In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits.A bank's profit or a loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non- Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters: The Logistic regression model. The data is collected from the Kaggle for studying and prediction. Logistic Regression models have been performed and the different measures of performances are computed. The models are compared on the basis of the performance measures such as sensitivity and specificity. The final results have shown that the model produce different results. Model is marginally better because it includes variables (personal attributes of customer like age, purpose, credit history, credit amount, credit duration, etc.) other than checking account information (which shows wealth of a customer) that should be taken into account to calculate the probability of default on loan correctly. Therefore, by using a logistic regression approach, the right customers to be targeted for granting loan can be easily detected by evaluating their likelihood of default on loan. The model concludes that a bank should not only target the rich customers for granting loan but it should assess the other attributes of a customer as well which play a very important part in credit granting decisions and predicting the loan defaulters.

INTRODUCTION

This Problem is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to a particular person will be safe or not. We have implemented this loan prediction problem using Decision tree algorithm and data cleaning in Python as there are missing values in the dataset. We use map function for the missing values. The aim of this paper is to apply machine learning technique on dataset which has 1000 cases and 7 numerical and 6 categorical attributes. The creditability of a customer for sanctioning loan depend on several parameters, such as credit history, Installment etc.

**LITERATURE REVIEW**

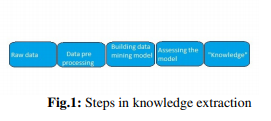
Logistic Regression is a popular and very useful algorithm of machine learning for classification problems. The advantage of logistic regression is that it is a predictive analysis. It is used for description of data and use to explain relationship between a single binary variable and single or multiple nominal, ordinal and ration level variables which are independent in nature. The model development for the prediction is taken in account using the sigmoid function in logistic regression as the outcome is targeted binary either 0 or 1 [11][15]. The dataset of bank customers has been divided into training and test data sets.. The train dataset contains approximately 600+ rows and 13+ columns whereas the test dataset contains 300+ rows and 12+ columns, the test dataset does not contain the target variable. Both the datasets are having missing values in their rows, and the mean, median or mode is used to fill the missing values but not removing the rows completely because the datasets are already small. Using the Feature Engineering techniques, the project is further proceeded and move towards the exploratory data analysis, where the dependent and independent variable is studied through statistics concepts such normal distribution, Probability density function etc. Study of the univariate, bivariate and multivariate analysis will give the view of the inside dependent and independent variable[13]14]. The model is focusing on to target those customers who are eligible for loans and therefore the logistic regression is enabled using the sigmoid function as it divided the probability into binary output. Therefore the Prediction model can be developed

**IN “PHILHYO JIN DO ,HO-JIN CHOI, “SENTIMENT ANALYSIS OF REAL-LIFE SITUATIONS USING LOCA- TION, PEOPLE AND TIME AS CONTEXTUAL FEATURES,” INTERNATIONAL CONFERENCE ON BIG DATA AND SMART COMPUTING (BIGCOMP), PP. 39–42. IEEE, 2015”** What emotion do l\re feel when we see a situation? Multimodal sentiment analysis has been used to answer this question, but most of the research considers only low-level perceptual information such as textual, acoustic, and visual features. Ho\\'ever, these features are not appropriate for the classification of situations as it is difficult to depict real-life complexities with low-level features. In this paper, we propose an emotion prediction framework which identifies polarity of emotion in situations using high-level contextual information, namely, location, people and time. Before predicting emotions, the framework structures data into 'situation' segments and labels each segment based on our carefully designed annotation guideline. Our approach is tested with various situations in TV sitcoms as a substitute for real-life situations. Experimental results indicate that contextual information is more effective than textual or acoustic features in determining emotions induced by situations Imagine a man proposing to his girlfriend in a restaurant while having a romantic dinner. When other people see this situation, what emotion would they normally have? In this paper, we try to find out the answer to the following question: ~~What type of emotion would people feel in this situation?" We investigate the polarity of emotion, positive, neutral, and negative, induced by real-life situations. In most cases, situations are often hard to describe and interpret without knowing the persons involved, and the context. This makes collecting good quality of data with detailed descriptions from a variety of situations almost impossible. To address this, we simplified our experimental settings to examine emotions induced by various situations found in TV situation comedies (sitcoms). We chose sitcoms because they allowed us to easily collect various real-life situations which happen between a small number of people in limited spaces. In this respect, our research is related to multimedia sentiment analysis. However, there are clear differences between emotions people feel when placed in real-life situations and emotions elicited by watching a video. We resolved this difference by using a carefully planned annotation guideline. We propose an emotion prediction framework which estimates the polarity of emotion induced by situations presented in sitcoms. The typical approach in multimedia sentiment analysis uses low-level information such as textual, acoustic, or visual features to predict induced emotion. However, these features are insufficient due to the gap between low-level perceptual features and high-level complexities in real-life situations. Thus, the novelty of our approach lies in exploiting high-level contextual features - location, people and time - in place of low-level textual and acoustic features. To the best of our knowledge, employing these contextual features to analyze multimedia sentiment is novel. To sum up, our main contribution is to define a new framework (Fig. 1) which has the following characteristics: • Introducing high-level contextual features in emotion prediction. We employ location, people and time information as contextual features to estimate the polarity of viewers' emotion. • Defining the concept of 'situation' and proposing a carefully designed annotation guideline. We use sitcoms as a substitute for real-life situations. We define situation segments in sitcoms, and introduce an annotation guideline. Our method is applicable in various research fields that deal with emotion. Humanoid robotics and human computer interaction are two examples, since people interact with machines more naturally and enjoyably when the machine's response accompanies emotion. In addition, our method can be used in many business domains, as when trying to predict possible responses when people are given a product or service in specific situations. For example, the method can provide information so that advertisements can be located in the right situation where particular emotions are likely to occur. Research on sentiment analysis has been active in recent years. Extensive research has been done on text corpora; Danisman et al. [1] classified 5 emotions from news headlines, Chesley et al. [2] determined the subjectivity and polarity of emotions by investigating verbs and adjectives in blog posts, and Zhe et al. [3] classified sentences from fairy tales by emotional affinity. Since sources other than text abound, particularly videos, new approaches which address other features, such as speech utterances [4] or facial expression [5] have been developed to identify emotions. Nowadays, multimodal sentiment analysis that combines various types of features has attracted a lot of attention. The movie domain [6,7] has mostly been exploited for that purpose. Besides movies, Morency et al. [8] presented a model to integrate visual, audio and textual features that can be used to recognize emotions in YouTube videos. Jiang et aI. [9] used visual, audio and attribute features to classify user-generated videos. Similar to our research, Xu [10] used vocabularies and audio features in sitcoms, however, sitcoms were not explored in depth compared to others domains. Moreover, our research is different from the previous multimedia sentiment analysis methods, since we are classifying 'situations' conveyed by videos, thus, video-specific features, such as camera angle, shot length, and incidental music between scenes, are not taken into account. There is a growing body of work addressing contextual information in sentiment analysis. The meaning of context varies according to the dataset. In text corpora, context typically means surrounding words [11] or sentences [12]. Using videos, Soleymani et al. [13] added genre as contextual information along with other audio, visual and textual features. Contextual information similar to our work has been explored in [14], which used social relationship information to detect the polarity of opinions in Twitter.

**IN “BING LIU, “SENTIMENT ANALYSIS AND OPINION MINING,” MORGAN & CLAYPOOL PUBLISHERS, MAY 2012”** With the rapid growth of social media, sentiment analysis, also called opinion mining, has become one of the most active research areas in natural language processing. Its application is also widespread, from business services to political campaigns. This article gives an introduction to this important area and presents some recent developments. Sentiment analysis or opinion mining is the computational study of people’s opinions, sentiments, appraisals, attitudes, and emotions toward entities and their aspects expressed in text. Motivation and Background Sentiment and opinion and their related concepts, such as evaluation, appraisal, attitude, affect, emotion, and mood, are about people’s subjective beliefs and feelings. They are key influencers of human behaviors. Whenever we need to make a decision, we often seek out others’ opinions. This is true for both individuals and organizations. The development of sentiment analysis coincides with the growth of social media (i.e., reviews, forum discussions, and blogs) on the Web. For the first time in human history, we now possess a huge volume of opinion data recorded in digital forms. These user-generated contents (UGC) are full of people’s opinions. Mining useful knowledge from these corpora gives rise to the task of sentiment analysis. Since the early 2000s, it has been one of the most active research areas in natural language processing (NLP) (Pang and Lee 2008; Liu 2012). The research and applications have also spread from computer science to management science and social sciences because of its importance to business and society as a whole. Sentiment analysis techniques have been widely applied in practice, from business services to political campaigns. In a nutshell, the task of sentiment analysis is to mine people’s opinions and emotions from text. The term opinion is used as a concept represented with a quadruple (s, g, h, t/ covering four components (Liu 2012): sentiment orientation s, sentiment target g opinion holder h, and time t. Sentiment is the underlying feeling, attitude, evaluation, or emotion associated with an opinion. Sentiment orientation can be positive, negative, or neutral. Sentiment target, also known as the opinion target, is an entity or an aspect of the entity that the sentiment has been expressed upon. Opinion holder is an individual or organization that holds an opinion. Time is when the opinion is expressed. We will discuss emotion specifically later. We use the following camera review as an example (an ID number is associated with each sentence for easy reference): Posted by John Smith Date: September 10, 2011 (1) I bought a Canon G12 camera six months ago. (2) I simply love it. (3) The picture quality is amazing. (4) The battery life is also long. (5) However, my wife thinks it is too heavy for her. Given the review, the task of sentiment analysis aims to extract the following opinion quadruples from sentences 2, 3, 4, and 5, respectively: (positive, Canon G12 camera, author, 2011/09/10) (positive, picture quality of Canon G12 camera author, 2011/09/10) (positive, battery life of Canon G12 camera, author, 2011/09/10 ) (negative, weight of Canon G12 camera, author’s wife, 2011/09/10 ) The opinion target can be an entity (Canon G12 camera) or an aspect of the entity (picture quality, battery life, and weight of the Canon G12 camera). An aspect can be explicit (e.g., battery life) or implicit (e.g., weight is indicated by heavy) (Hu and Liu 2004). In many applications, it is useful to decompose opinion target to entity and aspect for more fine-grained analysis. Then, the above quadruples become the following quintuples, where GENERAL represents the entity itself (Liu 2012): (positive, Canon G12 camera, GENERAL, author, 2011/09/10) (positive, Canon G12 camera picture quality, author, 2011/09/10) (positive, Canon G12 camera, battery life, author, 2011/09/10 ) (negative, Canon G12 camera, weight, author’s wife, 2011/09/10 ) An opinion from a single opinion holder is usually not actionable in an opinion mining application. The user often needs opinions from a large number of opinion holders, which leads to opinion summary. A summary of opinions is normally constructed based on positive and negative sentiments about opinion targets, which is called aspect-based opinion summary (or featurebased opinion summary) (Hu and Liu 2004). Figure 1 shows an opinion summary generated from product reviews of Apple iPad by Google products. Generally, opinion summary needs to be quantitative, which is reflected by the proportions or the numbers of positive and negative opinions for each sentiment target or aspect

**IN “LOGISTIC REGRESSION BASED LOAN APPROVAL PREDICTION”** As we know that now-a-days there is a rapid growth in banking sector, resulting lots of people are applying for bank loans. Finding out the applicant to whom the loan will be approved is a difficult process. In this paper, we proposed a model which predicts loan approval/rejection of an applicant using machine learning techniques. This can be done by training the model with the data of the previous records of the people applied for loan. Distribution of the loans is the main business part of almost every bank. The main portion of the bank’s asset is directly from the profit earned from the loans distributed by the banks. The prime goal in banking domain is to invest their assets in safe hands. Lending money to unsuitable loan applicants results in the credit risk. Today many banks approve loans after a long procedure of verification, yet there is no guarantee whether the picked candidate is the right candidate or not. Estimating the risk, which is involved in a loan application, is one of the most significant concerns of the banks in order to survive in the highly competitive market. Through our proposed model we can predict whether that specific customer is safe or not and the entire procedure of approval of features validation is automated by machine learning technique. Data mining algorithms are used to study the loan-approved data and exact patterns, which would help in predicting the reasonable defaulters, thereby helping the banks for making better choices in the future. Loan Prediction is extremely useful for employee of banks and for the applicant also. The main aim of this model is to provide a speedy, immediate and simple approach to pick the deserving applicants. The Loan Prediction System automatically calculates the weight of each feature involved in loan processing and on new test data same features are processed with respect to their associated weight. A period breaking point can be set for the applicant to check whether his/her loan can be approved or not. This model is solely for the managing authority of Bank/finance companies, entire procedure of prediction is done secretly that is, no stakeholders would able to alter the processing. Result against specific Loan Id can be send to different departments of banks in order to take an appropriate action on application In [1] the author acquaints a structure to successfully recognize the Probability of Default of a Loan applicant. The metrics got from the predictions reveal the high accuracy of the built model. In [2] an effective model was proposed for predicting the right customers who have applied for loan. Decision Tree is applied to foresee the traits significant for believability. The model proposed in [3] has been built using data from banks to predict the status of loans. This model uses three classification algorithms namely j48, bayes Net and naive Bayes. The model was implemented using Weka. In [4] a decision tree model was utilized as a classifier and for feature selection genetic algorithm is utilized. The model was tried utilizing Weka. In [5] two data mining models were created for credit scoring that helps in decision making of giving loans for the banks in Jordan. With the consideration of accuracy rate, the regression model is found to perform better than radial function model. The work in [6] analyses support vector machine based models for credit-scoring created using the different default definitions. The work inferred that the expansive definition models are better than the narrow definition models in their performance. In [7] financial data analysis was done by figuring out techniques like Decision Tree, Random forest, Boosting, Bayes classification, Bagging algorithm etc. Techniques like Support Vector Machine, Decision Tree, Logistic Regression, Neural Network, Perception model are combined in this model. The accuracy rate of each of these techniques is studied. The analysis results show the performance is extraordinary based on accuracy

**IN “LOAN PREDICTION BY USING MACHINE LEARNING MODELS”** With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this project we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this project is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections (i)Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing. In this paper we are predict the loan data by using some machine learning algorithms they are classification, logic regression, Decision Tree and gradient boosting This Problem is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to a particular person will be safe or not. We have implemented this loan prediction problem using Decision tree algorithm and data cleaning in Python as there are missing values in the dataset. We use map function for the missing values. The aim of this paper is to apply machine learning technique on dataset which has 1000 cases and 7 numerical and 6 categorical attributes. The creditability of a customer for sanctioning loan depend on several parameters, such as credit history, Installment etc Data mining is the process of analyzing data from different perspectives and extracting useful knowledge from it[3]. It is the core of knowledge discovery process. The various steps involved in extracting knowledge from raw data as depicted in figure-1. Different data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression etc. Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to classification technique. This approach frequently employs Decision tree based classification Algorithm. In classification, a training set is used to build the model as the classifier which can classify the data items into its appropriate classes. A test set is used to validate the model.



Data Mining in Banking Due to tremendous growth in data the banking industry deals with, analysis and transformation of the data into useful knowledge has become a task beyond human ability. Data mining techniques can be adopted in solving business problems by finding patterns, associations and correlations which are hidden in the business information stored in the data bases. By using data mining techniques to analyze patterns and trends, bank executives can predict, with increased accuracy, how customers will react to adjustments in interest rates, which customers are likely to accept new product offers, which customers will be at a higher risk for defaulting on a loan, and how to make customer relationships more profitable[5]. Globalization and the stiff competition had led the banks focus towards customer retention and fraud prevention. To help them for the same, data mining is used. By analyzing the past data, data mining can help banks to predict credible customers. Thus they can prevent frauds, they can also plan for launching different special offers to retain those customers who are credible. Certain areas that effectively utilize data mining in banking industry are marketing, risk management and customer relationship management. Marketing:It is one of the most widely used areas of data mining in the banking industry. The consumer behavior with reference to product, price and distribution channel can be analyzed by the marketing department. The reaction of the customers to the existing and new products can also be known. This information can be used by the banks to promote the products, improve quality of products and services, and gain competitive advantages. Bank analysts can also analyze the past trends, determine the present demands and forecast the customer behavior of various products and services, in order to grab more business opportunities Risk Management: It is widely used for managing risks in the banking industry. Bank executives need to know the credibility of customers they are dealing with. Offering new customers credit cards, extending existing customers’ lines of credit, and approving loans can be risky decisions for banks, if they do not know anything about their customers. Banks provide loans to their customers by verifying the various details relating to the loan, such as amount of loan, lending rate, repayment period etc. Even though, banks are cautious while providing loan, there are chances of loan repaying defaults by customers. Data mining technique helps to distinguish borrowers who repay loans promptly from those who default. Customer Relationship Management: Data mining can be useful in all the three phases of a customer relationship cycle such as customer acquisition, increasing value of the customer and customer retention. Customer acquisition and retention are very important concerns of any industry, especially the banking industry. Banks have to cater the needs of the customers by providing the services they prefer. This will ultimately lead to customer loyalty and customer retention. Data mining techniques help to analyze the customers who are loyal from those who shift to other banks for better services. If the customer is shifting from his bank to another, reasons for such shifting and the last transaction performed before shifting can be known, and this will help the banks to perform better and retain their customers.

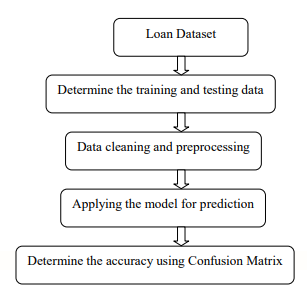
**IN “EXPLORING THE MACHINE LEARNING ALGORITHM FOR PREDICTION THE LOAN SANCTIONING PROCESS”** Extending credits to corporates and individuals for the smooth functioning of growing economies like India is inevitable. As increasing number of customers apply for loans in the banks and non- banking financial companies (NBFC), it is really challenging for banks and NBFCs with limited capital to device a standard resolution and safe procedure to lend money to its borrowers for their financial needs. In addition, in recent times NBFC inventories have suffered a significant downfall in terms of the stock price. It has contributed to a contagion that has also spread to other financial stocks, adversely affecting the benchmark in recent times. In this paper, an attempt is made to condense the risk involved in selecting the suitable person who could repay the loan on time thereby keeping the bank’s non-performing assets (NPA) on the hold. This is achieved by feeding the past records of the customer who acquired loans from the bank into a trained machine learning model which could yield an accurate result. The prime focus of the paper is to determine whether or not it will be safe to allocate the loan to a particular person. This paper has the following sections (i) Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance than other models in terms of loan forecasting. Finance raising and lending for real estate, consumer, mortgage and companies‘ loans is the central part of almost every bank‘s business model. Lending money to inappropriate customers forms the major source of credit risk. The major share of the bank‘s assets comes directly from the profit derived from the bank‘s loans. The banking companies‘ face, however dual challenge to distinguish the possible deliberate defaulters from the applicants and the biased nature of few bank employees who have been at the instigation of developers of defaulting companies for many years. The primary goal of the banking community is to safely invest their capital. In the current scenario, many NBFCs and banks approve loans after a clear verification and authentication process, however, it remains uncertain whether the candidate selected is the worthy correct of all the applicants. Through this method, we can predict whether or not that particular applicant is secure and the machine learning technique automates the entire process of authentication. The major Revised Manuscript Received on November 10, 2019. \* Correspondence Author Dr.E.ChandraBlessie, Department of MCA, Nehru College of Management, Coimbatore, Tamilnadu, India R.Rekha\*, Department of MCA, Nehru College of Management, Coimbatore, Tamilnadu, India, rekhaoct18@gmail.com disadvantage of this model lies in the fact that more importance is given in assigning weightages to each factor, but, in real-time a loan can be sanctioned solely on the basis of a single strong factor, which is not feasible through this method. This paper seeks to ensure that the deserving customers can be quickly selected with ease which offers various benefits to the bank itself. This method will measure the weight automatically of each criterion that participates in the loan processing and process the same with regards to the associated weight of the new test data. The borrower can set a time to check whether or not the loan is sanctioned. This system can skip the sequential verification process and could jump to a specific point to be verified on the basis of priority. This system of loan prediction is solely for the bank officials and completely foolproof that the private players and investors can not alter the data. Results on a particular loan ID may be sent to various bank departments to take adequate action and carry out other formalities. Evaluating the risk associated with a loan application is one of the most important concerns for the sustainability and profitability of the highly competitive market. On a daily basis, these banks receive multiple applications for loans from their clients and others. Not all of them are sanctioned. Many banks use their own credit scoring and risk analysis methods to review the request for loan and make credit approval decisions. Nonetheless, there are many instances occurring each year where borrowers do not repay the sums of the loan and they default, as a result of which these financial institutions incur tremendous losses. The following section depicts the existing loan prediction methods. Aboobyda Jafar Hamid and Tarig Mohammed Ahmed [1] presented a loan risk prediction model based on the data mining techniques, such as Decision Tree (J48), Naïve Bayes (NB) and BayseNet approaches. The procedure followed was (1) training set preparation, (2) building the model, (3) Applying the model and finally (4) Evaluating the accuracy. This approach was implemented using Weka Tool and considered a dataset with eight attributes, namely, gender, job, age, credit amount, credit history, purpose, housing, and class. Evaluating these models on the dataset, experimental results concluded that, J48 based loan prediction approach resulted in better accuracy than the other methods Vimala and Sharmili [2] proposed a loan prediction model using NB and Support Vector Machines (SVM) methods. Naïve Bayes, an independent speculation approach, encompasses probability theory regarding the data classification. On the other hand, SVM uses statistical learning model for classification of predictions. Dataset from UCI repository with 21 attributes was adopted to evaluate the proposed method. Experimentations concluded that, rather than individual performances of classifiers (NB and SVM), the integration of NB and SVM resulted in an efficient classification of loan predictions. Kacheria, Shivakumar, Sawkar and Gupta [3] suggested a loan sanctioning prediction procedure basded on NB approach integrated with K-Nearest Neighbor (KNN) and binning algorithms. The seven parameters considered were income, age, profession, existing loan with its tenure, amount and approval status. The sub-processes include, (1) Pre-processing (handling the missing values with KNN and data refinement using binning algorithm), (2) Classification using NB approach and (3) Updating the dataset frequently results in appropriate improvement in the loan prediction process. Experimentation put-forth the conclusion that, integration of KNN and binning algorithm with NB resulted in improved prediction of loan sanctioning process. Jency, Sumathi and Shiva Sri [4] proposed a Exploratory Data Analysis (EDA) regarding the loan prediction procedure based on the client‘s nature and their requirements. The major factors concentrated during the data analysis were (1) annual income versus loan purpose, (2) customer‘s trust, (3) loan tenure versus delinquent months, (4) loan tenure versus credit category, (5) loan tenure versus number of years in the current job, and (6) chances for loan repayment versus the house ownership. Finally, the outcome of the present work was to infer the constraints on the customer who are applying for the loan followed by the prediction regarding the repayment. Further, results showed that, the customers were interested more on availing short-tenure loans rather than long-tenure loans. Goyal and Kaur [5] suggested an ensemble technique based loan prediction procedure for the customers. The sub-processes in the present method includes, (1) data collection, (2) filtering the data, (3) feature extraction, (4) applying the model, and finally (5) analysis the results. The various loan prediction procedures implemented in the present method were Random Forest (RF), SVM and Tree model with Genetic Algorithm (TGA). The parameters considered for evaluating the models were (1) accuracy, (2) Gini Coefficient, (3) Area Under Curve (AUC), (4) Receiver Operating Curve (ROC), (5) Kolmogorov - Smirnov (KS) Chart, (6) Minimum Cost - Weighted Error Rate, (7) Minimum Error Rate, and (8) K-Fold Cross Validation parameters. Experimentation outcome concluded that the integration of three methods (RF, SVM and TGA) resulted in improved loan - prediction results rather than individual method‘s prediction. Goyal and Kaur [6] presented a loan prediction model using several Machine Learning (ML) algorithms. The dataset with features, namely, gender, marital status, education, number of dependents, employment status, income, co-applicant‘s income, loan amount, loan tenure, credit history, existing loan status, and property area, are used for determining the loan eligibility regarding the loan sanctioning process. Various ML models adopted in the present method includes, (1) Linear model, (2) Decision Tree (DT), (3) Neural Network (NN), (4) Random Forest (RF), (5) SVM, (6) Extreme learning machines, (7) Model tree, (8) Multivariate Adaptive Regression Splines, (9) Bagged Cart Model, (10), NB and (11) TGA. When evaluated these models using R Environment in five runs, TGA resulted in better loan forecasting performance than the other methods. Sudhamathy [7] suggested a risk analysis method in sanctioning a loan for the customers using R package. The various modules include data selection, pre-processing, feature extraction and selection, building the model, prediction followed by the evaluation. The dataset used for evaluation in this method was adopted from UCI repository. To fine tune the prediction accuracy, the pre-processing operation includes the following sub-processes: detection, ranking and removal of outliers, removal of imputation, and balancing of dataset by proportional bifurcation regarding testing and training process. Further, feature selection process improves the prediction accuracy. When evaluated, the DT model resulted in 94.3% prediction accuracy. Supriya, Pavani, Saisushma, Vimala Kumari and Vikas [8] presented a ML based loan prediction model. The modules in the present approach were data collection and pre-processing, applying the ML models, training followed by testing the data. During the pre-processing stage, the detection and removal of outliers and imputation removal processing were carried out. In the present method, SVM, DT, KNN and gradient boosting models were employed to predict the possibilities of current status regarding the loan approval process. The conventional 80:20 rule was adopted to split the dataset into training and testing processes. Experimentation concluded that, DT has significantly higher loan prediction accuracy than the other models. Arun, Ishan and Sanmeet [9] suggested a load prediction procedure using ML models. The sub-processes include data collection, feature selection, training, testing and analyzing the performance of the present model. The dataset with 10 features were employed for observation and loan prediction process. Various ML approaches used in the present method includes LM, DT, RF, SVM, NN and Adaboost methods. Further, authors suggested few significant parameters that plays a major role in loan prediction process for various ML models, such that, it helps to bankers in approval of loans to the customers based on their requirements

**IN “PREDICTION OF LOAN APPROVAL USING MACHINE LEARNING”** Loan approval is a very important process for banking organizations. The system approved or reject

the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. In recent years many researchers worked on loan approval prediction systems. Machine Learning (ML)techniques are very useful in predicting outcomes for large amount of data. In this paper three machine learning algorithms, Logistic Regression(LR), Decision Tree (DT) and Random Forest (RF)are applied to predict the loan approval of customers. The experimental results conclude that the accuracy of Decision Tree machine learning algorithm is better as compared to Logistic Regression and Random Forest machine learning approaches Now a day’s people rely on bank loans to fulfill their needs. The rate of loan applications increases with a very fast speed in recent years. Risk is always involved in approval of loans. The banking officials are very conscious about the payment of the loan amount byits customers. Event after taking lot of precautions and analyzing the loan applicant data, the loan approval decisions are not always correct. There is need of automation of this process so that loan approval is less risky and incur less loss for banksArtificial Intelligence AI is an emerging technology now a day. The application of AI solves many problems of the real world. Machine Learning is an AI technique which is very useful in prediction systems. Figure 1 is showing a basic model of machine learning. It creates a model from a training data. While making the prediction the model which is developed by training algorithm (which is machine learning) is used. The machine learning algorithm trained the system using a fraction of the data available and test the remaining data. The machine Learning techniques can be applied on a sample test data first and then can be used in making prediction related decisions. This paper applied the machine learning approaches in solving loan approval problem of banking sector. Next section discusses the literature survey. Then proposed work, results and analysis are discussed. Finally, conclusion and future scope is discussed which is followed by the references used in this paper.Figure 1 : Basic Machine Learning Model2.Literature SurveyA. Vaidyaproposed a method for approval of loan prediction using logistic regression [1]. Logical regression is a machine learning technique which is very useful in prediction system. The approval of loan is a very important processin banking system. A. Vaidyasolves the problem by applying machine learning in a sample data set for loan approval applications. It also opens other areas on which machine learning is applicable. A. Li and Q. Sun[2] find a method to calculate risk involved in loan approvals for SMEs. A concept of loan consuming radius was introduced which was based upon supply chain in consumer market. F. M. Isiket al. develop a loan approval system using Business Process ExecutionLanguage BPEL[3]. The concept of BPELis very useful in business firms. A reasoning engine was developed which removes some services from the BPEL process which are not necessary to complete a process. The system was applied on oan approval which involve many processes. [4]V. C. T. Chanet al. proposed a credit approval system using web services. The system approved credit for the customers. With credit application the customer submits some other useful information’s. This information’s are processed by Credit Approval System which finally give credit score to the applicant. The paper developed a web services based solution of this problem. J. Lohokareet al. [5] proposed a system which automatically collect data for an applicant and decides the credit score. The system work on the social media to collect information about the user. R. Yanget al. [6] analyzed that whether the credit default behavior of a SME depends upon credit features of its owner or not. The author concluded that features of the owner behaves as valuable parameters to calculate risk of a loan for SMEs. [7]M. Bayraktaret al. [7] proposed a method for credit risk analysis using machine learning. Boltzman machine was used to make the analysis for risk calculation of loan. H. A. P. Pérezet al. [8] introduced fuzzy model for calculation of credit score of the customer. The information collected by the system for calculation of the credit score was converted into gradual values using fuzzy sets. The fuzzy based method performs better for calculation of the credit score of the applicants. S. Yadav and S. Thakur[9] applied Big Data approach for loan analysis. The techniques of big data analysis was applied on customer data to calculate bank loan analysis. Hadoop based method was used in theloan analysis. Y. Lin[10] analysis of the effect of the political approaches effect the loans of state banks. The paper investigated that in state owned banks, the political relationship plays a considerable role. [11]Ruifen Zhaoworked on approval of college loans. Education loans are very common among students because of rise in the cost of education. The paper investigated the issues in loan approval of college students. M. Houshmand and M. D. Kakhki[12] proposed an expert system which evaluates the loan approvals. The system used rule base approaches for loan approval decisions. L. Hui-ling[13] analyze the relation between characteristics of the banks, firms and loans approval. The paper investigated that there is a strong relationship between approval of loans and characteristics of business firm who apply the loan and characteristics of the bank. C. Yin[14] apply fuzzy logic to calculate the bank loan risks. A new pattern recognition system using fuzzy logic was developed which evaluate the risks involves in the approval of bank loans for applicants. J. Ma and Y. Cheng[15] proposed Markov Chain based model for risk management of bank loans. A. V. Gutierrez[16] proposed a model for housing loan. The model was worked for green housing loans. J. Chen and W. Guo[17] worked on loan limit of the loan applicants. The model worked on supply chain for financing decision making. G. Arutjothi and C. Senthamarai, [18] used machine learning classifier for prediction of loan approval status in banks. The machine learning based prediction system was applied on commercial banks. The paper conclude that the machine learning approach is very useful in loan status prediction. Y. Shi and P. Song[19] proposed a method for evaluating project loans using risk analysis. The method evaluate the risk involved in loans of commercial banks. R. ZhangandD. Li [20] used machine learning approached in prediction systems. The machine learning approach was used for assessment of water quality. The paper concluded that machine learning is a very unimportant tool in prediction systems. C. Franket al. [21] used machine learning in prediction of smoking status. Different machine learning approaches were applied and investigated for finding the smoking status. From the results its was ensured that logistic algorithm performs better. R. Lopeset al. applied machine learning approach for the prediction of credit recovery [22]. Credit recovery is very important issue for banking system. The prediction of credit recovery is a challenging tasks. Different machine learning approach was applied to predict the credit recovery and gradient expansionalgorithms (GBM)outperformed the other machine learning approaches.After going through this literature it is found that loan approval predictionproblem is very important for banking system. Machine learning algorithm are very useful in predicting outcomes even when data is very big in size. This paper investigated some machine learning algorithms and applied ML on test data set of loan approvals.Next section discussed the three machine learning approaches.

**PROPSOED SYSTEM**

The architecture of the proposed model is shown in flow chart Fig.1. The major objective of this project is to derive patterns from the datasets which are used for the loan sanctioning process and create a model based on the patterns derived in the previous step. Classification data mining algorithms are used to filter out the probable loan defaulters from the list. For analysis purposes, essential inputs like gender, age, marital status, residential status, job, income, loan expectation, existing client, account balance, total debt, etc., are collected and used to find the appropriate attributes.



**ALGORITHM**

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

In a binary logistic regression model, the dependent variable has two levels (categorical). Outputs with more than two values are modeled by multinomial logistic regression and, if the multiple categories are ordered, by ordinal logistic regression (for example the proportional odds ordinal logistic model[2]). The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier. The coefficients are generally not computed by a closed-form expression, unlike linear least squares; see § Model fitting. The logistic regression as a general statistical model was originally developed and popularized primarily by Joseph Berkson,[3] beginning in Berkson (1944), where he coined "logit"; see § History.

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.

For example,

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

Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time.

From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

***Model***

Output = 0 or 1

Hypothesis => Z = WX + B

hΘ(x) = sigmoid (Z)

***Sigmoid Function***



Figure 2: Sigmoid Activation Function

If ‘Z’ goes to infinity, Y(predicted) will become 1 and if ‘Z’ goes to negative infinity, Y(predicted) will become 0.

***Analysis of the hypothesis***

The output from the hypothesis is the estimated probability. This is used to infer how confident can predicted value be actual value when given an input X. Consider the below example,

X = [x0 x1] = [1 IP-Address]

Based on the x1 value, let’s say we obtained the estimated probability to be 0.8. This tells that there is 80% chance that an email will be spam.

Mathematically this can be written as,

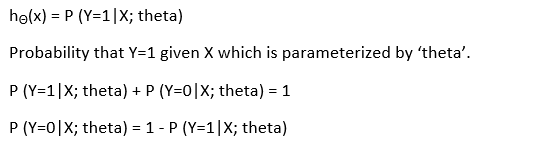


Figure 3: Mathematical Representation

This justifies the name ‘logistic regression’. Data is fit into linear regression model, which then be acted upon by a logistic function predicting the target categorical dependent variable.

***Types of Logistic Regression***

1. Binary Logistic Regression

The categorical response has only two 2 possible outcomes. Example: Spam or Not

2. Multinomial Logistic Regression

Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)

3. Ordinal Logistic Regression

Three or more categories with ordering. Example: Movie rating from 1 to 5

***Decision Boundary***

To predict which class a data belongs, a threshold can be set. Based upon this threshold, the obtained estimated probability is classified into classes.

Say, if predicted\_value ≥ 0.5, then classify email as spam else as not spam.

Decision boundary can be linear or non-linear. Polynomial order can be increased to get complex decision boundary.

***Cost Function***

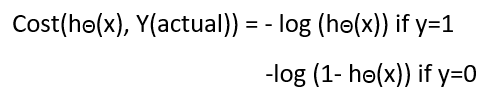


Figure 4: Cost Function of Logistic Regression

Why cost function which has been used for linear can not be used for logistic?

Linear regression uses mean squared error as its cost function. If this is used for logistic regression, then it will be a non-convex function of parameters (theta). Gradient descent will converge into global minimum only if the function is convex.

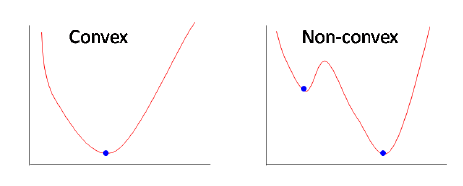


Figure 5: Convex and non-convex cost function

***Cost function explanation***

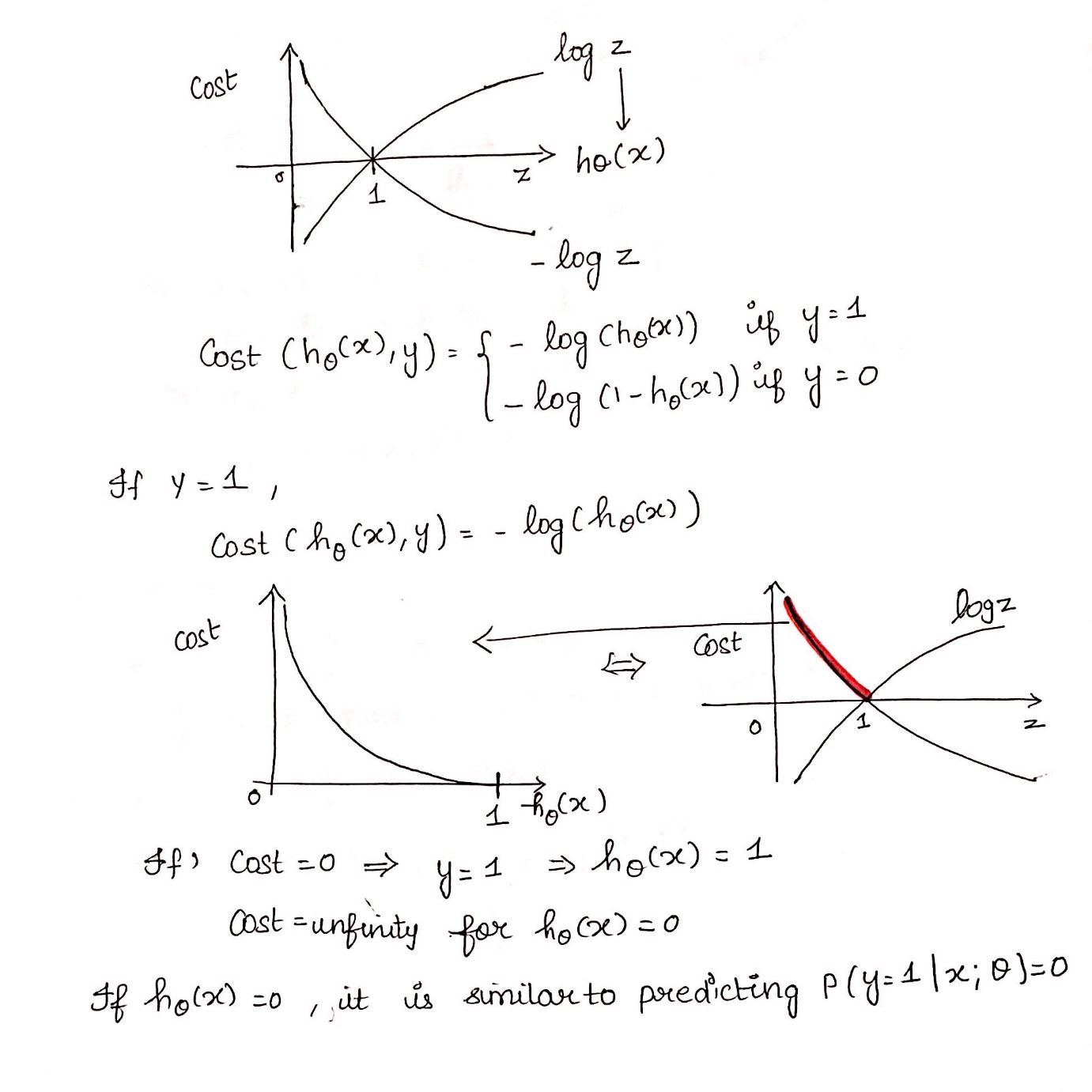


Figure 6: Cost Function part 1

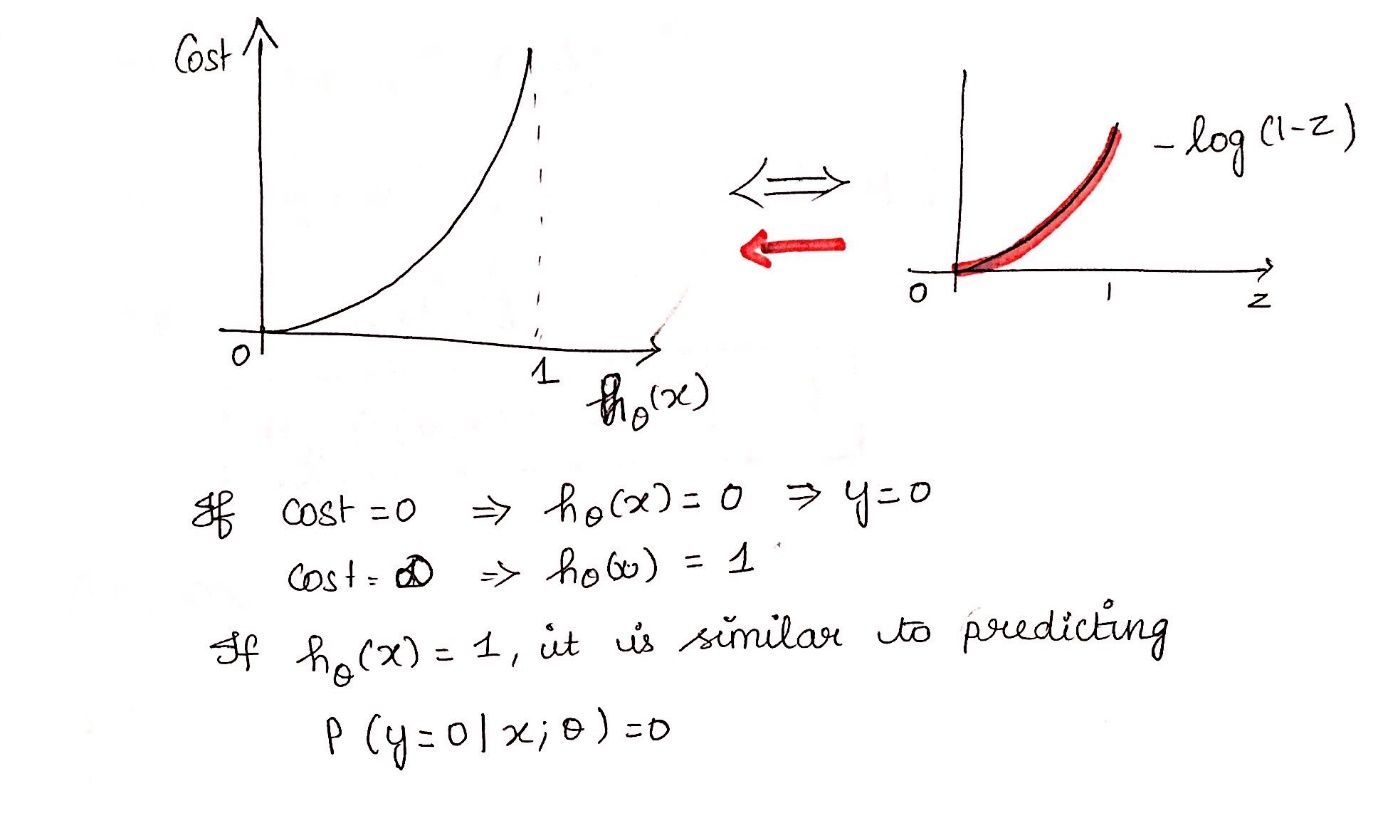


Figure 7: Cost Function part 2

***Simplified cost function***

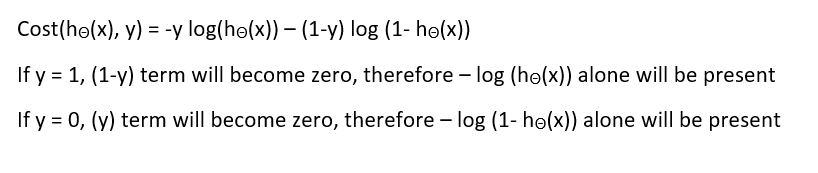


Figure 8: Simplified Cost Function

***Why this cost function?***

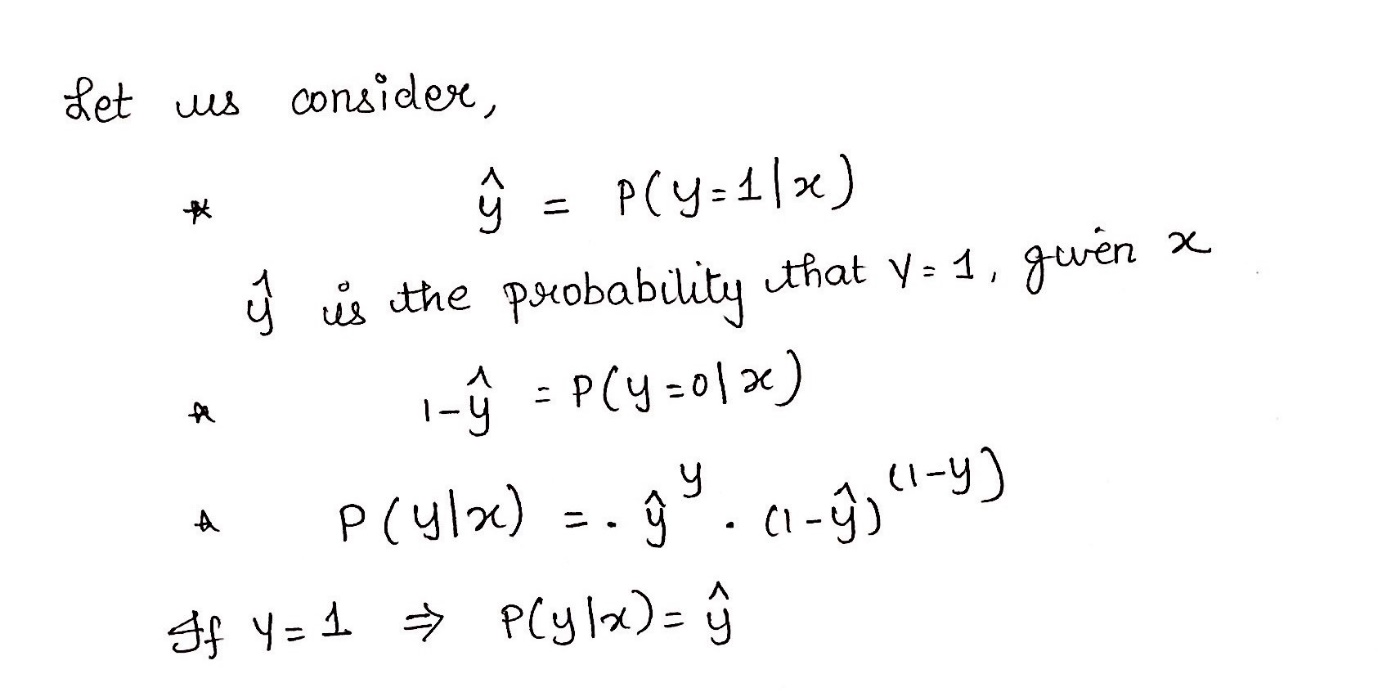


Figure 9: Maximum Likelihood Explanation part-1

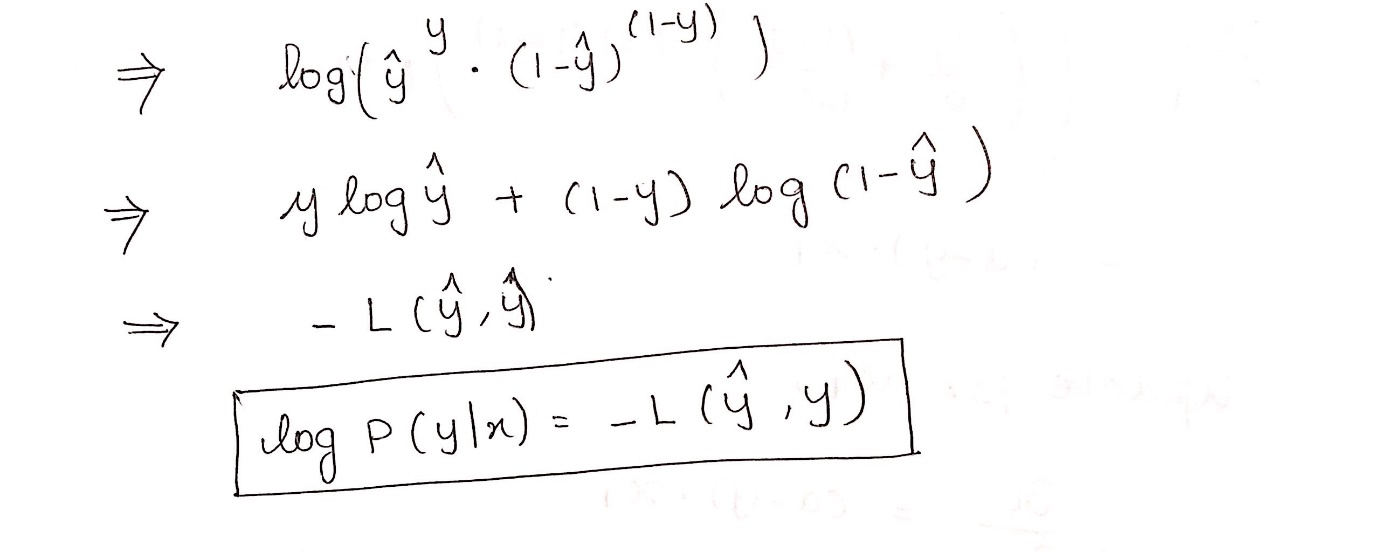


Figure 10: Maximum Likelihood Explanation part-2

This negative function is because when we train, we need to maximize the probability by minimizing loss function. Decreasing the cost will increase the maximum likelihood assuming that samples are drawn from an identically independent distribution.

***Deriving the formula for Gradient Descent Algorithm***

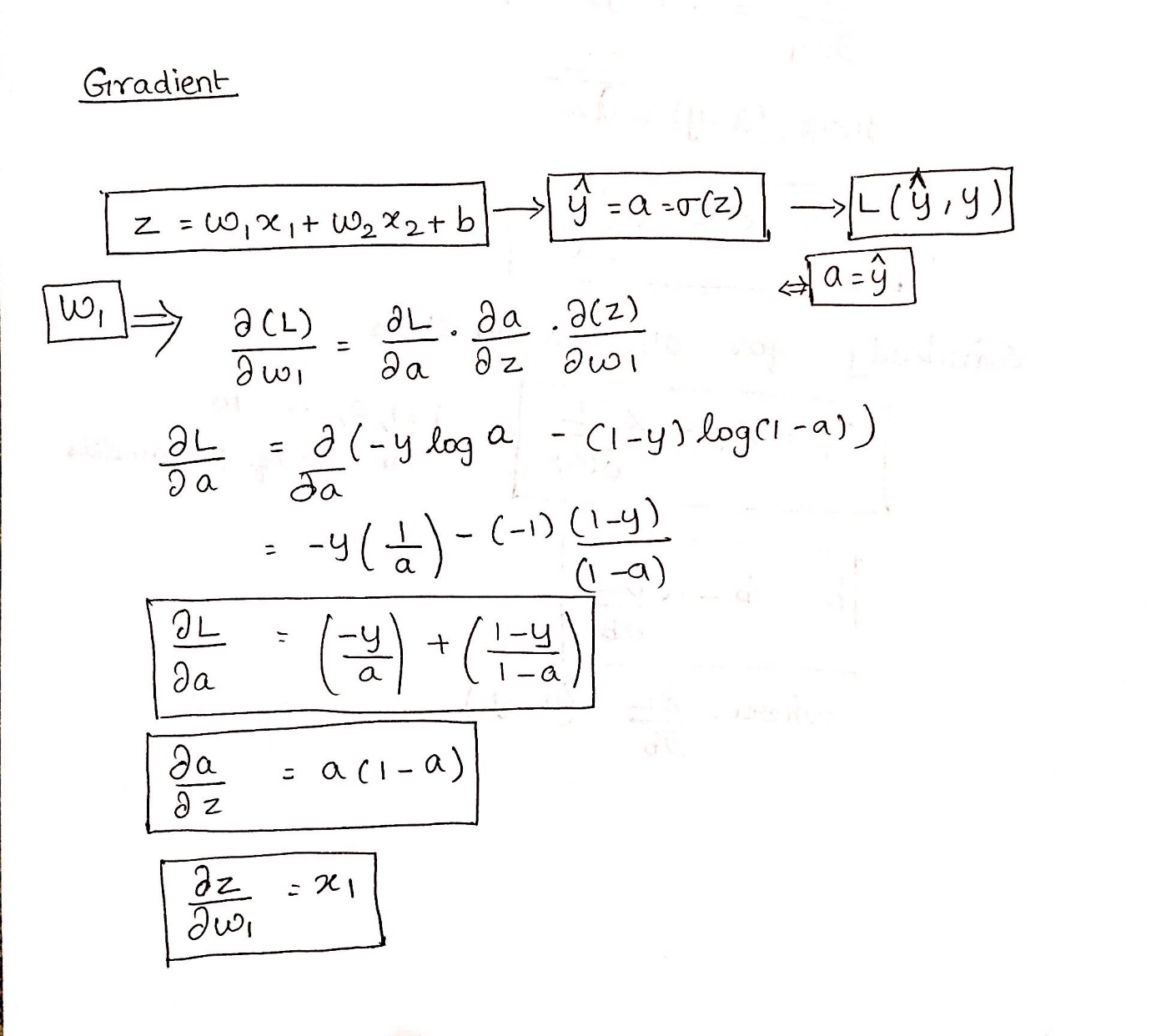
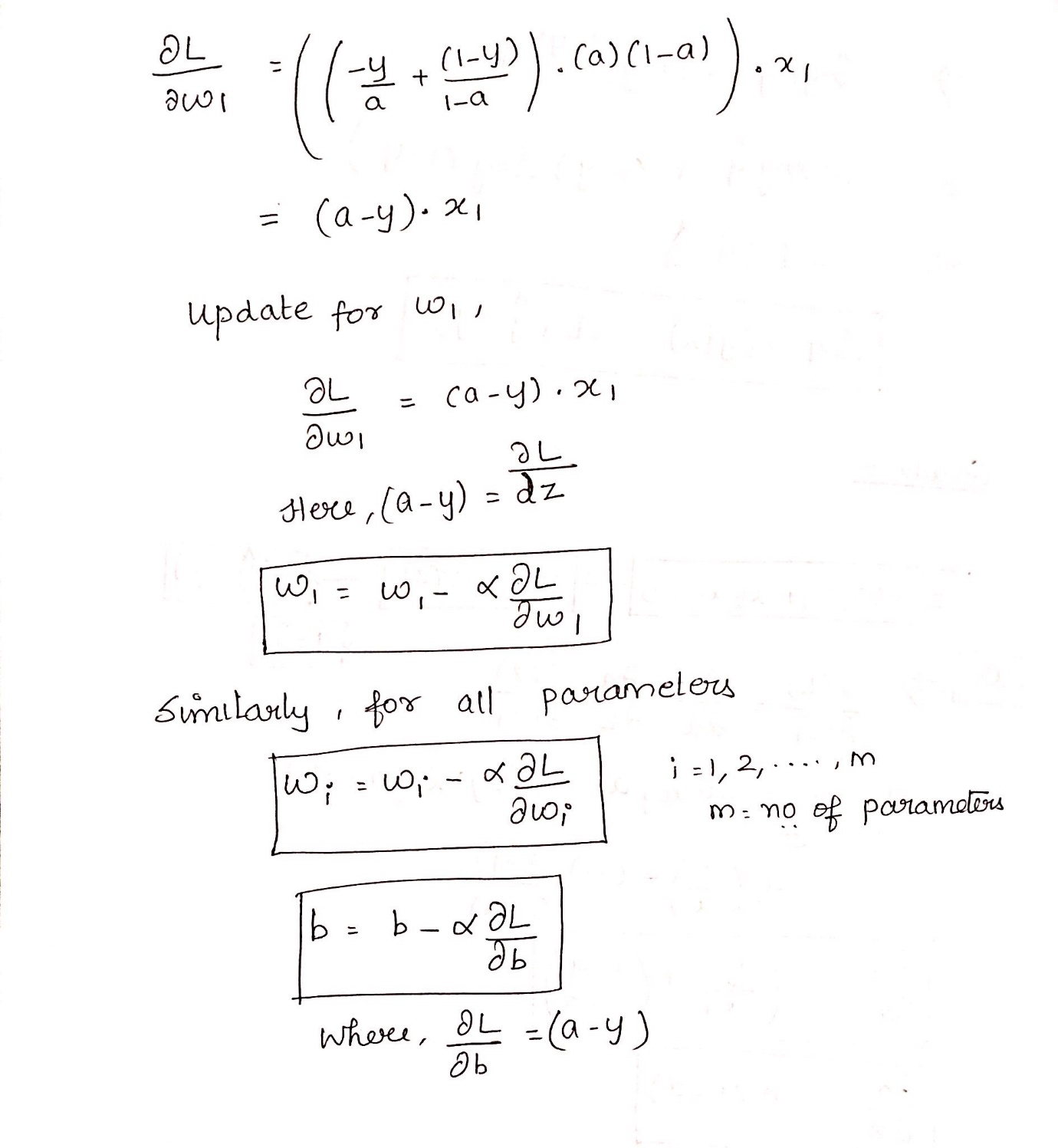


Figure 11: Gradient Descent Algorithm part 1



**MACHINE LEARNING BASICS WITH THE K-NEAREST NEIGHBORS ALGORITHM**

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. Pause! Let us unpack that.



ABC. We are keeping it super simple!

Breaking it down

A **supervised machine learning** algorithm (as opposed to an unsupervised machine learning algorithm) is one that relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data.

Imagine a computer is a child, we are its supervisor (e.g. parent, guardian, or teacher), and we want the child (computer) to learn what a pig looks like. We will show the child several different pictures, some of which are pigs and the rest could be pictures of anything (cats, dogs, etc).

When we see a pig, we shout “pig!” When it’s not a pig, we shout “no, not pig!” After doing this several times with the child, we show them a picture and ask “pig?” and they will correctly (most of the time) say “pig!” or “no, not pig!” depending on what the picture is. That is supervised machine learning.



“Pig!”

Supervised machine learning algorithms are used to solve classification or regression problems.

A **classification problem** has a discrete value as its output. For example, “likes pineapple on pizza” and “does not like pineapple on pizza” are discrete. There is no middle ground. The analogy above of teaching a child to identify a pig is another example of a classification problem.

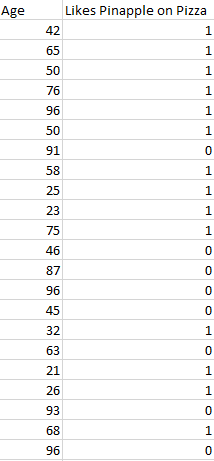


Image showing randomly generated data

This image shows a basic example of what classification data might look like. We have a predictor (or set of predictors) and a label. In the image, we might be trying to predict whether someone likes pineapple (1) on their pizza or not (0) based on their age (the predictor).

It is standard practice to represent the output (label) of a classification algorithm as an integer number such as 1, -1, or 0. In this instance, these numbers are purely representational. Mathematical operations should not be performed on them because doing so would be meaningless. Think for a moment. What is “likes pineapple” + “does not like pineapple”? Exactly. We cannot add them, so we should not add their numeric representations.

A **regression problem** has a real number (a number with a decimal point) as its output. For example, we could use the data in the table below to estimate someone’s weight given their height.

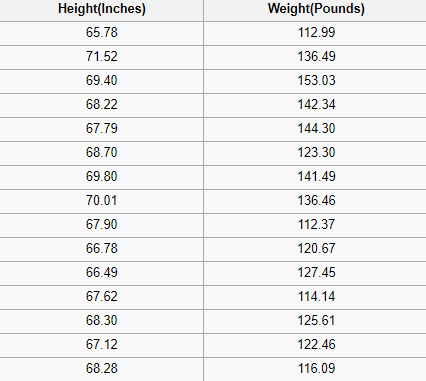


Image showing a portion of the[SOCR height and weights data set](http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights)

Data used in a regression analysis will look similar to the data shown in the image above. We have an independent variable (or set of independent variables) and a dependent variable (the thing we are trying to guess given our independent variables). For instance, we could say height is the independent variable and weight is the dependent variable.

Also, each row is typically called an **example, observation, or data point**, while each column (not including the label/dependent variable) is often called a **predictor, dimension, independent variable, or feature.**

An **unsupervised machine learning**algorithm makes use of input data without any labels —in other words, no teacher (label) telling the child (computer) when it is right or when it has made a mistake so that it can self-correct.

Unlike supervised learning that tries to learn a function that will allow us to make predictions given some new unlabeled data, unsupervised learning tries to learn the basic structure of the data to give us more insight into the data.

K-Nearest Neighbors

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

“Birds of a feather flock together.”



[Image showing how similar data points typically exist close to each other](https://commons.wikimedia.org/wiki/File:Map1NNReducedDataSet.png)

Notice in the image above that most of the time, similar data points are close to each other. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood— calculating the distance between points on a graph.

***Note:****An understanding of how we calculate the distance between points on a graph is necessary before moving on. If you are unfamiliar with or need a refresher on how this calculation is done, thoroughly read “*[*Distance Between 2 Points*](https://www.mathsisfun.com/algebra/distance-2-points.html)*” in its entirety, and come right back.*

There are other ways of calculating distance, and one way might be preferable depending on the problem we are solving. However, the straight-line distance (also called the Euclidean distance) is a popular and familiar choice.

The KNN Algorithm

1. Load the data
2. Initialize K to your chosen number of neighbors

3. For each example in the data

3.1 Calculate the distance between the query example and the current example from the data.

3.2 Add the distance and the index of the example to an ordered collection

4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances

5. Pick the first K entries from the sorted collection

6. Get the labels of the selected K entries

7. If regression, return the mean of the K labels

8. If classification, return the mode of the K labels

The KNN implementation (from scratch)

**CONCLUSION** The process of prediction starts from cleaning and processing of data, imputation of missing values, experimental analysis of data set and then model building to evaluation of model and testing on test data. On Data set, the best case accuracy obtained on the original data set is 0.811. The following conclusions are reached after analysis that thoseapplicants whose credit score was worstwill fail to get loan approval, due to a higher probability of not paying back the loan amount. Most of the time, those applicants who have high income and demands for lower amount of loan are more likely to get approved which makes sense, more likely to pay back their loans. Some other characteristic like gender and marital status seems not to be taken into consideration by the company.

**REFERENCES**

[1] Toby Segaran, “Programming Collective Intelligence: Building Smart Web 2.0 Applications.” O’Reilly Media.

[2] Drew Conway and John Myles White,” Machine Learning for Hackers: Case Studies and Algorithms to Get you Started,” O’Reilly Media.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman,”The Elements of Statistical Learning: Data Mining, Inference, and Prediction,” Springer ,Kindle

[4] PhilHyo Jin Do ,Ho-Jin Choi, “Sentiment analysis of real-life situations using loca- tion, people and time as contextual features,” International Conference on Big Data and Smart Computing (BIGCOMP), pp. 39–42. IEEE, 2015.

[5] Bing Liu, “Sentiment Analysis and Opinion Mining,” Morgan & Claypool Publishers, May 2012.

[6] Bing Liu, “Sentiment Analysis: Mining Opinions, Sentiments, and Emotions,” Cambridge University Press, ISBN:978-1-107-01789-4.

[7] Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng, “CNN for situations understanding based on sentiment analysis of twitter data,” Procedia computer science, 111:376–381, 2017.CrossRef.

[8] K I Rahmani, M.A. Ansari, Amit Kumar Goel, “An Efficient Indexing Algorithm for CBIR,”IEEE- International Conference on Computational Intelligence & Communication Technology ,13-14 Feb 2015.

[9] Gurlove Singh, Amit Kumar Goel ,”Face Detection and Recognition System using Digital Image Processing” , 2nd International conference on Innovative Mechanism for Industry Application ICMIA 2020, 5-7 March 2020, IEEE Publisher.

[10] Amit Kumar Goel, Kalpana Batra, Poonam Phogat,” Manage big data using optical networks”, Journal of Statistics and Management Systems “Volume 23, 2020, Issue 2, Taylors & Francis.