Predicting the Adoptability of Pit Bulls or mix Using Machine Learning and Data Visualization

(Simultaneously checking the factors representing high Euthanasia rate of PitBulls)

Ritsumeikan Asia Pacific University

Data Mining Individual Project

Professor Claster William B. (Phd)

DASH Soujanya - 12418113





**Statement of Authorship**

I certify that this assignment is my own work and contains no material which, to the best of my knowledge and belief, has been previously published or written by another person, except where due reference is made in the text of the assignment.

DASH Soujanya 12418113



**Contents**

|  |  |  |
| --- | --- | --- |
| Sl. No. | Description | Page No. |
| 1. | Introduction | 3 |
| 2. | Literature Review | 4 |
| 3. | Brief about the Dataset | 5 – 7 |
| 4. | Data Pre-processing and cleaning | 8 - 10 |
| 4.1 | - Importing data | 8 |
| 4.2 | - Dealing with Missing Data | 9 |
| 4.3 | - Removing Attributes | 10 |
| 5. | Derived Variables | 10 - 11 |
|  | * Boolean | 10 |
|  | * Normalization | 10-11 |
|  | * Binning | 11 |
|  | * New attribute | 11 |
|  | * Mapping | 11 |
| 6. | Data Visualization with Tableau | 12 - 17 |
| 7. | Class Imbalance | 18 |
| 8. | Attribute Selection (3 methods) | 18 - 20 |
|  | * Using Correlation AttributeEval Method | 18 - 19 |
|  | * Using Wrapper Method | 19 |
|  | * Using CfsSubsetEval Method | 20 |
| 9. | Testing and Training | 20 |
| 10. | Classification | 20 - 23 |
| 10.1 | - Baseline Classification (ZeroR) | 21 |
| 10.2 | - J48 Decision Tree | 22 - 23 |
| 11. | Experimenter | 23 - 25 |
| 12. | Testing on Neural Network (Multi-layer Perceptron) | 25 - 26 |
| 13. | Ensemble Learning | 26 - 27 |
| 13.1 | - Bagging | 26 - 27 |
| 13.2 | - AdaBoost | 27 |
| 14. | CV Parameter Selection | 28 |
| 15. | Boundary Visualization | 29 |
| 16. | Cost Sensitive Learning/ Cost-Benefit Analysis | 29-30 |
| 17. | Knowledge Flow Interface and ROC curve | 31-32 |
| 18. | Conclusion | 33 |

*The greatest problem that no-kill shelters face is more intakes as compared to outcomes so any new animal has to be transferred to another shelter which may euthanise it*

**1.0 Introduction**

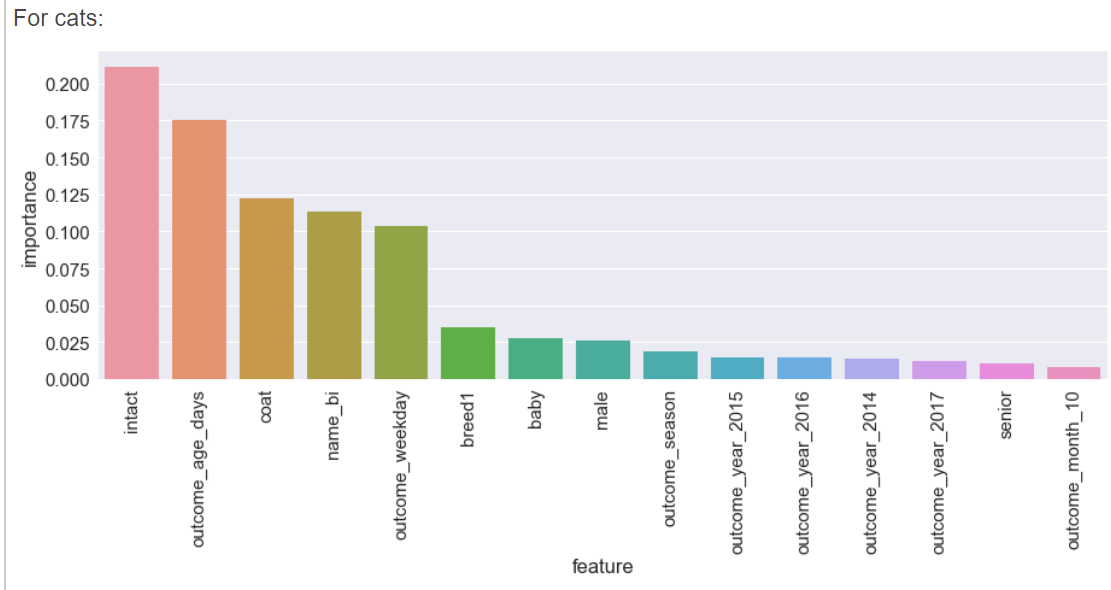
Austin animal shelter is in the city of Texas and is one of the largest animal rescue centres (no-kill shelter) which provides home to more than 16000 animals annually plus animal protection and pet resource services to Austin. Before coming to the main point, what do we exactly mean by an animal shelter? An animal shelter is an establishment maintained by local government or supported by charitable contributions, that provides a temporary home for dogs, cats, and other animals that are offered for adoption and the unwanted ones are put to sleep, but there are mainly 3 types of shelters.

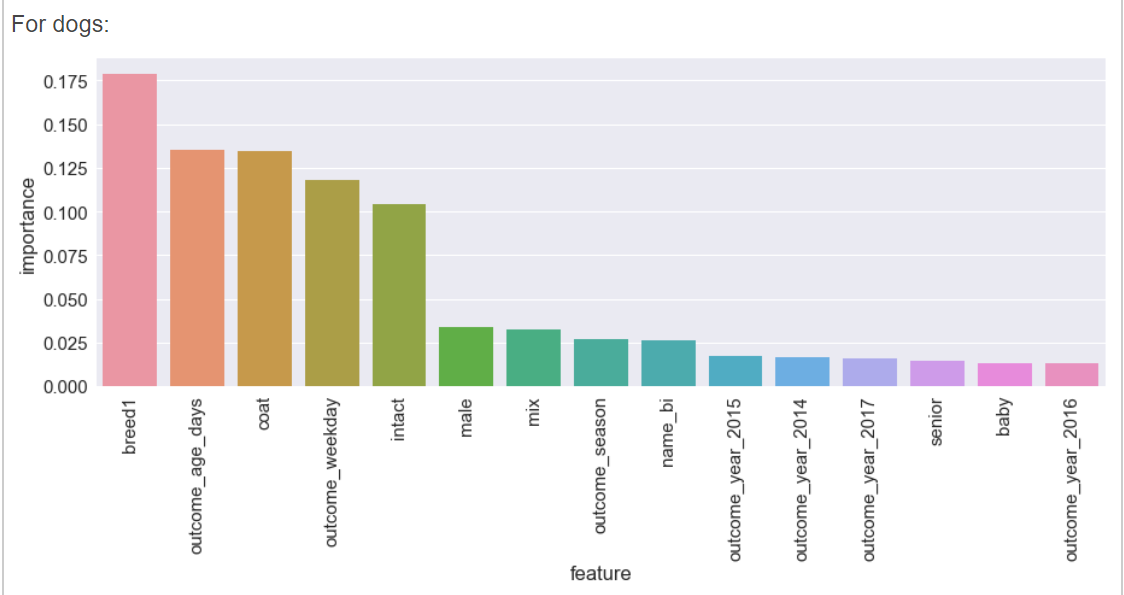
*In a typical shelter, Pit Bulls/ Pit Bull mix have a high euthanasia rate just because their breed is infamous for vicious attacks & not a family-friendly dog and the aim of this project is to optimize the adoption process with the help of Data visualization and machine learning algorithms.*

The three types of shelters are no-kill, typical and private. Our focus will be on the no-kill shelter where every animal is treated with the equal amounts of care and medical help. No-kill shelters do not euthanize an animal unlike its counterpart where each animal irrespective of behavioural issues, age, type is treated as a number and euthanized. Sometimes even perfectly healthy animals are put to sleep just because of their bad representation as a pet. For instance, **Pit-Bulls** are generalized based on their breed and not given many chances at any shelter. Pit Bulls have been infamous for their vicious attacks and hard to control behaviour but it seems the same is actually not true.

This project will be useful for no-kill shelters who can opt the given methods to optimize adoption of Pit Bulls and thus, can accommodate more needy animals for further adoption process. In order to conduct this project, Weka, a knowledge analysis software; developed at the University of Waikato (by Prof. Ian Witten); is used extensively, supported by Tableau, Microsoft Excel and Rapid Miner for Data Visualizations and Python programming for writing codes.

**2.0 Literature Review**:





Previous research on California animal shelters have already suggested that there are few features with relative importance ( we are considering only the positive correlation calculated using excel and visualized using Tableau) that heavily affects the adoption procedure for a cat and a dog respectively. Both have a common issue which is the seniority which means that as the age increases the probability of people adopting it decreases which may be due to health issues and extra-care required etc. Similarly, for a dog, breed is the utmost priority during an adoption. So, in our project we’ll be focussing on finding compatibility with the breed, Pit bulls[[1]](#footnote-1).

**3.0 Brief about the Dataset**:

The dataset is retrieved from **Kaggle** (who retrieved it from Austin Animal shelter official site). Austin Animal shelter has made its data available for the public and is up-to-date even during this pandemic. The dataset name is **Austin Animal Shelter Outcomes** and it was updated 2 years back which is still pretty new. But why should we take data from 2 years ago? The reason is that it contains information about around **80,000 intakes** and there were maximum positive outcomes during this time. The number of attributes and instances in the dataset are **39 and 79673** respectively.

<https://www.kaggle.com/c/shelter-animal-outcomes>

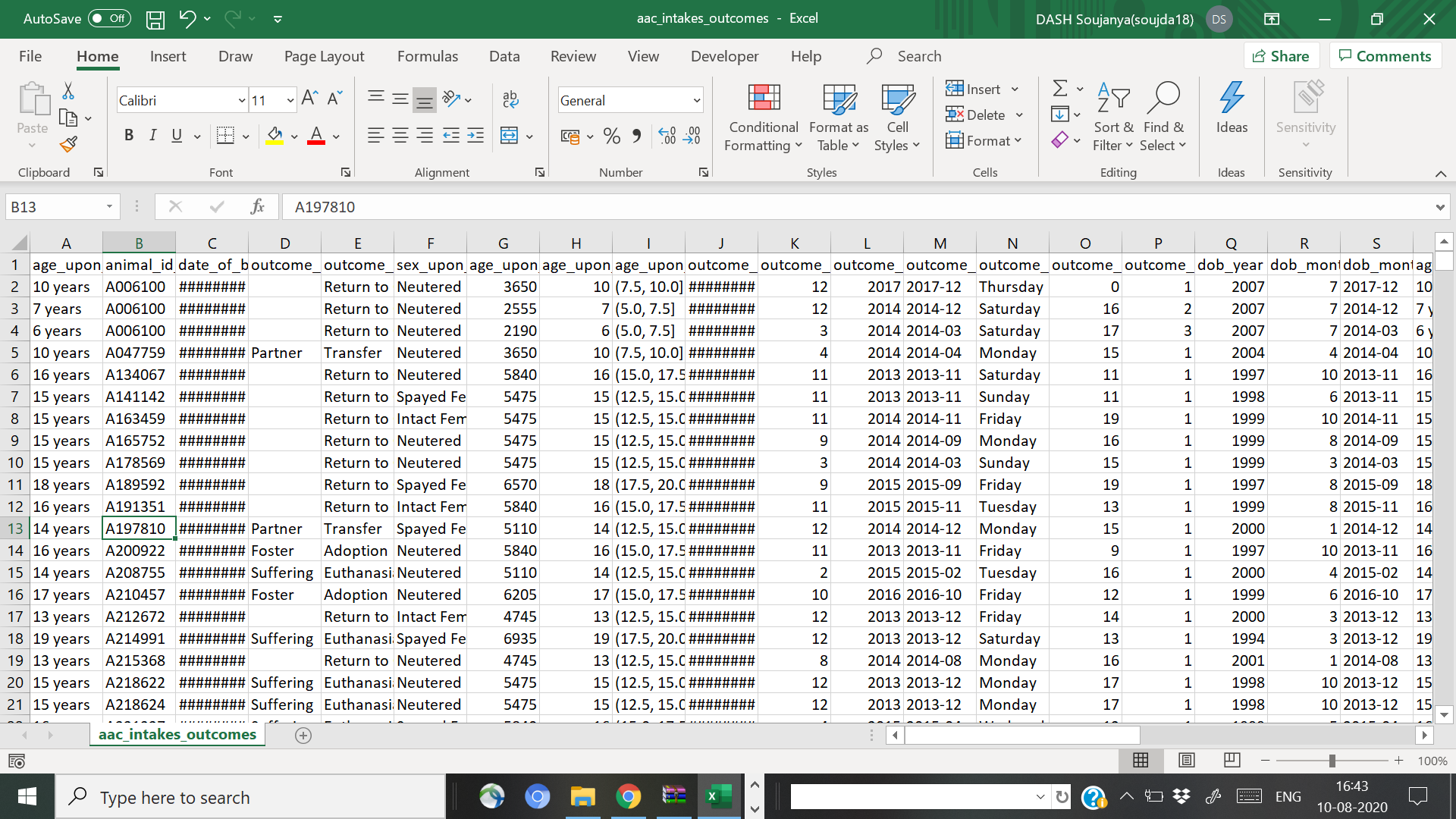


Figure Austin Animal Shelter Outcomes Data Set retrieved from Kaggle.com

**Brief about some of the concerned variables**: Some of the abbreviations used here might not be that known for all of us. Therefore, I planned to describe those here.

|  |  |  |
| --- | --- | --- |
| age\_upon\_outcome | Numeric Variable (Discrete) | The age of the animal upon outcome |
| animal\_id\_outcome | Numeric Variable (Discrete) | The animal's outcome ID. Should match the intake ID column |
| date\_of\_birth | Numeric Variable (continuous) | Date of birth of the animal. Estimated if exact birthdate is not known |
| outcome\_subtype | Categorical Variable | More specific outcome type corresponding to the outcome type where appropriate |
| outcome\_type | Categorical Variable | Adoption, Transfer, Return to owner, Euthanasia, Rto-Adopt, |
| sex\_upon\_outcome | Categorical Variable | The gender of the animal and if it has been spayed or neutered at time of outcome: Intact Male, Spayed Female, Neutered Male, female. |
| age\_upon\_outcome (days) (years) | Numeric Variable (continuous) | The age of the animal upon outcome represented in days, years etc. |
| age\_upon\_outcome\_age\_group | Categorical Variable | Grouped bins of the animal ages upon outcome. Goes by 2.5 year increments. |
| outcome\_datetime | Numeric Variable (continuous) | Date and time when the outcome occurred. |
| Outcome\_month (year, day) | Numeric Variable (continuous) | Month and year of outcome represented as a datetime |
| outcome\_number | Numeric Variable | Numeric value denoting if an animal has been released from the shelter more than once. Values higher than 1 indicate the |
| dob\_year (month, month year) | Numeric Variable (continuous) | Date and time of the birth date |
| age\_upon\_intake | Numeric Variable (continuous) | The age of the animal upon intake. |
| animal\_id\_intake | Numeric Variable (Discrete) | The unique ID given to the animal upon intake. Should match with the animal outcome ID. |
| animal\_type | Categorical Variable | Type of animal. May be one of 'cat', 'dog', 'bird', etc. |
| Breed | Categorical Variable | Breed of the animal |
| Color | Categorical Variable | Color of the animal |
| found\_location | Nominal Variable | Street address or general area in which the animal was found |
| Intake\_condition | Categorical Variable | The intake condition of the animal. Can be one of 'normal', 'injured', 'sick',etc. |
| Intake\_type | Categorical Variable | The type of intake, for example, 'stray', 'owner surrender', etc. |

|  |  |  |
| --- | --- | --- |
| sex\_upon\_intake | Categorical Variable | The gender of the animal and if it has been spayed or neutered at the time of intake |
| age\_upon\_intake (day, years) | Numeric Variable (continuous) | The age of the animal upon intake represented in days and years |
| age\_upon\_intake\_age\_group | Nominal Variable (ordinal) | Age group of the animal upon intake. Groups are in increments of 2.5 years |
| Count | Numeric Variable | Helper column for tabulating counts. All rows in this column are 1. |
| intake\_datetime | Ordinal Variable | Date and time when the intake occurred |
| Intake\_month (yea, monthyear) | Ordinal Variable | Month and year of intake as datetime |
| Intake\_weekday | Ordinal Variable | The day of week when the intake occurred |
| intake\_hour | Ordinal Variable | Hour represented as value from 1-24 denoting the hour in which the intake occurred |
| Intake number | Numeric Variable | The intake number denoting the number of occurrences the animal has been brought into the shelter. Values higher |
| Outcome\_hour | Ordinal Variable | Hour of the outcome represented as a numeric value from 1-24 |

Also, in the **outcome** attribute (Class variable in this case), there are a few labels that might be misleading.

* **Adoption:** the animal was adopted to a home
* **Barn Adoption:** the animal was adopted to live in a barn
* **Offsite Missing:** the animal went missing for unknown reasons at an offsite partner location
* **In-Foster Missing:** the animal is missing after being placed in a foster home
* **In-Kennel Missing:** the animal is missing after being transferred to a kennel facility
* **Possible Theft:** Although not confirmed, the animal went missing as a result of theft from the facility
* **Barn Transfer:** The animal was transferred to a facility for adoption into a barn environment
* **SNR:** SNR refers to the city of Austin's [Shelter-Neuter-Release](http://www.austintexas.gov/blog/changes-made-shelter-neuter-return-cat-program-reflect-community-stakeholder-input) program. I believe the outcome is representative of the animal being released.
* **Disposal:** The animal died while in the shelter and its body was disposed of.
* **Missing**: The animal went missing from the shelter for unknown reasons and has not been located.
* **Relocation:** The animal was transferred to a different location (usually the associated outcome subtype will have more information as to what facility the animal was moved.
* **Rto-Adopt**: It represents the animal was returned to its owner through the shelter's adoption process.

For Intake attribute:

1. **Feral:** The animal was taken to the shelter as a feral (fearful and possibly aggressive to people)
2. **Euthanasia Request:** The animal was brought to the shelter to be euthanatized for unknown reasons.
3. **Public Assist:** the animal was brought in through an outside rescue organization.

For Outcome subtypes:

* **In Kennel**: The animal was transferred to a kennel
* **SCRP**: Stray Cat Return Program. The cat was likely moved to be placed with that program rather than the shelter.
* **SNR:** Shelter, neuter, return. A variant of TNR (trap, neuter, release). More information can be found at Maddie's Fund[[2]](#footnote-2).

**4. Data Pre-processing and Cleaning**:

This is an indispensable step that aids in enhancing the quality of data to promote the extraction of meaningful insights from the data. This is usually the step before data visualization. With this process it is possible to transform raw data into an understandable and readable format. The steps are as follows: acquiring data, importing all crucial libraries, importing the dataset, identification and handling of missing data, encoding the categorical data, splitting and feature scaling.

Here I will be using a mix of Python programming and rapid miner to pre-process the dataset and then check on Weka for further investigations.

**4.1 Importing data:**

**Using RapidMine**r: Rapid Miner is an easy tool to perform this task and can be used by an amateur as it also provides troubleshooting for errors.

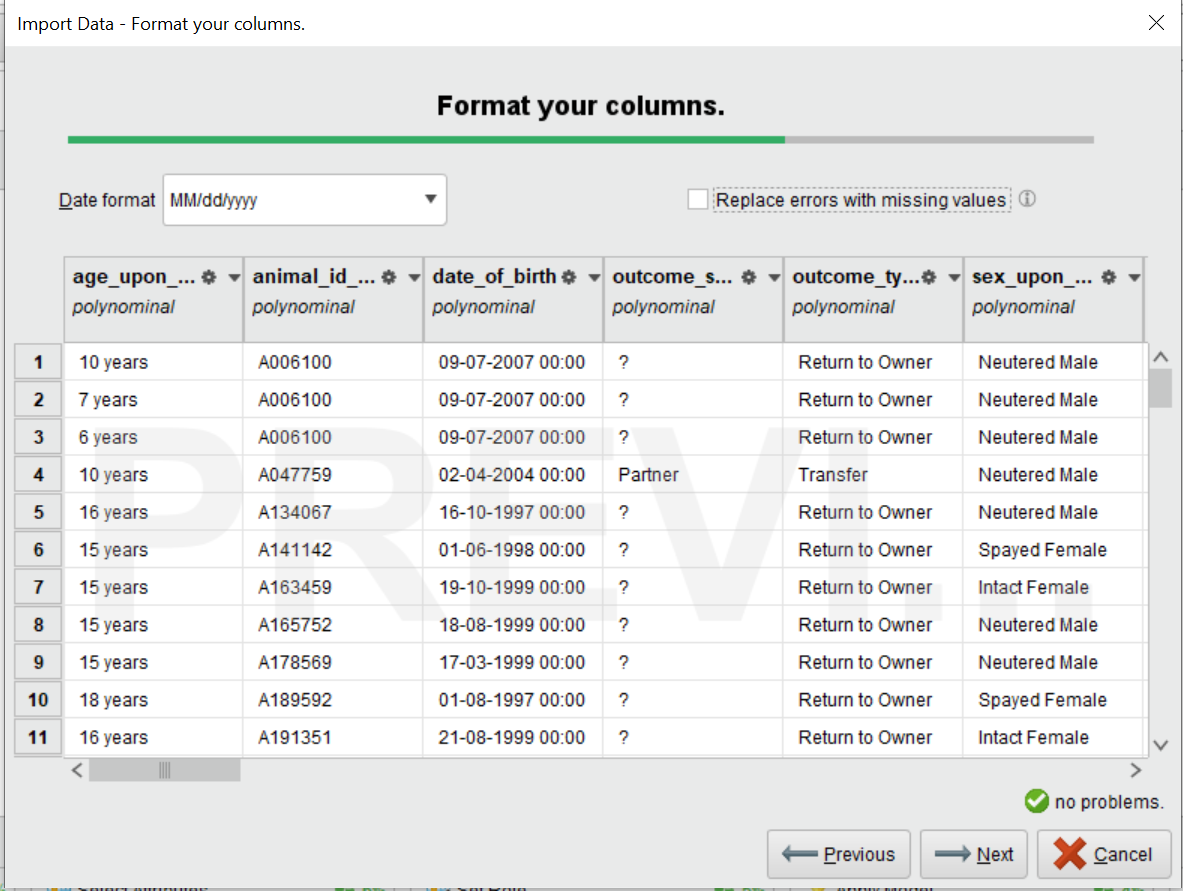


Figure Data Pre-processing using Rapid miner software (Importing data)

The missing values can also be dealt here when we use import data (first by making a new repository and directly importing data). Another way of **pre-processing** is to click on **Auto Model** and the in-built algorithms will clean the data for you so that you can customize it to your taste later.

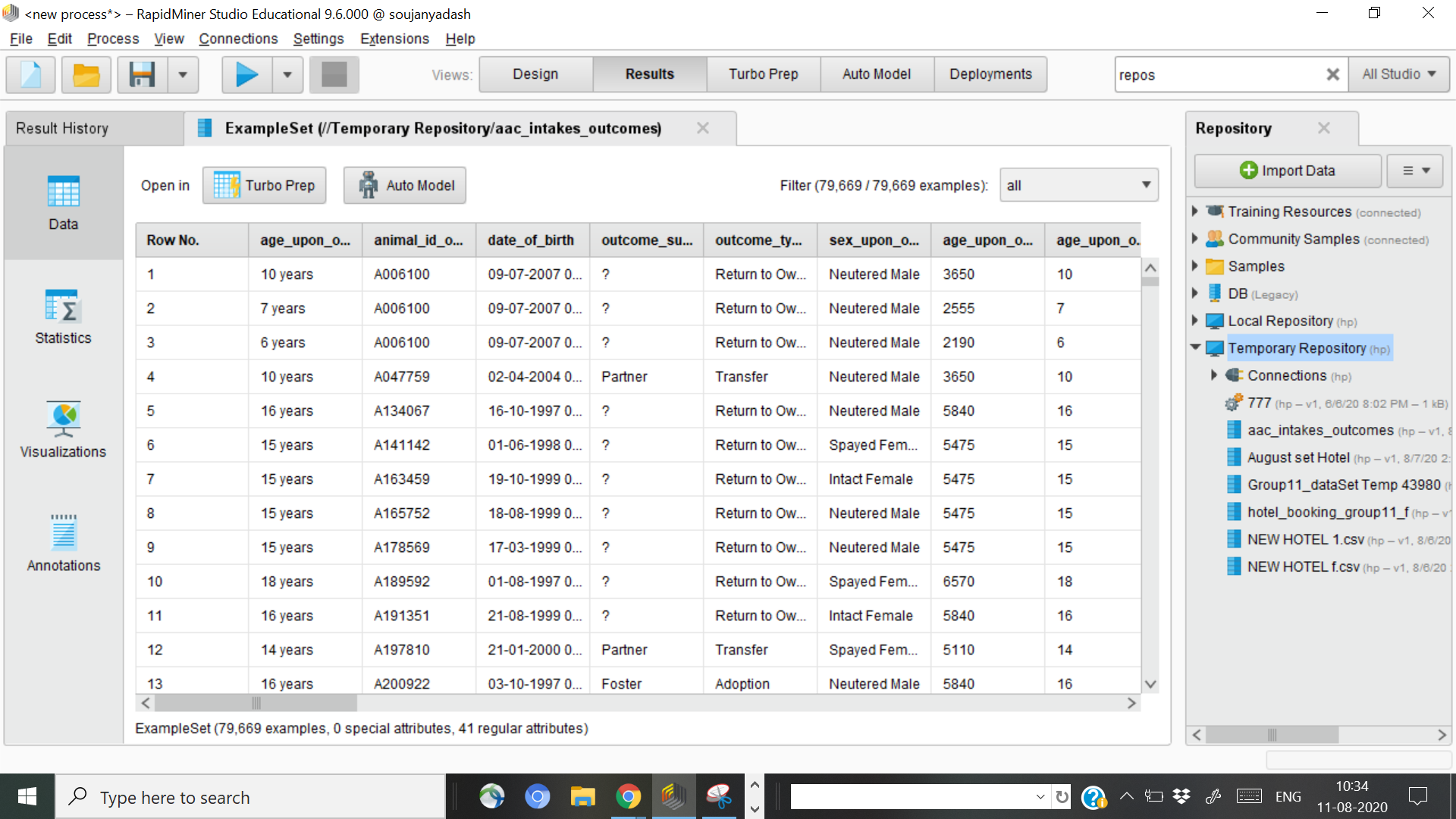


Figure Data Pre-processing can be done automatically by clicking on Turbo Prep option in Rapid Miner

**Using Python using Pandas and NumPy**: Using the following code we can import the data for further work.

*Import pandas as pd*

*mydata1 = pd.read\_csv("C:\\Users\\Animal Shelter\\Documents\\file1.csv", header = None)*

Then we sorted it (the breed) out in ascending order and only selected the ones that were Pit Bulls, Pit Bull Mix breed or Pit Bull ambiguity such as Pit Bull/ Labrador etc. As mentioned earlier, we will only focus on this breed for our predictions.

**4.2 Dealing with Missing data:**

Missing data is a problem because the absence of **data** reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. The lost **data** can cause bias in the estimation of parameters. Finally, it can reduce the representativeness of the samples.

Like I mentioned before, I will be using Rapid-Miner for preliminary cleaning where we found that missing data is in the attribute **outcome\_subtype** and so we filled the missing data with **NA (Not applicable)**. Then cross-checking it with Python. To cross-check, we also used Python and Pandas (as libraries). The code used to check:

*# Importing Libraries*

*import pandas as pd*

*Import numpy as np*

*#Read csv file into a pandas dataframe*

*Df = pd.read\_csv(“aac\_intakes\_outcomes\_Pit\_Bull\_only”)*

*# Take a look at the first few rows*

*Print df.head()*

*# Looking at the ST\_NUM column*

*from sklearn.impute import SimpleImputer*

*imputer = SimpleImputer(missing\_values=np.nan, strategy=’mean’)*

*imputer.fit(Training\_Set[:,39:79673]) Training\_Set [:, 39:79673] = imputer.transform(Training\_Set [:, 39:79673]) print df [‘ST\_NUM’] print df [‘ST\_NUM’].isnull()*

The result was the Boolean response True and this suggests that there aren’t any missing values after we corrected it with Rapid Miner.

**4.3 Removing attributes**:

There are some attributes which have really low significance and we are better off without those. Therefore, animal ID, breed (as all are PitBulls) and animal (all are dogs) are removed from the attributed section.

**5.0 Derived Variables**:

**5.1 Boolean Variables**:

Booleans can have only two possible values: **True (1)** or **False (0)**. Instead of the outcome number, it is better to make it as **“returned”** which would mean more than one visit or one visit only. If it is just 1 then False otherwise True.

Formula: IF(P2>1,” True”,” False”)

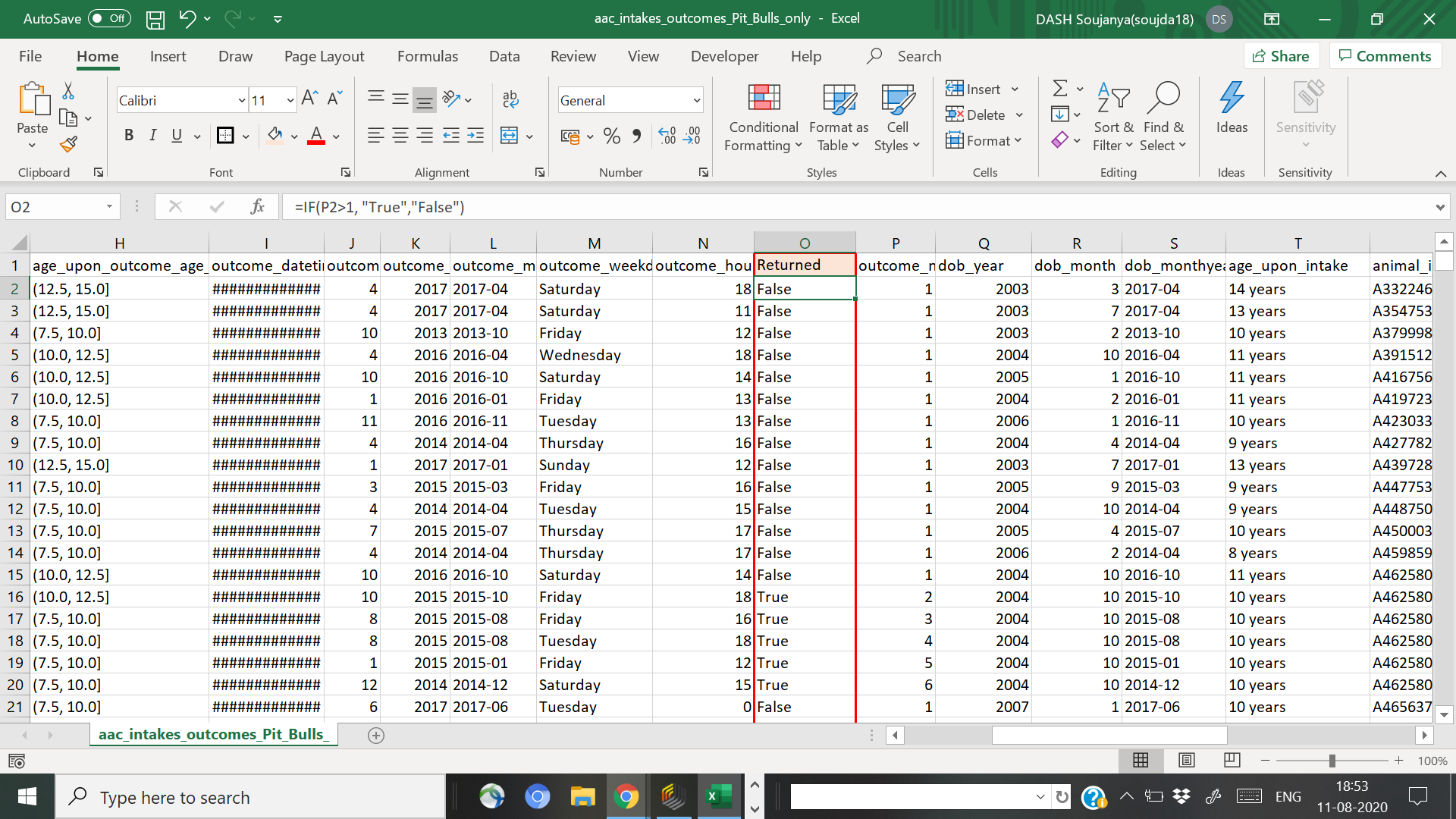
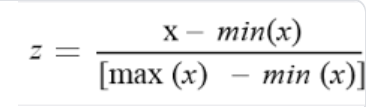


Figure New Boolean Variable: Returned

**5.2 Normalization**:

This is a data preparation technique for machine learning with the goal to change the values of numeric columns in the dataset to use a common scale, without distorting the differences in the ranges of values or losing information. The formula is provided below:



This was done using Rapid Miners in-built function after the cleaning process (Using the Turbo-Preparation option). This was used specifically for adoption days.

**5.3 Binning**:

Binning is a way to group several continuous values into smaller numbers called “bins” (retrieved online). Binning also results in converting numerical values to categorical values. For the **age groups** for outcomes and intakes, a bin size of 2.5-year increments was used.

**5.4 Time in shelter**:

This is the difference between Intakes and outcomes dates and times.

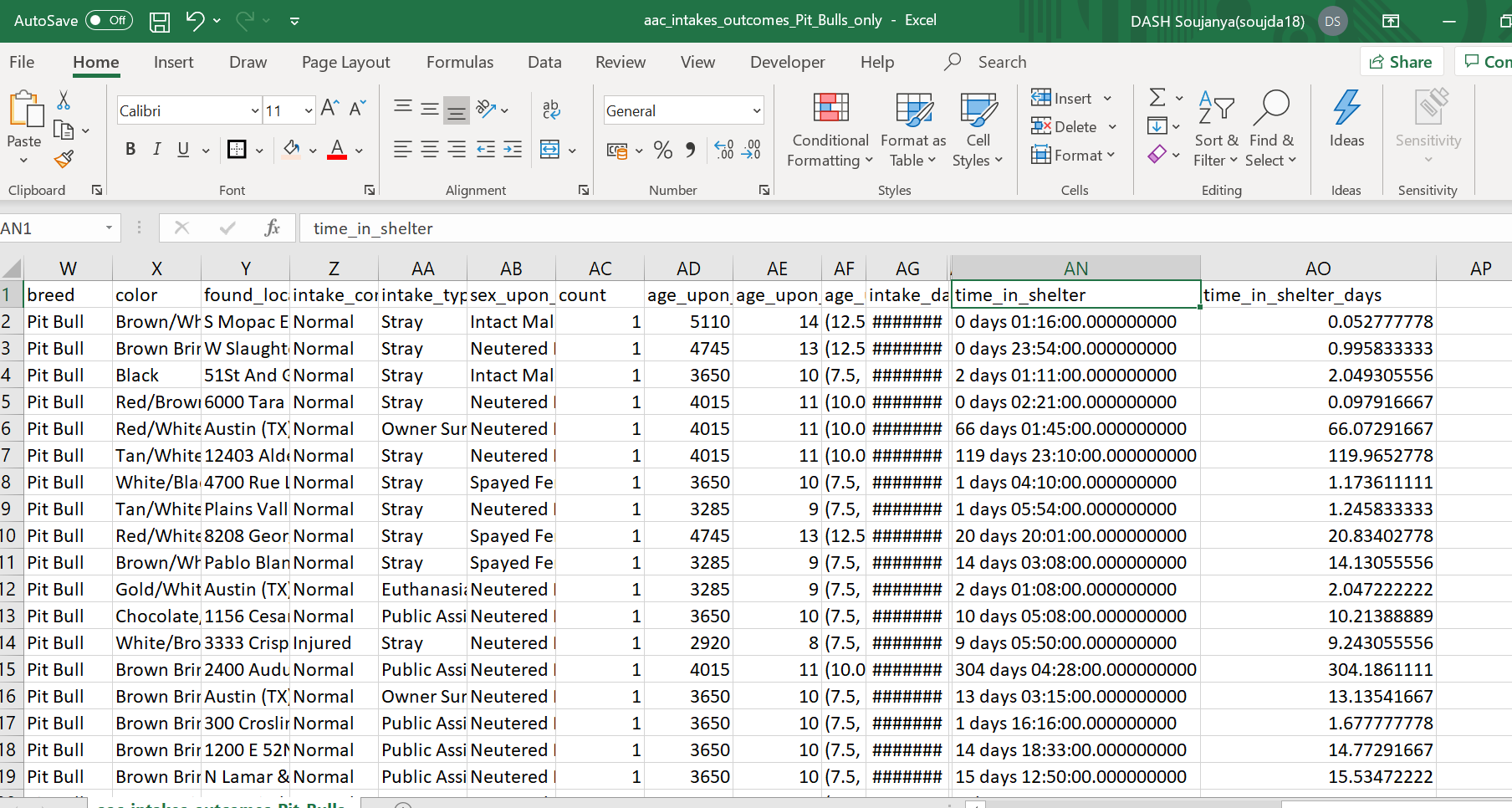
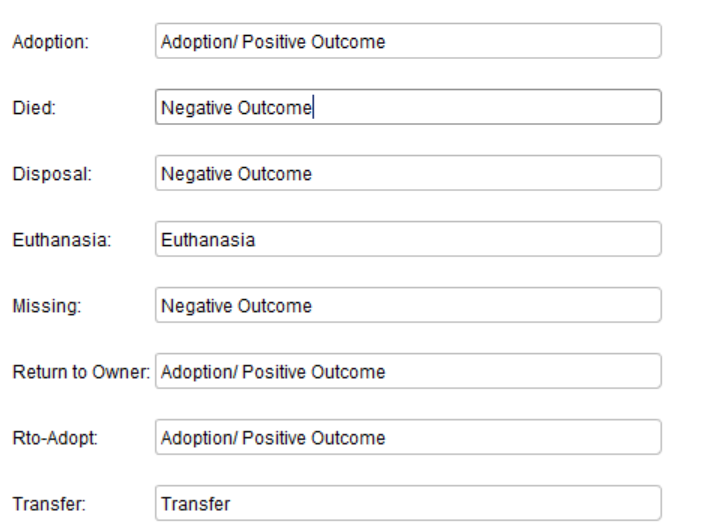


Figure New Attribute: Time in shelter

**5.5 Mapping (for class variable)**:

As there are too many different values for the class variable, the outcome, we mapped the result to the closest meanings using Rapid miner. These are as follows:



Euthanasia/ Negative Outcome

Euthanasia/ Negative Outcome

Euthanasia/ Negative Outcome

Euthanasia/ Negative Outcome

Euthanasia/ Negative Outcome

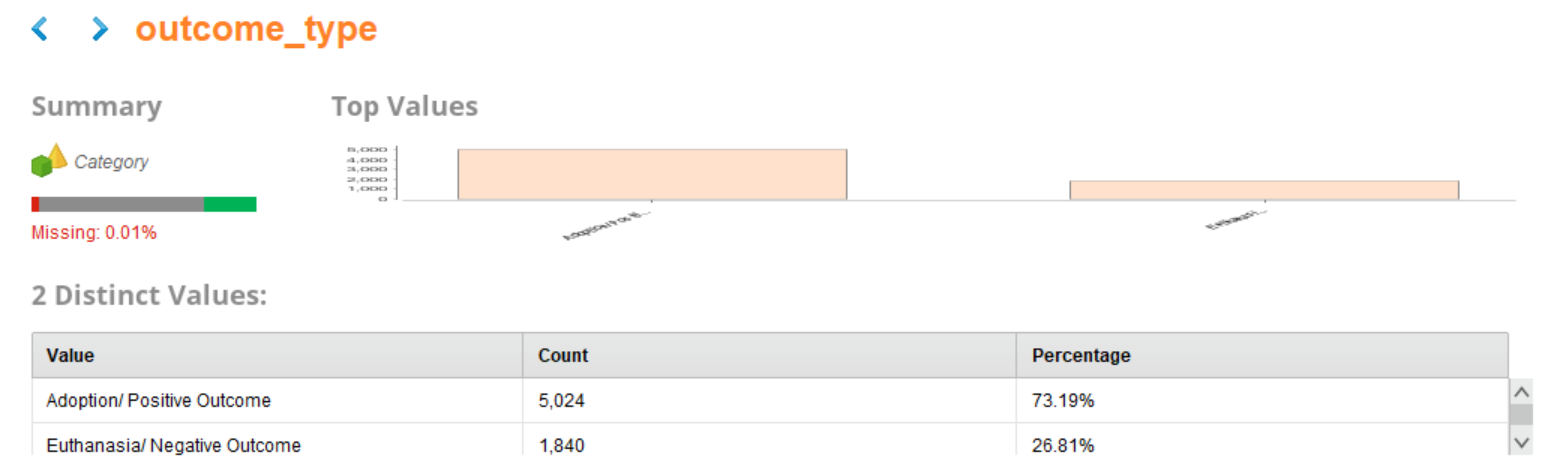
For comfortability of working with data, it was also assumed that Transfer, Euthanasia, Missing, Died and Disposal will be mapped as Euthanasia/ Negative Outcome.

**6.0 Data Visualization (using Tableau and other software):**

Data visualization is an essential part of mining process which is indispensable before formulating various business strategy. Visualization process helps the stakeholders to understand various loopholes in their business and the factors which might affect it. Various data visualization techniques which depend on the audience, context, purpose, dynamics, and content include charts, plots, Maps, diagrams etc.

The various tools which are used for this process are Tableau, SQL, Rapid Miner, Weka etc. for coding beginners otherwise Plotly (more complex tool used by coders) is widely used. In this project a mix of Python, Weka, Rapid Miner and Tableau will be used for visualization.

First, about the class variable, we can see an opposite trend as compared to what we expected. It seems Pit Bull and Pit Bull Mix have more adoption rate as compared to Euthanasia rate. For adoption rate, we can decipher that it is around 73.2% whereas for euthanasia rate, it is about 26.8%. This might be a case of class imbalance and we will deal with it later.



Euthanasia / Negative Outcome

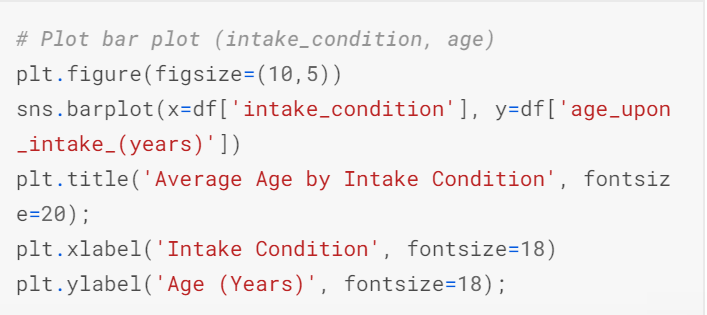


Adoption / Positive Outcome

Missing=0.0%

Figure Outcome\_type as the Class Variable

First our question should be what is the average age by intake condition? What is the age when the animal is reached out to shelters if it is injured? To answer such questions, we will look at this bar graph. The bar graph was made using python and the code is also provided below. The height of bars provides us with central tendency for a numeric variable (here it is age) while the black lines show the error. It seems that most of the aged dogs who reach out to such shelters are usually nearing their lifespan.

The maximum error is found in the other and pregnant category. Also, most of the injured and sick animals are dropped in the shelter when they are just 3 years old. From the data it is also inferred that PitBull’s lifespan is between 8- 15 years old.

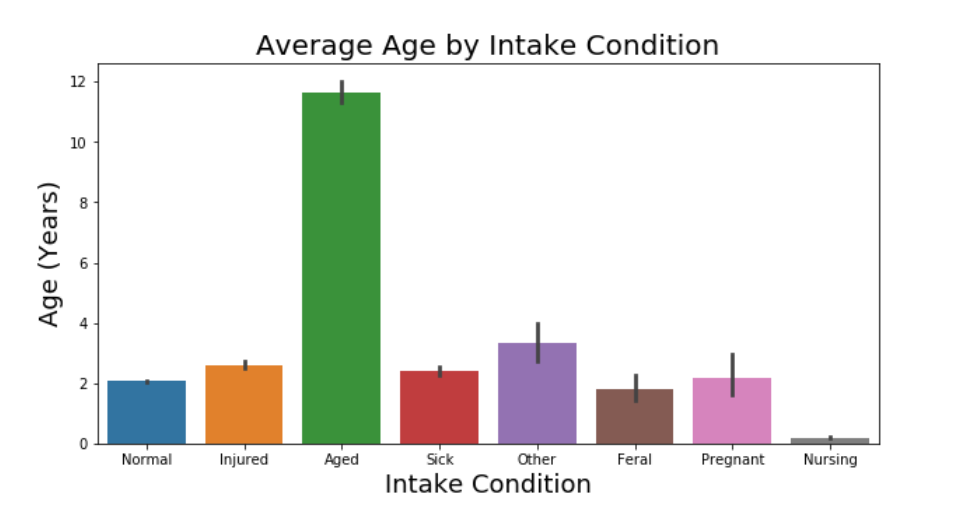


Figure Age vs. Intake Condition: Bars represent age whereas the black lines are extrapolated as mean errors

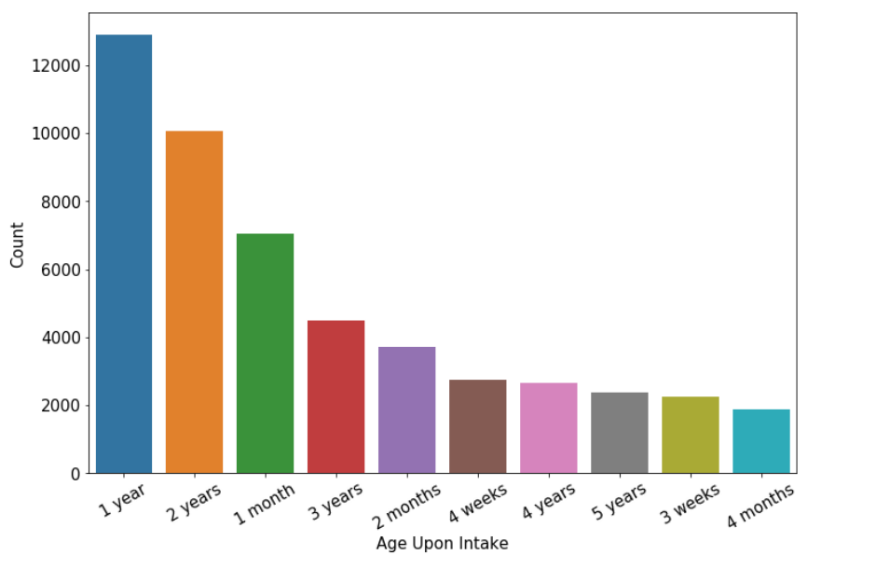
Obviously, the ideal intake age of any dog is 1 year when the dog has passed the socialization stage and either is active or calm and most reactive to any forms of dog training. Usually one month is not ideal but here it accounts for around 15% of the data. The reason might be because people want to adopt dogs which are younger. Also, as the age increases the adoptability decreases and if the age is as less as 3 weeks, it is not preferable as one has to put effort to toilet-train it and pass through the delicate times of contracting any kind of diseases.

Figure Count vs. Age upon Intake Bar graph

Next is the box plot graph of the same result in order to find the inter-quartile range of different conditions. It is obvious that the age of intake for the aged category is mainly between 10-15 years whereas for all other categories are less than 5 years old.

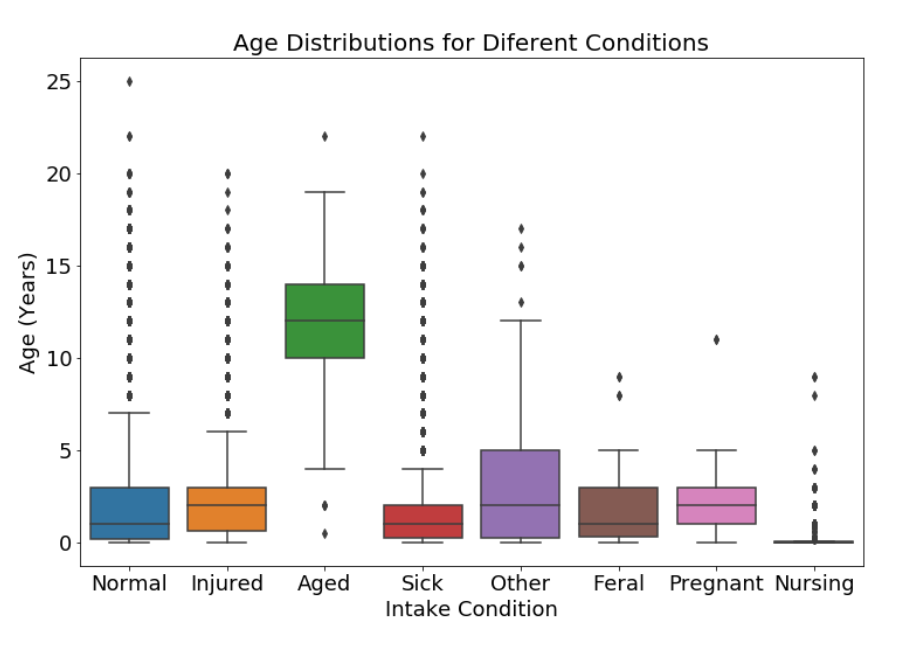
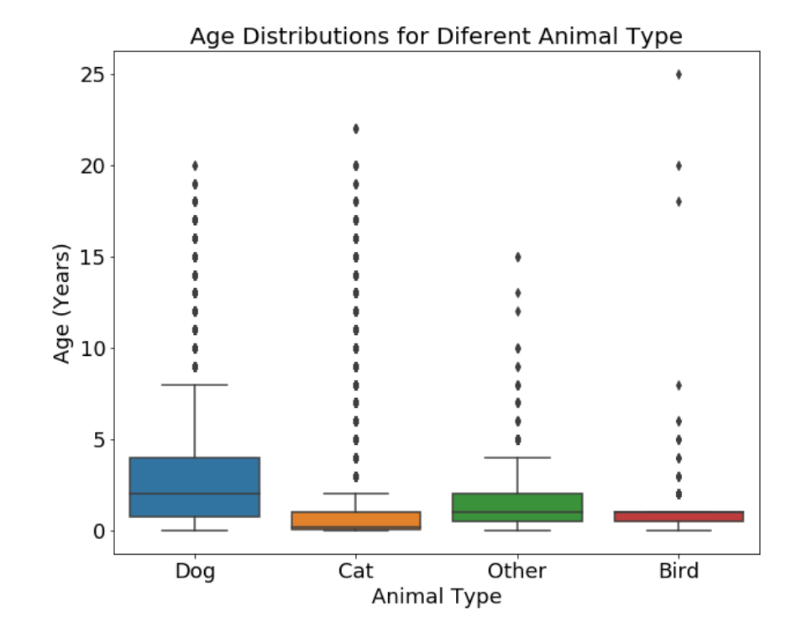
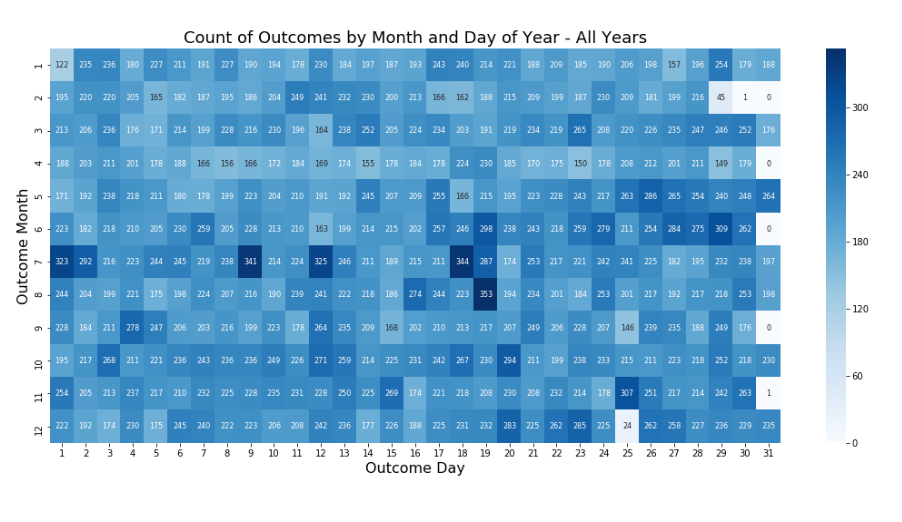


Figure Box Plot representing Age vs. Intake conditions

Even though our focus is on Pit Bulls, I would also like to compare the age of different animal during intakes with that of a dog. Obviously, for a dog, there seems to be more scope as compared to other animals due to the width of the box plot (around 5 years).

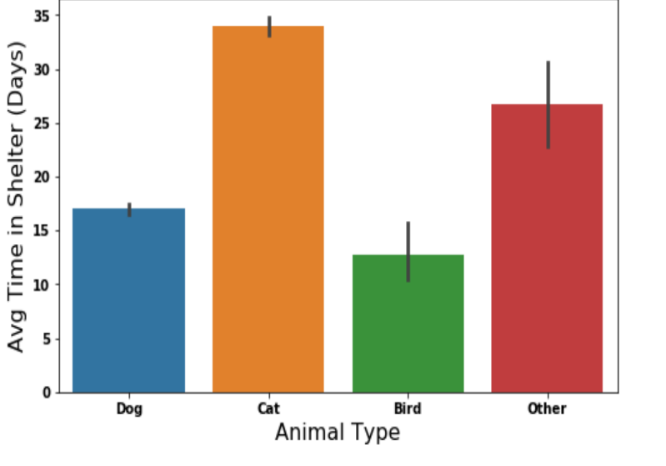
The following figure was very difficult to make so a combination of tools was considered in the process. Most positive outcomes were during July especially during the first half of the month. Overall, July is the month with the highest adoptions. But to understand why it is so, we need to study the data more.

Figure Box Plot: Age vs. Animal Type



**July**

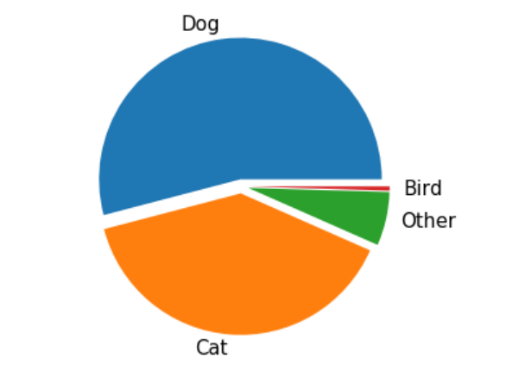
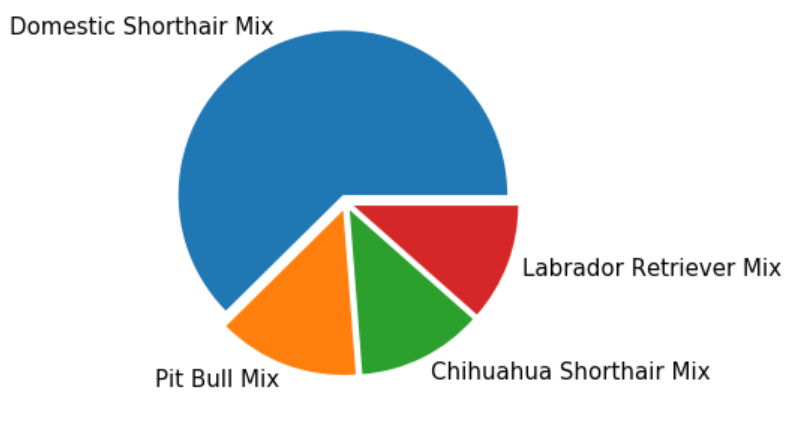
Data for positive Outcome

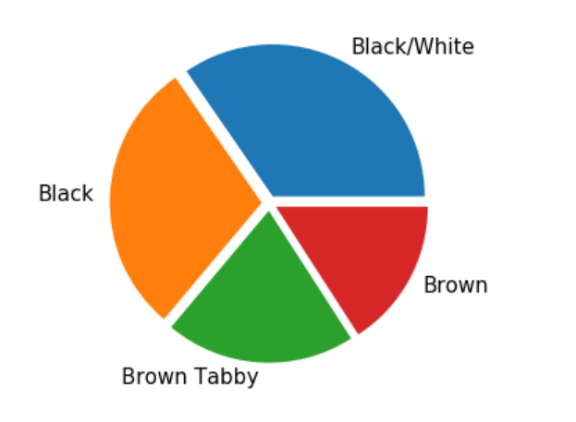


According to the data on the left, it takes more days for a cat to be adopted as compared to a dog and a bird.

Next is the pie chart which perfectly describes that adoption is more in dogs that to for the breed, Domestic Shorthair Mix, backed by Pit Bull Mix.

Figure Avg. Time in Shelter (Days) vs. Animal Type

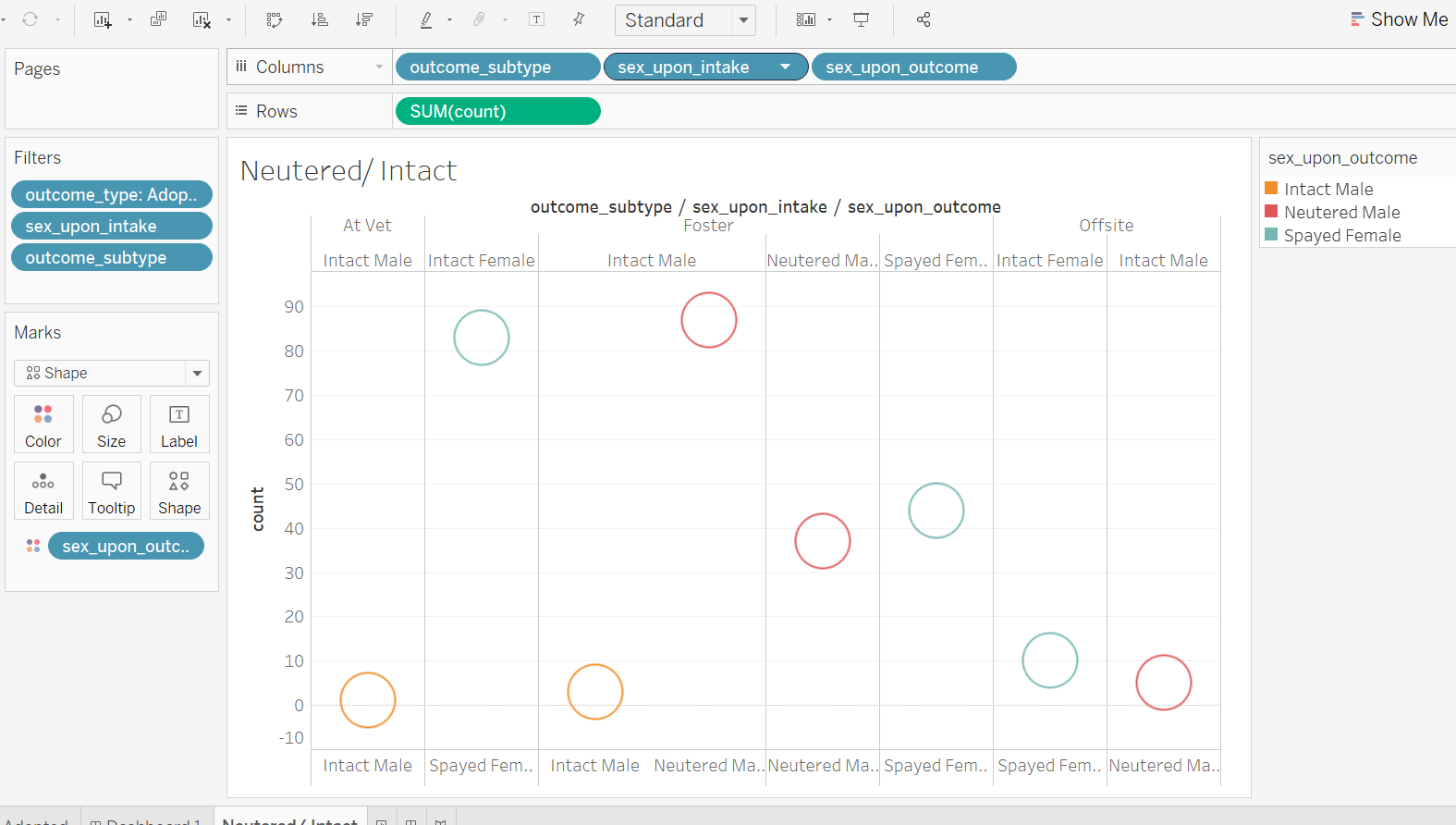


The pie chart at the left side depicts the popularity of colour of the breed Pit-Bull.

Black/ white variety seems to account for around 40% of adoptions and black alone accounts for 35% of adoptions. Brown and Brown Tabby in general seems to be less popular as compared to the former result.

The following depicts the intake type vs intake condition of Pit Bulls and Pit Bulls Mix (irrespective of gender) who had one or more return to the shelter. Most of the intake type seems to be stray (around 40%). For the adoption outcome, most intakes are brought-in with the help of public-assist and for euthanized those are owner surrender.





Sex upon outcome

Sex upon intake

Count for Adoption

Sex upon intake: Intact Female

Sex upon outcome: Neutered

This graph (Scatterplot) is to check whether a neutered Pit-Bull is more adoptable as compared to an intact one. According to the graph, it is clear that intact are less adoptable as compared to neutered or spayed one. If we check the subtype Vet, they would prefer adopting a spayed female or neutered male which is suggested by the high count (around 85 cases).

Similarly, when going for foster home, the same is the case for male Pit-Bulls which might indicate that they are more aggressive and difficult to train. There is rather a strange variation when it comes to that of female. Foster homes probably do not care much when it comes to female pit bulls ( they do not mind if it is spayed or not which is suggested by the two scatterplots above).

For Offsites, overall, they avoid pit bulls irrespective of gender which is quite a problem.

**Conclusion in short:** Adoptability of Pit-Bulls depends on whether they are spayed/ neutered or not. Colour also plays a great impact during adoption as shown by the pie chart that dogs whose breed is Pit Bull or mix can be made adoptable if it is of the colors Black/white, Black, Brown/ Tabby and Brown (priority and order maintained). Also, maximum intakes which are euthanised or adopted are actually stray dogs which might suggest that the previous owner abandoned them or the dog was lost.

**7. Class Balancer**:

It is essential to work on a balanced dataset for better predictions. We can make our algorithms work better, better in the sense, better accuracy, cost or less error and true positive rate etc. Our dataset is heavily imbalanced when we assume the outcome type to be our class variable. Positive accounted for around 73% whereas the other one accounted for only 27%. Therefore, we need to balance it and thus, weka’s **Class Balancer filter** was used (which is located in the pre-processing tab under the supervised instances section.)

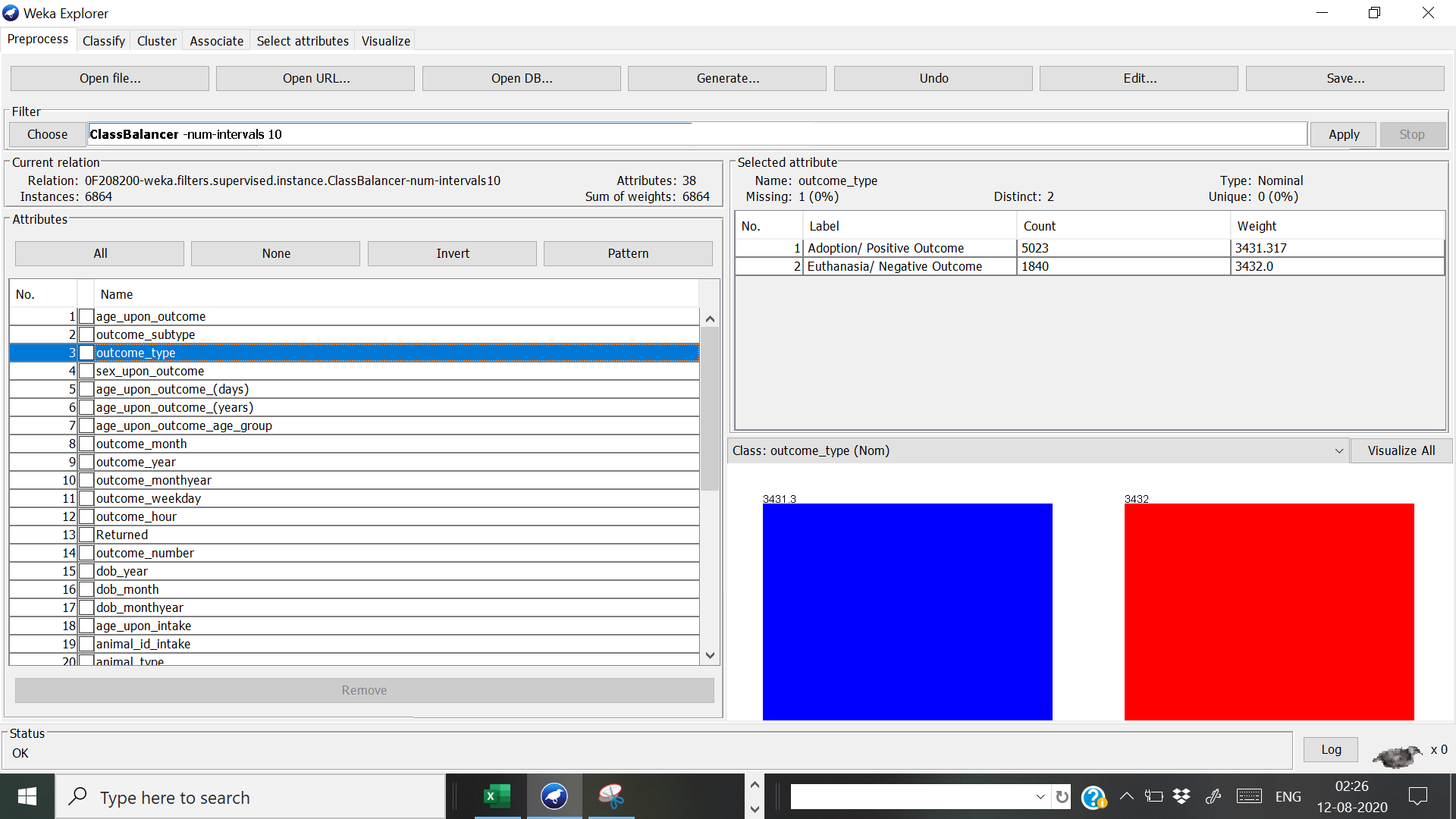


Figure Dealing with class imbalance using Class Balancer filter

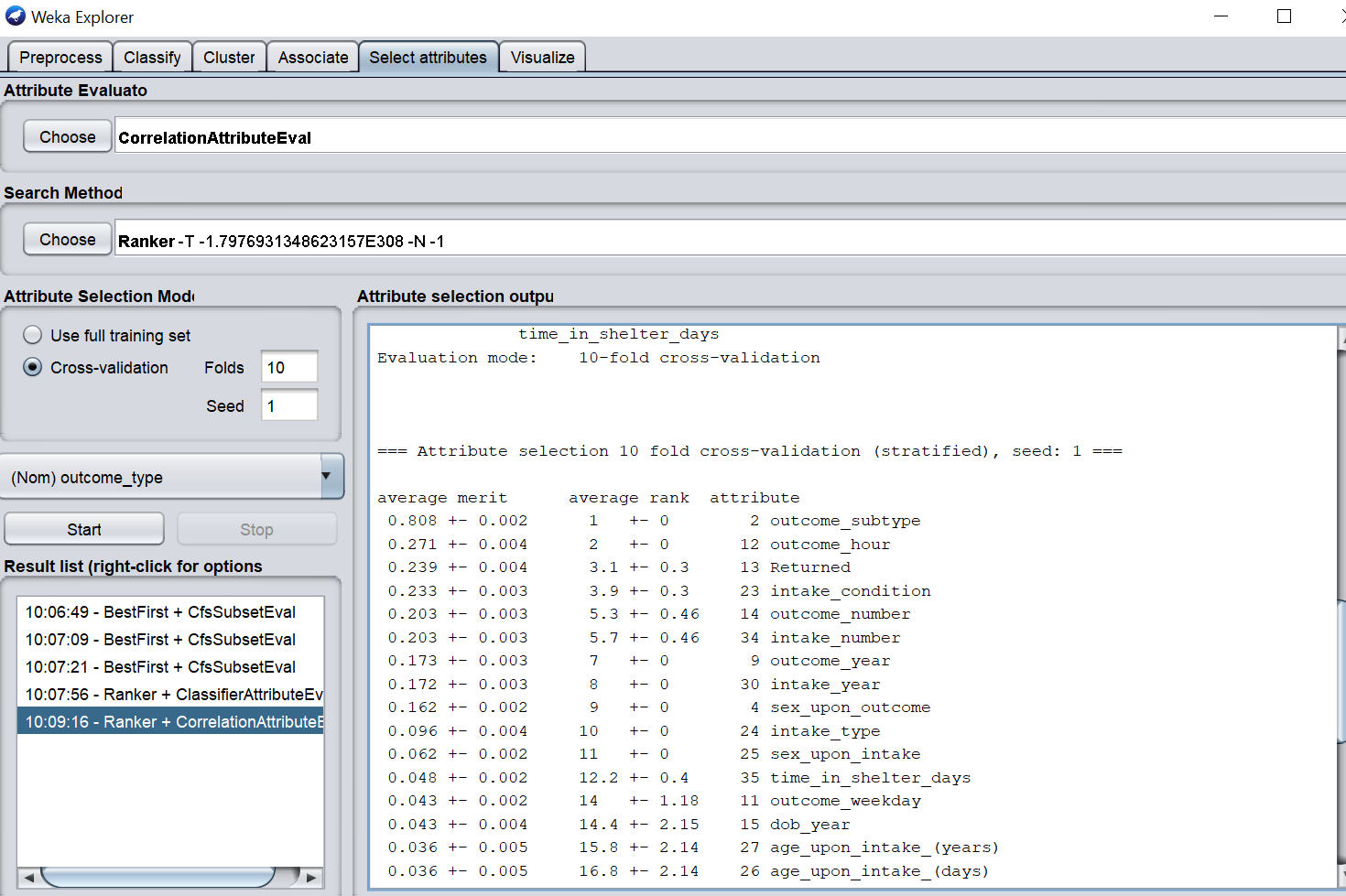
Now the distribution is 49.99% and 50.01% respectively.

**8.0 Attribute Selection**:

In Weka, there are three options of performing attribute selection from the commandline (and GUI): native approach, using a meta-classifier, filter approach. Attribute selection is essential to find the attributes that are most useful when classifying. During attribute selection different cleaning techniques were tested such as discretization vs. normalization, derived vs. not including, transformed vs. not transformed etc. (all done using Rapid Miner and Weka). The next step was to test the newly created subsets with the experimenter to search for the subset and classifier that provide us with better accuracy. The following is the process:

**8.1. Using Correlation AttributeEval method**:

To find the most relevant attribute, this method was used and it is in the ‘select attribute panel’ in Weka which evaluates the worth of an attribute by measuring the correlation. The search method used was ‘Ranker’ and selection mode is ‘Cross Validation’ 10 folds.



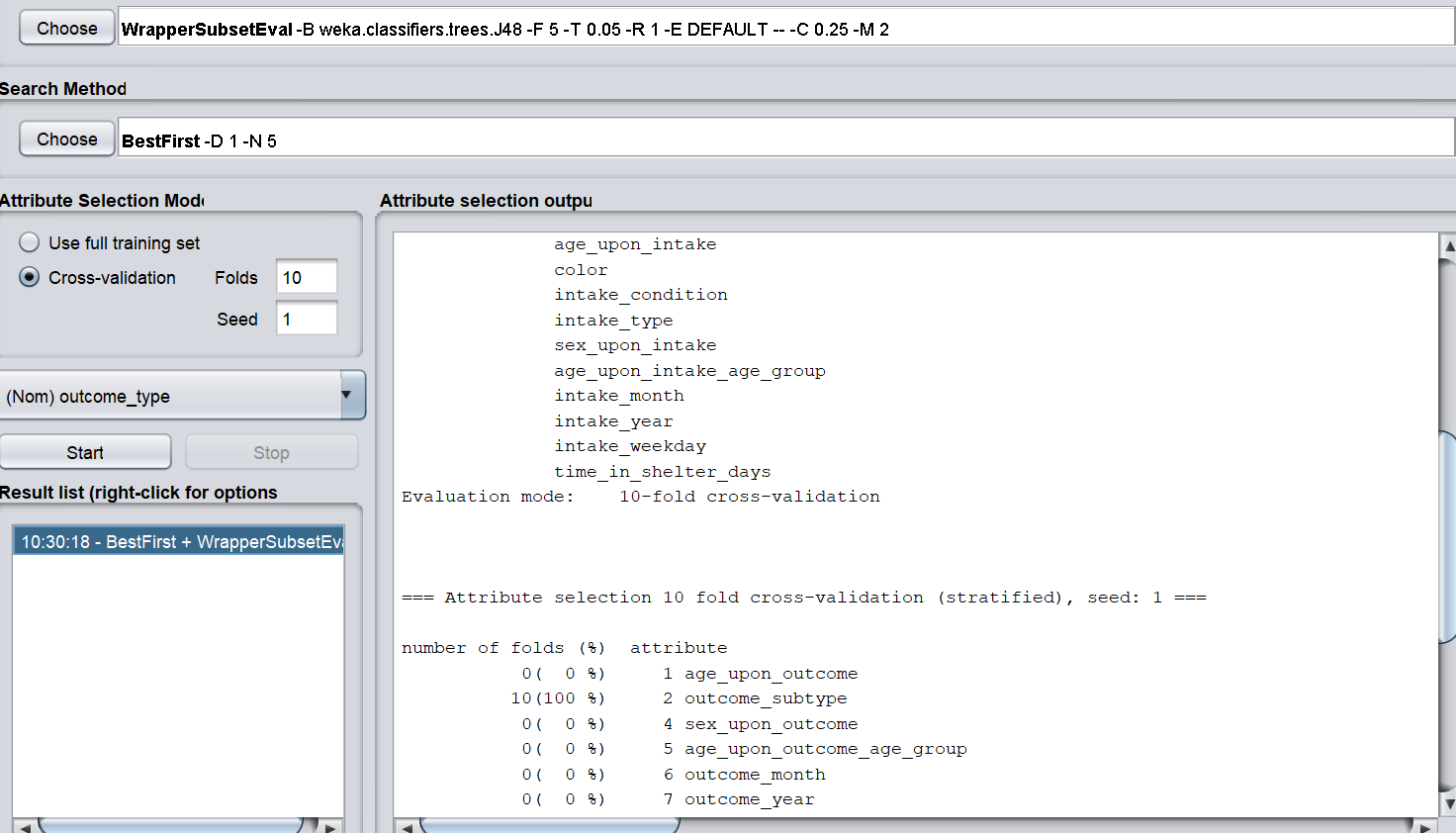
Obviously, outcome subtype is highly correlated with outcome type

Figure Attribute selection using CorrelationAttributeEval Method

According to the figure, the attribute with the highest correlation (positive correl coefficient close to 1) is Outcome\_subtype followed by outcome\_hour, Returned etc.

**8.2 Using Wrapper method**:

The wrapper method is used to evaluate attribute by utilizing the learned scheme of the data. Here the method was BestFirst (Bi-directional) and selection mode was Cross Validation 10 folds.





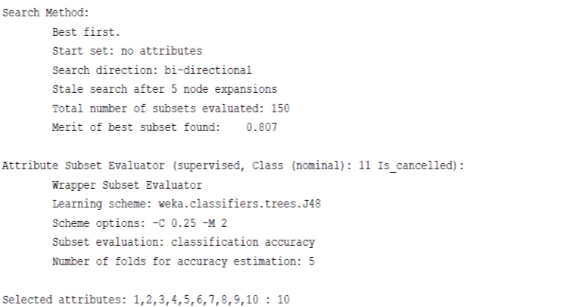
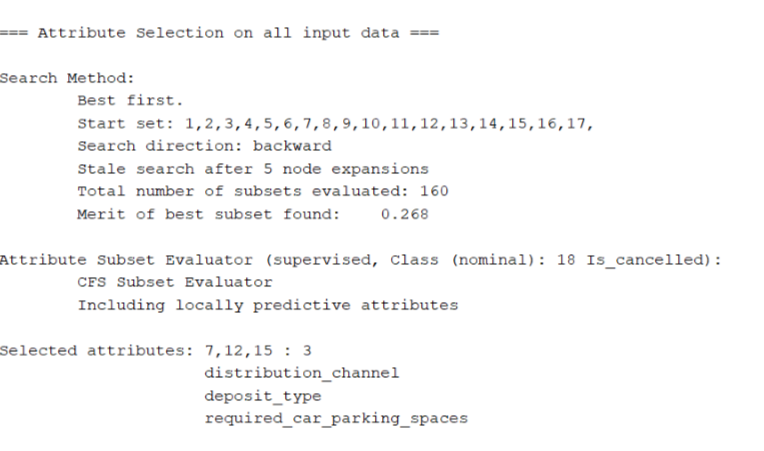


Figure Wrapper method: WrapperSubsetEval for Attribute Selection

Here the selected attributes are Outcome\_age, outcome\_subtype, Returned etc.

**3. Using CfsSubsetEval Method**:

Evaluates the worth of the subsets of attributes by considering the individual predictive ability of each feature along with the redundancy between them. Search method is Best first and the search direction is backward.



According to this method, the attributes 7, 12 and 15 are selected which are color, intake condition and intake type.

So, overall, the most important attributes seem to be Returned, outcome sub type, intake condition and intake type.



Figure Attribute Selection using CfrSubsetEval Method

**9. Testing and Training:**

The Train-Validate-Test Paradigm:

According to **Brownlee (2017)**, the train-validate-test paradigm is to divide the original dataset into 3 subsets which are training set, validation set, and test set. Training set is the sample of data to fit the model while testing set is the sample of data used to provide an unbiased evaluation of the final model fit on the training dataset.

In this project, two main methods will be used for this; **Cross-Validation** and **Percentage split** method (90%training and 10% testing) We used cross-validation to divide the data set into 10 folds in which 9 folds were training while last fold was testing (randomized for 10 times), the 11th fold is for deploying. But why do we need a validation set? It is to prevent the model from overfitting to the training set. Overfitting is when the model becomes too good at being able to classify the data in the training set but is unable to generalize and make accurate classifications on the test data. Weka does not provide a direct function to split the original dataset into 3 different datasets, so we will not be using the train-validate-test paradigm.

**10.0 Classification**:

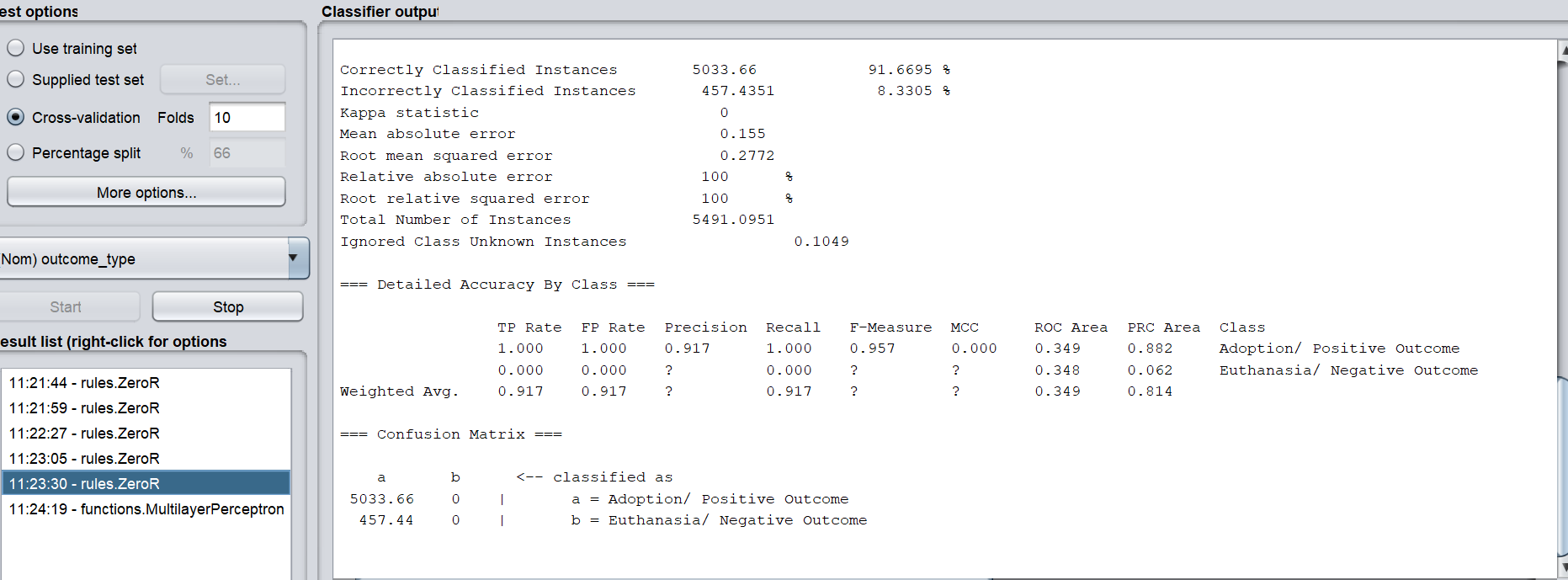
Classification in machine learning is a type of supervised learning which specifies the nominal class (here Outcome Type) to which data elements belong to and is best used when the output has finite and discrete value predicts a class for the input variable as well. Types of classification algorithms are:

➢ Linear models: Logistic Regression, SVM (Support Vector Machines)

➢ Nonlinear models: K-nearest neighbour, Naïve Bayes, Decision tree (J48), Random forest etc.

**10.1 Baseline Classifier:**

A baseline classification always classifies to the largest class. In Weka, there is the classifier known as ZeroR which serves the same purpose and finds the baseline accuracy for us so that we can compare it with all other classifiers and find the most suitable one for our prediction.



The average of these two will be considered as the Baseline Accuracy

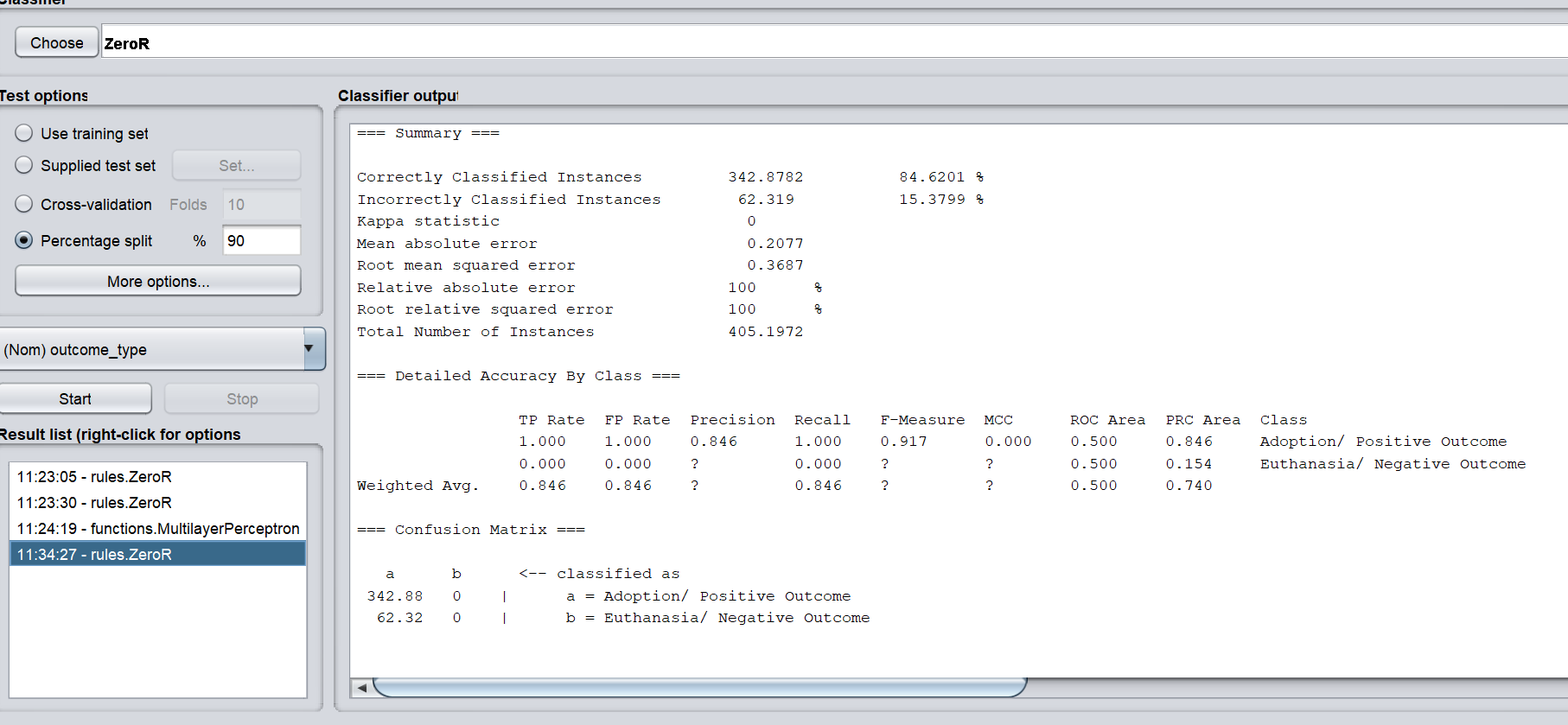


Figure Baseline Accuracy Top: using Stratified Cross-Validation; Bottom: Using Percentage Split

**The Baseline accuracy** for Percentage Split is 84.6201% whereas for Cross Validation, it is 91.6695%. So, as an average we’d consider the baseline accuracy to be 88.1448%. Only the classifiers with an accuracy more than this will be considered for further work.

*J48 algorithm did much better as compared to the Baseline accuracy which was 88.1448% thus, we will proceed with this algorithm for research, modelling etc.*

**10.2 J48 Decision Tree:**

Before starting up with anything else, what exactly is a decision tree?

A decision tree is a decision support tool that uses a tree like model of decisions (every node leads to a certain outcome) and their possible consequences where each internal node represents “test” on an attribute, each branch represents the outcome of the test and each leaf node represents a class label (Cravit, 2019). Similarly, (C4.5)J48 is an algorithm used to generate a decision tree (made in Java script) and is the best ML algorithm to examine the data categorically and continuously (Ian Witten).

The figure below shows that when we use this algorithm, our accuracy is 99.979% which is way better than the baseline accuracy and thus, we can research on the same and use for other machine learning works. Also, if we check the confusion matrix (will be discussed later), the error rate is as low as 0.019% for the adoption rate which is ideal for us.

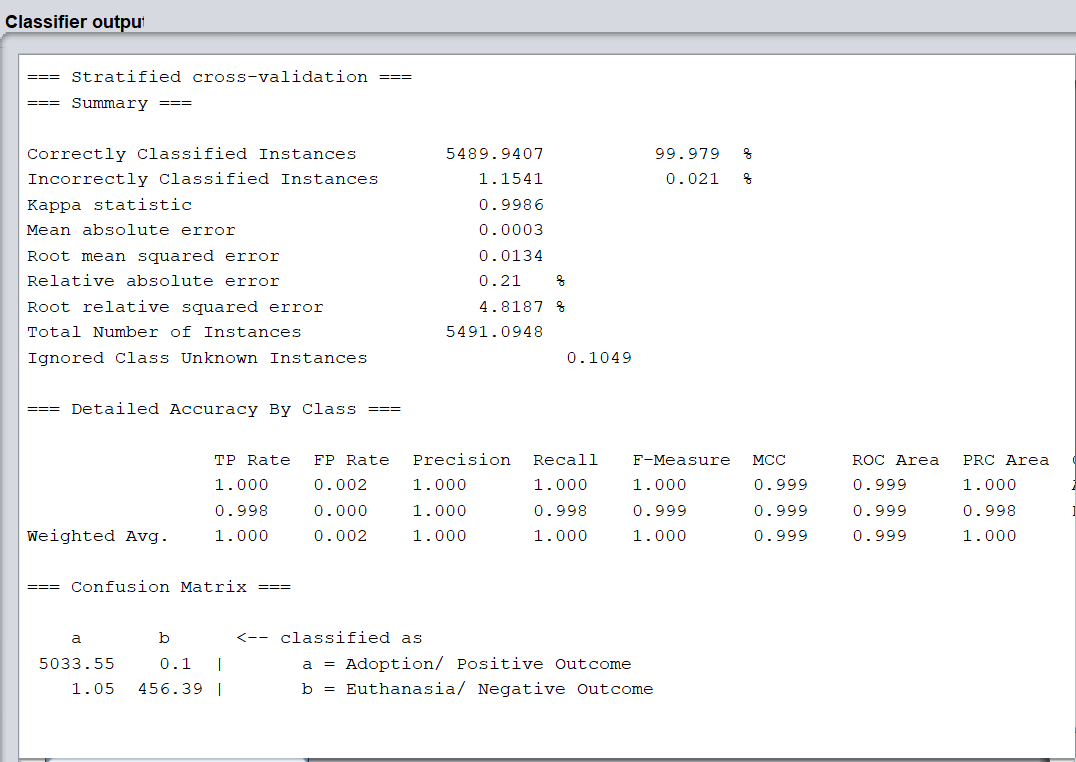


Figure Classifier Output: Performance accuracy for J48 Decision Tree is 99.979%

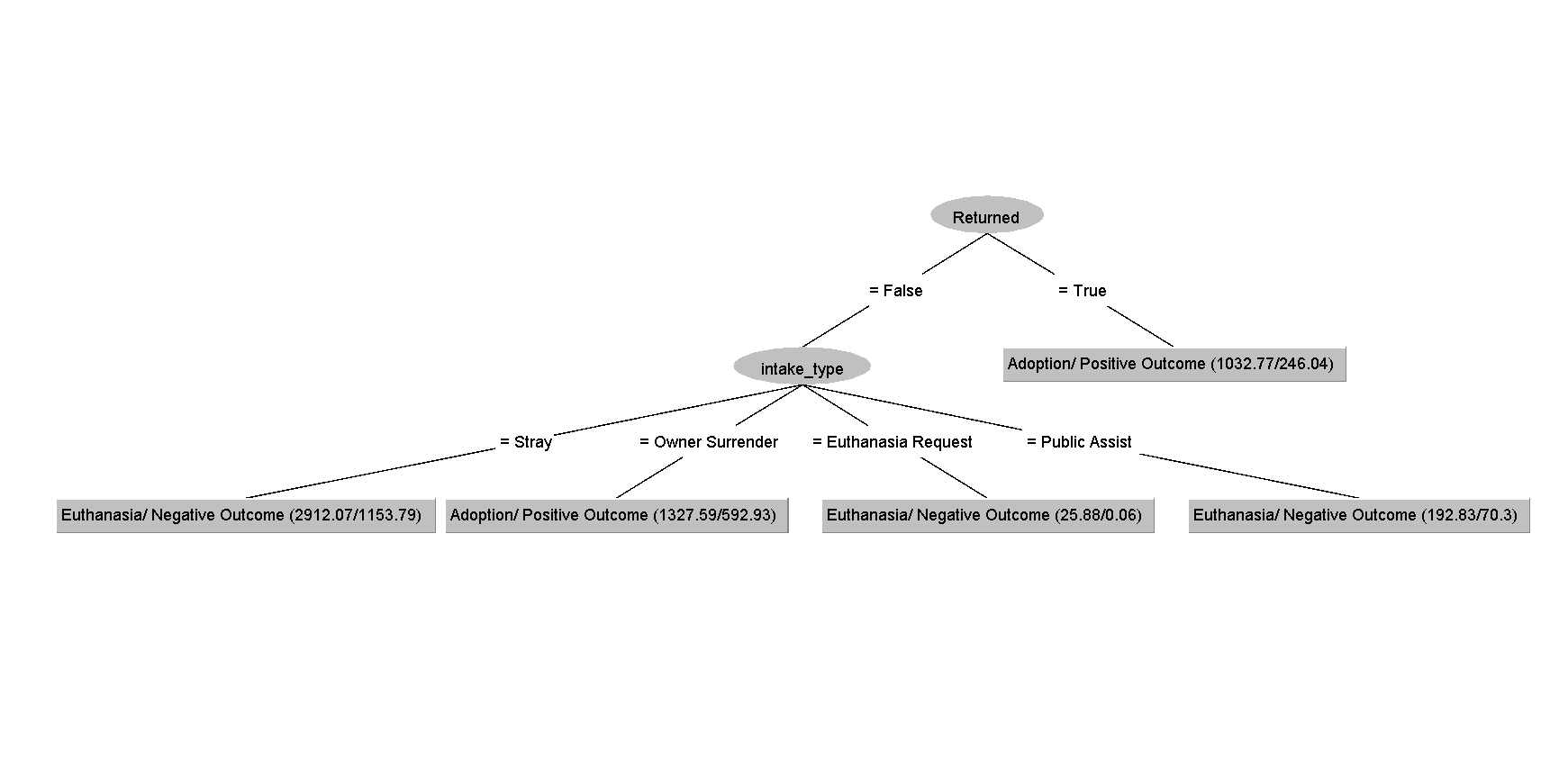
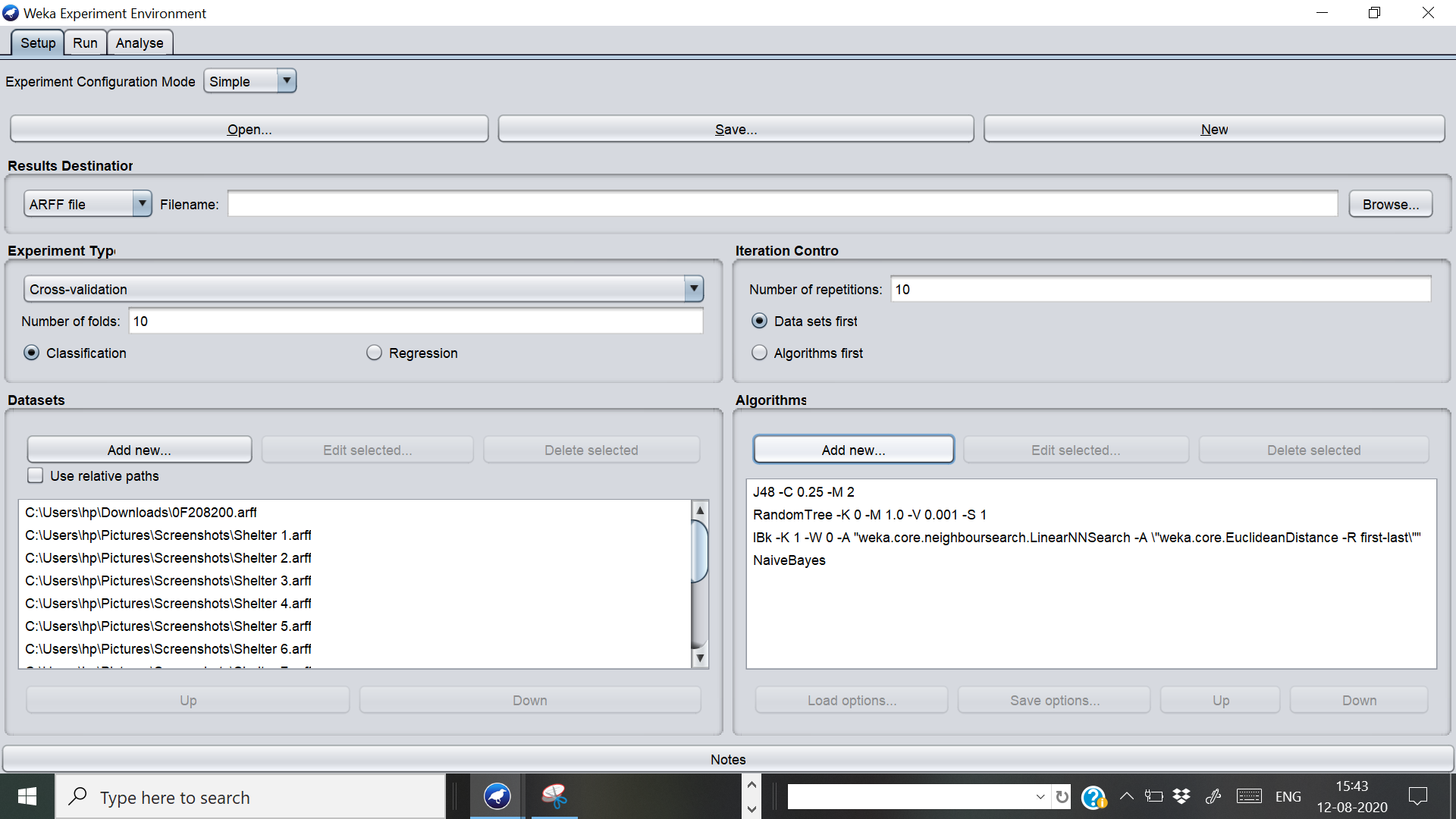


Figure J48 Tree visualization

**11. Experimenter**:

The Weka Experimenter allows one to design one’s own experiments of running algorithms on datasets, run the experiments and analyse the results. It’s a powerful tool.

In this interface, Experimenter Type was used: Stratified Cross-Validation 10 folds was chosen for classification. The 4 algorithms used were J48 Decision Tree, Ibk (K Nearest Neighbour), Naïve Bayes and Random Tree and these were used on 15 variations of the same dataset (extracted using attributes selection method discussed above, the derived variables and normalised instances).

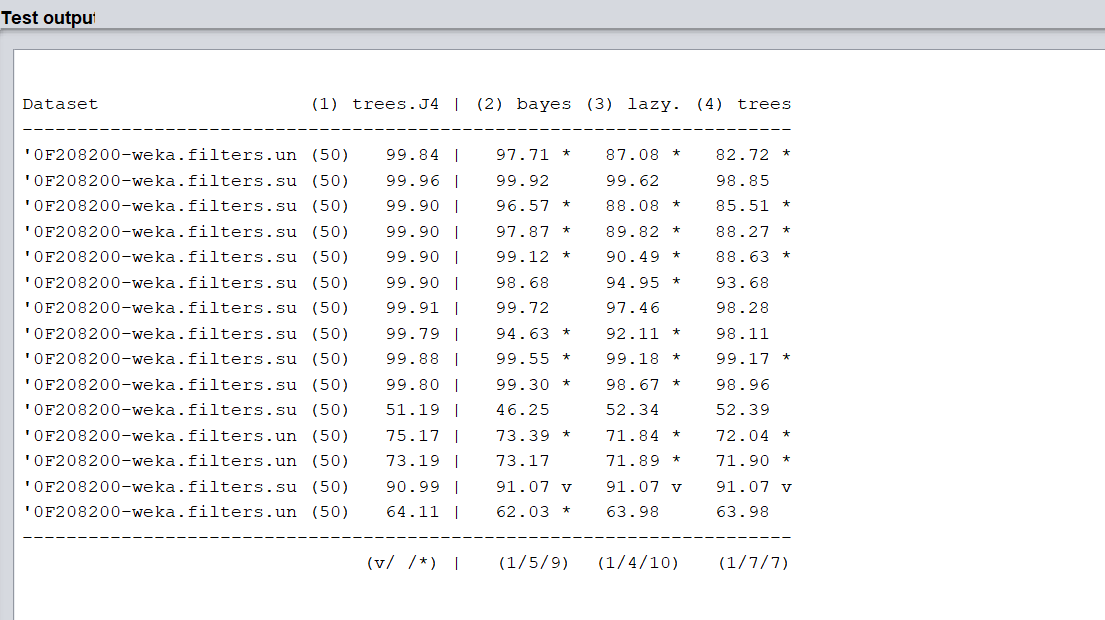


15 variations of dataset is used

Algorithms(4 classifiers) used are J48 Decision Tree,IBk,Naïve Bayes, Random Tree

Figure Experimenter: Left is the 15 variations of the Dataset; Right is the 4 algorithms used

The first result is used to select the algorithm and dataset with the best performance. Here the dataset 2 (also called Shelter\_1) seems to have overall great performance with an average of around 99.58%. Similarly, overall best performing algorithm is J48 tree which has an average accuracy of about 99%. The second-best classifier seems to be Naïve Bayes and Ibk in itself is a good classifier but has a big standard deviation so it is better not to use the same.



According to this result, the best dataset is the **2nd one** and two best performing algorithms are **J48 Tree and Naïve Bayes**

Figure Figure 20 Comparing Performance accuracy for 15 datasets on J48, Naïve Bayes, Ibk, Random Tree

**Classification: True vs. False and Positive vs. Negative**

A true positive is an outcome where the model correctly predicts the positive class. Similarly, a true negative is an outcome where the model correctly predicts the negative class. A false positive is an outcome where the model incorrectly predicts the positive class. And a false negative is an outcome where the model incorrectly predicts the negative class.

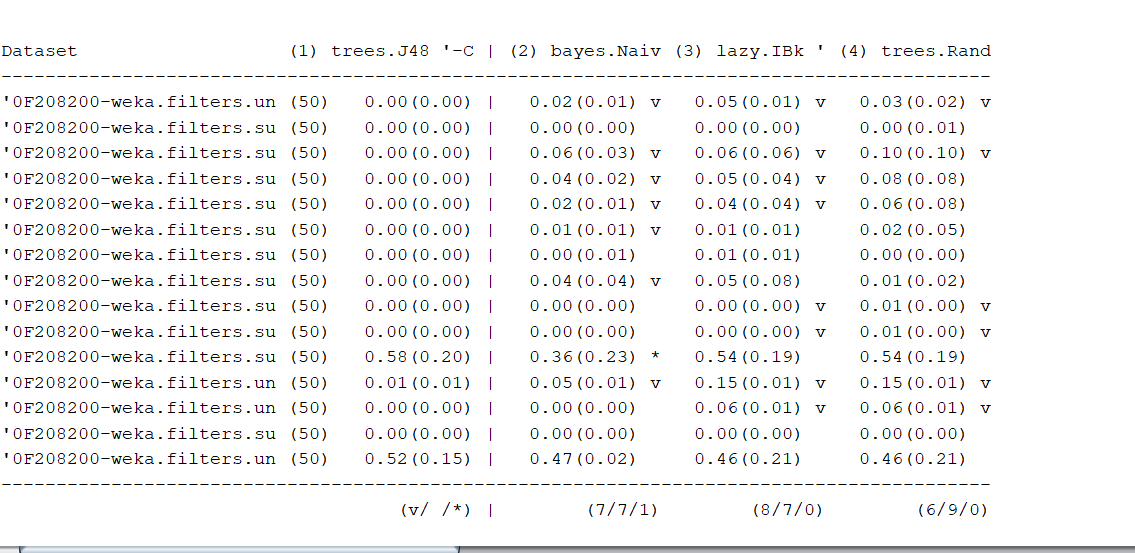


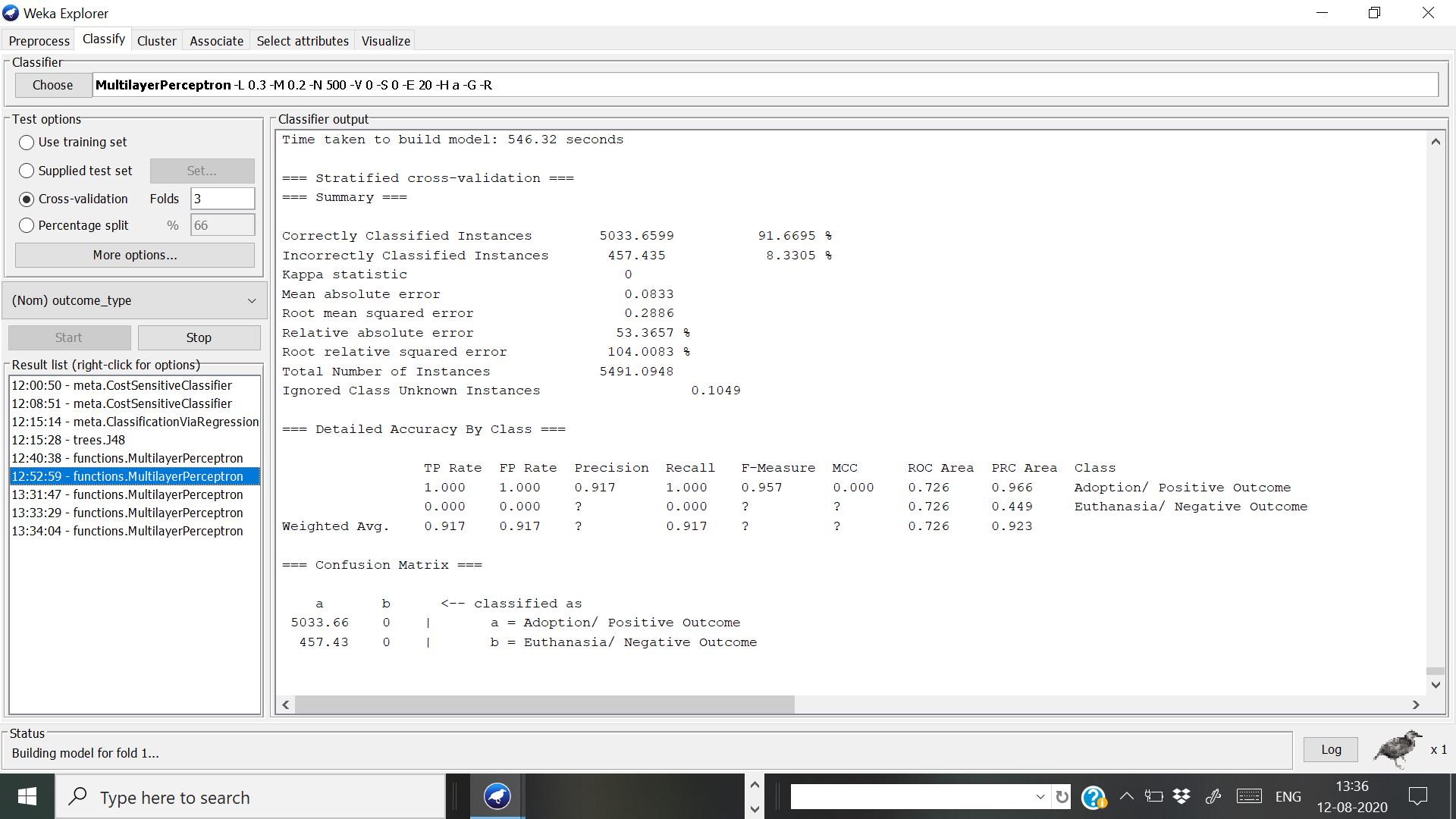
Figure Comparing False Negative Rate for 15 datasets on J48, Naïve Bayes, Ibk, Random Forest

In this case, as Pit Bulls are most vulnerable to negative rumours and thus, easily euthanised. It is true that they indeed have a rough behaviour when they aren’t trained enough and when they aren’t spayed. False Positive would mean that they were supposed to be euthanised but instead got adopted whereas False Negative would mean that they were supposed to be adopted but got euthanised instead which is absolutely unethical and inhumane. So, we would definitely want to reduce the false negative rate as close to 0 as possible. Here again the second dataset performed the best with only the last one having a standard deviation of 0.01 (which is negligible). Overall, J48 did great again, so we would prefer using J48 tree for our predictions.

**12. Testing on Neural Network (Multi-Layer Perceptron**):

To make our dataset more effective, after going through neural networks and its several properties, we concluded that, to make our visualization and our dataset more productive, we are using neural network for the neural network. For our dataset we used the Multi-layer perceptron which is a class of feedforward artificial neural network which sometimes refer to networks composed of multiple layers of perceptron (with threshold activation).

According to the figure below, the accuracy is 91.6695%. This means that it performs better than the baseline whereas when we compare it to J48, the performance decreases by a trivial number.



Didn’t do much better as compared to J48 Decision Tree

Figure Testing Neural Networks: Multilayer Perceptron, Cross validation 5 folds

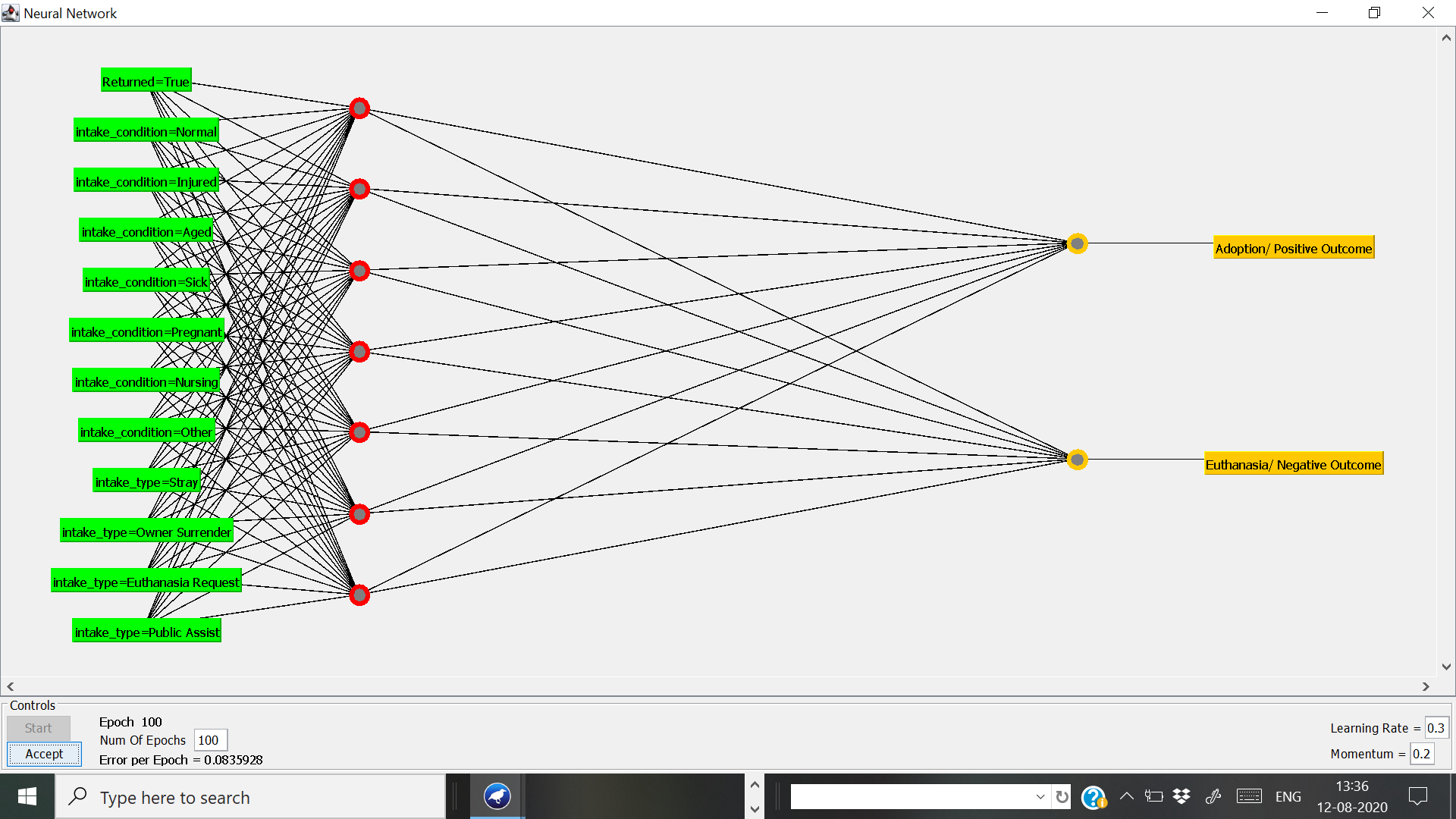


Figure Neural Network Visualization ( GUI: True); 100 Epochs used; 0.3 Learning Rate

Epoch is a term used in ML that indicates the number of passes of the entire training dataset the machine learning algorithm is completed. **Here 100 epochs** were used because as the number of epochs increases, a greater number of times the weight are changed in the neural network and the curve goes from underfitting to optimal or overfitting. Here our error per epoch is **very low (0.083).**

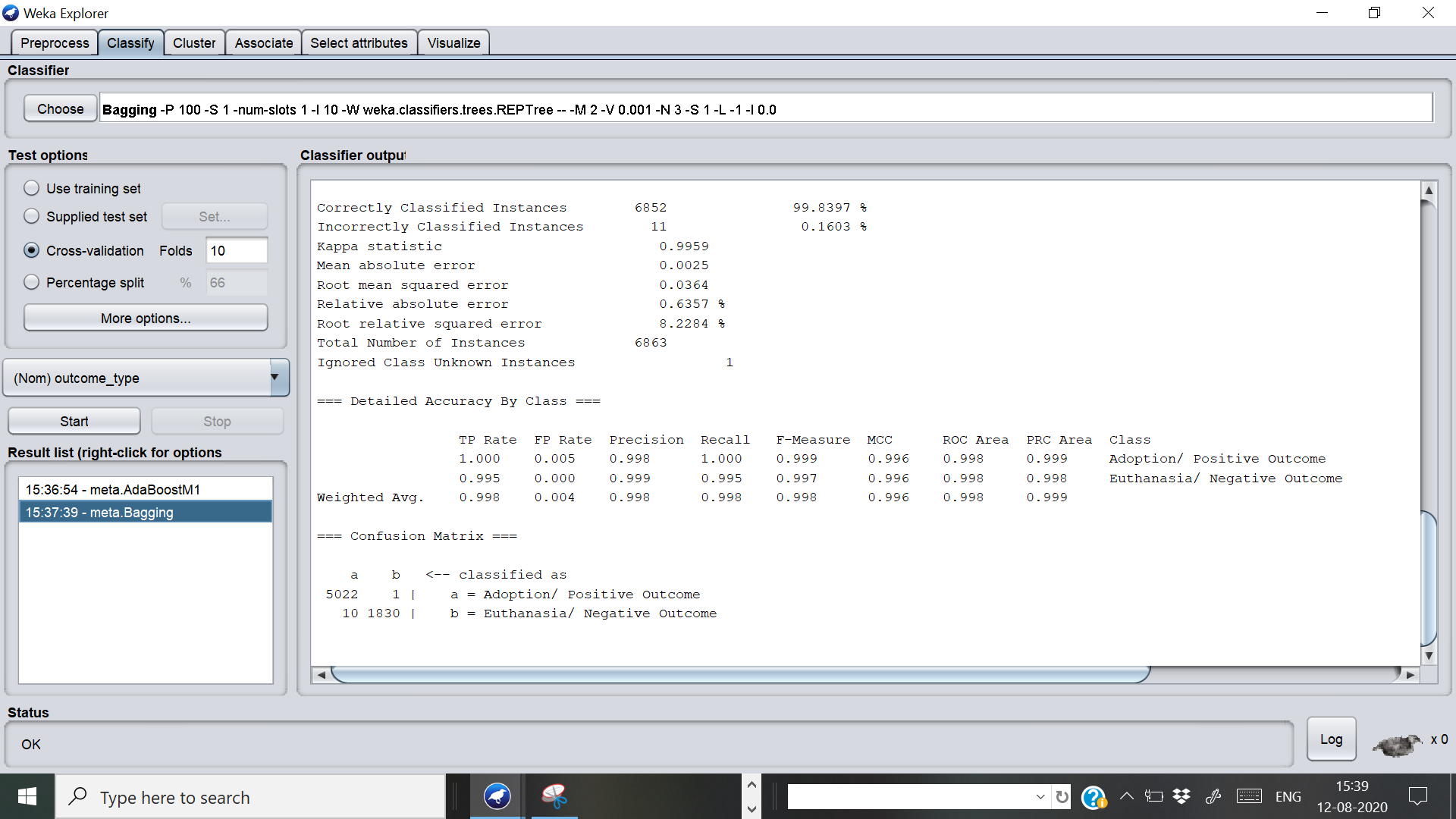
Learning rate: Amounts of weight that are updated during training are referred to as learning rate. They are a kind of hyper parameter which are configured and are used in the training of neural networks. It’s value ranges between 0-1. Here the model’s **learning rate is 0.3** which can be considered as really good.

**13.0 Ensemble Learning:**

Ensemble learning is a famous method in the field of data mining that can help algorithm that perform “weak” when working individually to do better when we have a group of models. This “weak” is to indicate unstable algorithm whom by making small changes in the training data can result in big changes in the entire model. Ensemble learning is used to improve the performance of a model or reduce the likelihood of an unfortunate selection of a poor one. Methods are Bagging, Randomization, Boosting and Stacking. For this project, we will be using **AdaBoosting and Bagging** methods to perform this with the help of Weka.

**13.1 Bagging Method**:

The bagging method samples the dataset to get different training set to build models that corresponded to that training set and then, combines the results of all the model by voting to find the best performed one. **Accuracy we got before bagging (using J48 tree) was *99.96%*** and **the bagging one is *99.8397%.*** Even though it didn’t prefer better but it is still a respectable score.



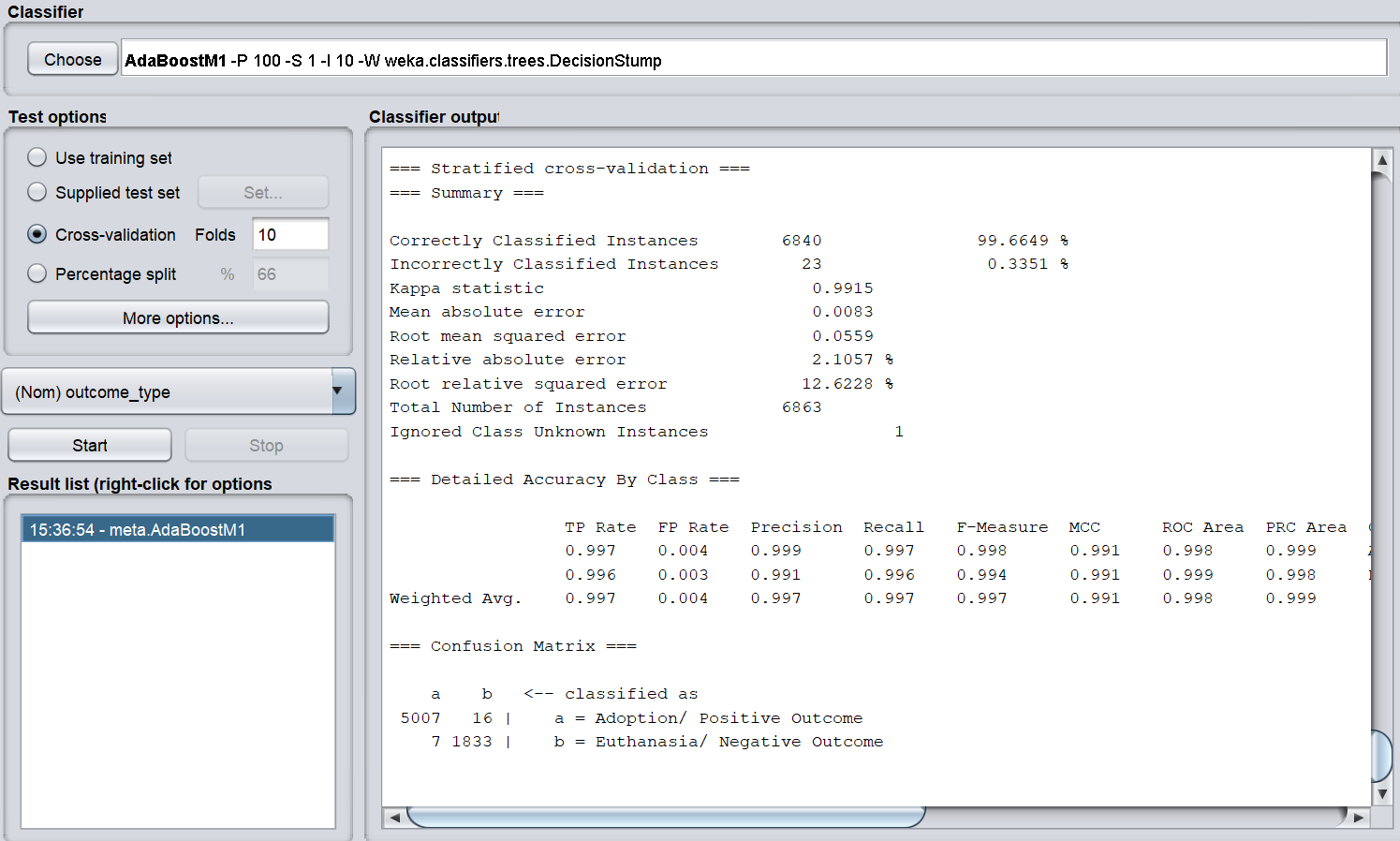
99.8397%

Figure Ensemble Learning: Bagging method; Performance accuracy: 99.8397

**13.2 AdaBoost**:

Boosting is another method of ensembling and it builds model one after another (a very tedious process especially if one is using decision trees). Every time a new model is made, there will be some mistakenly classified instances and it will take those instances into account and put more weight into it. After that, the models will also be averaged like in bagging, but boosting will also weight some model heavier than others based on their performance.

**AdaBoost in our case has an accuracy of 99.6649%** which is even worse than the previous result. But the difference in accuracy (0.2951%) is not that significant. Given that we want to focus on getting the highest accuracy and low false negative rate, we would prefer to keep the former model.



99.6649%

Figure Ensemble Learning: AdaBoost; Performance accuracy: 99.6649%

**14. CV Parameter Selection**:

CV parameter in weka, is one of the wrappers meta classifiers that selects the optimal value for the parameters. It has some analogy with the cross-validation and has one drawback that one cannot optimize on the nested options, only direct options of the base classifier. Before starting up, let us learn some terminologies. Confidence Factor: In the pruning process, confidence factor represents a threshold of allowed inherent error in data while pruning the decision tree. By lowering the threshold, one is applying more pruning and consequently generates more general models (Mustafa, 2012). It ranges from 0.1 to 1 in 10 steps. minNumObj ranges from 1 to 10 in 10 steps. The minimum instances per leaf justifies that at each split, at least 2 of the branches will have the minimum number of instances.

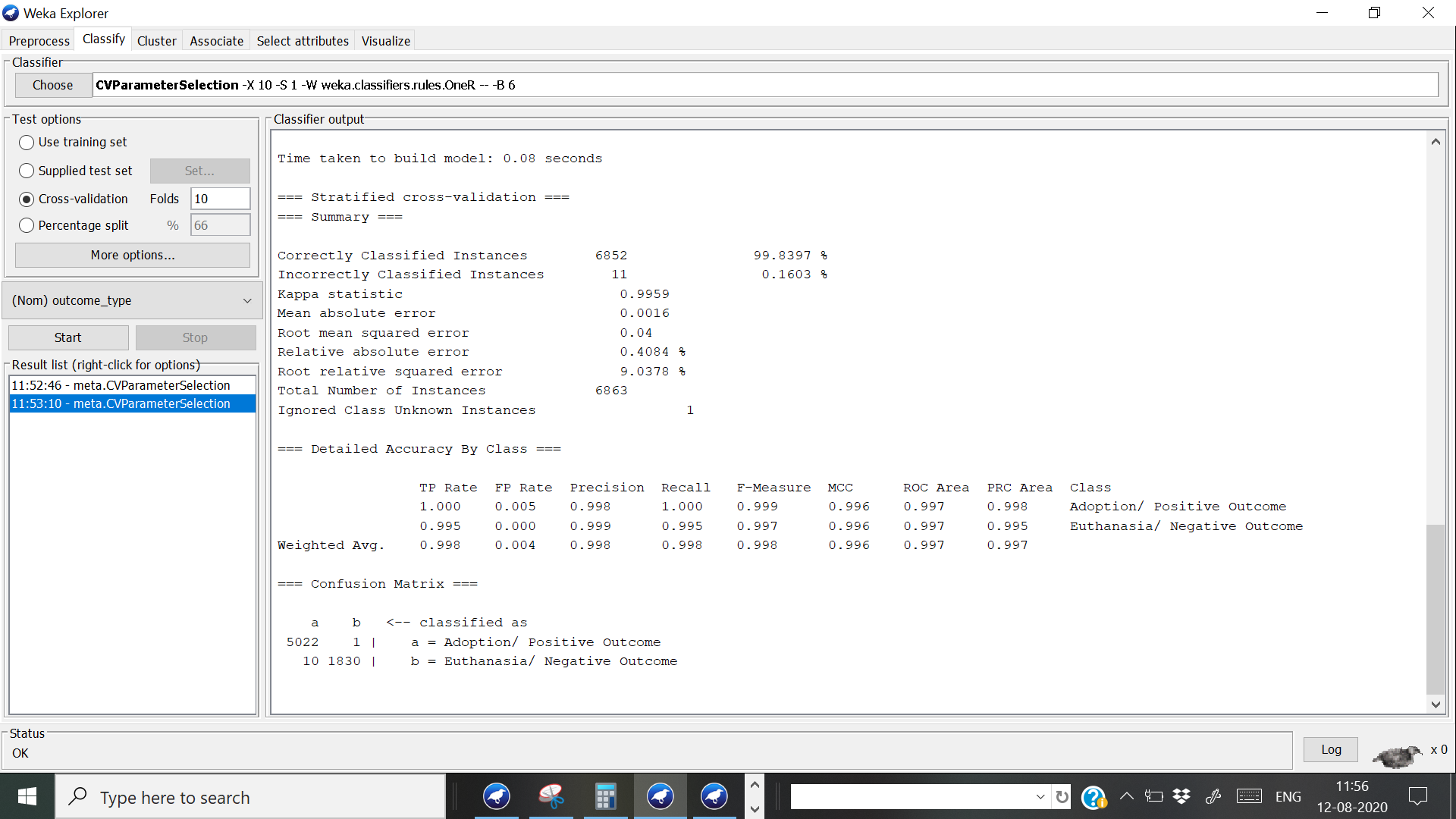


Figure CV Parameter Selection: 99.8397% accuracy

After we tried this parameter tuning meter, we did get an exceptional true positive rate for the adoption outcome (100%). Also, the false positive indicating healthy animal to be euthanised was reduced close to 0 which is great but the overall accuracy of the model became 99.8397%. This is actually trivial reduction so we can actually consider this model on the basis of high true positive rate and low false positive rate. This method proved to be better than ensemble learning, we applied earlier. But sometimes, parameter tuning results in hyper-optimization which might lead to overfitting of the data.

**15.0 Boundary Visualization:**

Boundary visualizer (a powerful tool offered by Weka) can help in letting the readers see the different classification boundary through the help of graphs that the machine algorithm makes. With this we will be able to distinguish all the class variable and some of the best algorithms to do so are Logistic Regression, Knn, SVM, decision trees like J48 etc.

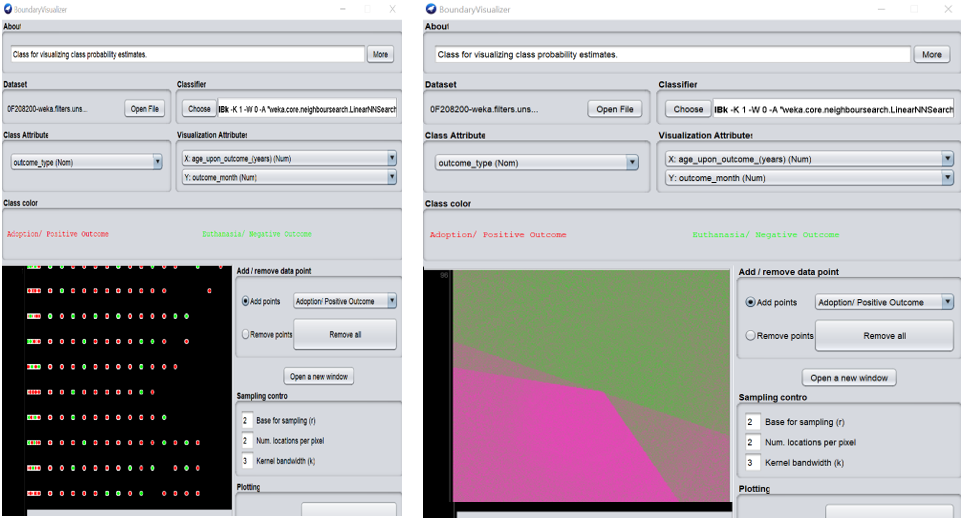


Figure Boundary Visualization: outcome month vs age upon outcome

The weka’s boundary visualization allow us to only work with 2-dimensional numeric data and thus, we worked with “age\_upon\_outcome” and “outcome\_month” (both are numeric attribute). The class is adoption or euthanised.The graph shows the two classes,Adopted in pink and Euthanised in green. The colour separation seems to be a boundary for the machine to distinguish between the colors. **Bias:** The error assumptions in our dataset is known as bias. The higher amount of bias results in the missing of important relations between the features and the underfitting output. Thus, bias is the accuracy of our predictions. **Lower Bias** means the predictive performance will fail to meet the assumptions of our biased dataset.

**16.0 Cost sensitive learning & Cost/Benefit Curve**:

Cost sensitive learning is a type of learning that takes the misclassification costs into consideration. The Cost/Benefit analysis component is useful for the analysis of predictive analytic outcomes. Here the aim is to minimise the cost.

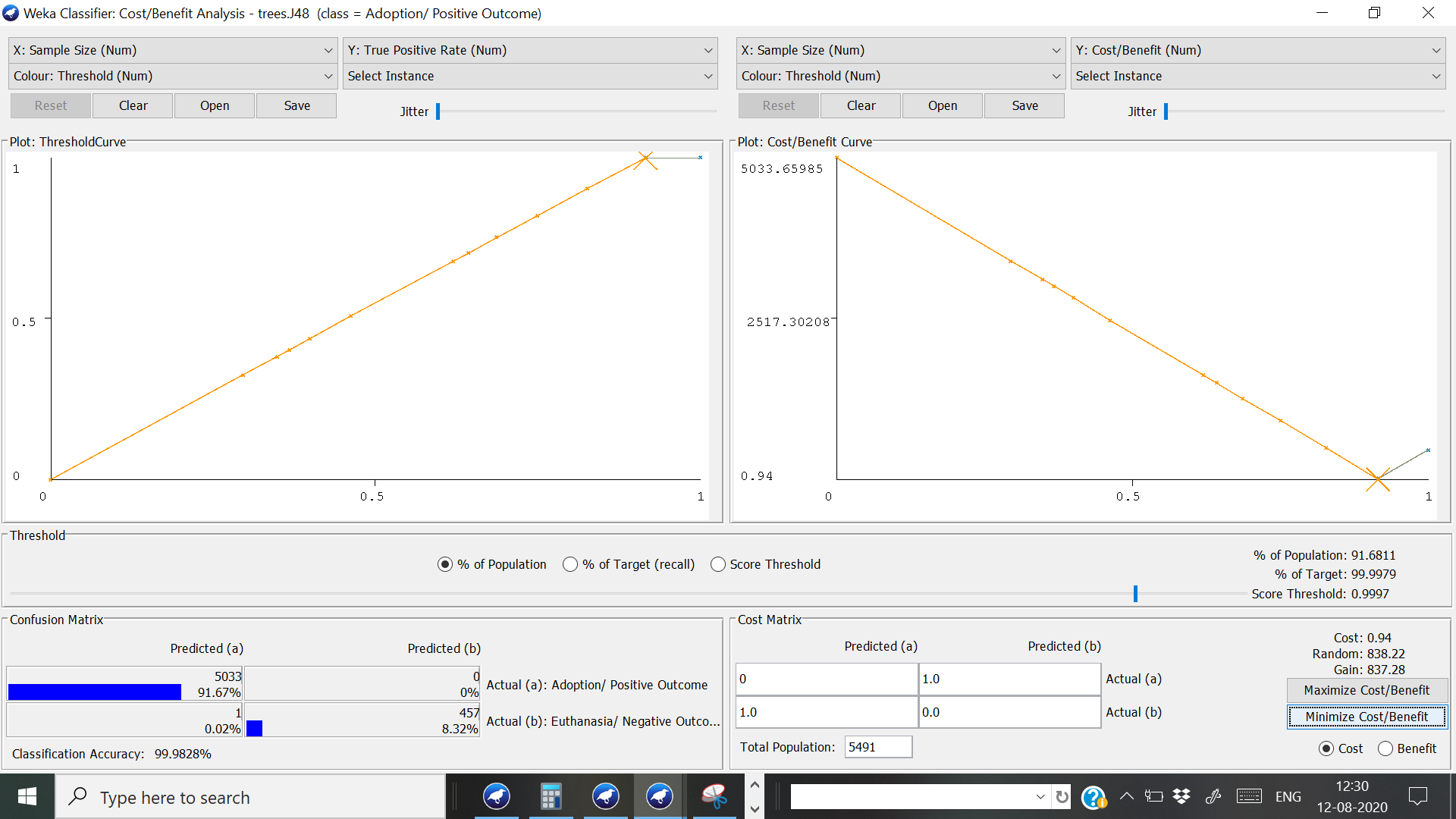
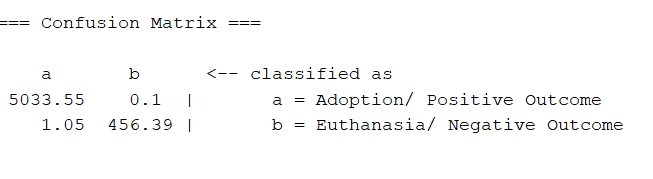


Figure Cost Sensitive Analysis for J48 Decision Tree

The algorithm tested for Cost/Benefit Analysis(Cost Sensitive Analysis) is J48 Decision Tree. Overall accuracy of the model is 99.979%. Here the total cost is 0.94. With cost matrix: Cost = 1\*0+ 1\*1 = 1. The cost is very low and it is good for our model.



**Confusion Matrix**:

In predictive analytics, it is a table with two rows and two columns that reports the number of false positives, false negatives, true positives and true negatives. It describes the performance of a classification model on a set of test data for which true values are known. Here the correctly classified **‘Euthanasia/ Negative Outcome’ (Specificity)** accuracy is **around 99.98%** whereas correctly classified ‘**Adoption/ Positive Outcome**’ (Sensitivity) is **99.9791%.** So, it classifies the retaining of customers better than cancellation itself by a small difference of 4.8%. Specificity is the ability of the test to correctly identify true negative rate (TN/TN+FP) while Sensitivity is the ability of a test to correctly identify true positive rate (TP/TP+FN).

**17. Knowledge Flow Interface**:

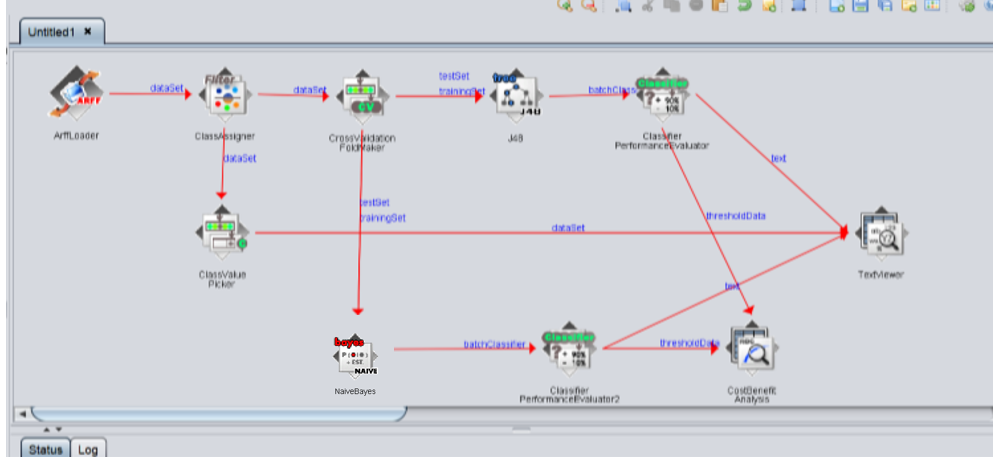




Figure Knowledge flow Interface: Best two algorithms chosen (J48 and Naive Bayes)

According to Professor Ian Witten (2013), Knowledge Flow interface is an alternative to the Explorer. It allows one to lay out different kinds of functions and connect them with different kinds of connectors. The Knowledge Flow environment model helps one to visualize the data and evaluate the process. One can also compare or find the connections between the different kinds of classifiers one plans to utilize.

Arff file is loaded into the workspace using ArffLoader. The dataset is connected with ClassAssigner (Assign a column to be the class for any data set, training set or test set). Then the dataset is connected with CrossValidationFoldMaker and ClassValuePicker. From ClassValuePicker the dataset is directly connected to TextViewer. Furthermore, training and testing set from CrossValidationFoldMaker (10-fold random seed) is connected with the Tree Classifiers called J48 and OneR. J48 is connected with ClassifierPerformanceEvaluator through batchClassifier (considering the batch training of all the training sample that pass through the learning algorithm in one Epoch before the weights are updated. As the datasets is too big, we took Epoch as 100). ClassifierPerformanceEvaluator evaluate the performance and give the output to TextViewer. In similar way, OneR is connected with ClassifierPerformanceEvaluator through batchClassifier and ClassifierPerformanceEvaluator evaluate the performance and give the output to TextViewer. Also, it connected to CostBenefitAnalysis through ThreshholdData.

**ROC curve** **(Receiver Operating Characteristics Curve**) and **AUC (Area under the curve)**:

The Area under the curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary for the ROC curve where ROC curves are used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests (Bhandari, 2020). ROC is a cost sensitive measure to evaluate classifier performance and thus, it will be used for the same.

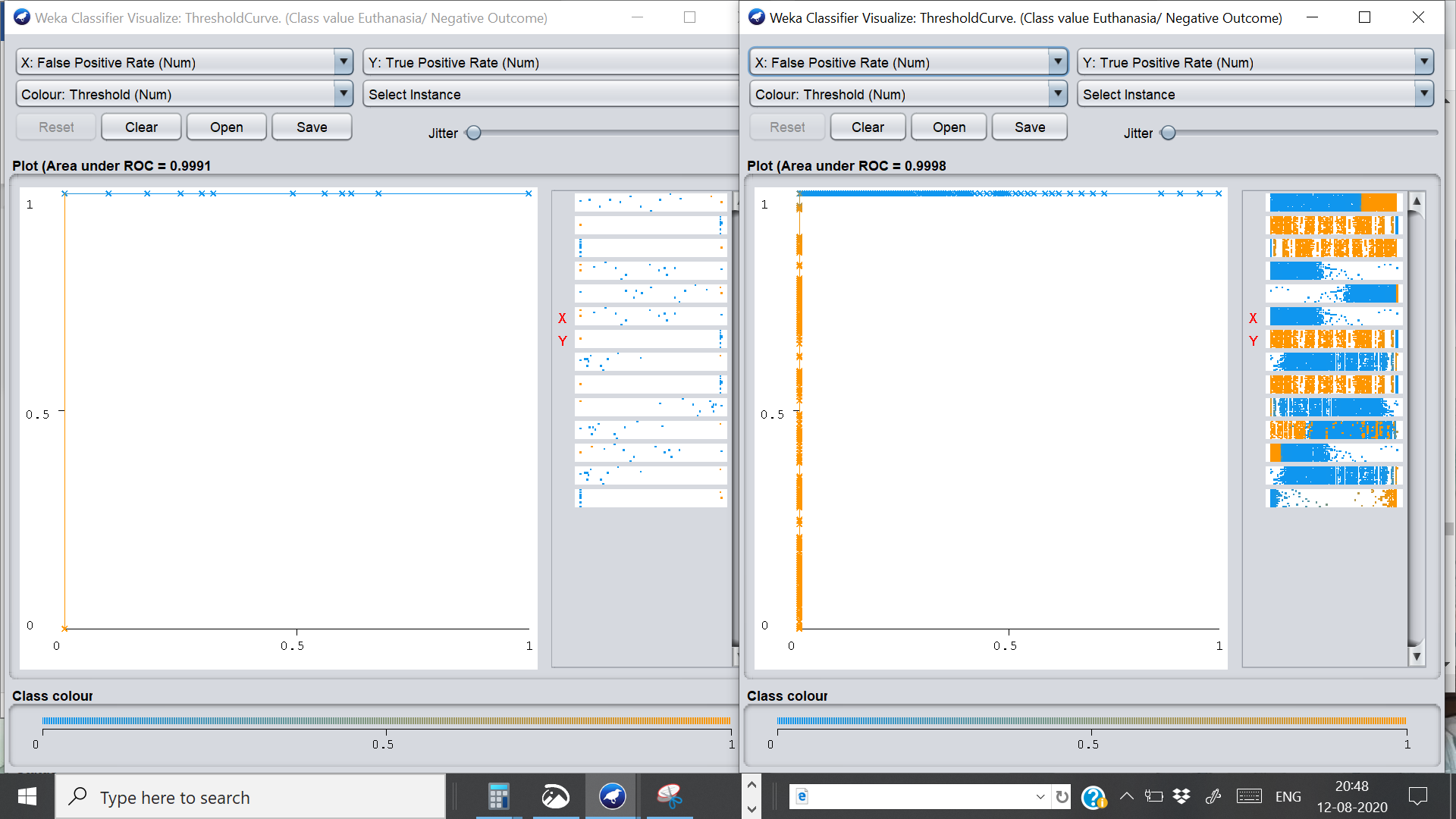


Figure ROC curve and AUC for J48 tree and Naive Bayes respectively

As both of the ROC curve is towards top left (close to 1), both algorithms are performing extremely well (minimising the cost). The AUC for J48 is 0.9991 whereas for Naïve Bayes is 0.9998. An AUC as high as this and the graph angle is closer to the top-left side which represent the optimal separability of the model.

*By analysing both, ROC curve and AUC, it can be concluded that both algorithms are doing equally good (almost) but we would still prefer to use J48.*

**18. Conclusion:**

Every no-kill shelter faces two common problems: First is accommodating new intakes and the second one is marketing of new intakes. If the target doesn’t know about it how will they manage to help the helpless animals? Also, according to our visualization, most of the shelter intakes are actually stray animals which may mean that they lost their owner, were born on the streets (full of diseases) or maybe went missing and made their way into the shelter.

*For better adoption of Pit Bulls or mix, it is better to get them clinically treated & neutered / spayed and for reducing the Euthanasia rate, it is better to get the Pit Bulls trained and tested for Rabies. This way adoption can be maximized.*

Here, the project’s focus was mainly on optimizing the adoption process so that most of the Pit Bulls or Pit Bull Mix intakes get adopted and the result was stupefying. Austin shelter has around 80% of their Pit Bulls adopted. Also, for further improving the adoption process, they can consider neutering the dog as according to the data, almost 90% of the neutered animal were adopted. Scientifically proven, neutered animals are less aggressive and can be a good family- friendly pet. Furthermore, to reduce the intakes which are supposed to be euthanised, maybe the shelters can consider appointing a dog trainer to fix any behavioural issues as most behavioural issues are due to PTSD which can be treated with proper time and dedication. Also, the Pit-Bulls should be tested for Rabies or any similar health issues so that the healthy (more or less) ones could be made ready for adoption.

In this project, 15 different variations of datasets were tested on a multitude of algorithms all capable of a certain task, to find the best model and algorithm. Algorithms used were J48 decision tree,Naïve Bayes, IBk (K Nearest Neighbour), ZeroR, Random Forest, Multilayer Perceptron etc. The complexity of our project highly affects the accuracy rate. After repeated testing, performance evaluation, comparing, ensemble learning, cost sensitive analysis and boundary visualization, it can be concluded that J48 Decision Tree performed the best in terms of accuracy, low false negative rate, negligible standard deviation, minimum cost etc. and thus, properly distinguishing between Boolean Variable.

(6955 words)

**References:**

1. Austin Animal shelter official information. Retrieved from,

<https://www.austintexas.gov/page/no-kill-plan>

1. Dataset retrieved from,

<https://www.kaggle.com/c/shelter-animal-outcomes>

1. Leo, B. (2018). Exploratory Analysis. Retrieved from,

<https://www.kaggle.com/lbarbosa/exploratory-analysis-austin-animal-shelter>

1. Andrews, A. (2020). How Data Science can be used to improve Animal Shelter Outcomes. Retrieved from,

<https://www.switchup.org/blog/how-data-science-can-be-used-to-improve-animal-shelter-outcomes>

1. Definitions from our group project. Hotel Cancellation prediction using Machine learning and Data Visualization. (DASH Soujanya’s part)
2. Cravit, R. (2019). What is a decision tree and how to make one. Venngage. Retrieved from,

<https://venngage.com/blog/what-is-a-decision-tree/>

1. Moustafa, A. (30 March, 2012). Comprehensive Decision Tree Models in Bioinformatics. PMC. NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3316502/#:~:text=Another%20setting%20influencing%20 the%20pruning,consequently%20generates%20more%20general%20models.
2. Brownlee, J. (2017). What is the Difference Between Test and Validation Dataset. Retrieved from,

<https://machinelearningmastery.com/difference-test-validation-datasets/>

1. The following has been retrieved from,

   <https://www.datasciencecentral.com/profiles/blogs/predicting-animal-adoption-with-random-forest-svm>

   **Caption for the picture:** 1. Importance vs. Feature for a cat’s adoption

   2. Importance vs. Feature for a dog’s adoption [↑](#footnote-ref-1)
2. This data is retrieved from Austin shelter’s official website whose reference is provided later. [↑](#footnote-ref-2)