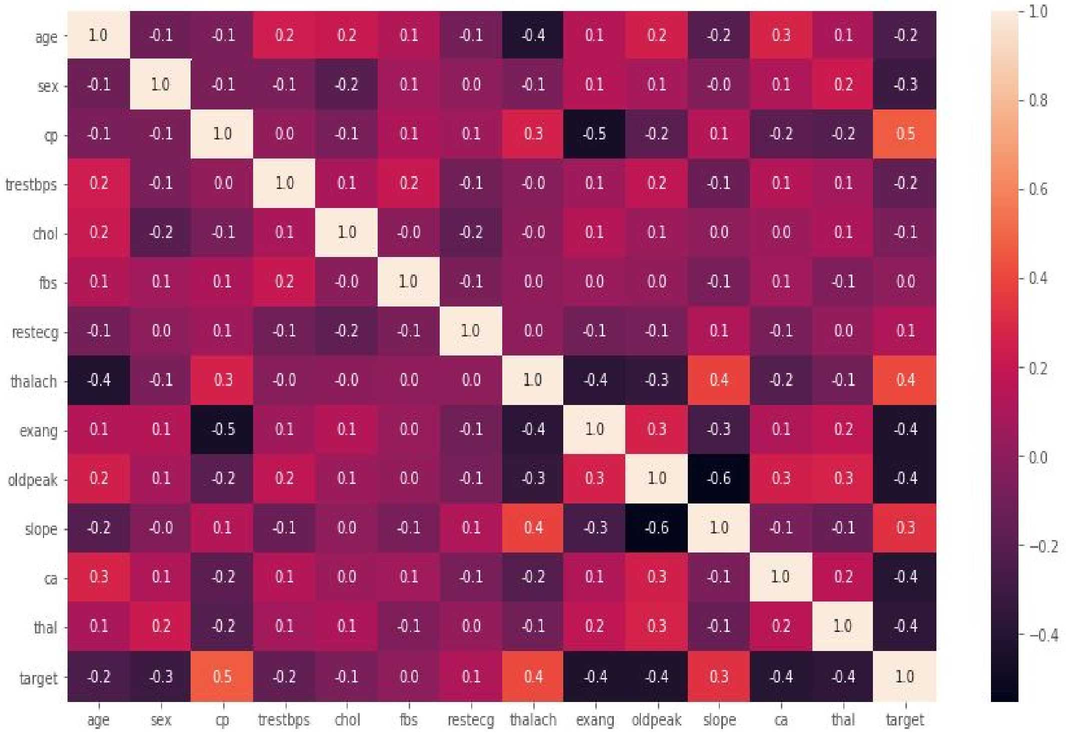


Fig.: The above correlation plot shows that chest pain type "cp" and "thalach" have highest positive correlation whereas "exang","oldpeak","ca","thal" have highest negative correlation with heart disease.

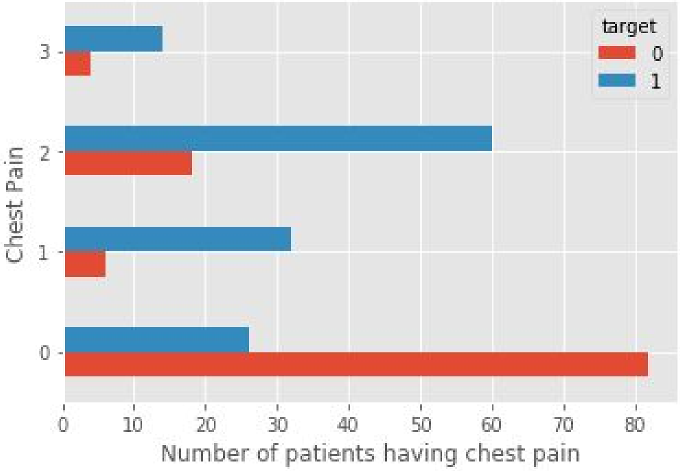


Fig.: This plot shows that people having chest pain of type 2 are more likely to have heart disease.

Now let’s analyze relation of blood pressure and cholesterol with heart disease:

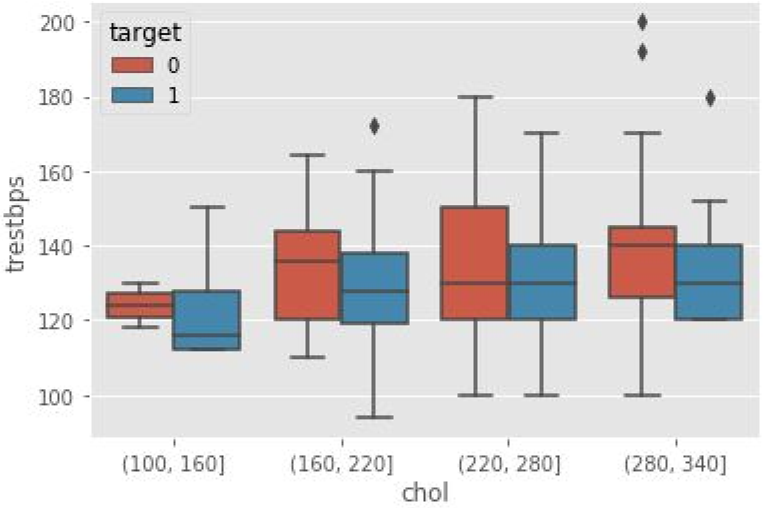
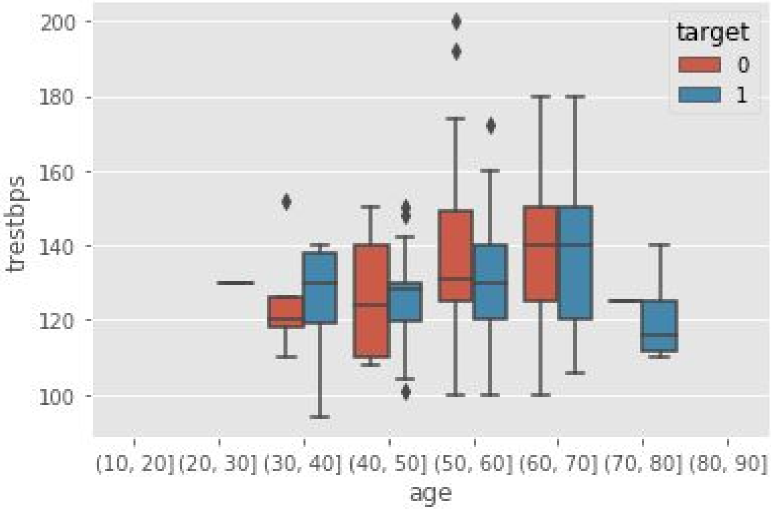


Fig.: The box plot shows people having cholesterol higher than 160 and blood pressure higher than 120 are more likely to have heart disease.

Fig.:This box plot shows people of age group 60-70 with bps in range 120-150 are more likely to have heart disease and people of age group 30-40 with bps range 120-140 are also more vulnerable to heart disease



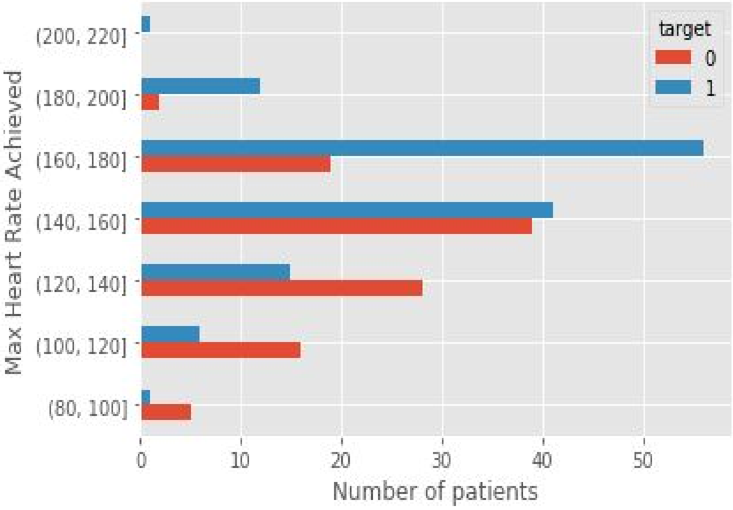


Fig.: This plot shows that patients whose max heart was between 160-180 had higher number of cases of heart disease than others.

2. FEATURE EXTRACTION AND PREPROCESSING

a. Sentiment Analysis Dataset

- Stopping characters like ‘,’ , ‘.’, ‘--’,etc were filtered out from the phrases as part of the data cleaning process.
- There were no NaN values in both training and test dataset so dropping rows with NaNs was not necessary.
- Before feeding the data to model for training, it was necessary to vectorize the data. So, for this purpose ***TfIdfVectorizer*** was used. Both training and test sets were vectorized.
- Since the number of examples by sentiment types were unbalanced, ***SMOTE*** interpolation method was used to upsample the dataset. After upsampling, the number of Phrases for each type were 7033.
- The train dataset was then splitted in X_train,y_train and test dataset into X_test, y_test.

b. UCI ML Heart Disease Dataset

- Four columns('cp', 'exang', 'slope', 'thal') with categorical values in the training dataset were encoded using ***OneHotEncoder.*** After encoding, the total number of columns became 23. The same process was carried out in the test dataset too.
- There were no NaN or missing values in the dataset.
- The train dataset was then splitted in X_train,y_train and test dataset into X_test, y_test.

3. GRID SEARCH

Models were trained on two datasets and grid search was carried out using following parameters :-

a. SVM

kernel	(‘linear’, ’rbf’)
Regularization (C)	(1, 20)

b. Decision Tree Classifier

max_depth	(3, 5, 7, 9, 11, 13)
min_samples_split	(2, 4, 6, 8, 10)
scorer	macro

c. Random Forest Classifier

n_estimators	(5, 8, 11, 13)
max_depth	(3, 5, 7, 9, 11, 13)
min_samples_split	(2, 4, 6, 8, 10)
random_split	(1, 42)

d. Gradient Boosting Classifier

n_estimators	(5, 8, 11, 13)
learning_rate	(0.05, 0.1, 0.001, 0.5)
max_depth	(3, 5, 7, 9, 11, 13),
min_samples_split	(2, 4, 6, 8)

e. AdaBoost Classifier

n_estimators	(5,8,11,13)
learning_rate	(0.1, 0.001, 0.5)
base_estimator	Decision Tree Classifier

The best models selected from Grid Search and their parameters by dataset are shown below :-

Classifier	*Dataset*	*Best Params*	*Best Score*
SVM	Sentiment Analysis	{'C': 10, 'kernel': 'rbf'}	88.92%
	Heart Disease	{'C': 20, 'kernel': 'linear'}	82.64%
Decision Tree	Sentiment Analysis	{'max_depth': 13, 'min_samples_split': 2}	43.17%
	Heart Disease	{'max_depth': 5, 'min_samples_split': 10}	78.26%
Random Forest	Sentiment Analysis	{'max_depth': 13, 'min_samples_split': 6, 'n_estimators': 13, 'random_state': 42}	49.69%
	Heart Disease	{'max_depth': 11, 'min_samples_split': 8, 'n_estimators': 13}	84.32%
Gradient Boosting	Sentiment Analysis	'learning_rate': 0.5, 'max_depth': 13, 'min_samples_split': 4, 'n_estimators': 8	71%
	Heart Disease	{'learning_rate': 0.5, 'max_depth': 13, 'min_samples_split': 10, 'n_estimators': 13}	79.81%
Adaboost	Sentiment Analysis	{'learning_rate': 0.5, 'n_estimators': 13}	49 %
	Heart Disease	{'learning_rate': 0.5, 'n_estimators': 13}	80.20%

4. MODEL EVALUATION AND COMPARISON

Each model with the best params selected from Grid Search was used to train the model and evaluated using test sets. The results are as follows:

		Macro Avg			
Model	Dataset	Precision	Recall	F1-Score	Accuracy
SVM	Sentiment Analysis	0.70	0.65	0.67	0.71
	Heart Disease	0.83	0.83	0.83	0.84
Decision Tree	Sentiment Analysis	0.34	0.35	0.28	0.32
	Heart Disease	0.70	0.70	0.70	0.70
Random Forest	Sentiment Analysis	0.38	0.39	0.30	0.31
	Heart Disease	0.77	0.77	0.77	0.77
Gradient Boosting	Sentiment Analysis	0.70	0.65	0.67	0.71
	Heart Disease	0.72	0.72	0.72	0.72
AdaBoost	Sentiment Analysis	0.41	0.39	0.35	0.36
	Heart Disease	0.75	0.75	0.75	0.75

5. CONCLUSION

The two datasets were trained on 5 classifiers. Among all the models, SVM produced better predictions with accuracy of 71% Sentiment Analysis and 84% on Heart Disease dataset. Hence SVM was chosen as the best model.

All the code and analysis are available here.