Automatic Detection of Money donors using Supervised Machine Learning Models

Naresh Vurukonda[1]

Department of Computer Science and Engineering,Koneru Lakshmaiah Education Foundation (KLEF), Deemed to be University, Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India-522302.

naresh.vurukonda@kluniversity.in

V.Vidyasagar[2]

School of Technology Management and Engineering, SVKM's Narsee Monjee Institute of Management Studies Deemed to be University,Hyderabad Campus, Mahbubnagar, Telangana,509301, India.

Email: vidyasagar24@gmail.com

DOI : 10.48047/IJFANS/11/S7/007

Abstract: Our goal is to correctly predict whether a given individual makes a profit of greater than 50,000 or less than 50,000 based on a set of attribute features which are already provided. So, with the data available we can come to conclusion that an individual can be a donor or not. And this model can help non-profit organizations which certainly depends on donation to correctly predict the donation that the organization has to request an individual based on the individual records. The is a hypothetical case study to identify potential donors to a charity that offers funding to people. It was found that every donor was making more than $50,000 annually. My task was to use machine learning algorithms to help this charity identify potential donors.

Keywords: Donor prediction; Non-profit organizations; Machine learning algorithms; Income classification

1. Introduction

We will utilize various diverse regulated calculations to definitively foresee people pay utilizing information gathered from the 2020 U.S Statistics census. We will then, at that point, pick the best applicant calculation from fundamental outcomes and further improve this calculation to best model the information. Our target with this execution is to build a model that definitively predicts whether a particular makes more than $50,000. Understanding a people pay can assist a non-benefit with bettering of a gift to ask for, or whether or not they should contact start with. As from our previous research we have found out that the individuals who are probably going to give cash to a foundation are the ones that make more than $50,000.

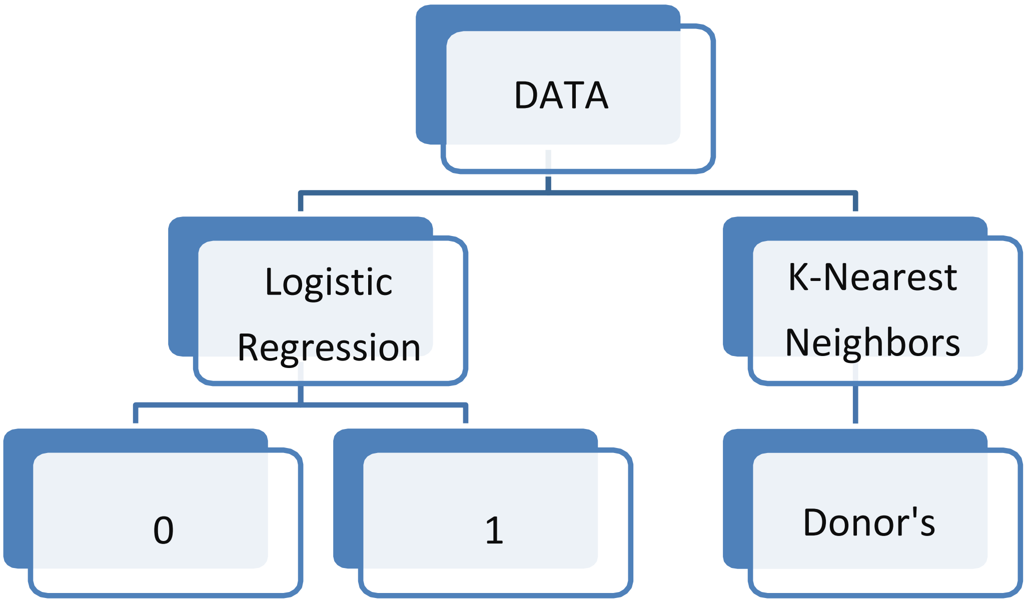


Figure 1 Data is divided into LR & K-NN

2. Design/Methods/Modelling

2.1 Logistc Regression:

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable (target) is categorical [1].

For example[2]

To predict whether an email is spam (1) or (0) Whether the tumor is malignant (1) or not (0)

2.2 K-Nearest Neigbors:

K-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent differentphysical units or come in vastly [3] different scales then normalizing the training data can improve its accuracy dramatically.

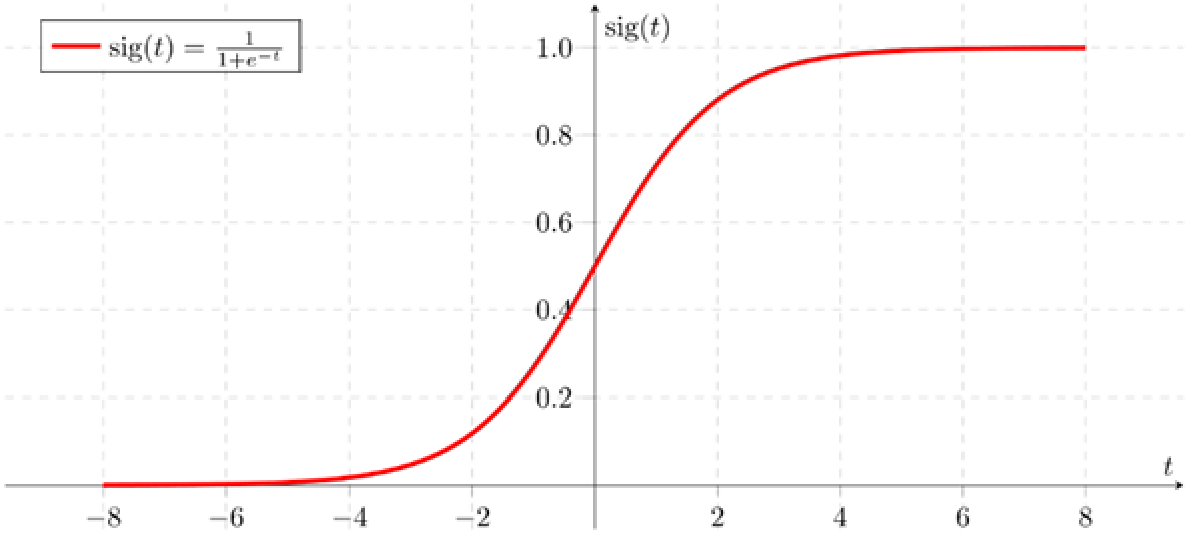
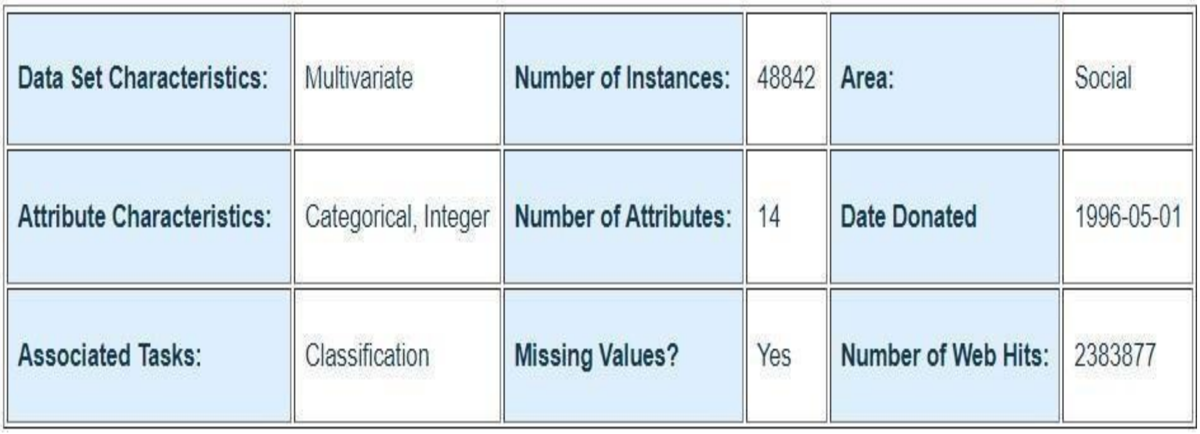


Figure 2 Displaying the sample Graph of K-NN

2.3 Data:

The changed statistics dataset comprises of roughly 32,000 relevant items, with each data-point having 13 highlights.[5] This dataset is a changed version of the dataset appropriated in the paper Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid, by Ron kohavi. You may track down this paper on the web, with the first data set facilitatedon UCI.



Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	2383877

Figure 3 Displaying the Versions of Data

2.4 Features:

i. age: Age

ii. workclass: Represents the employment status of individuals, including categories like Private, Self-employed (not incorporated), Self-employed (incorporated), Federal government, Local government, State government, Without pay, and Never worked.

iii. education_level: Indicates the highest level of education completed, ranging from options like Bachelors, Some school, 11th grade, HS-graduate (High School Graduate), Prof-school (Professional School), Assoc-acdm (Associate Degree, Academically Oriented), Assoc-voc (Associate Degree, Vocationally Oriented), 10th grade, 7th-8th grade, 9th grade, 12th grade, Masters, 1st-4th grade, Doctorate, 5th-6th grade, and Preschool.

iv. education-num: Represents the number of educational years completed by an individual.

v. marital-status: Indicates the marital status of the individual, such as Married (civilian spouse), Divorced, Never married, Separated, Widowed, Married (spouse absent), and Married (armed forces)

vi. occupation: Refers to the type of occupation or job held by the individual, including categories like Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct (Machine Operation Inspector), Adm-clerical (Administrative Clerical), Farming-fishing, Transport-moving, Priv-house-serv (Private House Service), Protective-serv, and Armed Forces.

vii. relationship: Describes the relationship status of the individual, including Wife, Own-child, Husband, Not-in-family, Other-relative, and Unmarried.

viii. race: Represents the racial background of the individual, with options like White, Asian-Pac-Islander (Asian Pacific Islander), Amer-Indian-Eskimo (American Indian Eskimo), Other, and Black.

ix. sex: Indicates the gender of the individual, with options being Female and Male.

x. capital-gain: Refers to the monetary capital gains of the individual.

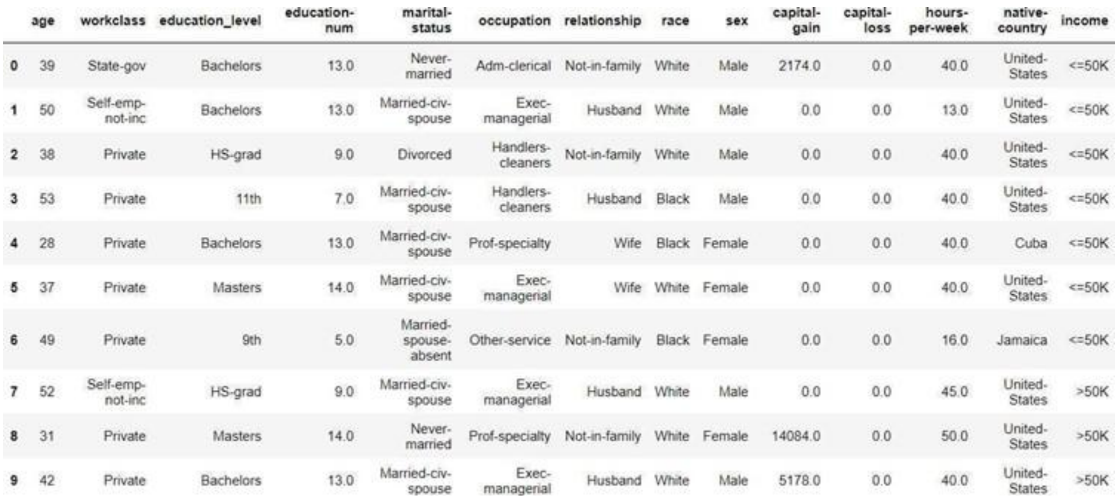
xi. capital-loss: Represents the monetary capital losses of the individual

xii. hours-per-week: Indicates the average number of hours worked per week by the individual.

xiii. native-country: Specifies the country of origin or the country where the individual is from. Options include United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US (Guam-USVI, etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Target Variable:

income: Income Class (<=50K, >50K)



	age	workclass	education_level	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	Bachelors	13.0	Never-married	Adm-clerical	Not-in-family	White	Male	2174.0	0.0	40.0	United-States	<=50K
1	50	Self-emp-not-inc	Bachelors	13.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0.0	13.0	United-States	<=50K
2	38	Private	HS-grad	9.0	Divorced	Handlers-cleaners	Not-in-family	White	Male	0.0	0.0	40.0	United-States	<=50K
3	53	Private	11th	7.0	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0.0	0.0	40.0	United-States	<=50K
4	28	Private	Bachelors	13.0	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0.0	0.0	40.0	Cuba	<=50K
5	37	Private	Masters	14.0	Married-civ-spouse	Exec-managerial	Wife	White	Female	0.0	0.0	40.0	United-States	<=50K
6	49	Private	9th	5.0	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0.0	0.0	16.0	Jamaica	<=50K
7	52	Self-emp-not-inc	HS-grad	9.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0.0	45.0	United-States	>50K
8	31	Private	Masters	14.0	Never-married	Prof-specialty	Not-in-family	White	Female	14084.0	0.0	50.0	United-States	>50K
9	42	Private	Bachelors	13.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178.0	0.0	40.0	United-States	>50K

Figure 4 Displaying the sample data records.

Table 1 Displaying the Authors References according to research.

Authors	ProposedSchemes	Services	Possibility	Availability
Papandreou, George, et al.	Weakly-and semi-supervised learning (2015)	IEEE international conference on CV	✓	✓
Fabris, Fabio, João Pedro De Magalhães, and Alex A. Freitas.	Supervised ML applied to ageing research (2017)	Biogeronto-logy 18.2	✓	✓
Bonica, Adam.	Inferring roll- call scores from contributions (2018)	Political Science of American Journal	✓	✗
Yan, Ke, et al.	Deep lesion graphs in the wild (2018)	IEEE Conference on Computer Vision and Pattern Recognition	✗	✓

2.5 Models and Algorithms:

The Logistic regression algorithm is a predictive analysis algorithm based on probability. It is used for classification problems. [3] The hypothesis of logistic regression [10] tends to limit the cost function between 0 and 1. The K-Nearest Neighbors is another machine learning algorithm used in solving this problem of predicting [5] donors. It can be useful for both regression and classification problems. KNN model implementation is done simply in few steps as:

I. Load the data and initialize the value of K.

II. For getting predicted class, [11] iterate from 1 to the total number of training data points.

III. Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it’s the most popular method. [4] The other parameters that can be used are Chebyshev, cosine, etc.

IV. Sort the calculated distances in ascending order based on distance values. [6]

V. Get top K rows from the sorted array.

VI. Get the most frequent class of these rows [8] and return the predicted class.

Applying both models will give a chance of choosing the best fit model for the source data [9].

2.6 Approach:

I. Load dataset from the source – Adult income dataset.

II. Data preparation [7] and visualization.

III. Feature encoding and normalization.

IV. Splitting data [3] into training data and testing data.

V. Naive predictor performance checking. [2]

VI. Dividing batches and training supervised learning models using two supervised algorithms/classifiers as: [13]

   i. Logistic Regression

   ii. K-Nearest Neighbors

VII. Predicting the testing data using the above classifiers.

VIII. Drawing the confusion matrix and measuring [12] the accuracy of models.

3. Results and Discussion

3.1 Proposed System:

Supervised algorithms used in learning: [4]

i) Logistic regression (LR)

ii) K-Nearest Neighbours (KNN)

3.2 Python Requirements:

i) Python libraries and modules required [6]

ii) Python libraries to visualize the data

3.3 List of features encoded:

i) One-hot encoding [5]

3.4 Data splitting: [7]

i) Train data

ii) Test data

3.5 Matrix model used:

i) Sklearn Confusion Matrix [9]

3.6 Dataset: Adult income dataset

i) The adult income dataset is a [8] multivariate dataset.

ii) Total number of records: 45000 (After removing null valued records)

iii)Total attributes = 14

3.7 Data splitting: [11]

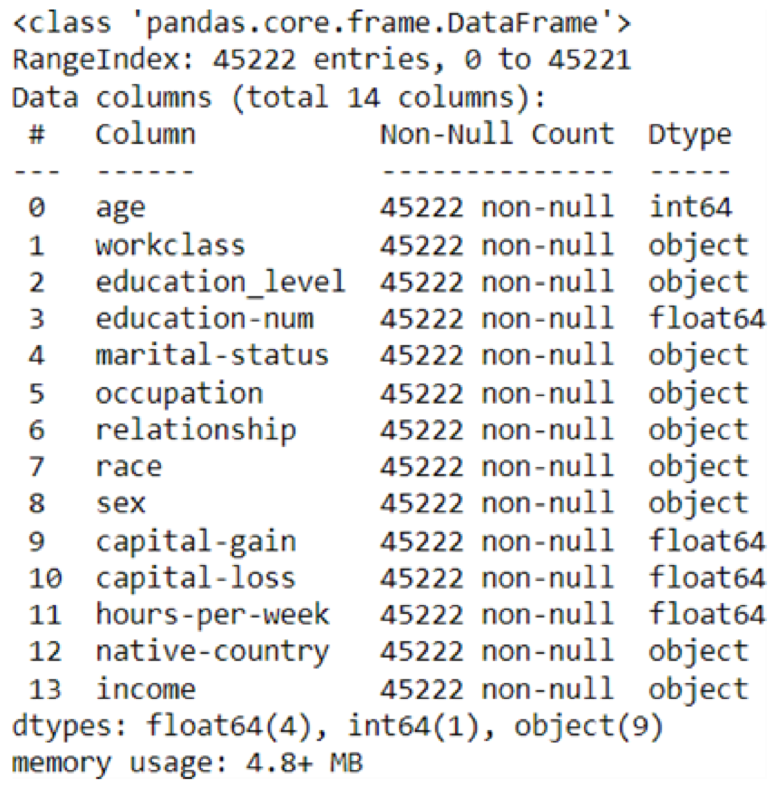
i) Training data – 80%

ii) Testing data – 20%

3.8 Data splitting:

i) Feature variables = 13

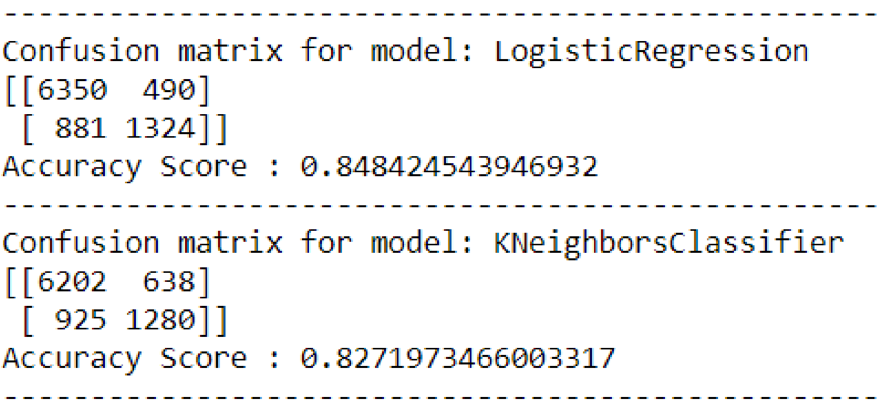
ii) Target variable = 1



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45222 entries, 0 to 45221
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   age              45222 non-null  int64
 1   workclass        45222 non-null  object
 2   education_level  45222 non-null  object
 3   education-num    45222 non-null  float64
 4   marital-status   45222 non-null  object
 5   occupation       45222 non-null  object
 6   relationship     45222 non-null  object
 7   race             45222 non-null  object
 8   sex              45222 non-null  object
 9   capital-gain     45222 non-null  float64
 10  capital-loss     45222 non-null  float64
 11  hours-per-week   45222 non-null  float64
 12  native-country   45222 non-null  object
 13  income           45222 non-null  object
dtypes: float64(4), int64(1), object(9)
memory usage: 4.8+ MB
```

Figure 5 Displaying the Information about Features

Based on the source data and machine learning algorithms, successfully training and testing model for finding donors using supervised learning has done with resulting accuracy scores for both the models. And “Logistic regression” has the best accuracy score when compared to the “K-Nearest Neighbor” algorithm accordingly to the data splitting.



```
--------------------------------------------------
Confusion matrix for model: LogisticRegression
[[6350  490]
 [ 881 1324]]
Accuracy Score : 0.848424543946932
--------------------------------------------------
Confusion matrix for model: KNeighborsClassifier
[[6202  638]
 [ 925 1280]]
Accuracy Score : 0.8271973466003317
--------------------------------------------------
```

Figure 6 Visualizing the confusion matrix for each classifier.

4. Conclusions

Both results are the same in terms of the five features which is not expected since the first features before data normalization.

The above results confirm that both our prediction using data exploration and expecting feature importance are confirming. This could be due to data correction. The F-score accuracy using the reduced data is 0.8271 which is less than the full data 0.8484. However, the difference is marginal. With regards to the training time, if the time is a constrain, we should eliminate or replace SVM with something else.

Acknowledgements

Provide acknowledgements accordingly. List here those individuals or institutions who gaves help, assistance during the research

References

[1] Jordan, Michael I., and David E. Rumelhart. "Forward models: Supervised learning with a distal teacher." Cognitive science 16.3 (1992): 307-354.

[2] Hosmer, David W., et al. "A comparison of goodness ‐ of ‐ fit tests for the logistic regression model." Statistics in medicine 16.9 (1997): 965-980.

[3] Zadrozny, Bianca, and Charles Elkan. "Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers." Icml. Vol. 1. (2001).

[4] Guo, Gongde, et al. "An kNN model-based approach and its application in text categorization." International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Berlin, Heidelberg, (2004).

[5] Chawla, Nitesh V., and Grigoris Karakoulas. "Learning from labeled and unlabeled data: An empirical study across techniques and domains." Journal of Artificial Intelligence Research 23 (2005): 331-366.

[6] VanRossum, Guido, and Fred L. Drake. The python language reference. Amsterdam, Netherlands: Python Software Foundation, (2010).

[7] Papandreou, George, et al. "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation." Proceedings of the IEEE international conference on computer vision. (2015).

[8] Fabris, Fabio, João Pedro De Magalhães, and Alex A. Freitas. "A review of supervised machine learning applied to ageing research." Biogerontology 18.2 (2017): 171-188.

[9] Bonica, Adam. "Inferring roll ‐ call scores from campaign contributions using supervised machine learning." American Journal of Political Science 62.4 (2018): 830-848.

[10] Yan, Ke, et al. "Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018).

[11] Mittal, Pooja, and V. K. Srivastava. "A Review of Supervised Machine Learning Algorithms to Classify Donors for Charity" International Journal of Advanced Research in Computer Science 12.1 (2021).

[12] Velde, Venkateshwarlu, et al. "Enterprise based data deployment inference methods in cloud infrastructure." Materials Today: Proceedings (2021).

[13] Vurukonda, Naresh, et al. "Simplified ciphertext-policy attribute- based encryption scheme with attribute level collusion resistance for cloud storage." Journal of Physics: Conference Series. Vol. 2089. No. 1. IOP Publishing, 2021.