**Prediction of fundraising goal using machine learning**

**INTRODUCTION**

Crowdfunding is a process of funding an individual, business and organizations projects with small amount of donations from many people. It can be said that instead of getting funds from one or two investors to get from large volume of people. There are 4 types of crowdfunding: debt, donation, rewards and equity. Most of the time the necessary funds are collected via online platforms (Johnson, 2019). It usually involves three players: the business which needs to be funded – fundraising team, the people who provide funding for the project, the website who acts as a mediator in between them.

Usually, the most common way is to raise funds as money. But there are other types like initial coin offerings (ICO). In initial coin offerings direct money is not exchanged. ICOs can be used to raise funds when a company needs money by generating and selling digital coins. These coins are called cryptocurrency token. These tokens have some value related to the organization or product. Hence, this can be used by the individual purchased. (Frankenfield, 2016).

The main objective of this coursework is to predict whether the business/company can reach its fundraising goal through ICOs successfully within the time allotted by using various machine learning model.

The purpose of the study is to apply different machine learning algorithms in the given dataset. This problem only consists of classification. Different models are predicted with different evaluation techniques to get the best model.

**DATA EXPLORATION**

It is always important to learn about the variables as it will help us to determine the correct variables for doing machine learning.

The given dataset contains 2767 observations with 16 variables.

* **ID** – It is the unique number given to each of the project present in the dataset. This column does not provide any insights. Hence, we remove the first column.
* **Success** – This parameter explains whether a fundraising project achieved its goal are not.

A picture containing text, font, typography, graphics

Description automatically generated

N indicates that the company was not successful in achieving its goal and Y is vice versa.

* **BrandSlogan** - It is the slogan for each fundraising organisation.
* **Has\_Video** – This variable tells us whether a company has a video in its campaign page or not.

A picture containing font, typography, graphics, text

Description automatically generated

0 indicates that there is no video and 1 is vice versa.

* **Rating** – This variable says about the quality of the project by investment experts. It ranges from 1 to 5 with 5 being very good. The distribution of the rating is shown as below.

A picture containing diagram, text, screenshot, line

Description automatically generated

* **PriceUSD** – It is the price of each blockchain coin in dollars. The distribution of the price consists of outliers and missing value as shown below which needs to be removed. It lies between the range between 0 to 39384.

A picture containing text, screenshot, diagram, font

Description automatically generated

* **Country Region** - The region or country in which the organisation is situated. This column consists of 115 countries with some countries misspelled and 71 blank rows.
* **Start Date and End Date** – These two columns indicate the start date and end date of the fundraising campaign.

* **TeamSize** – This column consists of the number of people in the organisation. It ranges from 1 to 75 people.

A picture containing text, diagram, screenshot, line

Description automatically generated

* **hasGithub** – This column tells that the company has provided the link for its GitHub page in their campaign page.

A picture containing font, text, typography, screenshot

Description automatically generated

0 indicates that the company has not provided the link in their page and 1 is vice versa.

* **hasReddit** – This column tells that the company has provided the link for its reddit page in their campaign page.

A picture containing text, font, typography, graphics

Description automatically generated

0 indicates that the company has not provided the link in their campaign page and 1 is vice versa.

* **Platform** – It is the blockchain platform of the fundraising company. We have 101 unique platforms in our data. This column also consists of misspelled and 6 blank rows.
* **CoinNum** – It is the total number of coins issued by the organisation. It is set by the organisation. It ranges from 12 to 2.261908e+16.

A picture containing text, screenshot, display, diagram

Description automatically generated

* **MinInvestment** – This tells whether the organisation has initial amount to invest in the project or not.

A picture containing font, text, typography, graphics

Description automatically generated

0 indicates there is no minimum amount to be invested and 1 is vice versa.

* **Distributed Percentage** – This tells us about the amount coin distributed with reacts to total number of coins. It ranges from 0 to 869.75.

A picture containing text, screenshot, diagram, display

Description automatically generated

**CORRELATION AND MISSING VALUE**

A picture containing screenshot, square, pattern, rectangle

Description automatically generated

Here we use Corrgram which is used to check the relationship between the variables using correlation matrix. The dark red represents there is a weak correlation between ID and other columns. Since it is only ID, we don’t consider it for our analyses. As we can see success has a positive correlation with seven of the variables and negative correlation with two variables.

A picture containing screenshot, rectangle, line, plot

Description automatically generated

The above figure suggests that there is total of 334 missing values in the numerical columns in which 154 rows of team size and 180 rows of priceUSD is missing. Also, there are 9 rows where both team size and price USD is missing.

A picture containing screenshot, design, art

Description automatically generated with medium confidence

This figure suggests which row consists of the missing value. We can see that most of the team size is NA towards the end of the dataset. The priceUSD’s missing value is evenly across the whole column.

We also have 71 blank rows in the country region and 6 rows in the platform. Since this is a categorical data, the blank data is considered as a variable. So, we see these blank data using the table format.

**DATA PRE-PROCESSING**

Data pre-processing is carried out in order to convert the raw data into useful data. The real-life data is always messy, unstructured and noisy as it is processed by variety of humans for their needs. Hence, it contains missing value spelling mistakes, wrong entry of data and data redundancy. (Lawton, 2022)

First, we remove ID and brand slogan columns as ID does not indicate any thing it is used for identification of the project and brand slogan consists of sentences. Hence, we can use brand slogan for text analysis.

**Text Analysis**

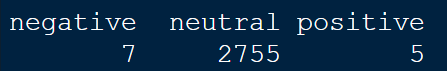
When we have column of sentences, instead of removing them we can create useful insights from the text by doing some prepossessing for text. We remove the space, past tense, future tense. We only keep the correct word.

**A close-up of words

Description automatically generated with low confidence**

By looking at the word cloud, we can see that blockchain, platform, decentralized and cryptocurrency are the most common words. We set minimum frequency of 50.

As we can see that most of the slogan does not interpret any sentiment. 99% of the text is neutral.



* **Platform –**

First, we remove the unwanted space in the column by just leaving one space between each word using the function str\_squish (). Then convert all the first letter of the word to upper case and remaining in lower case in order to keep all rows in the same manner using str\_to\_title (). As this helps to reduce the redundancy.

Next, we recode all the misspelled words in platform. E.g., Eth to Ethereum. We also remove the 6 blank rows as platform is considered one of the important columns in prediction. Removing the blank data will help to get better accuracy.

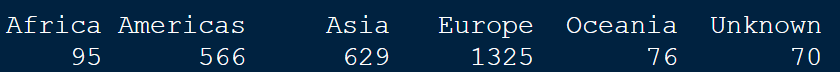
Finally, there are some platforms which occur only once. It will be difficult for the model to predict them. Hence, we group them based on their occurrence. In this case we group platforms which is less than 5 as “Others”.

A blue screen with white text

Description automatically generated with low confidence

* **Country Region –**

We convert all first letter of the word to upper case and convert the misspelled words. In order to use them in model we group them based on their continents. We also replace the NA values as “Unknown”. Finally, we remove the Country as there may redundancy in columns.



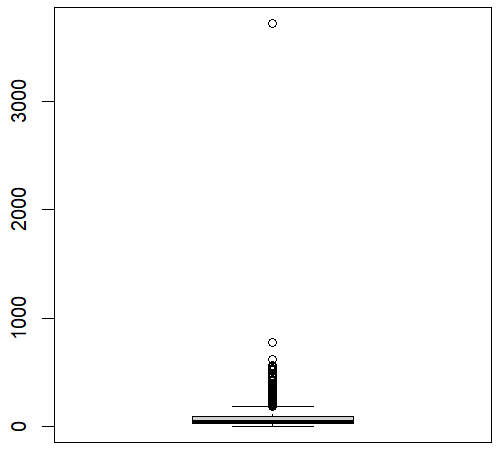
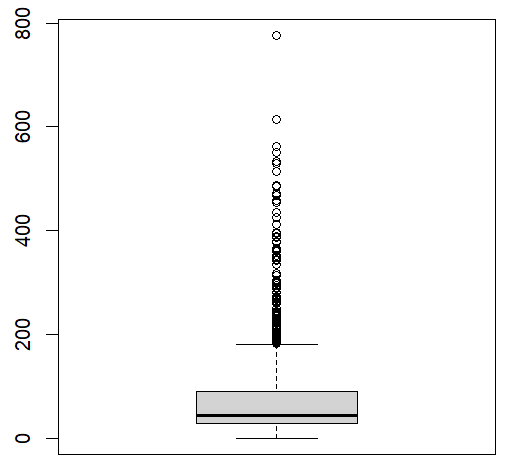
* **Start date and End date –**

Firstly, we change the format from text to date. Then we create a new column to calculate the duration of the fundraising project. This column is in a format called difftime which cannot be processed by the model.

Hence, we convert them into numeric and make its unit as days. Also, we remove negative data as no fundraising project cannot end before it starts. Hence, we remove the negative data.

Since, we converted the start date and end date as duration which can be used in the model to predict. We remove both the start date and end date. We also maintain a 99% confidence interval and remove 1 of the observations instead of all the outliers. This is done for the model to increase its predictability even in extreme values.

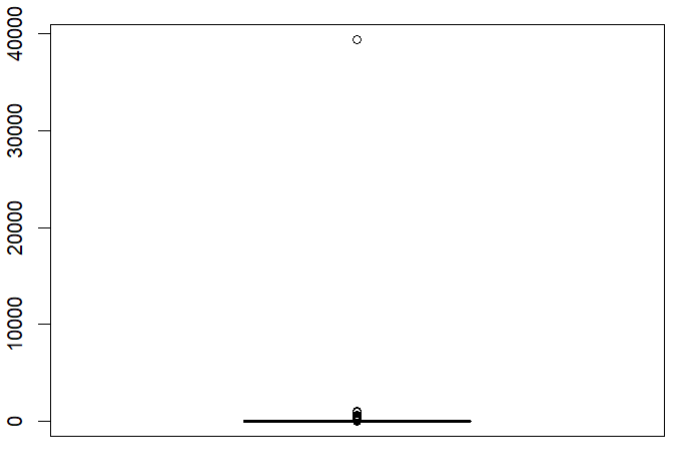
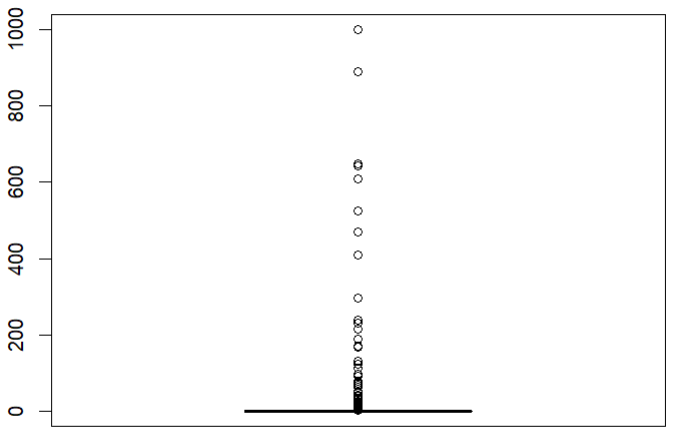
a) before removing outliers in duration b) after removing outliers in duration

* **Team Size and PriceUSD –**

As discussed earlier we need to do imputation in order to remove missing values. As there are two methods to do imputation: Simple imputation and Multiple imputation. Here we consider simple imputation in terms of median. Since doing imputation based on mean and multiple imputation involves mean. As our data consists of outliers, we might get a skewed distribution. Hence, we use simple imputation in terms of median.

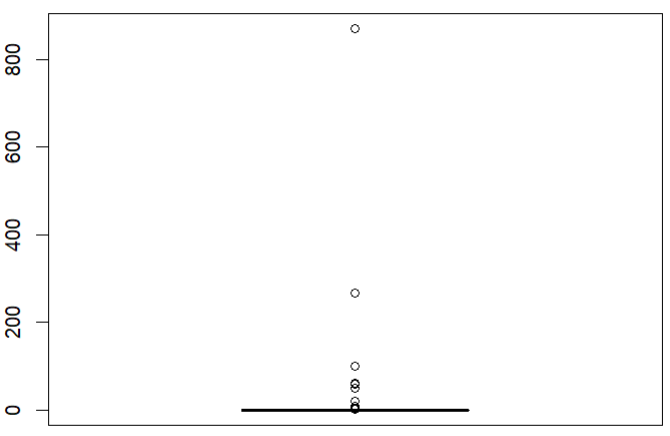
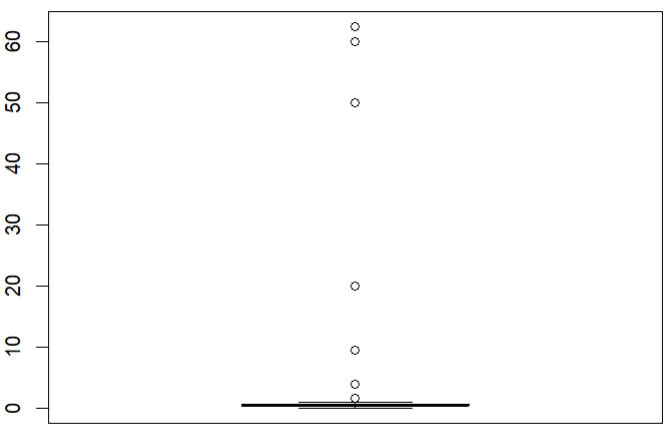
We also remove outliers from PriceUSD with a 99% confidence interval. We remove the row with value 39384.

a) before removing outliers in PriceUSD b) after removing outliers in PriceUSD 

* **Distributed percentage –**

For e.g., if we have 100 coins and sold all of them then we tend to say that we sold 100% of the products. Hence the percentage cannot cross more than 100. So, we remove values which are greater than 100.

a) before removing outliers in percentage b) after removing outliers in percentage

**MODELLING**

After all the preprocessing is carried out, now the data is in usable format. The dataset contains 2744 observations with 13 variables.

A screen shot of a computer code

Description automatically generated with low confidence

Before we investigate various models. We split the data into training data (90% of the data i.e., 2469 observations) and remaining as testing data (275 observation). These rows are chosen at random using the sample function.

**Decision Tree –**

It is a supervised machine learning algorithm used for classification and regression. Decision tree is one of the simplest methods to use. As it does not require the conversion of categorical data into numerical format. It also ignores the NA values. It creates a model in the form of a tree where the root node is the starting point and consists of the main attribute used. The decision node consists of each of the features we have derived. Finally, the leaf node is the predicted class or the outcome (Lantz, 2019). The split in the decision tree is based on the best information gain in the training data. Decision tree requires the label to be in a factor format.

A picture containing text

Description automatically generated

As we can see our root node here is rating and all the other columns are decision nodes and success column is our outcome. The size of the is 50. Also, in the rating the best split is at 3. Hence two nodes are created, one greater than 3 and the other less than 3.

A screen shot of a computer

Description automatically generated with low confidence

From the above figure we can see that for deciding whether the fundraising goal can be achieved is mainly dependent on three variables: rating, duration, teamsize. C5 .0 algorithm was used for the prediction. The decision tree was performed with including all the 13 variables. Decision tree is used as it is simple to interpret, it includes all kinds of data format.

**Ensemble methods –**

These methods are used to improve accuracy by improving the factors used for predictions. It basically combines many models instead of using one model. In this way it is less biased and less sensitive to data. Ensemble method is distinguished into three types (Trinidad, 2020).

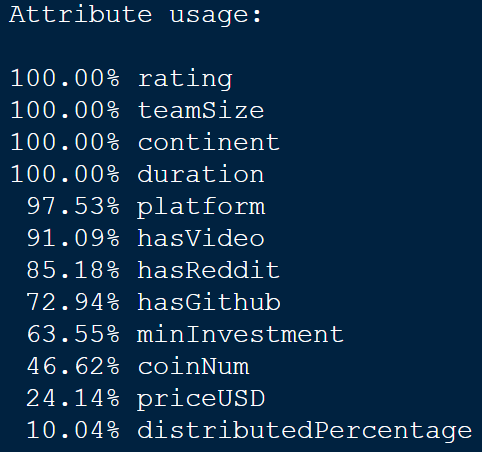
**A picture containing text, diagram, line, screenshot

Description automatically generated**

In our coursework we are gone use adaptive boosting and random forest

1. **Adaptive Boosting –**

Boosting always learns from the mistakes made by past predictions. Hence, we iterate the prediction model. Adaptive boosting works by iterating the set of weak learners from the previous iterations. These weak learners learn from the next (Trinidad, 2020).In our case we iterated the model 10 times in order to improve our performance. The size of the tree is reduced to 15.



In our previous model, it only considered mainly 3 attributes. But in adaboost the model learns each iteration. As a result, we can see that the model predicts the outcome mainly based on involving 9 attributes. This gives a better result than the previous decision tree model.

1. **Random Forest –**

Bootstrap aggregating is the abbreviation of bagging. In bagging the several models

are trained parallel with a random subset of the data. There is no correlation between these models like boosting. Once the result is obtained, we aggregate the result from each model. In classification we use voting. Here the class which receives the most votes in all models is the final predicted class. In random forest various decision trees are created and features are selected at random. (Trinidad, 2020).

In Random Forest, we decide the number of trees. It is always better to have a greater number of trees. In our case we set the tree size to 500

We have performed two ensemble methods with Decision tree. It is clearly visible that both the ensemble methods performed better than the normal decision tree.

**Support Vector Machines –**

Support vector machine is usedin high dimensional data and where we can differentiate between classes. The main purpose of SVM is to find the most suitable hyperplane to distinguish between different classes. It is always important to choose the plane with maximum distance from the data points in classes. This is done so that more points can be added in future with a better confidence interval. (Gandhi, 2018). The dimension of the hyperplane is set according to the features we have. In our case it is 28-dimensional as we have 28 different features.

The concept of soft margin is introduced in order to get a more flexible classification. As soft margin allows a very less amount of data points on the wrong side of the plane. Here we keep C=1 which implies a balanced way. As we don’t want perfect margin or a wrong classification (Mandot, 2017).

Kernal function is used to convert the data into required data using mathematical formula. There are many types of kernels available. But we have used two of them:” vanilladot”,” polydot”.

In SVM, we convert the dataset into numeric using the fastdummies library. This library is used to convert the

categorical value into 1s and 0s for each unique value in the column. SVM also needs the data to be normalized

but ksvm function used in this model does it for us.

We can see that both the model gives us almost the same accuracy.

**KNN –**

KNN stands for K-nearest neighbours. KNN can be used for both classification and regression. It

classifies the unlabelled data by assigning them with class of labelled data. It is classified based on the distance

between the nearest neighbours. K refers to the number of points we include while identifying the class of the

data by majority (Subramanian, 2022).

In KNN, the K value is usually the square of the training data available with us. The distance calculated for the

neighbours are mostly Euclidean distance. But other distance like manhattan, hamming, minkowski distance

can also be used (Subramanian, 2022).

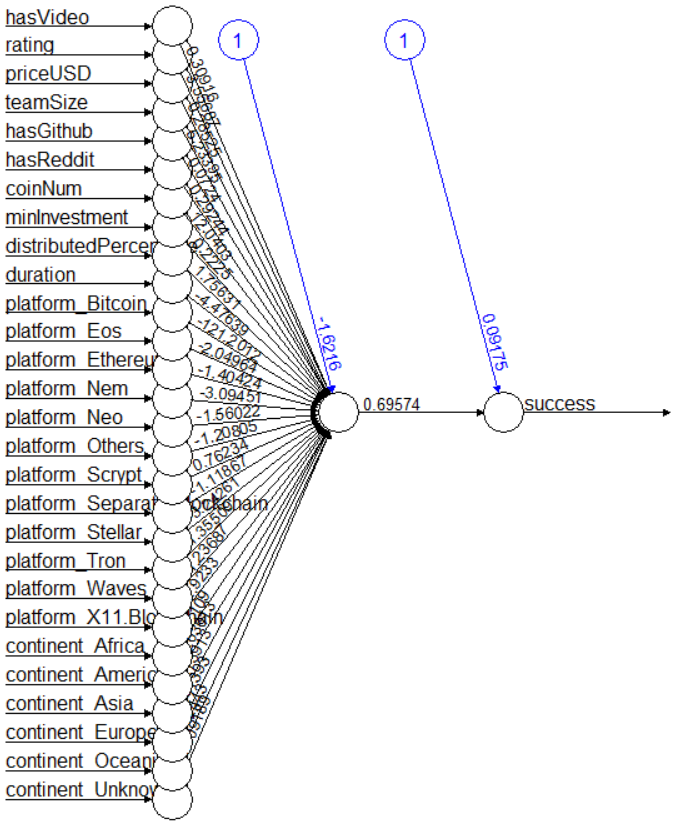
The class of the test data is decided by voting. The K value decides the number of neighbours it is compared. K number of nearest data to test data is selected. The test data’s class is selected by voting to majority of the class present.

KNN only takes in numerical data and the data is normalized. There two types of normalization: Minmax and Z score. Here we use Minmax and set the k value to 49. Since we get the maximum accuracy at this point.

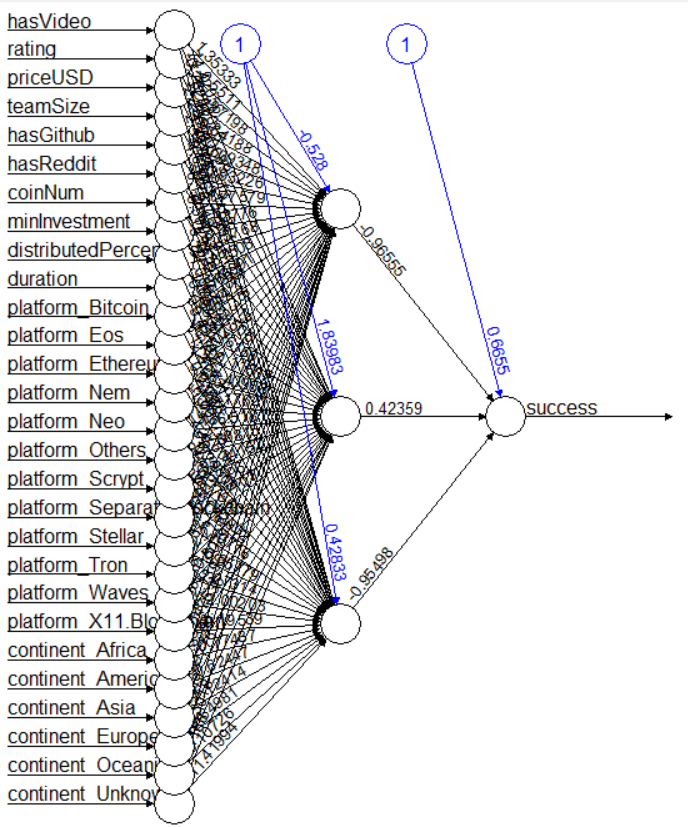
**ANN –**

ANN stands for Artificial Neural Networks. It depends on the structure of biological neural networks like the brain. It forms a link between each node. The ANN consists of three layers. The input layer, hidden layer and output layer (Rouse, 2011).

The basic algorithm behind ANN is that the data is passed through the input layer. Each neuron performs weighted sum and applies activation function and provides the output through the output node. The weighted sum is the result of learning from the previous node and is assed with more learning to the next node.

****

The above structure shows there is no hidden layer in this ann model.

****

Here the network consists of 2 hidden layers.

As we tried both the methods, we have presented the method with 2 hidden layers as it had better accuracy.

**EVALUATION**

This is the important step in the whole process as it tells us how each of the model performances including whether the model is overfitting or underfitting. Also, helps us to determine the best model using different evaluation techniques.

There are several types of evaluation. Here we have used 8 evaluation methods as follows: -

* **ROC CURVE –** It is the graphical representation of performance of classification model at each threshold level.
* **AUC –** It represents the area under the curve for each threshold. It evaluates the overall performance of the models.
* **ACCURACY –** It is a performance metric which measures how correctly it predicts the outcome or the labels.
* **SENSITIVITY –** It tells how many of the true positives are correctly predicted.
* **SPECIFICITY –** It tells how many of the true negatives are correctly predicted.
* **PRECISION –** It tells the proportion of true positive from the total predicted as positive.
* **KAPPA –** It is used to evaluate the performance between the observed classification and predicted classification.
* **K – FOLDS** - It is a cross validation technique used when we a finite amount of data and want to achieve correct accuracy. The data is separated into k – sized and iterated k times. As we performed other performance metrics everything is performed in the loop. But for a small group of data. Finally, the average of the performance metrics is calculated. It provides an accurate performance and reliable measure. (Lyashenko, 2022) In our case we have used 10 folds cross validation.

**A picture containing text, diagram, line, plot

Description automatically generated**

The curve tells that all the models have performed better than the 50 – 50 chances.

**Performance metrics before K folding**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MODEL** | **ACCURACY** | **AUC** | **SENSITIVITY** | **SPECIFICITY** | **PRECISION** | **KAPPA** |
| Decision Tree | 0.6909 | 0.6623 | 0.4253 | 0.8138 | 0.5139 | 0.2507 |
| Ada Boost | 0.6873 | 0.6640 | 0.3563 | 0.8404 | 0.5082 | 0.2139 |
| SVM (kernel-Polydot) | 0.6836 | 0.6959 | 0.3793 | 0.8245 | 0.5 | 0.2179 |
| SVM (kernel-vanilladot) | 0.6836 | 0.6961 | 0.3793 | 0.8245 | 0.5 | 0.2179 |
| KNN | 0.6945 | 0.6624 | 0.4023 | 0.8298 | 0.5224 | 0.2474 |
| Random Forest | 0.6764 | 0.6467 | 0.4023 | 0.8032 | 0.4861 | 0.2155 |
| ANN  (Hidden layer = 3) | 0.6836 | 0.6653 | 0.4253 | 0.8032 | 0.5 | 0.238 |

This tells us that the accuracy of all the models does not vary a lot. Moreover, the kappa value is just fair despite having high accuracy. It might be because of biased learning of the models.

**Performance metrics in K folding**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MODEL** | **ACCURACY** | **SENSITIVITY** | **SPECIFICITY** | **PRECISION** |
| Decision Tree | 0.6559 | 0.4007 | 0.8076 | 0.5602 |
| Ada Boost | 0.6661 | 0.3949 | 0.8274 | 0.5752 |
| SVM (kernel-Polydot) | 0.6774 | 0.3470 | 0.8738 | 0.6260 |
| SVM (kernel-vanilladot) | 0.6789 | 0.3118 | 0.8971 | 0.6448 |
| KNN | 0.6617 | 0.3421 | 0.8517 | 0.5809 |
| Random Forest | 0.6665 | 0.3855 | 0.8355 | 0.5798 |
| ANN  (Hidden layer = 3) | 0.6748 | 0.4212 | 0.8277 | 0.5891 |

**CONCLUSION**

We can see that in deciding whether a fundraising project is successful or not. **Rating, teamsize, duration, continent** are the features which are important to consider. Even after k folding SVM produces better results than any other model. ANN can also be used to predict the success of the company. By increasing the hidden layers, we can increase the accuracy even better.

For future analysis, all these methods can be tested by changing different parameters. Also, it is better to collect more data as the dataset is quite small and the classes are biased as class No contains 1000 data more than yes. Hence the model’s predictability is biased.

**REFERENCES**

Johnson, S. 2019. The Basics of Crowdfunding for Your Business - businessnewsdaily.com. *Business News Daily*. [Online]. [Accessed 11 May 2023]. Available from: <https://www.businessnewsdaily.com/4134-what-is-crowdfunding.html>.

Frankenfield, J. 2016. Initial coin offering (ICO): Coin launch defined, with examples. *Investopedia*. [Online]. [Accessed 12 May 2023]. Available from: <https://www.investopedia.com/terms/i/initial-coin-offering-ico.asp>.

Lawton, G. 2022. Data preprocessing: Definition, key steps and concepts. *Data Management*. [Online]. [Accessed 17 May 2023]. Available from: <https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing>.

Lantz, B. 2019. *Machine Learning with R: Expert techniques for predictive modeling* 3rd ed. Birmingham: Packt Publishing

Trinidad, C. 2020. Ensemble methods. *Corporate Finance Institute*. [Online]. [Accessed 17 May 2023]. Available from: <https://corporatefinanceinstitute.com/resources/data-science/ensemble-methods/>.

Gandhi, R. 2018. Support vector machine — introduction to machine learning algorithms. *Towards Data Science*. [Online]. [Accessed 17 May 2023]. Available from: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>.

Mandot, P. 2017. What is the Significance of C value in Support Vector Machine? *Medium*. [Online]. [Accessed 17 May 2023]. Available from: <https://medium.com/@pushkarmandot/what-is-the-significance-of-c-value-in-support-vector-machine-28224e852c5a>.

Subramanian, D. 2022. K-nearest neighbor algorithm: An introduction. *Built In*. [Online]. [Accessed 18 May 2023]. Available from: <https://builtin.com/machine-learning/nearest-neighbor-algorithm>

Rouse, M. 2011. Artificial Neural Network. *Techopedia*. [Online]. [Accessed 18 May 2023]. Available from: <https://www.techopedia.com/definition/5967/artificial-neural-network-ann>

Lyashenko, V. 2022. Cross-validation in machine learning: How to do it right. *neptune.ai*. [Online]. [Accessed 18 May 2023]. Available from: https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right.

**APPENDIX**

* Decision Tree

A screenshot of a computer screen

Description automatically generated with low confidence

* Ada Boost

A screenshot of a computer screen

Description automatically generated with medium confidence

* SVM

A screenshot of a computer screen

Description automatically generated with low confidence

* SVM (VANILLADOT)

A screenshot of a computer screen

Description automatically generated with low confidence

* KNN

A screenshot of a computer screen

Description automatically generated with low confidence

* Random Forest

A screenshot of a computer screen

Description automatically generated with low confidence

* ANN

A screenshot of a computer screen

Description automatically generated with medium confidence