Abstract

Cruise vacation is getting more and more popular nowadays due to the pandemic situation, predicting the previous cruise accident survival would help prevent death if there is a cruise accident. Therefore, the Titanic iceberg accident is taken into the investigation. In the project, the titanic survival dataset is used to make predictions with machine learning by using the programming language, Python, to get useful insights from various attributes in the data so that the relevant cruise companies and the government from different countries would do some measure and make some regulations to help reduce the death rate in case there is an accident. It is found that there are strong relationships between by finding out the correlations among different variables. Logistic Regression, Decision Tree, Random Forest and Linear Regression are implemented for making predictions of the titanic iceberg accident survival while Decision Tree has the best accuracy with the lowest root mean squared error(0.0937).

---

1. Introduction

Due to the recent pandemic situation, most of us are not able to travel abroad. Instead, having a cruise vacation will be a good alternative option. Therefore, we are interested in exploring the survival rate of cruise passengers if an accident occurs. And, Titanic would be a great sample for us to work on.

April 15, 1912 at 2:20 a.m.,

The 'unsinkable' Titanic sank in the North Atlantic Ocean with nearly 1,500 killed. Titanic symbolises state-of-the-art shipping technology 1910s. Before the accident, the titanic had already travelled 1,451 miles. It was 852 feet long, displaced 52,310 tons and was equipped with systems of seal bulkheads, believed to cause

Ships not sinking. During Titanic's maiden voyage, it hit an iceberg, creating a 300-foot-long rift on its side and flooded five bulkheads. Three hours after Titanic hit the iceberg, its bulkheads were gradually cultivated with water and then sunk.

In this fatal and grave accident, plentiful people think that the chances of survival depend greatly on the class couriers of the passengers. Moreover, the other information, ranging from

age to gender, from height to weight are also determining factors whether the passengers can survive in the accident.

In this project, there are three objectives. The first one would be the recognition of Titanic Survivors. The second objective would be finding out the indicative features of Titanic survivors. The third one would be figuring out the classification method of the lowest error and highest accuracy.

2. Data

2.1 Dataset

The dataset used in the project is extracted from kaggle. It consists of 1,309 passengers information of 7 variables, including embarked, passenger class, fare, parch, number of siblings/spouses abroad Titanic, age and sex. The information is used to indicate whether the passenger can or cannot survive from the iceberg accident.

2.2 Data Cleaning

By using the function isnull().sum(), we know that there are missing values in the column age, cabin and embarked. 891 values should be the complete values in each column but there are 177 missing values in "Age" column, 687 missing values in "Cabin"column and 2 missing values in "Embarked"column.

For the "Age"column, we cannot just delete this column directly from our dataset because only about 20% values are missing. We decided to find the mean value in the "Age" column and replace the 177 missing values with this mean value.

For the "Embarked"column, we cannot delete this column directly from our dataset because only about 0.2% values are missing. Also, we cannot calculate the mean value in this column because it is a categorical column and the value is in the form of text. In this case, we decided to find the mode in the "Embarked"column and replace the 2 missing values with the mode.

For the categorical column like "Sex"column and "Embarked"column, we will also convert them into numerical columns in order to train our models.

Finally, we will separate the features and target columns. For the features part, we will delete the "passenger id"column, "name"column, "ticket"column and "survived" column. For the target part, we only select the " survived" column.

2.3 Data Reduction

Most of the values in the "Cabin"column are missing, only about 200 values are present. Since most of the values in this column are missing, it's meaningless for us to do the data cleaning process like find the mean or mode of this column. Therefore, we decided to use a data reduction method that deletes this column from our dataset.

3. Methodology

3.1 Logistic Regression

Logistic Regression is a machine learning model used for predictive analytics, modelling and extends to application. It solves classification problems. In the analytics method, the dependent variable is finite or categorical. There are three types of logistic regression. Firstly, binomial logistic regression is used to solve the situations where the target variables have only two possible values, 0 or 1. The result can be Yes/No, True/False or other categorical value. Secondly, ordinal logistic regression classifies the problem into 3 or more values. The figures aren't quantitatively significant. Like 'cat', 'dog', 'cow'. Lastly, multinomial logistic regression is similar to ordinal logistic regression. The difference is the values are well ordered. Such as the result of MBI value as 'High', 'Average', 'Low'.

By using logistic regression to estimate the probabilistic values which lie between 0 and 1, we recognize the relationship between one or more independent variables and the dependent variable. The logistic function, which is also known as sigmoid function, is used to squeeze the output of a linear equation between 0 and 1 and a S-form curve formed.
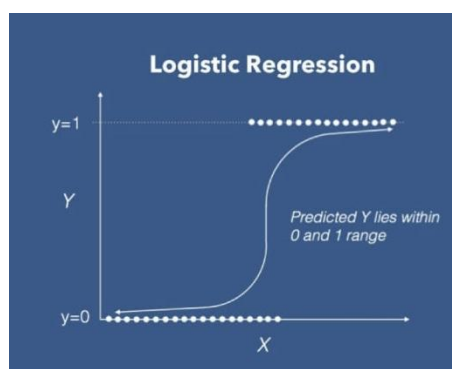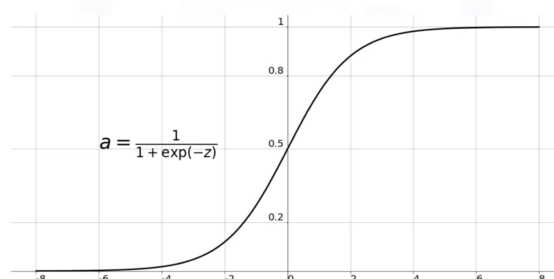

In the case of categorical values, logistic regression used the sigmoid function to create predictions. It usually sets 0.5 as the cut-off point value and gives the dataset's projected output in the form of the category to which it belongs. For example, if the output is below the cut-off point, the prediction output will be given as No, otherwise Yes if the output exceeds 0.5.

In our project, binomial logistic regression is used to indicate whether the passenger can or cannot survive from the iceberg accident, the result is represented as 0 (cannot survive) or 1 (survive).

3.2 Decision Tree

A Decision Tree is a machine learning model in which is to create a training model that can be used to predict the class of a given dataset by applying distinct decision rules. There are two types of decision trees, namely categorical variable decision trees and continuous variable decision trees. They solve classification and regression problems respectively.
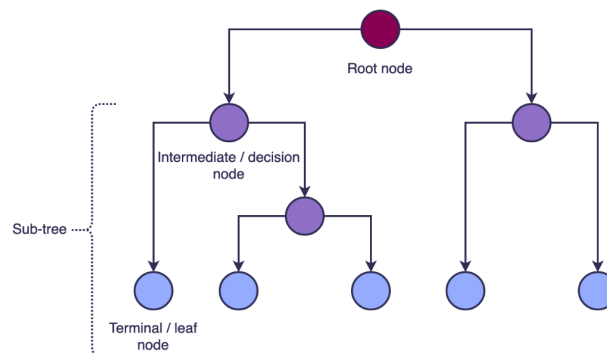
Starting from top to bottom:

Step1: Begin the tree with the root node, we select the decision rule which has the largest information gain to split the node

Step2: Split child node with considering decision rule never selected before in the current branch

Step3: Stop splitting if there are no decision rules left / group impurity equals zero

Note: To prevent overfitting / numerous splitting, pruning and early stop may be applied

Gini impurity

$$I_G(p) = \sum_{i=1}^{J} \left(p_i \sum_{k \neq i} p_k \right) = \sum_{i=1}^{J} p_i(1 - p_i) = \sum_{i=1}^{J} (p_i - p_i^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=1}^{J} p_i^2$$

If the impurity tends to zero, the more ideal the situation.

Information gain (IG)

$$\overbrace{IG(T,a)}^{\text{information gain}} = \overbrace{H(T)}^{\text{entropy (parent)}} - \overbrace{H(T \mid a)}^{\text{sum of entropies (children)}} = -\sum_{i=1}^{J} p_i \log_2 p_i - \sum_{i=1}^{J} -\Pr(i \mid a) \log_2 \Pr(i \mid a)$$

In our project, we select categorical variable decision trees to solve the classification problems. We apply the decision rules (e.g. Is he/she aged above 30? Is he/she male?) based on distinct categories (e.g. ages, sex, embarked, etc) that could optimize our training model. In each attribute, we calculate entropy using the equation:

$$\text{Entropy}(s) = -P_{yes} \log_2 (P_{yes}) - P_{no} \log_2 (P_{no})$$

where P_{yes} = probability of yes, P_{no} = probability of no
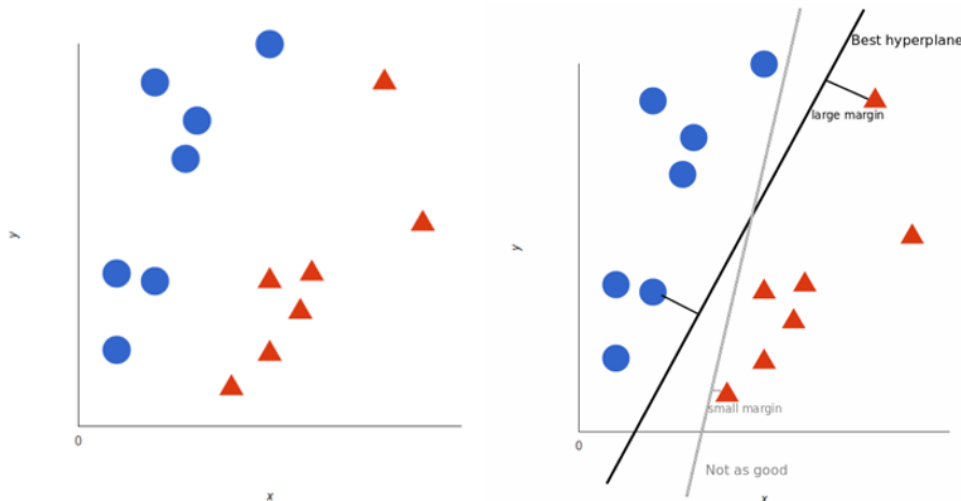
3.3 Random Forest
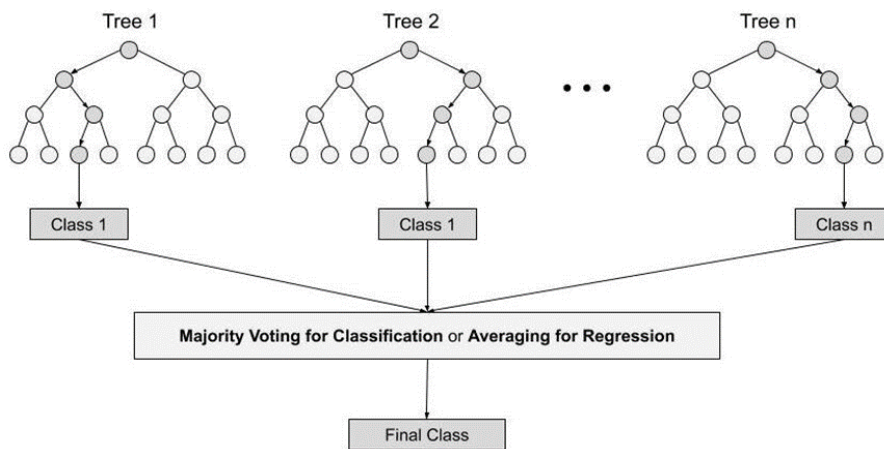
3.4 Support Vector Machine (SVM)

The basic principle of SVM is as follows.

For simplicity, the setup for the example data will include only two labels, blue and red, determined by two factors, x and y, respectively.

The required classifier would be a machine that takes these two factors in and outputs a classification label.

SVM would take the data and create the hyperplane that best separates the labels. Imagine that the hyperplane is the decision boundary. Data that fall on one side will be defined as blue, and data that fall on the other side will be defined as red.

The best hyperplane maximises the margin of the two labels.

In our project dataset, diabetics and nondiabetics are not classified using one or two factors, but two labels (diabetic and nondiabetic) are defined by a broad range of factors - such as passenger class, age, gender, fare, embarked, parch as well as number of spouse/siblings. To make it simple to understand, the data studied in the project are non-linear data.

### 3.5 Neural Network Classifier

Neural network is a type of deep learning model which results in a sequence of layers. By transmitting signals in between different neuron layers based on the connected network, it simulates the decision making process in the human brain's synapses' interaction.

In our project, we apply sigmoid as our activation function since its function outputs are compressed within 1 and 0 (0 < output < 1) and the graph presents as a S-shaped curve. We may define 0 as the not survived and 1 as the survived.

Sigmoid activation function:
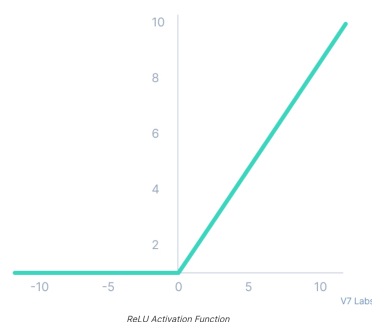
$$f(x) = \frac{1}{1 + e^{-x}}$$

Sigmoid activation function:

$$f(x) = max\,(0, x)$$

Sigmoid/Logistic Activation Function
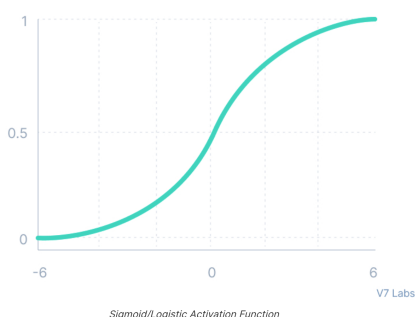
ReLU Activation Function

Also, we apply two layers, "sigmoid" with nodes and "rule" with nodes consecutively. The neurons within the network interact with the neurons in the next layer and eventually classify them as our desired outcomes(0 or 1).

For the optimization, we selected Adam (Adaptive Moment Estimation) which works with momentums of first and second order.

<u>First order momentums</u>

<u>Second order momentums</u>

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}.$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}.$$

The advantages are it converges rapidly and decelerates the learning rate.

For the loss function, we select binary-entropy loss, appropriate in binary classification settings to get one of two potential outcomes.
Loss function :

$$L = - (y_i \log (y^i) + (1- y_i) \log (1 - y^i))$$

3.6 Naive Bayes Classifier
Naive Bayes is a classifier that uses Bayes' theorem. It predicts probabilities for each class, such as the probability that a record or data belongs to a specific class. The class with the highest probability is defined as the class with the highest likelihood. The membership probability is well-known as maximum a posteriori probability (MAP).

The assumed MAP is:
map (H)
= max(P(H|E))
= max((P(E|H)*P(H))/P(E))
= max(P(E|H)*P(H))

P(E) is the probability of evidence used to normalise the results. It remains the same, removing it has no effect.

Naive Bayes classifiers assume that all features are independent of each other. The presence of one factor would not affect the presence or absence of any other feature.

In our datasets, we test hypotheses given multiple evidences (features). Therefore, the calculation would be complicated. To make the work simply, feature independence methods are used to "decouple" multiple evidences and treat each evidence as independent.

P(H|Multiple Evidence) = P(E1|H)* P(E2|H) …*P(En|H) * P(H) / P(Multiple Evidence)

In our project, a person might be considered with high probability to be a survivor if it is a first class, female, young passenger. Even those passenger features upon the existence of the other features, the classifier would consider all of the features independently to contribute to the probability that the person is a survivor.

3.7 AdaBoost

AdaBoost(Adaptive Boosting) seeks to develop a powerful classifier by merging many weak classifiers. It implements boosting, which entails connecting a sequence of weak classifiers in such a way that each weak classifier attempts to improve the classification of samples misclassified by the preceding weak classifier. During boosting, the decision trees used are called 'stumps'. One stump alone can only use one variable to make a decision. However, a fully formed tree can predict the target value by combining the decision from all variables. Here are the steps that show how adaboost works:

Step1: Weak classifiers (like decision stumps) are created based on weighted samples. All sample equal weights.

Step2: Create a decision stump for each variable and discover how each stump arranges the samples into their target classes.
Step3: More weight is assigned to correctly classifying the previously misclassified sample.

Step4: Repeat from step 2 until all data points are correctly classified or reach the maximum iteration level.

Weak learner #1　　Weak learner #2　　Strong learner
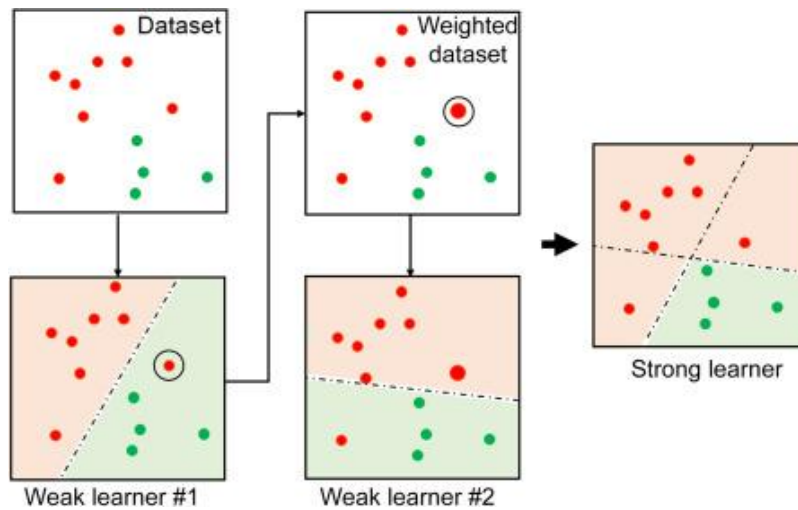
As more weak classifiers(weights) are added, The classification accuracy increases.

$$x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}$$

In a binary classification problem, y is the target variable which is either -1 or 1, indicating the first or the second class (e.g.survive or not survive in our project).

In our project, AdaBoost is used to determine the data correctly which classifies whether the passenger can survive or not in the titanic accident.

3.8 Multi-layer Perceptron

4. Results

4.1 Correlation between variables

From the scatter matrix plot above, whether a person is survived or not is correlated with different variables. We will find out which variables are more correlated with the "Survived" variable

The following are 3 of the most correlated variables among all the variables, since some of the attributes are removed in the data reduction part. Sex is the most related variable with the survived.

1. Sex (0.54)
2. Fare (0.26)
3. Embarked (0.11)

The following are 3 of the most negatively correlated variables among all the variables. Pclass is the most unrelated variable with the survived.

1. Pclass(-0.34)
2. Age(-0.07)
3. Number of siblings or spouses aboard the Titanic(-0.035)

4.2 Define accuracy Between Different Algorithm by Root Mean Square Error

We measured different model's Root Mean Square Error(RMSE) on the whole training set to account for the accuracy of the data set. The reason why Root Mean Square Error is chosen to measure the accuracy of the Machine learning algorithm, instead of the other method such as measuring the f1 score and the number of correctly predicted data points out of all the data points, is that RMSE of all the algorithm can be calculated, while Logistic Regression f1 score cannot be calculated out.

However, after discussion with our project advisor Dr Cheung King Chau, we found that RMSE may not be suitable due to the algorithm we use. Therefore, we use the other two methods to define the accuracy for the classification algorithm, which are entropy loss and ROC Curve respectively.

Algorithm	RMSE
LogisticRegression	0.4387
RandomForestRegressor	0.1597
DecisionTreeRegressor	0.0937
Support Vector Machine	n/a
Neural Network Classifier	n/a

The RMSE is largest for LogisticRegression and smallest for DecisionTreeRegressor. The larger RMSE means the less accuracy of the model. According to the RMSE, DecisionTree is the best model in RMSE case.

4.3 Define accuracy Between Different Algorithm by Precision Score, Recall Score, Accuracy Score, F1-Score

Confusion matrix is useful in machine learning because it can help to visualise the performance of different algorithms.

Actual Values

	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

(Predicted Values on the left axis)

a confusion matrix

TP : Observation is Positive and truly predicted as positive

FN : Observation is Positive and falsely predicted as negative

FP : Observation is Negative and falsely predicted as positive

TN : Observation is Negative and truly predicted as negative

a. Precision Score

Precision Score = TP/(FP+TP)

We usually use precision scores to measure the success prediction when the classes are very imbalanced. It is the measure of exactness or quality.

Classifier	Precision score
Logistic Regression	79.6%
Decision tree	74.8%
Random forest	81.3%
Support vector machine	64.9%
Naive Bayes (Bernoulli)	76.9%
Ada boost	80.6%
Multi-layer Perceptron	77.2%

b. Recall Score

Recall Score = TP/(FN+TP)

Recall score means the ability of classifiers to correctly predict the positives out of the actual positives.

Classifier	Recall score
Logistic Regression	79.5%
Decision tree	75%
Random forest	81.3%
Support vector machine	64.9%
Naive Bayes (Bernoulli)	76.9%
Ada boost	80.6%
Multi-layer Perceptron	77.2%

c. Accuracy score

Accuracy Score = (TP+TN)/(TP+FN+TN+FP)

Accuracy score is defined as the ratio of true positives and true negatives to all positive and negative observations. That means it can tell us how often the classifier can correctly predict an outcome out of the total number of times it made predictions.

Classifier	Accuracy score
Logistic Regression	79.5%
Decision tree	75%
Random forest	81.3%
Support vector machine	64.9%
Naive Bayes (Bernoulli)	76.9%
Ada boost	80.6%
Multi-layer Perceptron	77.2%

d. F1-Score

F1-Score = (2*Precision Score*Recall Score)/(Precision Score + Recall Score)

F1-score is used to choose either of precision or recall score can result in compromise if the classifiers give high false positives and false negatives respectively.

Classifier	Accuracy score
Logistic Regression	80%
Decision tree	75.9%
Random forest	81.7%
Support vector machine	70.1%
Naive Bayes (Bernoulli)	77.4%
Ada boost	80.8%
Multi-layer Perceptron	77.7%

4.4 Define accuracy Between Different Algorithm by ROC Curve and AUC Score

ROC for different classifier
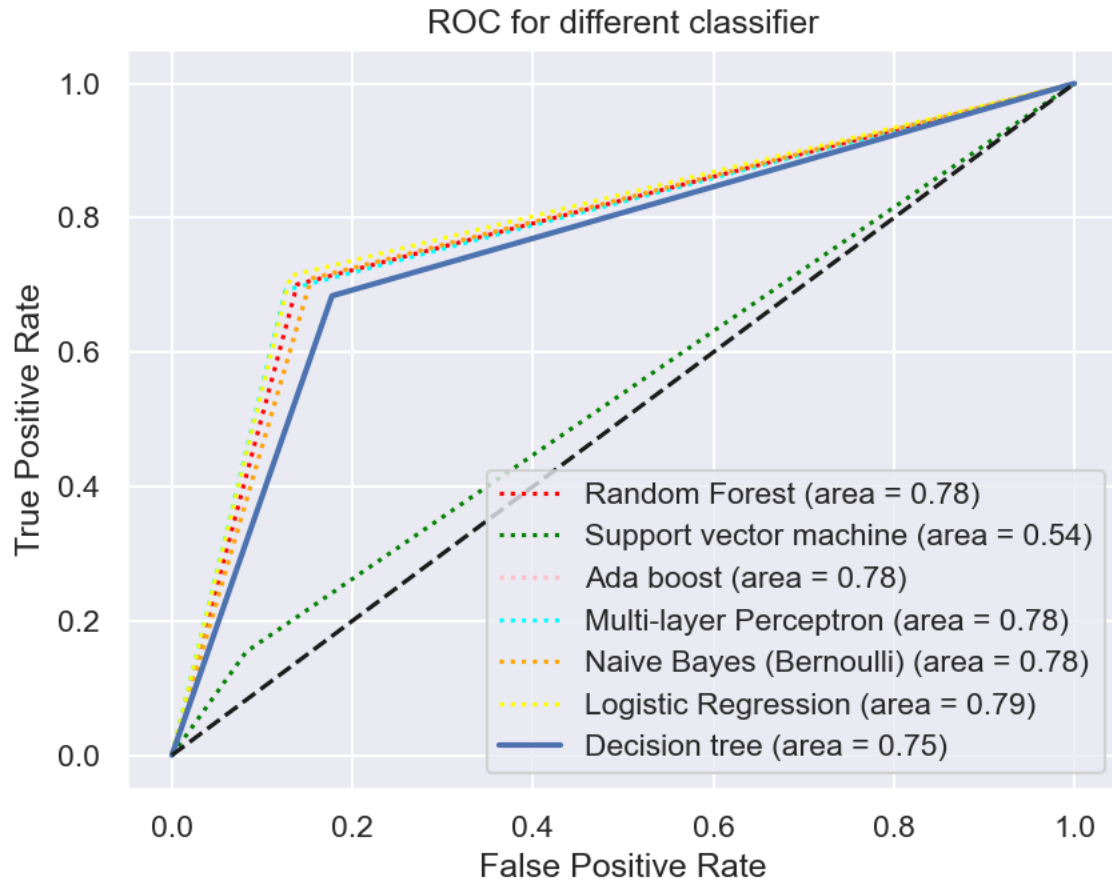
After the ROC curve is produced, we use the area under the ROC curve (AUC) and the shape of the curve to examine the predictive power of different algorithms. For SVM, the area is almost equal to 0.5 and the shape is nearly diagonal of the graph, which shows that SVM is just randomly guessing the result of the prediction. It means that SVM has nearly no predictive power in predicting the survivors in the project. For the other algorithms (Logistic Regression, Decision Tree, Random Forest, Naive Bayes Classifier, Adaboost, and Multilayer Perceptron), the areas are nearly 0.7-0.8, which shows that they are acceptable classifiers for predicting the survival of passengers in Titanic. All in all, logistic regression is the algorithm with the highest accuracy to make predictions in our project.

5. Features Analysis of Titanic Survival

5.1 Survival and Sex

From the scatter matrix plot, the correlation between survived and sex is 0.54 which is the largest compared to other variables.
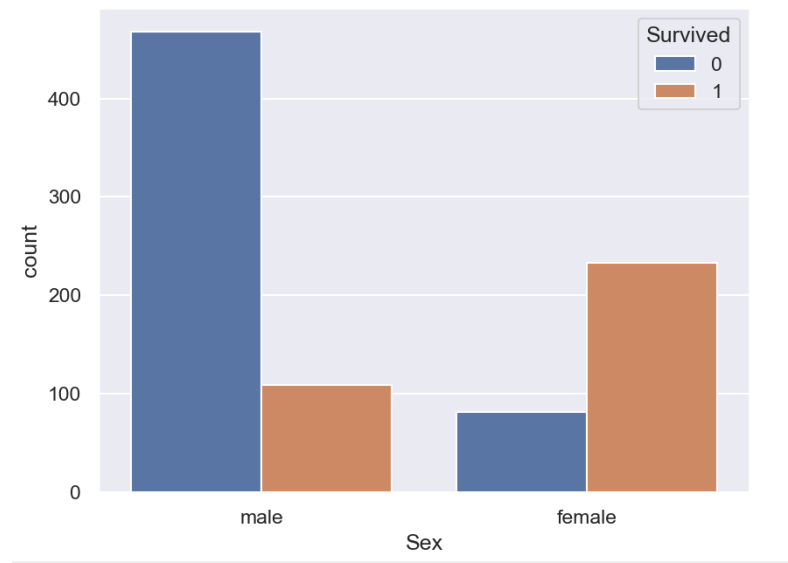
0 means dead and 1 means survived in the above bar chart. We can see that the survival rate of female is much higher than that of male. Because men let women and children leave first due to the lack of lifeboats in the past. The lifeboats are not able to accomodate everyone on the ship.As a result, women and children have the priority to take the lifeboat.While men help them escape from adversity and choose to sacrifice themselves. Also, women can survive longer on the boat than men in the sea. That's why sex is the main variable which correlated to the survived.

5.2 Survival and Fare

From the scatter matrix plot, the correlation between survived and embarked is 0.26 which is the medium compared to other variables.
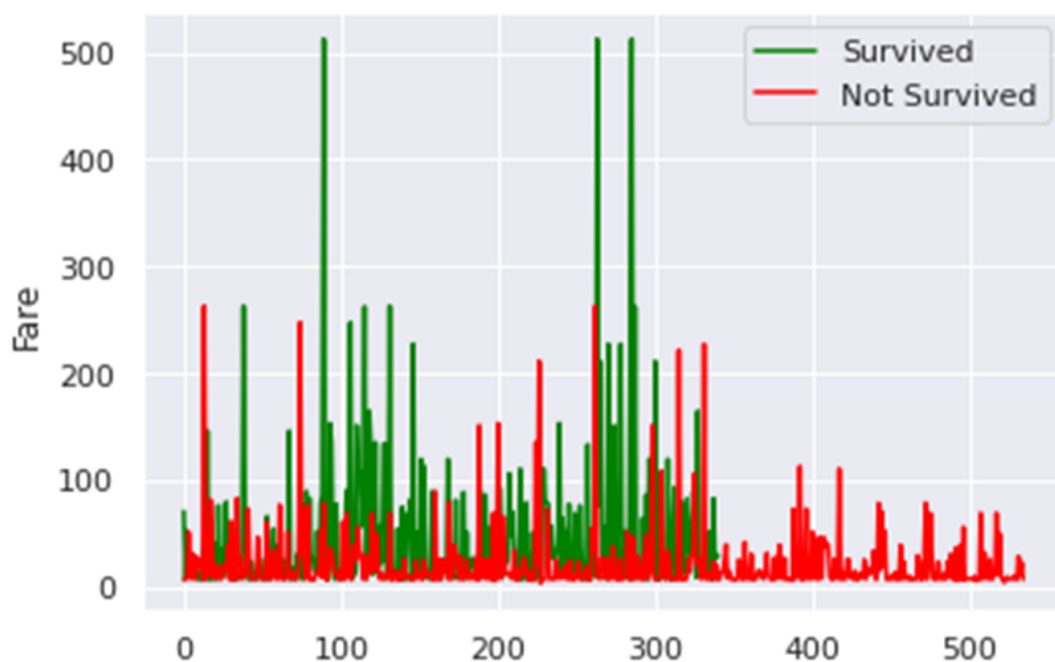
From the graph, we could see that people who had paid higher fare prices have a higher survival rate. While, the death rate of people who had paid low-priced fare(third ticket class) is relatively high too. One of the possible explanations is that passengers who spent a higher price had a better position for them to take the lifeboat. And the condition of the sinking ship was adverse to the third ticket class passengers. They could not escape from the accident at once. Moreover, wealthier people who buy expensive tickets have the first priority to escape from titanic. While children and women are the second priority to take the lifeboat. Because most of the wealthier people are nobles and hierarchy exists. They have high social status which allows them to have the right to leave first.

5.3 Survival and Embarked

From the scatter matrix plot, the correlation between survived and embarked is 0.11 which is the lowest compared to other variables.
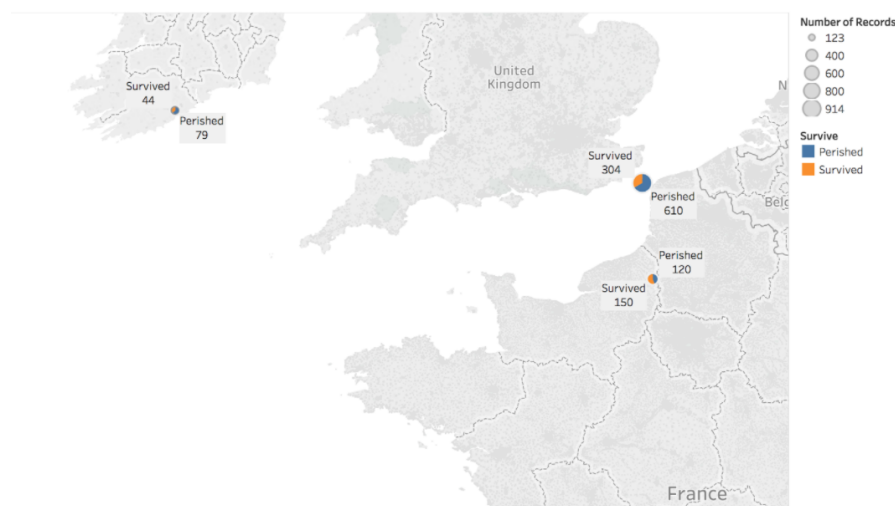

From the graph above, we could observe that the number of embarkation people is directly proportional to the survival people, with survival people embarked in Southampton > Cherbourg > Queenstown. Both Southampton and Queenstown have a smaller number of the survived than the not survived while Cherbourg has the reversed case. We are not able to

distinguish the potential caused by the relationship between embarkation and the survivors from the above.

5.4 Survival and Parch

Parch represents the number of parents or children aboard the Titanic and, parent includes mother and father; children include daughter, son, stepdaughter and stepson, other family relations are not counted in parch. From the scatter matrix plot, the correlation between survival and parch is 0.082, which is low compared with other variables.
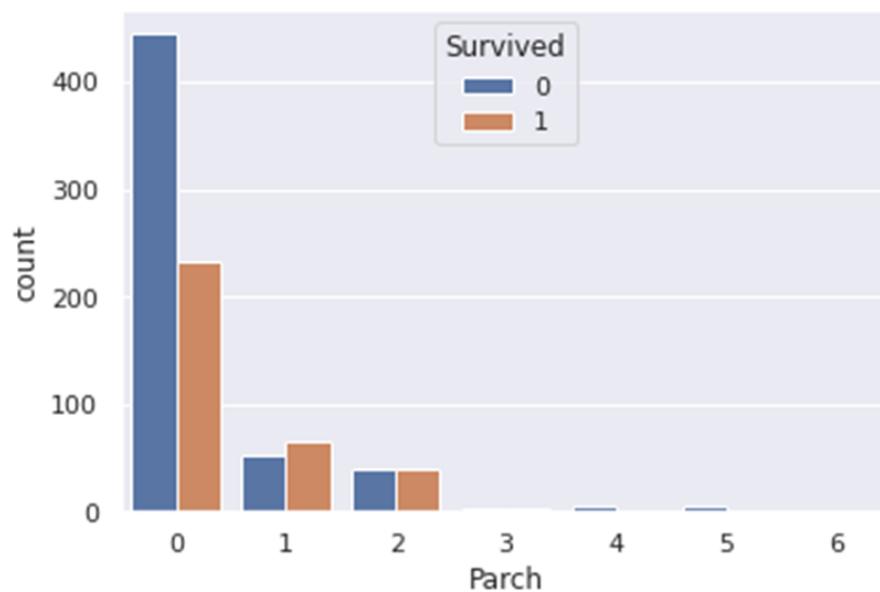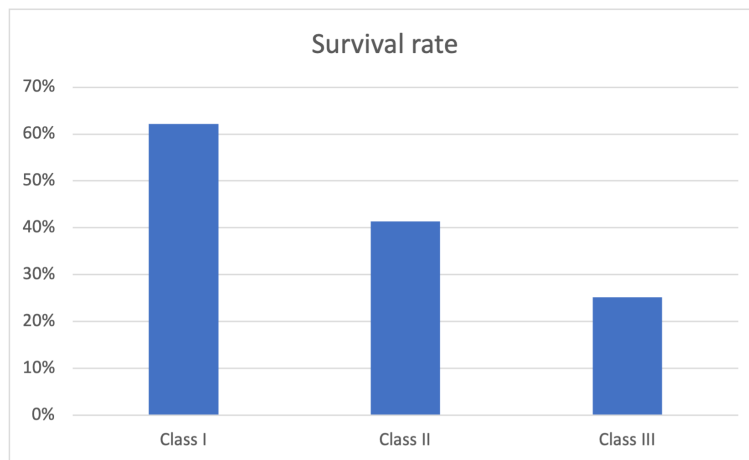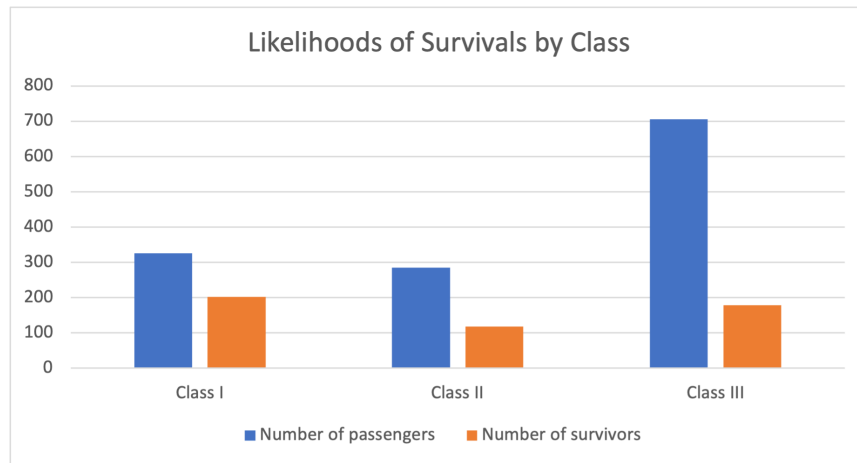

From the above histogram, for passengers with parch equal to 0, one third of them survived and, for passengers with parch equal to 1 and 2, half of them survived. Also, most of the passengers are with parch equal to 0. We are not able to distinguish the potential caused by the relationship between parch and survival.

5.5 Survival and Negatively Correlated Variables

Despite the negative correlation of passenger class, age and number of spouse/siblings, these do not imply that all these three variables are not related to survival rate.

Negative correlation indicates that two variables are having a statistical relationship that their values are moving in two opposite directions from the other one. In short, if variable I and variable II are negatively correlated, variable I will decrease with increasing variable II.

Relatively more negative correlation(-0.34) between passenger class and survival rate shows that class I passengers had the highest survival rate(62%) during the accident while class II would have smaller(41%), and class III would have the smallest survival rate(25%). It seemed that the more wealthier, the higher survival rate the passengers would have.

Likelihoods of Survivals by Class

Class	Number of passengers	Number of survivors
Class I	~325	~200
Class II	~285	~120
Class III	~705	~180

Survival rate

Class	Survival rate
Class I	~62%
Class II	~41%
Class III	~25%

The reason for this can be explained by several facts. First of all, the first class cabins were near the lifeboats. This might be the exclusive services provided for the first class passengers. The second reason would be the fact that there were mainly emigrants in the third class who did not know English. This meant that they did not even know what was happening at that time and the escape time for them would be less than the others. Also, they were not able to follow the instructions clearly from the English-speaker titanic crews.

Age is negatively correlated with the survival rate as well. This implies that the younger the passengers were, the higher the survival rate would be. The reason is similar to the variable "age". The children's survival rate(53%) was higher than men(20%) in the accident because the men let the children and women leave the boat first. This led to not only women having a higher survival rate, but also the children as well.
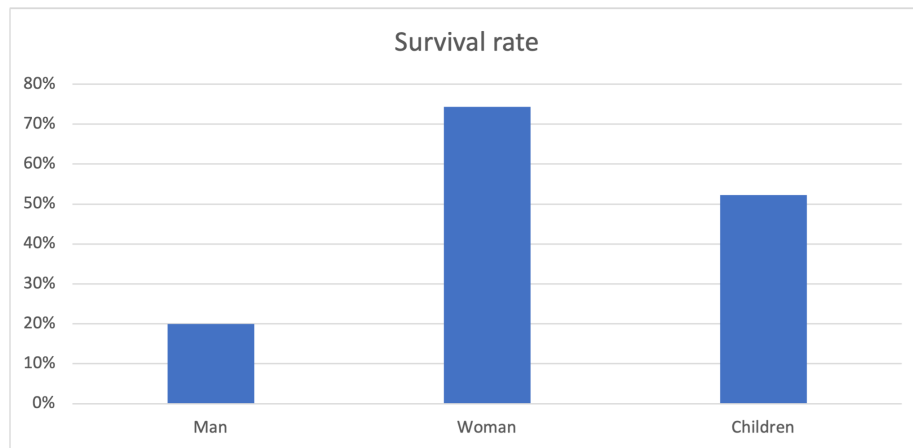
The correlation between the number of siblings/spouse and the survival rate is nearly 0 shows that they are not literally related.

6. Limitations

6.1 Quality of data

The better the quality of data, the more accurate of the prediction from the machine learning algorithms.The quality of a dataset can be determined by the Six data quality dimensions: Accuracy, Completeness, Consistency, Timeliness, Validity, and Uniqueness (Sarfin, 2021). The quality problem of our dataset regarding the six data quality dimensions are as follows.

●(Accuracy)

Accuracy means the similarity between the data and the reality. It is reliable for our data compared to reality.

●(Completeness)

Completeness means the completeness of our dataset. In our dataset, we have 3 columns in which values are missing: 177 values missing in the "Age"column, 687 missing values in the"Cabin"column and 2 missing values in the"Embarked"column. We fill in the missing value by the mean, the mode and drop the whole column to solve the problem of completeness.

●(Consistency)

We do not want any information conflicts in our data but it is not easy to check it out.

●(Timeliness)

Timeliness describes how new the data is and how it fits our prediction. The data was collected quite a long time ago. It may not be suitable in predicting today's survival rate.

●(Validity)

Validity means that the format of each column of the data is the same and correct. They should follow a specific format.

●(Uniqueness)
Uniqueness means that there is no repeated data in the dataset.

6.2 Features Captured
There are nine features captured in the dataset in order to discover whether passengers can or cannot survive from the iceberg accident. We found that sex, fare and embarked are the three features which closely related to the iceberg accident. Family number and size could be added as variables to determine whether passengers can stay alive or not. More features captured would result in high-accuracy outcomes.

7. Conclusion

The aim of the project is to make predictions of Titanic survival rates with machine learning by using the programming language, Python. First of all, the process of data cleaning and data reduction are implemented to obtain a more precise and suitable dataset. This helps the concentration on the research. Next, the four Machine Learning algorithms, including Logistic Regression, Decision Tree, Random Forest and Linear Regression, are used to train the dataset for making an accurate and precise prediction whether the individuals in the titanic iceberg accident would survive or not. The correlation between survival and the factors, including sex, fare, embarkedness, parch, passenger class, age as well as the number of spouse/siblings are also found during the data analysis process. The factors mentioned above are also used to help distinguish whether the individual would survive or not. In the last section of the project, the limitations and possible improvements are written down.
In conclusion, the project analyses the causes of Titanic passengers' survival from a wide range of quantitative and categorical variables and gives comprehensive suggestions to both people who want to study machine learning as well as the Titanic iceberg accident.

8. Special Thanks

Special thanks to our supervisor Dr Cheung King Chau, Simon. Before we started the project, he provided various materials to give us a clear path. During the making of the project, he also launched 2 meetings to discuss the project with us in order to help us work better. After his instruction, we tried a wide range of algorithms, including Adaboost and Naive Bayes Classifier, to make machine learning predictions in our project. Also, Dr Cheung also gave some advice on how to use ROC Curve to determine the accuracy of the algorithms. Finally, Dr Cheung suggested that we could use a better graph method to present our results.

9. Reference

Accuracy, precision, Recall & F1-Score - Python examples. Data Analytics. (2022, April 8).
Retrieved April 17, 2022, from
https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/

Adaboost. Adaboost - an overview | ScienceDirect Topics. (n.d.). Retrieved April 17, 2022,
from https://www.sciencedirect.com/topics/engineering/adaboost

Anesi, C. (n.d.). *Titanic Disaster: Official Casualty Figures. Titanic disaster: Official
casualty figures and commentary.* Retrieved April 10, 2022, from
https://www.anesi.com/titanic.htm

Bayesian network classifier - Christopher Sheaffe. (n.d.). Retrieved April 17, 2022, from
https://christophersheaffe.com/project/bayesnet

Brownlee, J. (2020, August 14). *Crash course on multi-layer perceptron neural networks*.
Machine Learning Mastery. Retrieved April 17, 2022, from
https://machinelearningmastery.com/neural-networks-crash-course/

Chuang, L. (2020, October 1). *Build a neural network in Python (binary classification)*.
Medium. Retrieved April 17, 2022, from
https://medium.com/luca-chuangs-bapm-notes/build-a-neural-network-in-python-binary-classification-49596d7dcabf

Decision tree algorithm, explained. KDnuggets. (n.d.). Retrieved April 10, 2022, from
https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Dhavalsays. Support Vector Machine Tutorial Using Python Sklearn. Github. (2018,
December 17). Retrieved February 16, 2022, from
https://github.com/codebasics/py/blob/master/ML/10_svm/10_svm.ipynb

Doshi, S. (2020, August 3). *Various optimization algorithms for training neural network.
Medium*. Retrieved April 17, 2022, from

https://towardsdatascience.com/optimizers-for-training-neural-network-59450d71caf6

Géron Aurélien. (2020). *Hands-on machine learning with scikit-learn, Keras, and tensorflow concepts, tools, and techniques to build Intelligent Systems* (2nd ed.). O'Reilly.

Hall, W. (2002, June 20). *Social class and survival on the S.S. Titanic. Social Science & Medicine*. Retrieved April 10, 2022, from
https://www.sciencedirect.com/science/article/pii/0277953686900419

How the naive bayes classifier works in machine learning. Dataaspirant. (2017, February 19). Retrieved April 17, 2022, from
https://dataaspirant.com/naive-bayes-classifier-machine-learning/

How to plot ROC curve in python. Stack Overflow. (1962, May 1). Retrieved April 17, 2022, from https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python

Independent Digital News and Media. (1998, April 10). *Did the third class passengers on the Titanic have a fair chance? The Independent.* Retrieved April 10, 2022, from
https://www.independent.co.uk/voices/did-the-third-class-passengers-on-the-titanic-have-a-fair-chance-1155678.html

Kayyath, A. (2021, October 28). *Confusion matrix : Let's clear this confusion. Medium.* Retrieved April 17, 2022, from
https://medium.com/@aatish_kayyath/confusion-matrix-lets-clear-this-confusion-4b0bc5a5983c

Kurama, V. (2021, April 9). *A guide to understanding AdaBoost*. Paperspace Blog. Retrieved April 17, 2022, from https://blog.paperspace.com/adaboost-optimizer/

Linear regression: Introduction to linear regression for data science. Analytics Vidhya. (2021, May 25). Retrieved April 10, 2022, from
https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-

machine-learning-model-linear-regression/#:~:text=In%20the%20most%20simple%20
words,the%20dependent%20and%20independent%20variable

Logistic regression: Beginner's guide to logistic regression using python. Analytics Vidhya.

(2021, April 25). Retrieved April 10, 2022, from

https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-logistic-regression-

using-python/

Machine learning decision tree classification algorithm. (n.d.). Retrieved April 10, 2022,

from

https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Magazine, S. (2010, March 2). *Titanic vs. Lusitania: Who survived and why?*

Smithsonian.com. Retrieved April 17, 2022, from

https://www.smithsonianmag.com/science-nature/titanic-vs-lusitania-who-survived-and

-why-24622866/

Picardo, E. (2022, February 8). *Negative correlation definition. Investopedia*. Retrieved April

10, 2022, from

https://www.investopedia.com/terms/n/negative-correlation.asp

Random Forest: Introduction to random forest algorithm. Analytics Vidhya. (2021, June 24).

Retrieved April 10, 2022, from

https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Receiver operating characteristic (ROC). scikit. (n.d.). Retrieved April 17, 2022, from

https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

Sarfin, R. L., & Editor, P. (2021, May 12). *Data Quality Dimensions: How do you measure

up?*. Precisely. Retrieved April 10, 2022, from

https://www.precisely.com/blog/data-quality/data-quality-dimensions-measure

Simplilearn. (2022, February 21). *An overview on multilayer perceptron (MLP) [updated]*.
Simplilearn.com. Retrieved April 17, 2022, from
https://www.simplilearn.com/tutorials/deep-learning-tutorial/multilayer-perceptron

Stephen J. Spignesi. (n.d.). *Titanic passenger survival rates*. dummies. Retrieved April 10,
2022, from
https://www.dummies.com/article/academics-the-arts/history/20th-century/titanic-passe
nger-survival-rates-180788/

Strom, A. (2021, December 18). *Belonging archives*. Re. Retrieved April 17, 2022, from
https://reimaginingmigration.org/tag/belonging/

Support Vector Machines (SVM) algorithm explained. MonkeyLearn Blog. (2017, June 22).
Retrieved February 16, 2022, from
https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

The Ultimate Guide to Linear Regression for Machine Learning. (n.d.). Retrieved April 10,
2022, from https://www.keboola.com/blog/linear-regression-machine-learning

12 types of neural networks activation functions: How to choose?. (n.d.). Retrieved April 17,
2022, from
https://www.v7labs.com/blog/neural-networks-activation-functions