ITEC3040 Final Project: Predicting Diabetes Based on Pre-existing Symptoms

Thuy Tran, Shenaz Islam, Nathan Didier

December 9, 2020

1 Introduction & Motivation to the problem

Diabetes is one of the most common diseases in the world, ranking in top 10 leading diseases which cause death globally according to the world health organisation. Diabetes is a disease which affects the way the body regulates blood sugar and reacts to insulin. Diabetes is broken into type 1 diabetes and type 2 diabetes, where type 1 identified patients' bodies do not produce insulin and type 2 patients do not respond to insulin as well as they should. Insulin is a key player in how our body fuels its cells, without it our body cannot use glucose to feed our cells. Both types of the disease can cause chronically high blood sugar levels which increase complications in the body. Some of these complications scale from cardiovascular disease, nerve damage, kidney damage and varying skin conditions the list of complications also go on. It is because of how common a disease diabetes is and the impact it has on the human body is why our group decided to pick the early stage diabetes risk prediction dataset to create a prediction model which can determine whether a patient is potentially in risk of developing diabetes or not based on their symptoms. If our prediction model is accurate enough we hope that with this model doctors could potentially use it to be able to quicker diagnosis patients and give them the necessary advice as needed.

2 Problem Definition

There are three major ways of diagnosing diabetes in patients these are Fasting Glucose Test, Random (anytime) Glucose Test and Haemoglobin A1c Test. These tests can sometimes be costly to the patient and requires the patients to specifically go into a doctor's appointment to have these test done to see if they have diabetes or not. Our hope with this research is to see if it is possibly to accurately predict whether someone has diabetes based on pre-existing symptoms. With hopes of being able to implement this model into the field which would help with more quickly identifying patients which could potentially become diabetic and get them the needed treatments that could help prevent the impact diabetes could have on their life. Another potential benefit of being able to implement this model would be a decrease prediabetic diagnostic test costs.

3 Pre-processing

3.1 Dataset and Features

The dataset we used was extracted from the UCI Machine Learning Repository. The dataset consists of 520 rows/tuples and 17 attributes collected from questionaries conducted with patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and were approved by a doctor.

Each row of the dataset corresponds to a patient. Out of the 17 attributes,

one of them is the class attribute denoting if a patient tested positive or negative for diabetes. Therefore, the class attribute is an asymmetric binary attribute, since it only has two possible values and we only care about the positive values.

The other 16 attributes are the age and gender of the patient, and the early diabetes symptoms they had.

The details of all attributes are listed below:

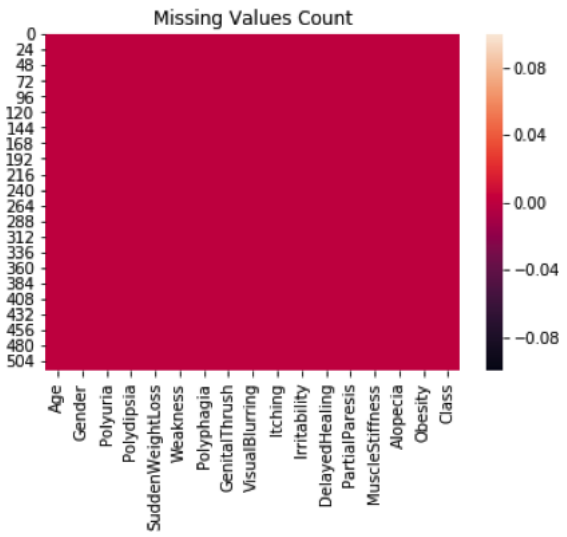
Attribute or Predictors	Definition	Data type and values
Age	Age of the patient.	Numeric, continuous, {16-90} age range
Gender	Gender of the patient	Categorical, symmetric binary, {Male, Female}
Polyuria	Excessive secretion of urine.	Binary, asymmetric, {Yes, No}
Polydipsia	The feeling of extreme thirstiness	Binary, asymmetric, {Yes, No}
sudden weight loss	Unexplained weight loss	Binary, asymmetric, {Yes, No}
weakness	The feeling of body fatigue or tiredness	Binary, asymmetric, {Yes, No}
Polyphagia	Excessive or extreme hunger	Binary, asymmetric, {Yes, No}
Genital thrush	A common infection caused by an overgrowth of the yeast	Binary, asymmetric, {Yes, No}

visual blurring	Lack of sharpness of vision	Binary, asymmetric, {Yes, No}
Itching	A sensation that causes the desire or reflex to scratch	Binary, asymmetric, {Yes, No}
Irritability	Quick excitability to annoyance, impatience, or anger	Binary, asymmetric, {Yes, No}
delayed healing	Slow wound healing	Binary, asymmetric, {Yes, No}
partial paresis	Slight paralysis of motor functions	Binary, asymmetric, {Yes, No}
muscle stiffness	When the muscles feel tight and difficult to move	Binary, asymmetric, {Yes, No}}
Alopecia	Partial or complete loss of hair	Binary, asymmetric, {Yes, No}
Obesity	An abnormal accumulation of body fat	Binary, asymmetric, {Yes,No}
Class	Tested or did not test positive for diabetes	Binary, asymmetric, {Positive, Negative}

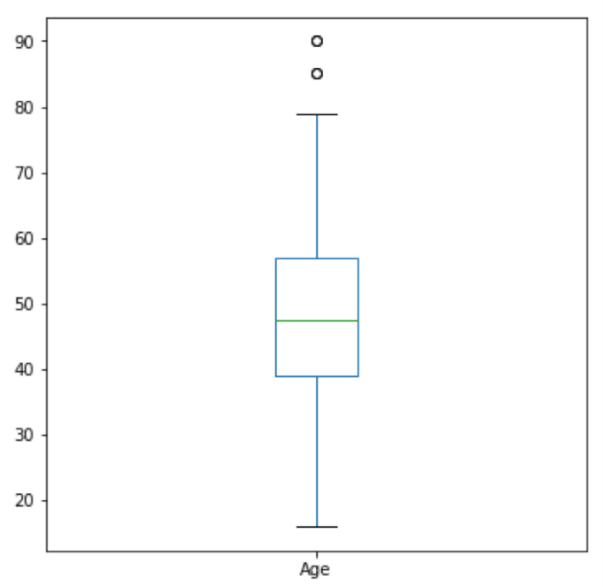
3.2 Data Cleaning

First, we scanned our dataset and found that it had no missing values. The dataset also had no inconsistent or noisy data since all the values fell within the specified range stated above.

```
    ....
    ...: rawData.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
Age                 520 non-null int64
Gender              520 non-null object
Polyuria            520 non-null object
Polydipsia          520 non-null object
SuddenWeightLoss    520 non-null object
Weakness            520 non-null object
Polyphagia          520 non-null object
GenitalThrush       520 non-null object
VisualBlurring      520 non-null object
Itching             520 non-null object
Irritability        520 non-null object
DelayedHealing      520 non-null object
PartialParesis      520 non-null object
MuscleStiffness     520 non-null object
Alopecia            520 non-null object
Obesity             520 non-null object
Class               520 non-null object
dtypes: int64(1), object(16)
memory usage: 69.1+ KB
```



The Age attribute had two outliers however, ž when reviewing the min-max range, Age is in an acceptable range.



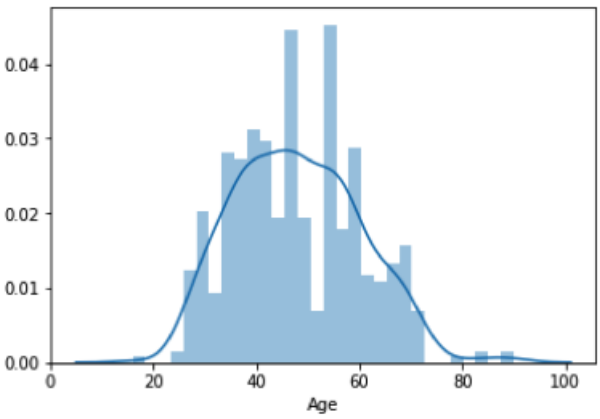
Age	count	520.000000
	mean	48.028846
	std	12.151466
	min	16.000000
	25%	39.000000
	50%	47.500000
	75%	57.000000
	max	90.000000

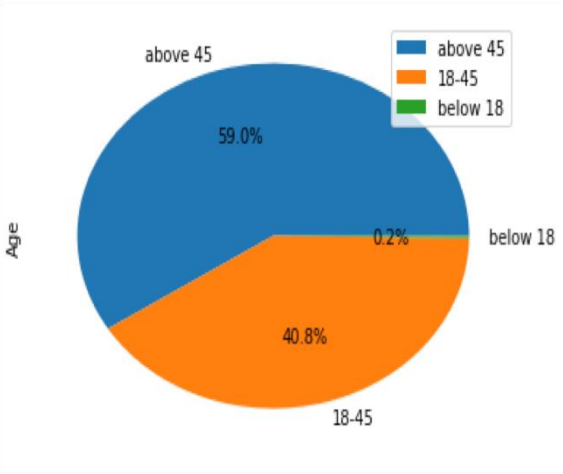
3.3 Data Transformation

We did not normalize our dataset since all attributes except the “Age” attribute are categorical variables.

For all binary attributes, we transformed their values to 1’s and 0’s, since machine learning algorithms cannot be used on categorical data directly. The age is known to be highly correlated with diabetes according to the CDC’s 2017 Report. Type 1 diabetes was once known as juvenile diabetes. That is because it is frequently diagnosed in children and young adults. While the fixed risk factors for Diabetes type 2 is being over 45 years of age. The age attribute values in the dataset are numerically much bigger than the values of other attributes (1’s and 0’s), and so if we used the age values as-is to construct our model, it would’ve dominated the calculation

and thus the prediction results. Therefore, we discretized the age values into three groups: <18 : ‘below 18’ | 18-45: ‘18-45’ |> 45: ‘above 45’.



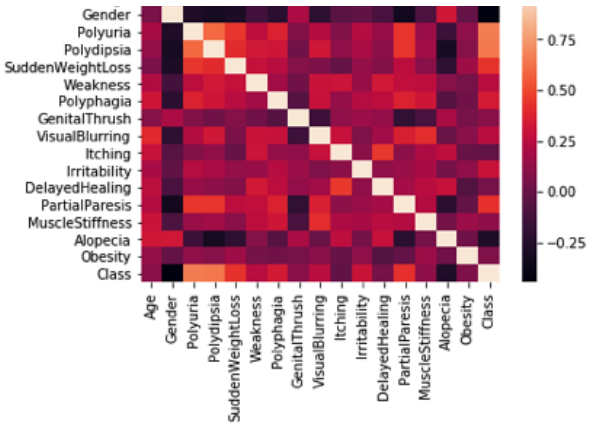


3.4 Data Reduction

We performed chi square test for each categorical attribute with the class attribute (positive or negative for diabetes test), to see if they are correlated.

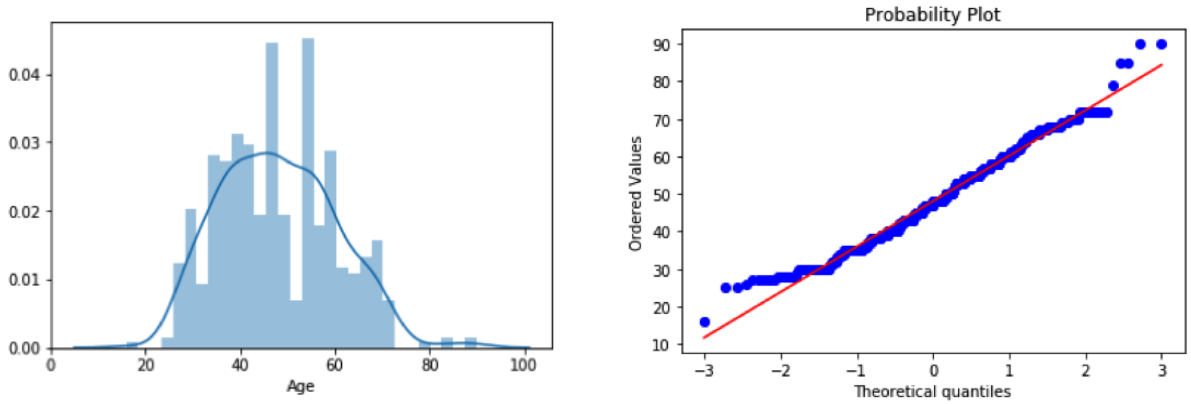
At 1% significance level the attributes ‘Itching’, ‘Delayed Healing’, ‘Obesity’, and ‘Genital Thrush’ were determined as not statistically significant. Therefore, we can conclude since these attributes are not related to the class attribute and so would not be helpful for constructing our prediction model. Therefore, we excluded these attributes for the model construction.

```
Gender----------3.289703730553317e-24
Polyuria----------1.7409117803442155e-51
Polydipsia----------6.1870096408863144e-49
SuddenWeightLoss----------5.969166262549937e-23
Weakness----------4.869843446585542e-08
Polyphagia----------1.1651584346409174e-14
GenitalThrush----------0.016097902991938178
VisualBlurring----------1.7015036753241226e-08
Itching----------0.8297483959485009
Irritability----------1.7714831493959365e-11
DelayedHealing----------0.32665993771439944
PartialParesis----------1.565289071056334e-22
MuscleStiffness----------0.006939095697923978
Alopecia----------1.9092794963634e-09
Obesity----------0.12710799319896815
Class----------3.498537810092432e-114
```

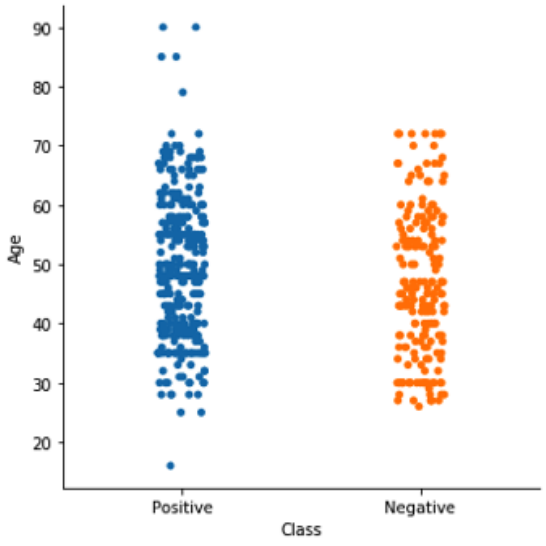


3.5 Exploratory Data Analysis

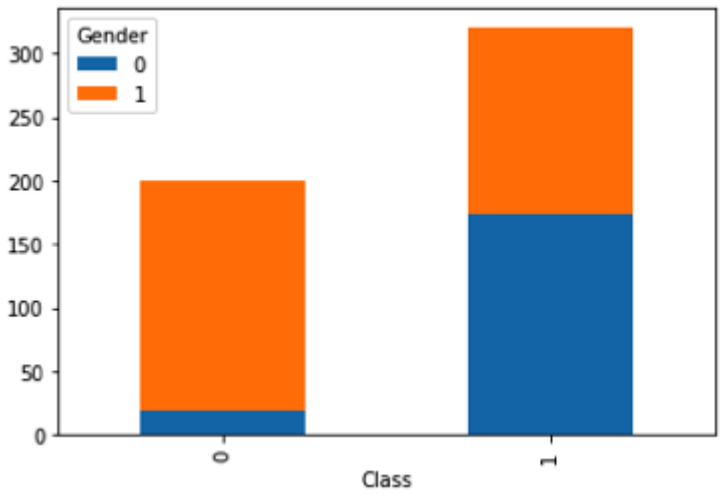
We found that the numeric Age attribute data was normally distributed and roughly symmetric, with a skewness of 0.33 and Kurtosis of -0.19. 50% of all age values were in the range of 39 to 57.



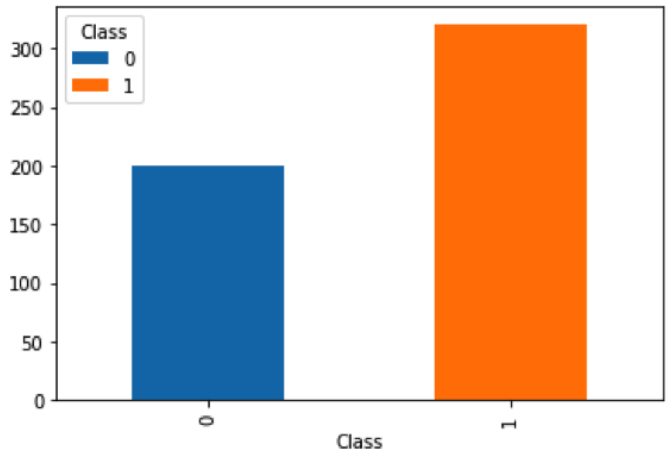
Based on the data we also found that patients above the age of 70 were more likely to be diagnosed with diabetes.



During the explorative data analysis for the categorical attributes, we found that even though the number of males are more than females, females are more likely to get diabetes.



We learnt that there were 120 more diabetic cases over non-diabetic cases, which may cause a class imbalance problem in our model construction.



Diabetic patients were also more likely to experience Polyuria, Sudden Weight Loss, Weakness, and Polydipsia.

It was also found that diabetic patients were less likely to experience Alopecia than non-diabetic patient and more likely to experience Visual Blurring, Polyphagia, Muscle Stiffness, and partial Paresis.

3.6 Data Splitting

In order to test out model we split the dataset into training and test sets where 20% was the test set and 80% the training set.

4 Method

As mentioned earlier, we would like to classify patients into groups of diabetic and non-diabetic patients based on the symptoms they are experiencing. This is defined as a classification problem. Classification is a 2 steps process, a learning step, and a prediction step. We split data into training data set which is used for the learning step, and test set which is used or prediction step.

We chose Decision Tree Classifier, KNN classifier, and AdaBoost classifier as described below. We later perform Hyperparameter Tuning and Cross Validation to improve the performance of our machine learning models which will be discussed in the next section.

4.1 Decision Tree

For our classification problem, we use Decision Tree as baseline model. Most of our attributes except for Age are binary nominal (yes or no), hence we chose Gini index as our attribute selection measure since it considers binary split for each attribute. Since the splitting point is predetermined, we calculate Gini Index to determine which attribute is used to further partition data set into its purest form.

We also discretized our numeric attribute (Age) into multiple bins as previously discussed. We later perform an experiment to leave Age as it is(discrete-valued attribute) and we would like to see how that affects that prediction power of the models. Details will be discussed in the next section. To do this, we used scikit-learn's decision tree classifier.

4.2 KNN

The second model we applied is KNN Classifier, which is an instance-based learning algorithm that uses different distance metrics (Minkowski, Manhattan or Euclidean) to calculate the similarity/dissimilarity between records. There is no need for training the model. In our model, we chose Euclidean as our distance metric and number of nearest neighbours to compare to is 3. Since most of our attributes are asymmetric valued attributes, only age is numeric, we will

later perform experiments with attribute discretization and attribute normalization to see how it affects the model performance.

To do this, we used scikit-learn's KNN classifier library.

4.3 AdaBoost

We have been looking at one individual classifier as a model to perform prediction. Another machine learning algorithm that we picked is AdaBoost Classifier, is an ensemble boosting algorithm. This can decrease variance using bagging approach (multiple classifiers to make prediction), and bias using boosting approach (the higher accuracy classifier helps the lower accuracy classifier to evaluate the misclassified records). We use 50 models of our baseline learning algorithm, decision tree classifier, to perform prediction. We also use scikit-learn 's AdaBoost Classifier to do this.

5 Experiments and Results

We performed four different experiments:

The first experiment, we include all the data attributes as it is (Age attribute is normalized, and number of attributes remained the same).

The second experiment, we include all the data attributes but perform Age discretization.

The third experiment, we performed attribute selection based on correlation/association level and attribute discretization. We performed Chi-square and rejected any attributes that have the level of significance > 1%.

The fourth experiment, we conducted hyperparameter tuning for decision tree and KNN. For decision tree, we focused on criterion (Ginni, Information Gain),split strategy(best, random best), max_depth. For KNN, we focus on tuning number of neighbours. For AdaBoost, we focus on number of estimators and learning rate.

For each experiment, we used 80% of the dataset for training and withheld 20% for testing. Our performance metric for all models was accuracy (fraction of correctly classified examples), and confusion metric to evaluate the misclassification rate, and ROC (AUC)(probability that model will be able to distinguish positive class and negative class)

5.1 Experiment 1 Result

	Algorithm	Accuracy	AUC	Error_rate
0	Decision Tree	0.97115385	0.971875	0.02884615
1	KNN	0.95192308	0.95625	0.04807692
2	AdaBoost	0.94230769	0.9296875	0.05769231

5.2 Experiment 2 Result

	Algorithm	Accuracy	AUC	Error_rate
0	Decision Tree	0.97115385	0.971875	0.02884615
1	KNN	0.99038462	0.9921875	0.00961538
2	AdaBoost	0.94230769	0.9390625	0.05769231

5.3 Experiment 3 Result

	Algorithm	Accuracy	AUC	Error_rate
0	Decision Tree	0.97115385	0.971875	0.02884615
1	KNN	0.98076923	0.9796875	0.01923077
2	AdaBoost	0.90384615	0.89375	0.09615385

5.4 Experiment 4 Result

	Algorithm	Accuracy	AUC	Error_rate
0	Decision Tree CV	0.98076923	0.9796875	0.01923077
1	KNN CV	0.99038462	0.9921875	0.00961538
2	AdaBoost CV	0.94230769	0.934375	0.05769231

5.5 Hyperparameter Tunning

Best Parameters:

Decision Tree: {'criterion': 'entropy', 'max_depth': 9, 'splitter': 'random'}

For decision tree, the modification in terms of attributes do not change the accuracy. The best splitting attribute criterion is Information Gain, with the maximum depth of 9. The best accuracy achieved is 0.98 which was improved from 0.97.

KNN: {'algorithm': 'auto', 'metric': 'minkowski', 'n_neighbors': 7, 'p': 2, 'weights': 'distance'}

For KNN, the best distance metric is Euclidean with the number of nearest neighbours is either 3 or 7. KNN works better with categorical attribute rather than normalized numeric attribute. Dropping uncorrelated attributes does not contribute to the prediction power of this model. Accuracy decreases from 0.99 to 0.98. But when performing cross-validation, the accuracy improves, and remains at 0.99.

AdaBoost: {'learning_rate': 0.9, 'n_estimators': 14}

Though Boosting algorithm was known to yield high accuracy prediction, AdaBoost did not perform well. The best accuracy score achieved is at 0.94. In fact, dropping columns with less correlation heavily impact the performance of the model. It tells us that even though the attributes may not directly correlated to the target class, all of them are important since it significantly contribute to the model performance in the end.

In our assessment, we will pick KNN without cross-validation as our prediction model based on the highest accuracy rate and the least error rate, and the resource-saving purpose. The best number of nearest neighbours is 3. We will not drop any attributes since all of them are important to the model performance. Any attribute that are numeric are highly recommended to be discretized before fitting into the model.

6 Conclusion and Further Discussion

In conclusion after our analysis of the early stage diabetes risk prediction dataset we believe it is possible to accurately predict whether a person may have diabetes based of their symptoms and age using our model.

Despite the high accuracy throughout all the test done we still believe that it would be beneficial to do more testing on a larger dataset to see if our results hold up.

Due to the higher accuracy of the KNN model in the final testing, this model would be what we would like to perform further testing mostly. If after further testing accuracy is upheld, the team believe this knowledge could be integrated into a medical machine learning system which could potentially alert doctors or nurses about patients who have been developing systems that could lead to diabetes and notify those patients in advance.

However, further research is still needed to be done on how integration of such knowledge could be implemented in a machine learning system and actually getting the information produced to the right people in time.

7 References

1.Diabetes. (n.d.). Retrieved from https://www.who.int/news-room/fact-sheets/detail/diabetes

2.Cherney, K. (2018, July 6). Age of Onset for Type 2 Diabetes: Know Your Risk. Healthline. https://www.healthline.com/health/type-2-diabetes-age-of-onset

3.Kivi, R. (2020, June 18). What Is Type 1 Diabetes? Healthline. https://www.healthline.com/health/type-1-diabetes-causes-symtoms-treatments

4.Centers for Disease Control and Prevention. (2020). National Diabetes Statistics Report, 2020. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

8 Contribution Table

Task	**Finished by**
Document Meeting Notes	Nathan Didier
Create GIT Hub Repository	Shenaz Islam
Data Pre-processing	Shenaz Islam
Create PowerPoint Presentation	Thuy Tran
Create Checkpoint Report	Nathan Didier
Decision Tree Implementation	Thuy Tran
KNN Implementation	Shenaz Islam
ADABoost Implementation	Nathan Didier
Exploratory Testing	Thuy Tran
Editing Final Report	Nathan Didier