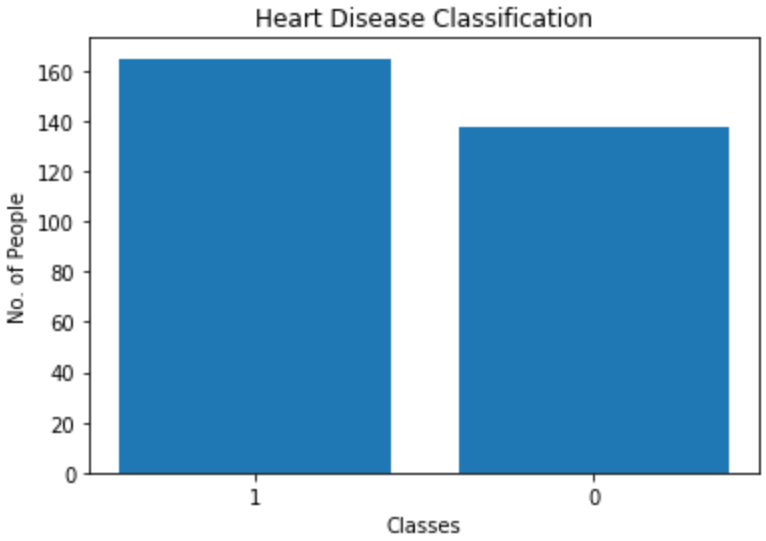
We explored five supervised machine learning algorithms on two datasets(heart disease dataset and 5 class sentiment analysis dataset). SVM and AdaBoosting classifier worked best on heart disease dataset and SVM classifier worked best on 5 class sentiment analysis dataset.

Heart Disease Dataset

1. Data Exploration

We loaded the data in “heart.csv” file. We had 303 examples in total. Among them, 165 belong to the positive class (labeled as 1) and 138 belong to the negative class (labeled as 0).
There were no any null values for any attributes. The data can be visualized as shown:



Our dataset contains a bit more positive examples than negative ones. As the number is not so different, we consider this to be balanced dataset. So we use accuracy as a measure here.

2. Feature Extraction and Preprocessing

We split our dataset into train set and test set in a 70:30 ratio. As a preprocessing step, we did a one-hot encoding of categorical features. Instead of representing them just by integers, we used one-hot encoding as our model will develop meaning from integers. We encoded both training and test sets.
We have used 22 features. Without one-hot encoding, there were 13 features provided in our dataset.

3. Grid Search

We performed a grid search on five algorithms:

a. SVM
b. Decision Tree
c. Random Forest
d. AdaBoost
e. Gradient Boosting

We selected a bunch of parameters for each algorithm and performed a grid search to find out the best parameters for each algorithm as shown in the table.

S.N.	Algorithm	Best parameters	Best Score
1	SVM	kernel = linear, C = 1	0.8142
2	Decision Tree	max_depth = 3, min_samples_split=2	0.7812
3	Random Forest	criterion=entropy, n_estimators=30	0.8264
4	AdaBoost	Learning rate =0.5, n_estimators = 10	0.8385
5	Gradient Boosting	max_depth = 3, min_samples_split=2	0.7933

```
<class 'sklearn.ensemble._weight_boosting.AdaBoostClassifier'>
0.8385204081632655 {'learning_rate': 0.5, 'n_estimators': 10}
<class 'sklearn.tree._classes.DecisionTreeClassifier'>
0.7812925170068027 {'max_depth': 3, 'min_samples_split': 2}
<class 'sklearn.ensemble._forest.RandomForestClassifier'>
0.8264455782312925 {'criterion': 'entropy', 'n_estimators': 30}
<class 'sklearn.svm._classes.SVC'>
0.8142857142857143 {'C': 1, 'kernel': 'linear'}
<class 'sklearn.ensemble._gb.GradientBoostingClassifier'>
0.79328231292517 {'max_depth': 3, 'min_samples_split': 2}
```

4. Model Evaluation and Comparison

On the test set, the SVM and the Adaboost algorithm gave highest accuracy while Gradient Boosting classifier gave the least accuracy. So on a heart disease dataset SVM and AdaBoost classifier works best with accuracy of 83.606%.

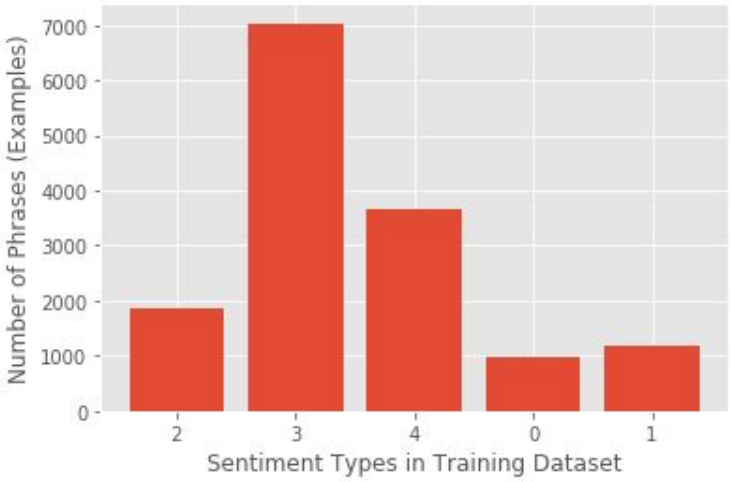
S.N.	Algorithm	Accuracy
1	SVM	**0.83606**
2	Decision Tree	0.8196
3	Random Forest	0.8033
4	AdaBoost	**0.83606**
5	Gradient Boosting	0.7541

	Classifier	accuracy
0	<class 'sklearn.ensemble._weight_boosting.AdaB...	0.836066
1	<class 'sklearn.tree._classes.DecisionTreeClas...	0.819672
2	<class 'sklearn.ensemble._forest.RandomForestC...	0.803279
3	<class 'sklearn.svm._classes.SVC'>	0.836066
4	<class 'sklearn.ensemble._gb.GradientBoostingC...	0.754098

5 Class Sentiment Analysis Dataset

1. Data Exploration

We loaded the data in the “sentiment_5_class.csv” file. We had 18389 examples in total. Among them 1235 belong to class 0, 1456 belong to class 1, 2345 belong to class 2, 8792 belong to class 3 and 4561 belong to class 4.
There were no any null values for any attributes. The data can be visualized as shown:



Our dataset is an imbalanced dataset.

2. Feature Extraction and Preprocessing

We split our dataset into train set and test set in a 70:20 ratio. As a preprocessing step, we converted our textual data into numerical representations using the TF-IDF vectorizer.We have used 7154 features as our vocab of unique words will be 7154 words long. So each row of our dataset will be 7154 elements long.

3. Grid Search

We performed a grid search on five algorithms:

f. SVM
g. Decision Tree
h. Random Forest
i. AdaBoost
j. Gradient Boosting

We selected a bunch of parameters for each algorithm and performed a grid search to find out the best parameters for each algorithm as shown in the table.

S.N.	**Algorithm**	**Best parameters**	**Best Score**
1	SVM	kernel = rbf, C = 10	**0.6578**
2	Decision Tree	max_depth = 19, min_samples_split=2	0.2824
3	Random Forest	criterion=gini, n_estimators=30	0.5611
4	AdaBoost	n_estimators = 100	0.3122
5	Gradient Boosting	max_depth = 15	0.5291

```
<class 'sklearn.ensemble._weight_boosting.AdaBoostClassifier'>
0.312232722062266 {'n_estimators': 100}
<class 'sklearn.tree._classes.DecisionTreeClassifier'>
0.28244318311721245 {'max_depth': 19, 'min_samples_split': 2}
<class 'sklearn.ensemble._forest.RandomForestClassifier'>
0.5611674782974916 {'criterion': 'entropy', 'n_estimators': 30}
<class 'sklearn.svm._classes.SVC'>
0.6578912079604808 {'C': 10, 'kernel': 'rbf'}
<class 'sklearn.ensemble._gb.GradientBoostingClassifier'>
0.5291466111927957 {'max_depth': 15}
```

4. Model Evaluation and Comparison

On the test set, SVM gave the highest f1 score while Decision Tree and AdaBoost classifier gave the least f1 score. So on a 5 class sentiment analysis dataset SVM works best with an accuracy score of 67.75%.

S.N.	Algorithm	F1 score
1	SVM	**0.6775**
2	Decision Tree	**0.2827**
3	Random Forest	0.5935
4	AdaBoost	0.3203
5	Gradient Boosting	0.5492

	Classifier	f1_score
0	<class 'sklearn.ensemble._weight_boosting.AdaB...	0.320368
1	<class 'sklearn.tree._classes.DecisionTreeClas...	0.282731
2	<class 'sklearn.ensemble._forest.RandomForestC...	0.593561
3	<class 'sklearn.svm._classes.SVC'>	0.677531
4	<class 'sklearn.ensemble._gb.GradientBoostingC...	0.549254