# The Study Of Three Machine Learning Models On Two Different Dataset

Oladipo Oladotun
20230532
x20230532@student.ncirl.ie

***Abstract*— Data is referred to raw facts that are collected and processed to generate useful information for a defined purpose. Most companies around the world are able to collect information to help their business grow or sustain its business model. The concept of Dataset is referred to a collection of instances that can be processed in a machine learning algorithm. These instances have similar attributes that allow for easy processing through Machine learning. This research will be focusing on two individual data set from different industries, The first data set is about YouTube's top trending videos in America while the second dataset is about Credit card applications for individuals in the financial industry.**
**YouTube is referred to be a video-sharing platform that allows individuals to share content with people globally. It is a world-famous video-sharing platform that allows users to comment, share, save and upload content. The purpose of this research is to use a model to show a linear relationship between views, likes, comment count, and dislikes to show correlation.**
**Credit card application focuses on payments made in different environment. It is a form of payment that can made in a shop, or over on the internet. Most individuals look to make use of credit cards to pay for goods. The credit card provider provides funds to a individuals for a particular period then afterward repayment would be made at the end of a month. Most credit card companies collect individuals' information to examine the ones that are likely to default in the future. The models that will be used are Decision tree and Logistic regression to categorize male and female looking to apply for a credit card.**

***Keywords—Decision Tree, Logistic Regression, Linear Model and Machine learning.***

## I. INTRODUCTION

YouTube is a platform that can be used by individuals, or business companies for the purpose of marketing or having a form of interaction with the public to create awareness. Most videos on the platform are able to trend globally because of number of viewership. The research question for this report would be based on what are the factors that have an influential impact on a video trending in a country? The main objective of this analysis is to focus on showing a corresponding relationship with factors such as Likes, Dislikes and Comment count having a positive impact on a videos views using a Linear model.

Credit Card application with the use of machine learning has been made it easy for people to ascertain. Before a credit card can be issued, a potential customer must submit their details to a financial company. The Credit card data set is analyzed based on understanding these information that is submitted by people looking to apply for credit card. The research question would be what are the number of male and female that applied for credit card ? The research objective is to categorizing Male and Female that looked to apply for credit card using classification models such as Decision tree and Logistic models. It focuses primarily on generating information about credit card applicants.

## II. RELATED WORK

A lot of researches have been conducted in relation to YouTube and Credit Card application. In this part of this report, a critical review would be examined to study past related works. The first section will focus on YouTube, while the second section will focus on Credit card.

### *A. YouTube*

Based on this report an evaluation conducted demonstrates that human sentiment can influence popularity of a video online[1].The use of sentiment analysis allows to evaluate viewer's feedback on a video post . The algorithms that was used is the Support vector machine algorithm that aids in ascertaining sentiment accuracy. The study focused on using Natural language processing models after using K nearest neighbors, Navies bayes failed. It is important to point out the human emotions can be misleading.
Twitter is a platform that is popular globally and allows people to express their opinion in form of words that is posted online. It is possible to monitor people' post and determine a trending topic through it [2] . This research work focuses on the twitter platform that looks into top trending topics by investigating user's comment. It is possible to classify people's sentiment using a machine learning program such as Support vector machine and Naïve Baye but unfortunately analyzing people's post can be misleading and bias. This implies such model can yield inaccurate results with an high degree.
The number of social network platforms that allows individuals to post videos online continues to increase, therefore implies that there are numerous ways videos go viral. This report focuses on understanding what makes a video go viral [3]. This analysis is useful for content creator and people looking to advertise their business on this type of platforms. In this research the use Support vector machine and Gaussian Radial basic function classification model was adopted analyze this prediction. Based on this research it was discovered to treat this as classification problem and best way to make an accurate prediction is to focus on predictors based on features of video characteristic to get relevant information.
The importance of video content in the global market, there is growing concern to understand the popularity characteristic of video content and find its implications to service design, advertisement planning and network planning is of great importance [4] .This research focused

on an online leading video platform in China called Youku to provide an understanding reason for present popularity and future popularity in a video. Based on the research it was discovered the linear and multi regression is useful tool to show a relationship, it was discovered that the popularity of video is likely to evolve and there exist a linear correlation between early view count and future view count of a video and therefore has an impact on the popularity pattern.

The use of Sentiment Analysis has played a major factor in social networking sites such as You-tube, Twitter and Facebook [5] . People are able to express their emotions by posting them on these platforms. This research work focus on the sentimental analysis used understand people's emotion to ascertain people reaction on the US presidential election. The use of this analysis model was conducted on more than 200 YouTube comment in the United State of America .

The importance of big data in organizations has made it difficult to manage, and adopt to business strategies and marketing plan [6]. Big data comes in different format such as structure and unstructured. This research focused primarily on data analysis on YouTube data, Five millions videos was extracted for the analysis. Based on the research, analysis was carried to analyze the likes, dislike, comment count with the use of new technologies such Hadoop, Spark and HBase helps in process of collecting large data.

Machine learning can be used to classify a set of data into a category, for this research work It is based on detecting whether a comment posted is spam or a normal message [7].The main focus of the research is to determine a spam message with the aid of Random forest for this classification problem on two different dataset with the K fold validation to test its performance. The algorithm can be applied to other classification problems such as attack detection, smart predictions.

The comment section in YouTube has allowed for user to be able to express their opinion about a video post, some of these might be negative while some can be positive [8].It is essential to be able to segregate the difference between the positive comment and negative comments in form of spam detection and comment classification. The implication of these negative comments on video post can act as a distraction to other users that interested and can easily send a wrong message. Malicious user can decide to fill comment section with negative remarks. The use of classification models such as Logistic regression, Decision tree and Random forest can be adopted to solve the issue.

In relation to previous related work defined above, the importance of human emotion is essential to determine how a video goes viral. The human emotions should not be underestimated. This research work will focus on how humans are able to express their feelings through Likes, Dislikes and Comments. It will show a relationship with these factors using a Multiple Linear Regression.

### *B. Credit Card*

The use of Support vector machine with the use multiagent ensemble learning can be used solve problem of credit risk that faces lending firms in the financial industry [9] . This model involves more than one different support vector machine paradigms to analyze a dataset. This research made use of credit dataset to test it effectiveness. The obtained result reveal that the proposed Support Vector Machine based multiagent ensemble learning mode can provide a promising solution to credit risk evaluation problem and implies that the proposed model based multiagent ensemble learning technique has great potential in its application to other classification problems.

Machine learning in the financial industry has developed to solve several problems, but the most common would be detecting fraudulent transaction. A fraudulent transaction can be referred to a situation where an owner of card loses sensitive information pertaining to their account to malicious attacker [10] . For this research the use machine learning model was administered to determine between a fraudulent and normal transaction conducted by users. The machine models used are Logistic regression, Naïve bayes, and Random forest tree to examine a credit card transaction. It was discovered that Random forest algorithm has the highest level of accuracy compared to other models.

There is currently a strong demand for credit cards amongst people in the market, but unfortunately people submit false information to get them. This implies that lending firms face a financial risk of providing credit card to people that are might not payback [11]. Based on this research the introduction of Fico credit score to asses an individual seeking a credit card was introduced in the first place. This model was able to provide an estimate that a person is likely to default payment but unfortunately it was based on a subjective judgment. The model never took into account factors that can be considered. Subsequently machine learning models was therefore introduced to evaluate people in an effective way. The use of Decision tree, Random forest that are able to provide an high level of accuracy better than Fico credit score.

Credit card application data is analyzed using different techniques from the of collection of exploring and identifying individuals that are likely default payment at a particular period of time [12]. The research focuses on improve method to analyze applicants using credit scoring models. The credit scoring models allow to for the grouping of applicants between two classes namely good credit or bad credit. This research made use of logistic regression for the classification problem.

This research focus on comparing two models in the aid of credit scoring for a credit industry, namely Random tree and Logistic regression[13]. Despite numerous classification models that can be adopted, Logistic regression is considered best. It is considered the best simply because of its stability and robustness. Other model are also considered good for credit scoring but it is difficult to explain and interpret.

This research is based on analyzing repayment failure of individuals in a Peer to Peer lending system [14]. A risk evaluation was carried out to analyze the reason why an individual will default payment with the use four several machine models such as Random forest, Extreme gradient boosting tree, Gradient boosting model, and Neural network. The data set used to examined this was derived Renrendai.com in China. It was discovered that individuals who were married, had mobile phones, had Income and Job stronger influence on repaying loan.

Credit card fraud is ever growing threat in the world because most prefer to use them to buy goods online [15]. This attacks have made financial companies to become vulnerable to this attack. The exposure of credit card allows malicious users to take advantage. This research focuses on how financial looking to improve security measure by adapting to an effective machine learning algorithm and also managing the use of big data. It uses previous financial transaction conduct an analyses. Models such as Decision tree, and Logistic regression.

Based on the related work defined above, it is obvious that financial companies do face issues regarding offering credit to people. Despite the introduction of machine learning algorithm, there is a still a gap that needs to filled to protect these firm. This research work would focus on analysing a credit card application dataset to categorise Male and Female that applied for Loan with use of Decision tree and Logistic regression

## III. Data Mining Methodology

The data methodology technique adopted for this project is the Cross Industry Standard Process for Data Mining. In short it is also called CRISP-DM and can defined as a structure that is adopted for the purpose of constructing a data mining project. This technique comprises of six key elements namely Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment.

### *A. Business Understanding*

It is important to understand that for this research two dataset would be used. The first dataset is based on the top trending videos in America in relation to the You-tube platform. The main importance for this dataset to understand factors that contribute making a video trend in a particular period of time. It can used to establish an improved mechanism on how to attract more users to their platforms. There are components that play vital role in establishing how a video is likely to trend based on number of views, comment count and likes and dislikes. This information can also be used by users of the platform such as youtubers to give assessment on how well a video is performing in the real world.

The Second dataset describes how a credit card application can either be approved or denied by a financial company. Before a credit card can be issued to a person, essential information must be submitted to determine a person's eligibility. Information such as name of applicant, gender, previous jobs, present job and many more. The dataset for this report can be used gives an idea on how likely a person will default future payment. Most companies are able to gather as much information for this purpose and rate their previous performance with the company based on their credit score ratings. It is essential that people's information are submitted for evaluation for the sake of protecting the interest of lending firms.

### *B. Data Understanding*

The YouTube dataset comprises of both 40881 rows and 16 columns that will be analyzed based on showing that there exist a relationship between the number of views for a video with other factors such dislikes, likes and comment count. The 16 columns are listed Video.Id, Trending date, Title, Channel_title, Category_Id, Publish time, Views, Likes, Dislikes, Comment Count, Thumbnail Link, Comments _disabled, Ratings disabled, Video_error_or removed and Description. The study for this research is find a relationship between how influential these factors are to making a video trend. It will focus on three independent variables which are Likes, Dislikes and Comment count to find a corresponding relationship with Views(Dependent variable). The other columns will be ignored for the purpose this research.

| Feature Name | Categories |
|---|---|
| Video. Id | Factors with 24427 levels |
| Trending Date | Factor with 205 levels |
| Likes | Integers |
| Dislikes | Integers |
| Comment Count | Integers |
| Comments Disabled | Factors with Two levels |
| Views | Integers |
| Ratings Disabled | Factors with Two levels |

The Credit card dataset comprises of both 537667 and 19 columns that will be analyzed to determine a possible outcome. The 19 columns comprises of Id.

| Feature Name | Explanation | Categories |
|---|---|---|
| ID | Client Number | Integer |
| Code Gender | Gender | Factor (Male and Female) |
| Flag_Own_Car | Is there a car | Factor(Yes/No) |
| Flag_Own_Realty | Is there a property | Factor(Yes/No) |
| Cnt_Children | Number of children | Factor with three levels |
| Amt_Income_Total | Annual Income | Numerical |
| Name_Income_Type | Income category | Numerical |
| Name_Education_Type | Education level | Factor with Five levels |
| Name_Family_Status | Marital status | Factor with Five levels |
| Name_Housing_Type | Way of living | Factor with Six levels |
| Days Birth | Birthday | Integers |
| Days Employed | Start date of employment | Integers |
| Flag Mobil | Is there a mobile phone | Integers |
| Occupation Type | Occupation | Factors with 9 levels |

The research in regards to this dataset will analyze the importance of columns such Occupation and Annual Income in the process of classifying male and female applicants looking to apply for a credit card.

### *C. Data Preparation*

It is important to understand how the YouTube dataset will be analyzed for evaluation . It is used for the purpose of showing a correlation between views of a video and other factors such as Comment count, Likes and Dislikes that makes a video to trend in America. It is used to solve a regression problem that provides an output that is continuous because it is numerical in nature . It requires one dependent variable and three independent variables. The first step was to remove the null value and unnecessary

column that can affect the prediction. Afterwards the dataset was cleaned with the use of tidy verse that can be installed on R studio and outliers were removed. In the process of cleaning the data, 1296 missing values was detected

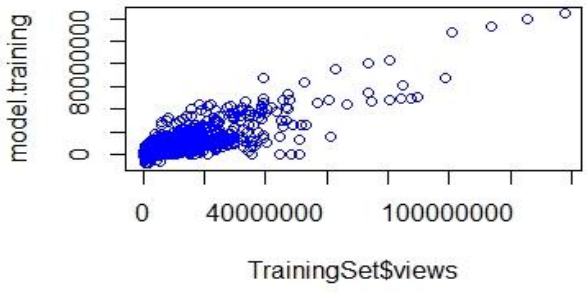
The Credit card dataset was analyzed to show how Occupation and Annual income can be used to classify male and female that apply for a credit card. It involves a classification output that categorized its output between male and female. It required one dependent variable and two independent variable. It uses the same process of removing irrelevant columns and null values as that of the YouTube dataset. it was discovered that there were Zero missing values but outliers were identified.

### *D. Modeling*

This research makes use of three machine learning technique which are Linear Regression, Logistic Regression and Decision tree that will be used to analyzed both datasets.

*1) Linear Regression/ Multi linear Regression*

This refers to machine learning model that can be used to establish a relationship between an independent and dependent variable. It can be used to show linear relationship between two possible variables by drawing a line across a graph. It can be used to show relationship between an input variable and an output variable. In some cases there can be more than one input variable, therefore it is referred to a multiple linear regression. For this research, it was to establish a relationship between views which is the dependent variable and three other independent variables in the names of Likes, Dislikes and Comment count. The main purpose define the purpose of these independent variables. Below is the formula Y=a+bx, where x represents independent variable, y represents dependent variable, b represents the slope and a represents the Intercept. The diagram below shows a relative a good correlation on training model using both independent and dependent variables

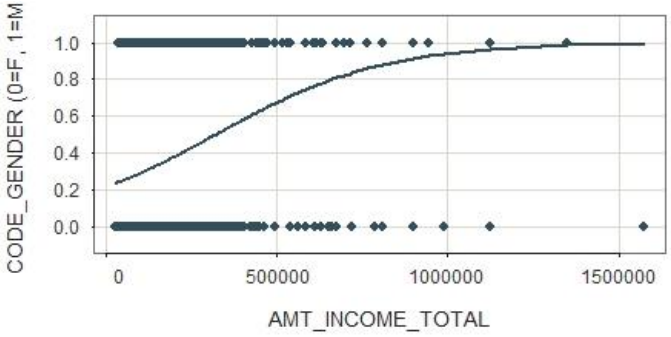



*2) Logistic Regression*

Logistic regression is a machine learning technique that is adopted to provide an output that is binary in nature. This implies that the dependent variable is always categorized in either 0 or 1. For this research work, the purpose of the logistic regression is to determine the number of Male/ Female that applied for a credit card using their Amount of income submitted. The dependent variable will be gender that is binary between Male and female, where Female is 0 and Male is 1 and the independent variable is the amount_income_total. Below is a formula that describes a logistic regression.

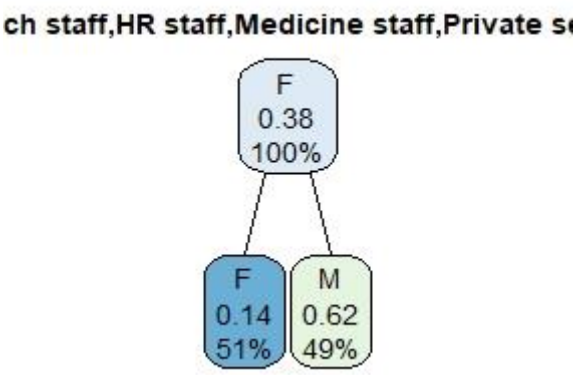
$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$

Y for the formular above represents the a predicted outcome, the b0 is the intercept on the graph, and b1 is considered to be the coefficient of an inputted variable. Below is a diagram that establishes a relationship between Amount of Income with Gender.




*3) Decision Tree*

This is a type of machine learning that can be used to solve categorical and regression problems. It is also considered to be a supervised type of learning. For this research it produces a categorical output that classifies male and female into a category. The top most part of a decision tree is considered to be a root nodes that that splits the data into two part, for this research the decision tree will be used to classify male and female credit card applicants based on their occupation . Based on the list of occupations that was registered, there are about 51% of female and 49% male that applied for a credit card. The list of occupation comprises of Managers, Laborers, Core staff, Drivers, and Medical staff.




### *E. Evaluation*

There are types of evaluation technique that is adopted for this research work. In regards to the YouTube dataset, the model that was used is a Linear regression model, and it was evaluated based on its performance. The technique used are, Multiple R squared and Correlation co-efficient. This was derived after the entire dataset was partitioned into a train and test set that allows easy evaluation. A training dataset allows for a dataset to be evaluated more than once for the

purpose of adjusting to its intended purpose. The test dataset allows that when a model is designed, it is important for such model to be able to confer an accurate prediction. The essence of evaluating a dataset is for the purpose of making predictions based on real world circumstances. This dataset was split between 80% and 20% for both the train and test set.

The Credit card dataset was split into two set for the Decision tree model that was used for its evaluation. In this case the train and test set is divided into 70% and 30% respectively. The performance of this model is evaluated using Accuracy, Kappa , Sensitivity and Specificity for the prediction. Another Machine learning model was used for this dataset is the Logistic Regression, that uses Accuracy and Precision for its evaluation.

### *F. Deployment*

This is the final stage of the CRISP Methodology that uses the outcome of the machine learning process. It is important to highlight the importance of every models that has been used for analysis in this research work.

In relation to the You-tube dataset, the essence of using Linear regression is to show that there exists a strong relationship between Views, Likes, Dislike and Comment count. All these factors play an important role for a video to trend in a country. This analysis can be used for the YouTube platform for the purpose of recommending videos for people to watch based on high number of views. It is very much certain that a video with a high number of views, is likely to attract other people's attention.

While the Credit card dataset uses both Decision tree and Logistic regression for the purpose of determining the number of Male and Female looking to apply for a credit card. This analysis allows financial companies to have an idea about the ratio of people in Male and Female looking to apply for a credit card . It can also be used for informative purpose for companies.

## IV. Evaluation methodology

The YouTube dataset is analyzed using Linear model for machine learning. The main purpose of this technique is to establish a relationship between Independent variables and Dependent variable. In this case out of several other columns, this research focused primarily on Views, Likes, Dislikes, and Comment Count columns for the evaluation. The linear model is used to show that there exist a relationship with Views(Dependent Variables and other columns used . The dataset is partitioned into two sets namely Training data and Test Data.

Based on evaluating the performance of linear model Pearson's correlation coefficient is used to measure that there exist a strong relationship between independent variable and dependent variable. According to the result derived running a coefficient correlation for the training set and test set shows a percentages 0.82% and 0.86% respectively which demonstrates that there exist a positive relation with Views(dependent variable) and Likes, Dislikes, and Comment count(Independent variable) . In other words there is a corresponding increase in both independent and dependent variable.

The second evaluation measurement is the Multiple R squared, which can be used to determine the proportion of variance of the dependent variable that can explained through the use of independent variable. It can be used to show the percentage how well independent and dependent variable fit together. For both the training set and test set shows 0.75% and 0.69% . This percentage reveals that about 75% and 69% of the data is able to fit the regression model successful. The formular is denoted below with the use

R-squared = Regression /SS total

The Credit Card dataset is analyzed using a Decision tree and Logistic Regression model for the credit card applications record. The Decision tree model makes use of measurement techniques to shows if the model is successful for prediction. The main purpose of this model is to categorize Male and Female by using Occupation column to make the prediction. This prediction will be used to determine the percentage of male and female that are looking to apply for a credit card. The performance technique used are Accuracy, Kappa, Sensitivity and Specificity. A confusion matrix was used to highlight the measures below

```
               Accuracy : 0.741
                 95% CI : (0.7388, 0.7431)
    No Information Rate : 0.6209
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.4802

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.6997
            Specificity : 0.8087
         Pos Pred Value : 0.8569
         Neg Pred Value : 0.6218
             Prevalence : 0.6209
         Detection Rate : 0.4344
   Detection Prevalence : 0.5069
      Balanced Accuracy : 0.7542

       'Positive' Class : F
```

The Accuracy demonstrates a 0.741 or 74% which when considered that there is just one predictor that is used for the Decision tree. To increase the level accuracy, more independent variable should be added to ascertain an higher accuracy level. It is important not to rely on the Accuracy percentage because it does not give a full description on how accurate the prediction is made.

Kappa is a type of accuracy measurement that is adopted to provide more accurate result based on random chance of the dataset. The outcome is 0.4802 or 48% , this is very low, unlike when compared to the accuracy level above. It gives a better result because it focuses on random selection of the dataset for analysis.

Sensitivity is a metric that can be used to evaluate whether a model is able to predict positives of a category. It is used for binary classification model that have two possible outcome. For this research, there are two possible outcomes while are male and female. According to the sensitivity level it was able to categorize Male and Female prediction correctly with 0.70 or 70% measurement . Specificity is a type of evaluation metrics that predicts true negatives of a

particular category. The specificity level is high which 0.8087 or 80% measurement that implies it was also able to predictions wrongly.

In relation to Logistic regression that is used for analysis for this dataset, is used to classify male and female looking to apply for a credit card based on their Annual amount of income. Based on this prediction, there is only one predictor which is Annual amount of income(Independent variable) and Gender column is regarded as an Dependent variable.
The use of precision will be used to evaluate how well the prediction is made. Precision is referred to the ratio that compares a true positive to other positive in a prediction. It helps to identify how correctly a model is able to fit, for this prediction, the precision is 52.22% which is low considering there is just one predictor.
The accuracy level for this prediction is 62.4% which is low as well considering there is just one predictor. The best way to ascertain high Accuracy and Precision would be to add more independent variable to help improve its accuracy.

```
                  Baseline        Predicted
-------------------------------------------------
                  Total  %Tot       0      1  %Correct
-------------------------------------------------
              0   333832  62.1   311180  22652     93.2
CODE_GENDER   1   203835  37.9   179076  24759     12.1
-------------------------------------------------
           Total 537667                          62.5

Accuracy: 62.48
Recall: 12.15
Precision: 52.22
```

## V. Conclusions and Future Work

Machine learning models has played various roles in different industries around the world. The Business world continues to change because of mass data that is circulating, and the need to find an effective way to manage them has become a major concern. The essence of this report is analyze how Machine learning models are able to operate. This reports focused on both classification and regression model to analyze two unique dataset. The first dataset highlight key elements that make video trend on YouTube . Based on previous research work conducted on this type of dataset, it was highlighted that human emotions play an important role and the use of sentiment analysis to categorize human emotions can provide a good understanding. The comment, like, dislike and view section of a video posted is very influential. The use of Linear model was able justify that there exist good relationship for a video to trend with number viewership, likes, dislikes and comment count. Based on the model more than one independent variable was just used to justify that there was great correlation between this factors. This sort of result is helpful for people that enjoy using social media platforms to connect with the world. It seems everyone around the world looking to the internet. The number of social media application around the world has made it easy for people to have freedom to post their story in form of pictures, videos, comments
In regards to the other dataset that was used for this research, it focused primarily on credit card application. People around the world are looking to apply for credit card to satisfy their buying needs. There is risk for financial companies that look to offer credit card to customers. There are certain customers that are likely to default payment. Based on related works concerning this dataset, it was established that firms make use of machine learning to prevent individuals that will likely default payment in the future. For this research, the main concept was to categorize the number male and female that applied for a credit card using a single independent variable. The decision tree model used got low Kappa of 48% and Accuracy of 74%
While the Logistic model also got low accuracy of 52.22% . These models were not able to justify the research objective. The best way to get a better result to use more than one independent variable for this classification problem.
Machine learning will continue to be used by companies in the future because it has introduced an effective way to manage a business operation.

## References


[1] G. M. Prabha, Madhumitha and R. P. Ramy, “Predicting the Popularity of Trending Videos in Youtube Using Sentimental Analysis,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 2278-3075, 2019.

[2] T. Rathod and M. Barot, “Trend Analysis on Twitter for Predicting Public Opinion,” *International Journal of Computer Applications* , vol. 180, no. 26, pp. 0975-8887, 2018.

[3] B. M. Dalmoro and S. R. Musse, “Predicting Popularity of Facebook Videos Through Visual Features Using Support Vector Machine Classifier,” *IEEE Transactions on Multimedia,,* p. 2561–2570, 2017.

[4] C. LI, J. LIU and S. OUYANG, “Characterizing and Predicting the Popularity of Online Videos,” *IEEE Access,,* vol. 4, p. 3026–3033, 2016.

[5] S. Singh and G. Sikka, “YouTube Sentiment Analysis on US Elections 2020,” in *2021 Second International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 2021.

[6] S. Amudha, V. R. Niveditha and D. P. R. Kumar,

"Youtube Trending Video Metadata Analysis Using Machine Learning," *International Journal of Advanced Science and Technology,* vol. 29, no. 2005-4238, pp. 3028-3037, 2020.

[7] T. Ghodke and V. M. Khadse, "Effective Text Comment Classification Using Novel ML Algorithm—Modified Lazy Random Forest," in *2nd International Conference on Data Science, Machine Learning and Applications, ICDSMLA 2020*, Pune, 2022.

[8] K. Kavitha, A. Shetty and B. Abreo, "Analysis and Classification of User Comments on YouTube Videos," *Procedia Computer Science,* vol. 177, pp. Pages 593-598, 2020.

[9] L. Yu, W. Yue, S. Wang and K. Lai, "Support vector machine based multiagent ensemble learning," *Expert Systems with Applications,* vol. 37, no. 2, p. 1351–1360, 2010.

[10] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," in *International Symposium on INFOTEH-JAHORINA (INFOTEH)*, 2019.

[11] Y. Yu, "The Application of Machine Learning Algorithms in Credit Card Default Prediction," in *Computing and Data Science (CDS), International Conference on*, Stanford, 2020.

[12] Ms.D.Jayanthi, "CREDIT APPROVAL DATA ANALYSIS USING CLASSIFICATION AND REGRESSION MODELS," *International Journal of Research and Analytical Reviews* , vol. 5, no. 3, pp. 2349-5138, 2018.

[13] E. Dumitrescu, S. Hué and C. Hurlin, "Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects," *European Journal of Operational Research,* vol. 297, no. 3, pp. 1178-1192, 2022.

[14] X. J.a, L. Z.a and X. Y, "Loan default prediction of Chinese P2P market: a machine learning methodology," *Scientific Reports,* vol. 11, no. 1, 2021.

[15] A. S, N. Sethumadhavan and H. N. AG, "Credit Card Fraud Detection using Apache Spark Analysis," in *International Conference on Trends in Electronics and Informatics (ICEI)*, 2021.

[16] Kaggle Inc, 2016. [Online]. Available: https://www.kaggle.com/rikdifos/credit-card-approval-prediction. [Accessed 26 December 2021].

[17] J. Davidson, B. Liebald and J. Liu, "The YouTube Video Recommendation System," in *ACM Conference On Recommender Systems*, Barcelona, 2010.

[18] W. Hoiles, A. Aprem and V. Krishnamurthy, "Engagement and Popularity Dynamics of YouTube Videos and Sensitivity to Meta-Data," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* vol. 29, no. 7, pp. 1426-1437, 2017.

[19] S. Chelaru, C. Orellana-Rodriguez and I. S. Altingovde, "How useful is social feedback for learning to rank YouTube videos?," *Web Information Systems Engineering - WISE 2012*, p. 552–566, 2012.

[20] ChunhuiWen, JinhaiYang and LiuGan, "Big data driven Internet of Things for credit evaluation and early warning in finance," *Future Generation Computer Systems,* vol. 124, pp. 295-307, 2021.

[21] A. Guarino and D. Malandrino, "An automatic mechanism to provide privacy awareness and control over unwittingly dissemination of online private information," *Computer Networks,* vol. 202, 2022.

[22] Y. C.a, N. C.a and S. X.a, "The Master of Detecting Deception: Machine Learning," in *Lecture Notes in Networks and Systems*, 2022.

[23] H. D, D. S and S. D, "Classification Framework for Fraud Detection Using Hidden Markov Model," in *1st International Conference on Cyber Intelligence and Information Retrieval, CIIR 2021*, Bengaluru, 2022.