PREDICTIVE MODELLING TO PREDICT ABSENTEEISM IN MNC'S

Project report submitted in partial fulfillment of the requirement for the degree of

BACHELORS OF TECHNOLOGY

IN

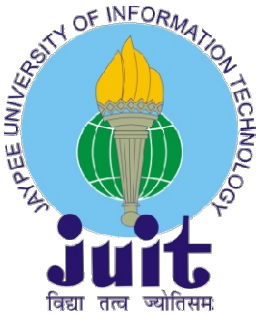
ELECTRONICS AND COMMUNICATION

By

SHRIYA VANDITA

Under the Supervision of

DR. SHRUTI JAIN



DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY WAKNAGHAT,
SOLAN ,173234, INDIA

MAY, 2018

TABLE OF CONTENTS

DECLARATION BY SCHOLAR

We hereby declare that the project work entitled "**PREDICTIVE ANALYSIS OF ABSENTEEISM IN MNCs**" submitted to the **Department of Electronics and Communication, Jaypee University of Information Technology, Solan** is a record of an original work done by us under the supervision of **Dr. Shruti Jain**. This project work is submitted as a part of partial fulfillment for the award of the degree of Bachelor of Technology under Jaypee University of Information Technology.

Name and Signature of Student-

Shriya Vandita (151092)

Department of Electronics and Communication Engineering

Jaypee University of Information Technology, Waknaghat, India

May, 2019

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
(Established by H.P. State Legislative vide Act No. 14 of 2002)
P.O. Waknaghat, Teh. Kandaghat, Distt. Solan - 173234 (H.P.) INDIA
Website: www.juit.ac.in
Phone No. (91) 01792-257999
Fax: +91-01792-245362

CERTIFICATE

This is to certify that the work reported in the B.Tech project report entitled **"Predictive Modeling to Predict Absenteeism in MNC's"** which is being submitted by Shriya Vandita in fulfillment for the award of Bachelors of Technology in **Electronics and Communication Engineering by the Jaypee University of Information Technology**, is the record of candidate's own work carried out by her under my supervision. This work is original and has not been submitted partially or fully anywhere else for any other degree or diploma.

Dr. Shruti Jain
Associate Professor
Department of Electronics & Communication Engineering
Jaypee University of Information Technology, Waknaghat

ACKNOWLEDGEMENT

We take this opportunity to express our gratitude to our supervisor Dr. Shruti Jain for her insightful advice, motivating suggestions, invaluable guidance, help and support in successful completion of this project and also for her constant encouragement and advice throughout our Bachelors program.

The in-house facilities provided by the department throughout the Bachelors program are also equally acknowledgeable. We would like to convey our thanks to the teaching and non-teaching staff of the Department of Electronics and Communication Engineering for their invaluable help and support.

Shriya Vandita (151092)

ABSTRACT

It is rightly said: “Data is the new oil.” The future will move around generation of enormous amount of data which needs to be critically analyzed for the benefit of mankind, industry, academics, defense, intelligence, disaster management and counter terrorism.

Managers and organizational practitioners need a detailed method for measuring absenteeism loss as well as other measures needed for managerial evaluation to decrease absenteeism rate and compare the effectiveness of absence/attendance policy from period to period.

Our aim is to predict the absenteeism for MNCs by the previous recorded datasets. For this we will use predictive analysis using machine learning. For faster processing of massive dataset, the data has to be analyzed efficiently so that we have the minimum response time and turn-around time, which is only possible when we use the right set of algorithms and by hard wiring of program. By looking at the results of each technique we can make some insights about the problem. The methods used in the project include Linear Regression and Logistic Regression.

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Absenteeism has become a severe problem for many organizations. Obviously, it has been an undeniable issue faced by companies which can result in serious financial and nonfinancial losses [1]. Because of the negative consequences of employee absenteeism, it is important that organizations keep the rate of absenteeism low [2]. Managers and organizational practitioners need a detailed method for measuring absenteeism loss as well as other measures needed for managerial evaluation to decrease absenteeism rate and compare the effectiveness of absence/attendance policy from period to period [2].

1.1 Background of Predictive Analytics

Considering the fact that 1980's, the agencies have amassed nicely endowed amounts of clients data understanding to be preserve in databases [2], because the corporations accumulate all of these expertise, they begin to think, however they may use this knowledge to enhance operations or to supply extra edges, this sort of thinking fashioned 'a natural development closer to exploitation of the data to decorate estimates, forecasts, decisions and in the end, efficiency'. these databases grew to this sort of massive length, in flip, became overlarge for human beings to research on their very own [2].Predictive analytics become a solution at the manner to deal with large databases. It's a process that comes with the usage of technique approaches to exercising important and useful patterns in statistics. It turned into devised from connected fields of examine pattern reputation, data, system learning, computer science and statistics processing. [1]

Predictive analytics is said to be driven by data in the experience that the methods that have been implement, generates a new model from the input samples and traits of the records by itself. Abbott believes that these models are prompted from the

records. The algorithms which are driven from data hired in the analytics may want to incorporate the “recognizable proof of factors to be enclosed inside the models, factors that outline the model, strengths or model complexity” [2]. The algorithms used affirm styles among the data.

It incorporates the usage of statistics, however it is a matter that calls for a special method which has specific beliefs other than records. The most important, facts is a area that works on a particularly described set of guidelines and a basis of theory, while within predictive analysis, it isn’t generally be similar case. For instance, the algorithms used are are from fields like AI and man-made brainpower, inside prescient investigation that don't have a quality feasible answer or this kind of solution which could also be proved. Moreover, experts that suggest it are more lenient on the subject models and less specific with model parameters [3]. That is, when becoming a predictive version to the facts, the focus is on the emphasis is on advancing prescient precision of some objective. In common, the techniques used have a tendency to be much less thorough in evaluation to many statistical evaluation techniques [3].

The calculations can be sorted as either supervised learning systems or unsupervised learning procedures. Supervised learning models reason to anticipate an objective variable, represented with the single section inside the set of information, by utilizing different factors or segments inside the dataset. Supervised learning is likewise alluded to as predictive modeling. The most widely recognized predictive modeling calculations are order while we manage a clear cut target variable. Unsupervised learning calculation does not have an objective variable, yet on the other hand assembles a model with utilization of bunches of the data.[4] Unsupervised learning models are alluded to as descriptive modeling [3].

Machine learning, the application and study of calculations that produces feeling of learning, is the best time field of all the computer science. We have a tendency to live in a age where information comes in abundance, exploitation of the self learning where we are able to flip this data into meaningful information so ,will extract information from it. Due to the numerous powerful open supply libraries that are created as of late, there has never been a greatly improved time to break into AI field and gain proficiency with the best approach to use ground-breaking

calculations to recognize designs in data and construct forecasts concerning future occasions.

During this chapter, we'll study the ideas and different types of machine learning along with basic introduction to the relevant word, we'll lay the groundwork using machine learning techniques for sensible problem solving. [2]

In this period of ongoing innovation, one asset we have in bounty a larger than usual amount of organized and unstructured data. Inside the half of twentieth century, AI has advanced as a subfield of processing that is worried about the occasion of self taking in calculations to pick up learning from the data in order to frame expectations instead of expecting people to physically determine principles and manufacture models from examining tremendous measures of information , AI offers an extra practical option for catching the data in information a tiny bit at a time to improve the exhibition of prescient models, and constructed information driven choices.[4] Machine learning helps in various applications which consolidates solid email spam channels, advantageous content and voice acknowledgment bundle, dependable web indexes. [3]

1.2 Types of Machine Learning

There are different types of machine learning. In this project, we have worked on three types of Machine Learning algorithms: supervised learning, unsupervised learning, and reinforcement learning.

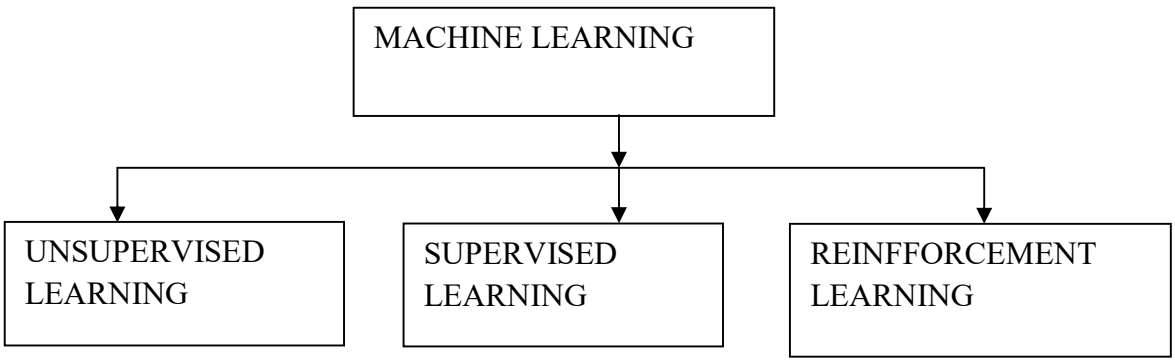


Figure 1.1: Types of Machine Learning Algorithms

The principle aim in supervised learning is to inspect a model from marked training information that enables us to make expectations about inconspicuous or future information. Here, the word supervised alludes to a lot of tests wherein the favored yield signals (marks) are now respected. We have two sorts of information:

- **Categorical data**: Categorical factors comprise kinds of information which might be partitioned into groups. Instances of categorical factors are sex, race, age groups, and instructive dimension. Indeed, even as the last factors can likewise be considered in a numerical way using explicit qualities for age and greatest grade finished, it's far often additional enlightening to classify such factors into few groups. Clear cut factors include a limited number of classes or unmistakable groups.
- **Continuous data**: Continuous factors are numeric factors which have an interminable number of qualities among any two qualities. A continuous variable might be numeric or date/time. For example, the term of a part or the date and time a charge is gotten.

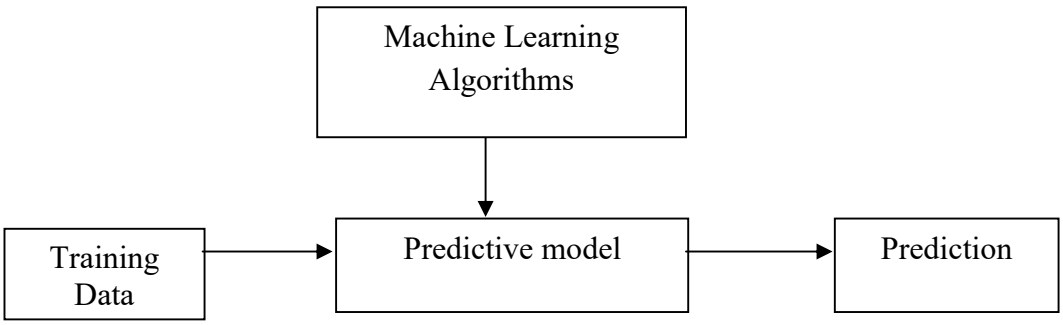


Figure 1.2: Illustrates How a Typical Machine Learning Algorithm Works

The Figure 1.2 signifies the flow of our project. Using our training data, we derived a predictive model for calculating absenteeism in MNCs. After calculating this model, we made our prediction as to how one can reduce the absenteeism.

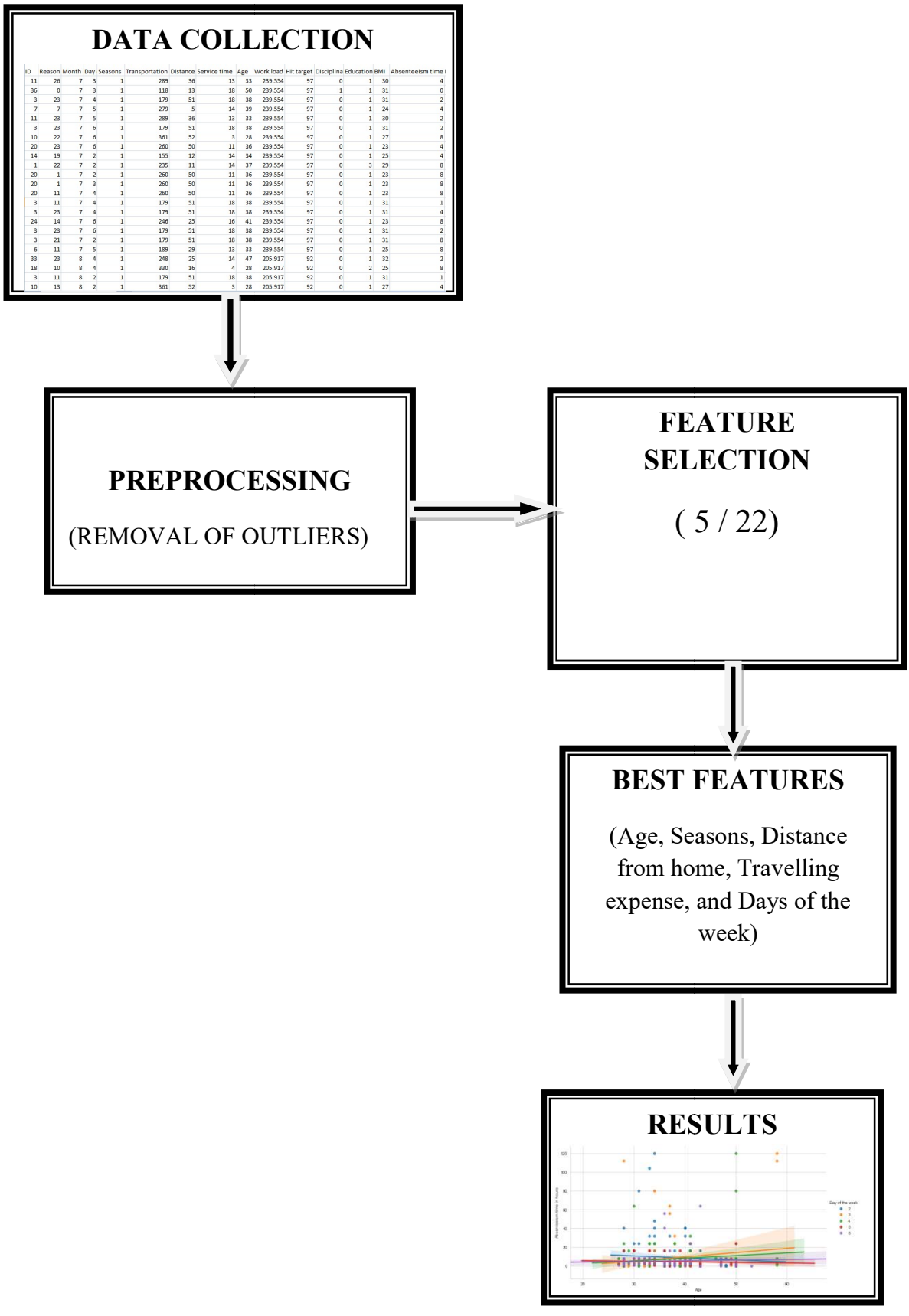


Figure1. 3: Predictive analysis model with results

In figure 1.3 it is shown the steps of predictive analysis that has been carried out in this project with the results at each step. The steps involve data collection which has been done from the site, pre processing methods that involve removal of outliers, feature engineering from which we select the best features and implementing regression models on it to obtain desired results.

The point of classification is to appoint categorical, unordered marks to occasions. A second sort of directed learning is the forecast of ceaseless outcomes, which is moreover called relapse investigation. In relapse investigation, we are given various indicator (informative) factors and a nonstop reaction variable (last outcomes), and we endeavor to discover a relationship among those factors that grants us to anticipate a result.

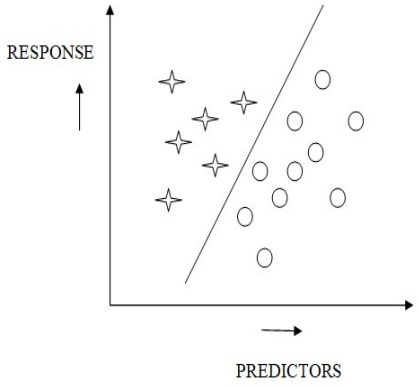


Figure1.4: Illustrates the concept of binary classification

Figure 1.4, outlines the idea of straight relapse. Given an indicator variable and a reaction variable, we fit a straight line to this information that limits the separation most usually the normal squared separation between the examples focuses and the fitted line.[5] We would now be able to utilize the capture and incline gained from this information to foresee the result variable of new information:

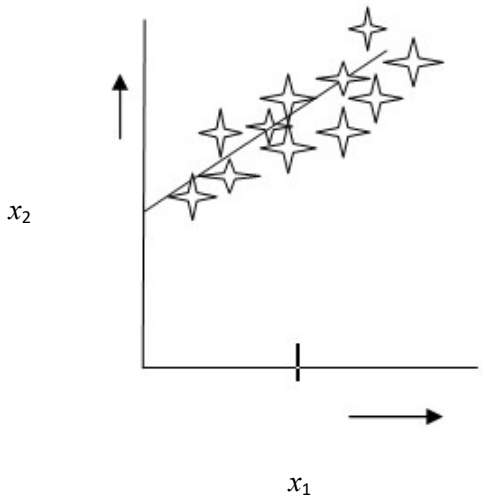


Figure1.5: Illustrates the concept of linear regression

In Figure 1.5, x1 denotes the predictor variable and x2 determines the response variable. Linear regression is a form of regression which helps us in establishing a linear relationship between two parameters.

In fortification learning, the point is to widen a framework (operator) that improves its exhibition upheld associations with the environment. Since the measurements in regards to this condition of the surroundings generally likewise incorporate an expected reward signal, we can review support learning as a field related with managed acing.[5] However, in support learning this remark isn't generally the correct ground reality name or worth, yet a level of the way well the activity progressed toward becoming estimated through a reward work. Through the transaction with the environment, an operator will at that point use fortification figuring out how to look into a chain of activities that amplifies this reward through an underlying preliminary and-blunders approach.[6]

1.3 Software Used

There are different types of software through which we can work on Machine Learning. However, in this project, we have used Spark and Jupyter notebook.

We have used Spark software because Apache Spark is a fast and general purpose computing system, providing its users with a rich set of higher-level tools such as

Flash SQL, for SQL and organized information, MLib, for AI, and GraphX for Graph preparing..

SPARK:

Apache Spark is an open source cluster computing framework. One of the most important features of Spark is to supply high speed which is the ability to run the system in a more efficient manner than Map scale back for complicated applications running on the disk.

This framework supports different programming languages like Java, Python and R programming languages. It also integrates closely with different huge data tools. Particularly, Spark will run in Hadoop clusters and may access any Hadoop information source including Cassandra. [6]

At high level, a Spark application comprises of a driver program that dispatches various parallel activities on a bunch. The driver program contains the most capacity of our application which would then be able to be appropriated to the bunch individuals for execution. The Spark Context object is utilized by the driver program to get to the processing cluster. [3]

Jupyter:

Records delivered by method for the Jupyter Notebook App incorporate both PC code (for example python) and content elements (passage, conditions, figures, joins, and numerous others...). Note pad records are both comprehensible reports containing the examination delineation and the results (figures, tables, thus on...) notwithstanding executable documents which might be raced to perform information investigation. [6]

Whilst a notebook document is opened, the related kernel is automatically released. While the notebook is achieved, the kernel performs the computation and produces the outcomes. depending on the sort of computations, the kernel may also consume tremendous CPU and RAM. [5]

1.4 Approach

Predictive analysis is used to predict unknown events or unobserved events by analyzing the existing data set with the help of data mining, statistical modeling, and Machine Learning techniques. For prediction analysis, first the objective is defined, and then the data set is prepared. Based upon the prepared data, a model is laid down for deployment and monitoring.[7] Predictive analysis identifies the cause-effect relationship across the variables from the given data set and discovers hidden patterns with the help of data mining techniques. It may apply observed patterns to unknowns in the past, present or the future.

Predictive models are supervised learning models, which try to predict certain values using the values in the input data set. The learning model establishes a relationship between the target figure, that is, the feature being predicted and the predictor features. The predictive models have a clear focus on what they want to learn and how they want to learn.[7]

The models which are used for the prediction of target features of categorical values are known as classification models. The target feature is known as a class and the categories in which the classes are divided into are known as levels. Some of the popular classification models include KNN and Decision Trees. Predictive models may also be used to predict numerical values of the target feature based on the predictor features. The models which are used for the prediction of the numerical values of the target feature of a data instance are known as regression models. They include, Linear Regression, Support Vector Regression, Lasso, Ridge regression and other forms of regression models.

Predictive analytics is said to be driven by data in the sense that the steps, codes and algorithms used, generate a model from the patterns and characteristics of the data alone. We would be applying data preprocessing techniques to a data set and reduce that data set into training data set and test data set. Later, we will generate algorithms using training data set and apply those to the test data set.

The algorithms derived and driven from data are used in the analytics may involve the identification of variables to be included in the model, parameters that define the model, weights or model complexity.[8]

Our aim is to predict the absenteeism for MNCs by the previous recorded datasets. For this we will use predictive analysis using machine learning.

1.5 Motivation

We were motivated to work in the area of prediction analysis field because data science is an inter-disciplinary field where processes and systems are analyzed to extract information and thereby, knowledge from data in any form, structured as well as unstructured. The availability of advanced machines and special tools has led to the analysis of big data, which may solve many social problems, including poverty and unemployment.

Through this project, we got the opportunity to learn about how the statistics can be used in various ways across various dimensions such as intelligence, defense and artificial intelligence. It also broadened our horizons of looking at an unknown data and trying to find useful features and patterns. We also got an opportunity to learn Python as we programmed our analysis on this language.

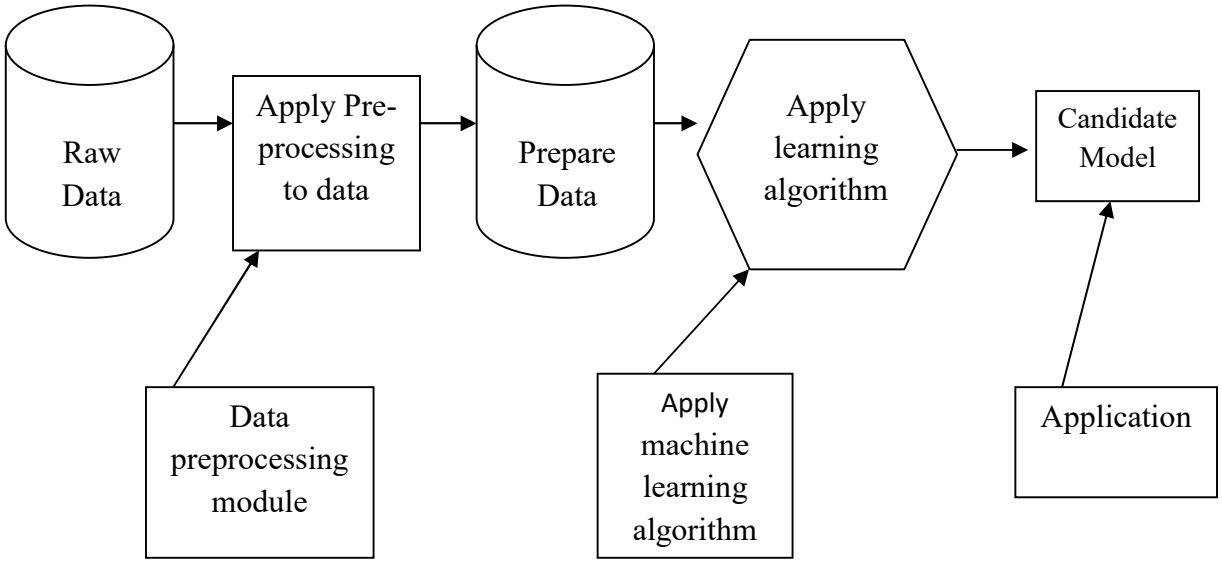


Figure1.6: Elaborate process of machine learning

Raw data is the data which is noisy, that is, containing errors and it is also incomplete. Hence, we need to apply pre-processing techniques like standardization and normalization, missing value replacement, and feature selection. We also imported basic libraries such as NumPy, Pandas, Matplotlib etc.

Below mentioned are the steps for pre processing data using libraries from Python. Preprocessing methods we tend to remove outliers and normalize the data.

Step 1: Import the Libraries: There are different libraries and its keywords in Python. Out of which we are using Numpy, Pandas, Matplotlib, and seaborn keywords in our project.[8]

i. **NumPy** is the fundamental pack for logical registering with Python. It contains amazing N-dimensional exhibit object, critic (broadcasting) capacities, apparatuses for coordinating C/C++ and FORTRAN code, and valuable direct polynomial math, Fourier redesign, and irregular wide variety.

ii. **Pandas** are for information control and investigation. Pandas are a publicly released, BSD-affirmed library exhibiting extreme generally speaking execution, simple to-utilize information frameworks and measurements examination gear for the Python programming language.

iii. **Matplotlib** is a Python 2D plotting library which produces distribution amazing figures in an assortment of printed copy designs and intelligent conditions all through frameworks. Matplotlib can be utilized in Python contents, the Python and Python shells, the Jupyter scratch pad, web application, and four graphical UI toolboxes.

iv. Seaborn is a Python data recognition library reliant on matplotlib. It gives an unusual state interface to drawing appealing and informational quantifiable outlines.

Warning note are commonly displayed in the situations where it's far useful to inform the user of a few situations in a application, wherein that condition (generally) doesn’t warrant elevating an irregularity and ending the program. For instance, one may need to difficulty a warning while a program makes use of an out of date module.

Step 2: Import the Data-set: With the usage of Pandas we import our data-set and the file that we have got used here is .csv file [note: It's not important each-time we address CSV file; sometimes we have to cope with HTML or Xlsx/xls (Excel document). But, to have a faster access to our document, we use CSV files because of their lights weights. After importing the dataset, we use head feature.

Step 3: Check out the Missing Values: The idea of missing qualities is critical to perceive so as to effectively oversee information. In the event that the missing qualities are not managed by the specialist, at that point he/she may likewise come to draw a wrong surmising around the insights. Due to unsatisfactory dealing with, the outcome gotten by the analyst will contrast from ones where the missing qualities are available.

Step 4: Check for Categorical Values: Machine Learning models are based on Mathematical equations and we can intuitively understand that it might motive some trouble if we are able to maintain the Categorical data within the equations due to the fact we'd simplest need numbers within the equations. So, we want to encode the Categorical Variable. For Instance, in our data set Employee name" column will cause trouble, so we are able to convert it into numerical values. To transform Categorical variable into Numerical records we will use Label Encoder () class from pre processing library.

There are four common methods to achieve Feature Scaling:

1. **Standardization**: **sklearn.preprocessing.scale** helps us to implement standardization in Python.
2. **Mean Normalization:** $x' = \frac{x - mean(x)}{max(x) - min(x)}$ **(4.1)**

   Standardization and Mean Normalization can be used for algorithms that assume zero centric data like Principal Component Analysis (PCA).
3. **Min-Max Scaling:** $x' = \frac{x - min(x)}{max(x) - min(x)}$ **(4.2)**
4. **Unit Vector:** $x' = \frac{x}{||x||}$ **(4.3)**

CHAPTER 2

LITERATURE SURVEY

Absenteeism is a vital issue that requires immediate attention by both, the employee and the employer. If the employees enjoy the work they do then they will enjoy their workplace and will not take leave. So employers are expected to keep their employees in good spirit and motivated so that the employees deliver their best to the organization for the benefit of the both.

Absenteeism is a habitual pattern of absence from duty. Absenteeism not only affects the cost but is also an indicator of the poor morale of the employees. The reasons behind absenteeism can be factors like depression, personal stress and anxiety which can cause the employee/workers to get detached, and have issues coping up with the workload of office, stressful meetings/presentations and feelings of being unappreciated.

We found this topic very interesting and read a number of articles and tools to see how today the pattern of absenteeism can be predicted using the available computing tools. We found that the phenomenon of absenteeism in organizations is complicated to take a look at due to the regularly coincidental, multiple multilevel and interacting causes. Despite the fact that many models describing this have already been offered, settlement is still very few.

We determined that among almost all studies till date, it's not been a common exercise to take account of of absenteeism over a specific length and in this way tally the absolute number of missing days as an aggregate measure or the assortment of estimated nonattendance periods, as a prompt measure (nonappearance recurrence).

Most importantly typically actualized measurable models were connected to break down the accessible information like Ordinary Least Squares (OLS) models and the Poisson relapse model (PRM).A new model turned into then recommended where absenteeism is considered to be a combined variable among the two and more processes.

Today, HR departments gives away a large amount of data on a every day basis : recruitment, leaves, departures, salaries and advantages, annual opinions, career evolution, social conflicts and so forth. This leads to huge size data, termed as Big Data and for its analysis we need specific tools which were not required earlier.

Big data and predictive analytics when combined together can open a huge opportunity for HR professionals and can generate extreme benefits for all stakeholders inside the company: HR enterprise partners, managers and personnel.

Efficiently making use of predictive analysis to HR data can provide new vision into present and future production optimization opportunities at multiple moments of the workers lifestyles cycle. For example:

- Prediction of absenteeism and risk of work accidents
- The measurement of employee commitment.
- Finding the most efficient on boarding technique to reduce the time to perform.

Predictive analysis of employee absenteeism ends in better employee retention through the organization. HR Big data analytics lets in to apprehend the employees' motivations very appropriately and what makes them sustain longer in a corporation or to give up and leave. Primarily based on some data items – effortlessly to be had through most HR departments – algorithmic models can extract effective evaluation that we could one determine about the right moves that could improve worker retention rates, like

a. Detecting the employees that have the potential to leave and taking a preventive measure against it.
b. Analyzing employee contribution by the factor of business unit or department.
c. Predicting the turnover cost and its growth for the short, mid as well as for long run.

Out of the various papers we looked at, we consulted the following papers to carry out our project work:

Paper	Author &Year	Conclusion
A practical guidance to count absenteeism: A Refined Quantitative Method.	**Wijaya N.,2000**	Absenteeism has been characterized in different ways. "Any inability to report for or stay at work as planned, regardless of reason." "The title given to a condition that exists when an individual neglects to come to work when appropriately booked to work."
Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods.	**Finlay,2014**	"a characteristic movement toward utilizing the data to improve forecast, estimates, choices, and at last, productivity" Predictive analysis is an answer on the most proficient method to deal with such huge databases. It is a methodology that joins the utilization of computational strategies to decide significant and helpful examples in huge data.
Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst	**Abbott,2014**	Machine learning is a logical control which centers around naturally perceiving complex patterns and settling on astute choices dependent on accessible information. This part of study creates calculations for computers to develop practices for the equivalent.
Spark 2.3.0 documentation, Apache Spark	**2013**	Apache spark is an open source cluster processing structure, which may be up to multiple times quicker than Hadoop MapReduce.

Remote sensing big data computing: Challenges and opportunities	**Y.Ma, 2015**	We give a concise outline on the Big Data and data intensive issues, including the examination of RS Big Data, Big Data challenges, current procedures and works for handling RS Big Data.
Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods.	**Finlay,2014**	Machine learning bases on the improvement of computer programs that can change when displayed to new data. It is the path toward changing over comprehension into capacity or learning.
A practical guidance to count absenteeism: A Refined Quantitative Method.	**Wijaya N.,2000**	Linear regression is a philosophy for showing the association between a scalar ward variable y and in any event one sensible components demonstrated as X. The occurrence of one educational variable is called simple linear regression.
A comparative analysis of machine learning systems for measuring the impact of knowledge management Practices.	**Delen, D., Zaim, H., Kuzey, C. and Zaim. S. (2013)**	Support Vector Machine, has every one of the highlights like the primary highlights through which the calculations (maximal edge) are describe .The Support Vector Regression (SVR) incorporates or utilizes indistinguishable standards from the SVM with just a few contrasts.
Data Mining: Practical machine learning tools and techniques.	**Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J. (2005).**	ML models rely upon Mathematical conditions and we can normally fathom that it would cause some issue in case in the event that we can keep the Categorical information in the conditions since we would simply require numbers in the conditions.

CHAPTER 3

REGRESSION ANALYSIS

In factual displaying, regression examination could be a quantifiable methodology for assessing the associations among components much expressly, multivariate examination grants one perceive how the standard variable esteem changes while each person of the free factor is moved, while the other autonomous elements are fixed. Normally, relapse appraisal assesses the prohibitive prospect of the needy variable given the free factors that is the standard significance of the reliant variable while the self-governing components are foreordained.[6] An associated yet better way to deal with vital condition assessment that evaluates the most extreme well worth of the needy variable for given extremely worth of the autonomous variable to clarify what estimation of the free factor is pivotal anyway no longer adequate for a given an incentive off the dependent variables.

Regression analysis is extensively used for expectation and assessing, wherein its usage has huge spread with the zone of Machine Learning. Comprehended frameworks like linear regression and standard least squares regression are parametric, in that regression work is portrayed out in articulations of a restricted grouping of cloud parameters that are typical from the data, estimation degeneration insinuates strategies that license the relapse ability to lie in a particular course of action of capacities, which may be perpetual dimensional.[7]

The exhibition of regression analysis methodologies practically speaking relies upon the type of the data creating process, and the manner in which it identifies with the regression approach being utilized.[5] In a few applications, especially with little impacts or questions of connection bolstered observational information, regression ways will give misleading outcomes.

In a narrower sense, regression may allude extraordinarily to the estimation of consistent response factors, rather than the discrete reaction factors used in class. The instance of a continuous yield variable might be additional uniquely known as metric regression to recognize it from related issues.[6]

There are different points of interest of using relapse investigation. They are according to the accompanying:

a. It exhibits the tremendous connections between autonomous factor and ward variable.
b. It shows the nature of effect of various free factors on a reliant variable.

Relapse examination likewise lets in us to analyze the impacts of factors estimated on totally various scales, similar to the aftereffect of significant worth changes and the sort of limited time exercises. Those points of interest help information experts/economic specialists/data researchers to wipe out and contrast the handiest arrangement of variables with be used for structure prescient models**.**

3.1 Types of Regression:

There are a wide range of sorts of regression utilized for forecast analysis model, for example linear regression, logistic regression, polynomial regression, stepwise regression, ridge regression, lasso regression, elastic net regression, and support vector regression.

1. Linear Regression

In measurements, straight relapse is a strategy for displaying the association among a scalar organized variable y and one or additional enlightening factors (or free factors) signified as X. The instance of single informational variable is named as straightforward direct relapse. For a few enlightening variable, the system is called as different direct relapses. (This term is magnificent from variable direct relapse, in which more than one corresponded based factors are foreseen, instead of one single scalar variable.)

In linear regression, the connections are sculptural, the use of linear predictor works whose obscure model parameters are estimated from the information. Such models are insinuated as immediate models. Most typically, the unforeseen mean of y given the cost of X is believed to be a relative typical for X .The middle or some other score of the restrictive dissemination of y given X is communicated as a direct capacity of X. Like a few sorts of regression analysis, direct regression centers around the

conditional possibility dissemination of y given X, as opposed to at the joint likelihood distribution of y and X, that is the area of examination.

Linear Regression was the first type of relapse investigation and to be utilized considerably in reasonable applications. This will be because of models that depend straightly on their obscure parameters are less entangled to fit in models which may be non-directly connected with their parameters and for the reason that factual homes of the resulting estimators are simpler to instructional meeting. Straight relapse has a few reasonable employments. [6] in the unlikely event that the point is desire, or deciding, or mix-up refund, straight relapse is used to fit a judicious structure to a discovered informational collection of y and X esteems. During the time spent building up this sort of adjustment, if an additional estimation of X is, by then given without its going with estimation of y, the fitted model is used to make figure of the estimation of y.

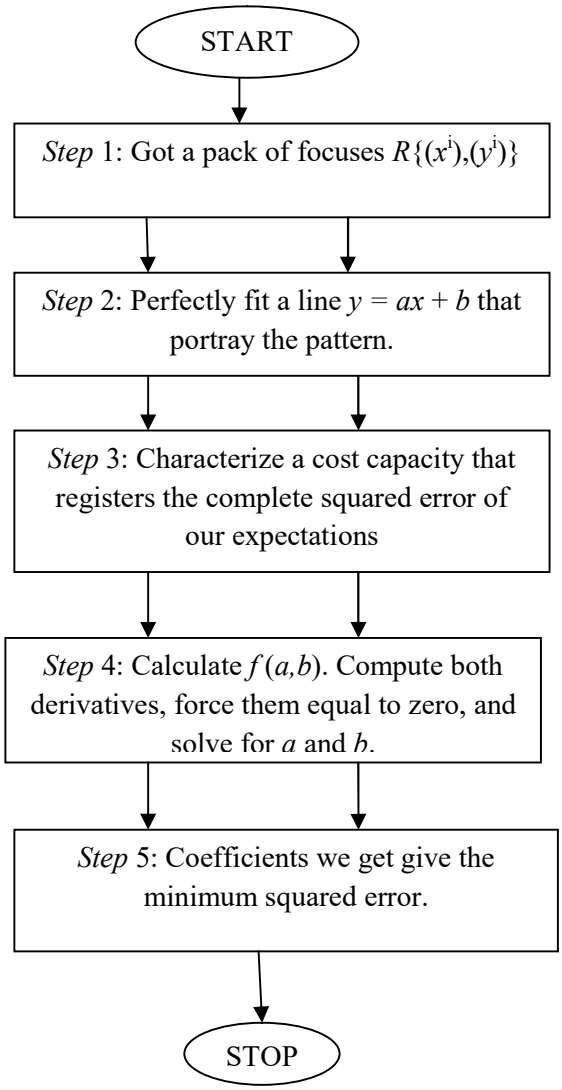


Figure3.1: Algorithm Illustrating Linear Regression Model

Given a variable y and an amount of factors X1, ..., Xp that may likewise be related to y, linear regression investigation can be completed to measure the intensity of the association among y and the Xj, to assess which Xj may have no association with y by any stretch of the imagination, and to choose which subsets of the Xj incorporate excess records around y.

It is the most popular and extensively used modeling approach. Linear regression is most of the primary strategies which are chosen by using human beings at the same time as learning predictive modeling. in this approach, the variable amount is non-stop, impartial variable may be non-stop or wonderful and the nature of regression is linear.[7] It establishes a relationship among one or greater unbiased variable (X) and one structured variables (Y) the usage of a high-quality fit directly line (additionally called regression line).

It's far spoken to by a condition, wherein slope of the road is b, the intercept is a and e is mistake term. This condition might be utilized to foresee the estimation of objective variable dependent on given indicator variables.

The distinction this is observed between linear and multiple regression is that during linear regression there may be best one unbiased variable whereas in multiple regression there are multiple impartial variable.

How to obtain best fit line (Value of a and b)?

Least square method is the fine manner to attain this task. It's far the maximum common technique used for fitting a regression line. With the guide of limiting the sum of the squares of the vertical deviations from each data factor to the line, it ascertains the best-fit line for the input data. Because of the reality the deviations are first squared, while conveyed, there's no offsetting among positive and negative qualities.

$$\min \ ||X\omega - y||_2^2 \qquad (3.2)$$

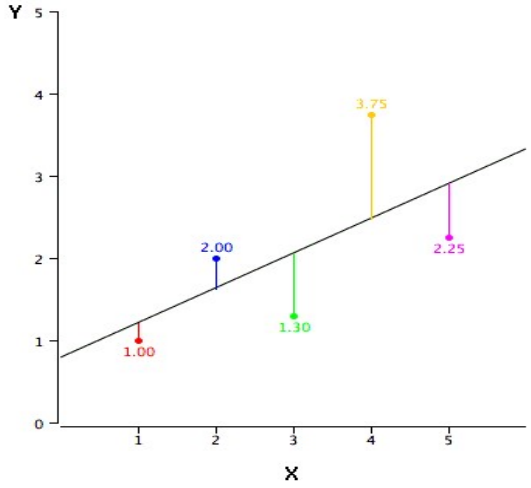


Figure3.2 : Illustrates Regression Error

We can evaluate the model performance using the metric **R-square**.

Important Points:

There ought to be linear relationship among established and impartial variables

Multiple regressions suffer from autocorrelation, multi-co linearity.

Outliers can terribly have an effect on the regression line and in the end the forecasted values.

In case of multiple of impartial variables, we are able to go along with backward removal, forward selection and step wise approach for selection of maximum independent variables.

2. Logistic Regression

Logistic regression is widely used to study the connection between a binary response and a collection of covariates.

In facts, logistic regression is a regression model wherein the dependent variable is specific. One of the case is in which it covers binary structured variables—that is, where it could take only binary values that are two values, "0" and "1", which constitute consequences which includes bypass/fail, win/lose, alive/lifeless or

healthful/ill. Multinomial logistic regressions are cases where the structured variable has more than outcome categories; classes are ordered, in ordinal calculated relapse. In different terms, calculated relapse is an instance of an emotional response/discrete tendency interpretation. Strategic relapse changed into made through expert David Cox in 1958. The parallel key structure is shown to evaluate the probability of a twofold reaction that depends completely on at any rate one marker factors which are independent factors or features. It licenses one to express that the proximity of a threat point extends the opportunity of a given last results by strategies for a specific rate.

An explanation of logistic regression can start with a clarification of the standard strategic capacity. The strategic capacity is helpful on the grounds that it can take any genuine information (t ϵ R), while the output dependably takes esteems somewhere in the range of zero and one and henceforth is interpretable as a likelihood.[9] The Logistic function is

$$h_a(x) = g\left(\sum_{j=0}^{n} a_j\, x_j\right) = g(xa) \tag{3.3}$$

$$g(z) = \frac{1}{1+e^{-z}} \tag{3.4}$$

$$0 \leq g(z) \leq 1, for\ all\ z\ \epsilon\ \mathrm{R} \tag{3.5}$$

$$\lim_{z\to-\infty} g(z) = 0\ and\ \ \lim_{z\to+\infty} g(z) = 1 \tag{3.6}$$

Logistic regression is employed to seek out probability of event=Success and event=Failure. We should reliably use calculated relapse when the dependent variable is parallel (0/1, True/False, Yes/No) in nature. Here is the approximation of Y that ranges from zero to one and it very well may be spoken to by following condition.

odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

$$\text{logit}(p) = \ln(p/(1-p)) = b_0+b_1x_1+b_2x_2+b_3x_3....+b_kx_k$$

Above, p is the probability of presence of the characteristic of interest.

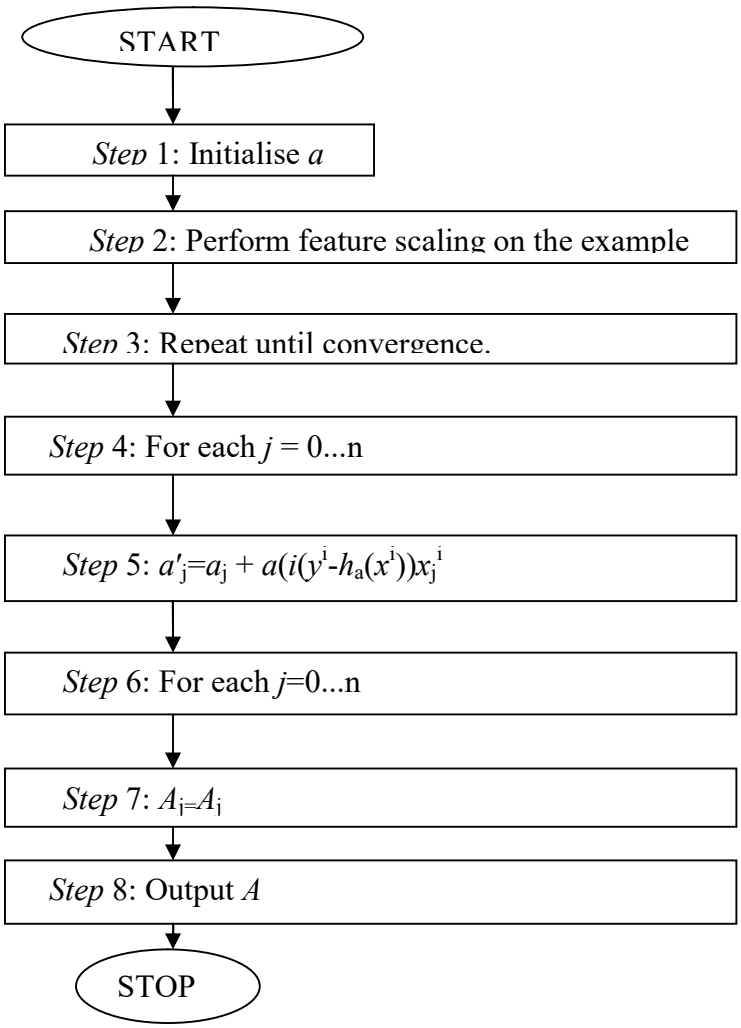


Figure 3. 3: Algorithm Illustrating Logistic Regression Model

Given that we're operating right here with a binomial dispersion (subordinate variable), we have to pick a connection trademark that is best material for this conveyance. Furthermore, it's miles calculated trademark. Inside the condition, the parameters are chosen to amplify the probability of gazing at the example esteems instead of limiting the total of squared mistakes (like in typical relapse).

Important Points

i. Logistic regression is generally utilized for grouping issues.

ii. Linear connection isn't required by logistic regression between dependent and predicted odds ratio it can take care of numerous types of relationships.

iii. To keep away from under fitting and over fitting, we should encompass each and every basic variable. To ensure this preparation is to use a phase astute procedure to assess the calculated relapse.

iv. As most extreme probability gauges are considerably less successful at low example sizes

v. There should be no multi co linearity. But, we've got the alternatives to include interaction consequences of express variables in the evaluation and inside the model.

3. Polynomial Regression

A regression condition is a polynomial regression condition if the quality of fair factor is more noteworthy than 1. The condition beneath speaks to a polynomial condition:

$$y = a + bx^2 \qquad (3.7)$$

In this regression technique, the best fit line is a curve that fits into the data points rather than a straight line. [4]

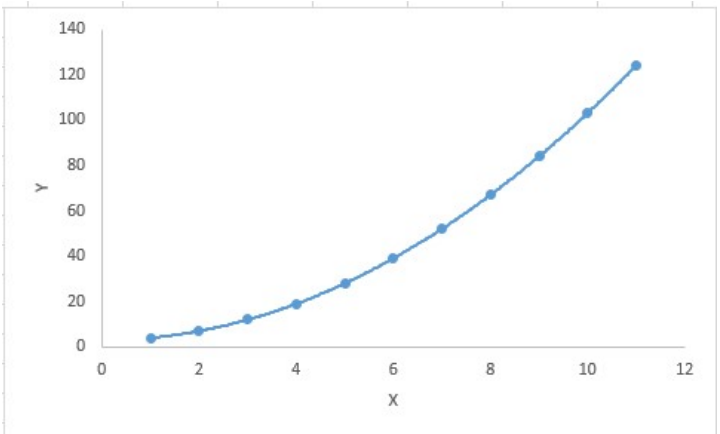


Figure3.4: Illustrates Polynomial Regression

Important Points:

While there may be a compulsion to fit as a fiddle a higher recognition polynomial to get lower errors, this can result in over-fitting. Continually one should plot the connections to see the fit and spotlight on ensuring that the bend suits the idea of the issue. Here is an occurrence of the way plotting can help.

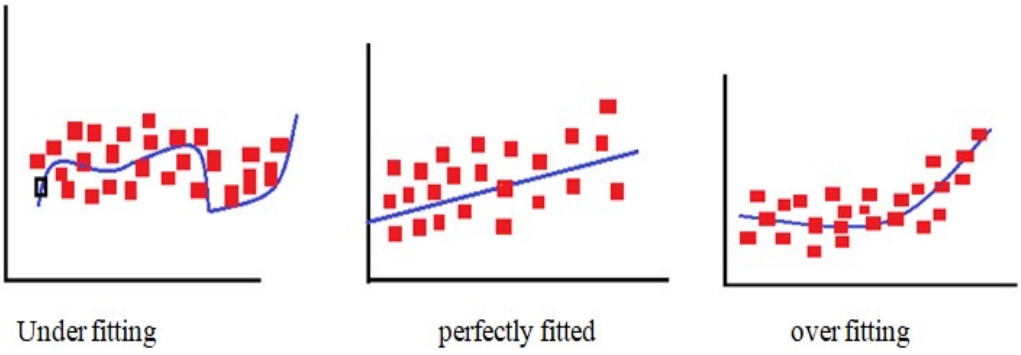


Figure3.5 : Illustrates Types of Plotting **[3]**

- Especially one should look out to bend towards the terminations and see whether those shapes and examples look good. Higher polynomials can wrap up conveying odd results on extrapolation.

5. Stepwise Regression

Even as handling a couple of independent variables, this form of regression may be very efficient. On this method, an automatic procedure, which includes no human intervention, is used for the selection of independent variables.

This accomplishment is practiced by method for taking a gander at measurable qualities like t-stats, AIC metric and R-square to observe fundamental factors. Stepwise regression fundamentally fits the regression form by methods for including/dropping covariates independently upheld an ostensible model. Some of

routinely used Stepwise relapse approaches are recorded underneath: Standard stepwise backslide finishes two things. [9] It incorporates and empties markers as required for every movement.

- When the choice begins with most huge indicator in the model and includes variable for each progression it is known as Forward Selection.
- When the procedure begins with all indicators in the model and expels the least noteworthy variable for each progression it is known as Backward Elimination.

The goal of this showing approach is to extend the conjecture control with unimportant number of pointer factors. It's far one of the techniques to address better dimensionality of informational index.[4]

5. Ridge Regression

Ridge Regression is a method that's widely used while the information suffers from multi-co dimensionality (unbiased variables are relatively correlated). In multi-co linearity, even supposing the least squares estimates (OLS) are independent, their variances are huge which deviates the decided fee some distance from the real value. With the aid of including a degree of bias to the regression estimates, ridge regression reduces the same old errors. [4]

As we know the linear equation can be described as-

$$y = a + bx \tag{3.8}$$

The equation also has an error term. Thus complete equation becomes:

$$y = a + bx + e \tag{3.9}$$

where "*e"* is error term.

Error term is the value required to correct an error between the observed and predicted value.

$y = a + y = a + b_1 x_1 + b_2 x_2 + \cdots + e$ for multiple independent variables. (3.10)

Prediction mistakes can be decomposed into two sub components whilst working on linear regression. First is because of the bias and second is because of the variance. Prediction error can occur due to anyone of those two or each additives. Here we can speak approximately the errors caused by variance.[9]

Ridge regression solves the multi-co linearity hassle through shrinkage parameter λ (lambda). Inside the equation below-

$$= argmin\ ||y - X\beta||_2^2 + \lambda||\beta||_2^2 \quad (3.11)$$

We have two segments. Initial one is least square term period and diverse one is lambda of the summation of β2 (beta-square) in which β is the coefficient. that is conveyed to least square term period with a reason to curtail the parameter to have an extremely low change.

Important Points:

i. The suspicion of this regression is indistinguishable as least squared regression aside from ordinariness isn't to be expected.
ii. It recoils the estimation of coefficients however doesn't accomplish zero, which demonstrates no capacity decision include.
iii. This is a regularization strategy and utilizations L2 regularization.

6. Lasso Regression

Lasso Regression (Least Absolute Shrinkage and selection Operator) also penalizes absolutely the length of the regression coefficients. further, It's capable to enhancing the accuracy of linear regression models.

$$argmin\left|\left|y - X\beta\right|\right|_2^2 + \lambda||\beta||_2 \quad (3.12)$$

Equation (4.12) shows that Lasso regression uses absolute values within the penalty characteristic, as opposed to squares and that's the way it differs from ridge regression in a manner that it. This often ends in penalizing values which reasons a number of the parameter estimates to show out precisely zero. The greater the penalty carried out, in addition the estimates get reduced closer to absolute 0. This results to selection of variables out of n variables.

Important Points:

i. In this regression normality can't be assumed, rest all the assumptions are same as least square regression.
ii. It helps in feature selection as it shrinks its coefficients to zero.
iii. This technique is a regularization strategy and utilizations L2 technique.
iv. If gathering of indicators are exceptionally associated, rope picks just a solitary one of the indicators if gathering of indicators are exceedingly co related and contracts the others to zero.

7. Elastic Net Regression:

Elastic Net is hybrid combination of Ridge and Lasso Regression techniques. It is trained with L1 and L2 prior as regularizer. When there are multiple options which are correlated Elastic Net play an essential role. While lasso might pick only one elastic net tends to pick both.

$$\beta = argmin(\left|\left|y - X\beta\right|\right|^2 + \lambda_2 \left|\left|\beta\right|\right|^2 + \lambda_1 \left|\left|\beta\right|\right|_1) \quad (3.13)$$

A reasonable favorable position of exchanging off among Lasso and Ridge is that, it enables Elastic-Net to acquire a portion of Ridge's dependability under revolution.

Important Points:

i. In case of especially correlated variables it encourages group effect.
ii. No limitations on the wide variety of selected variables.

iii. It could go through with double shrinkage.

Past these seven most normally utilized regression strategies, there are different models like Bayesian, Ecological and Robust regression.

8. Support Vector Regression:

Support Vector Machine, has every one of the highlights like the primary highlights through which the calculations (maximal edge) are describe .The Support Vector Regression (SVR) incorporates or utilizes indistinguishable standards from the SVM with just a couple of contrasts. Above all else, it turns out to be hard for the forecast of data within reach, since result is a genuine number which has unbounded conceivable outcomes. Be that as it may, the calculation is progressively convoluted consequently to be considered. In any case, the principle thought is: To limit the blunder, boosts the edge by individualizing the hyper-plane, keeping mistakes into record.

Support Vector Machine is applied as regression also and is not limited to classification only. Although it includes all the principle features that represent maximum margin algorithms. a non-linear function is leaned by means of linear machine learning of device mapping into high dimensional kernel prompted function. The device potential is controlled through parameters that do not depend on the dimensionality of feature space. in the same manner as with class approach the primary idea is to searching for and optimize the generalization bounds mentioned for regression. The primary concept is to depend on defining the loss function that ignores mistakes, that are located within the sure variety or distance of the authentic value.[9] This kind of characteristic is often referred to as – epsilon extensive – loss characteristic. maximum Margin classification with SVM- every other broadly used and powerful studying algorithm is help vector machine (SVM) which may be considered as an extension of the perceptron. using the perceptron set of rules from neural network, we reduce misclassification mistakes. In SVM our optimization goal is to maximize the margin. The margin is described as the gap between the isolating hyper plane and training samples which might be closest to hyper aircraft, that's called aid vectors.

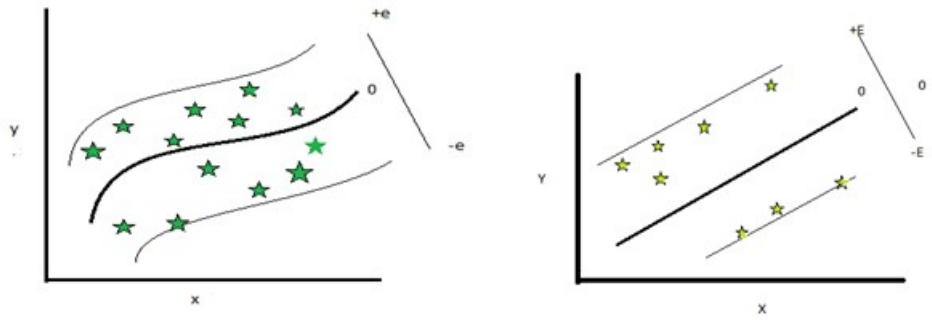


Figure 3. 6: Illustrates Non Linear and Linear SVM Classification

How to select the right regression model?

It is easier to work when we know only one or two techniques to apply as it cuts down the efforts of other possibilities. After studying different papers, we came across the idea of what results can be expected after applying different regressions. We can expect continuous graph after linear regression. If it is zero or one, or if we have to compute probability of an event logistic regression must be used. However, the more number of options available leads to more difficult judgment, to choose the right one. A similar case happens with regression models.

As there are more than one varieties of regression model, it's vital to pick the most appropriate approach based totally on kind of dependent and independent variables, dimensionality within the statistics and other important traits of the information. [8]

Below are the important thing factors that you ought to exercise to

i. Information examination is an unavoidable bit of structure judicious model. It should be the underlying advance before picking the most ideal model like recognizes the relationship and impact of variables.

ii. We can examine different estimations like accurate significance of parameters, R-square, BIC, AIC, Adjusted r-square and botch term in order to take a gander at changed models.

iii. The most perfect way to deal with survey models used for conjecture is to cross endorsement. here the assurances is part into social event (getting ready and test). A clear mean squared differentiation between the watched and anticipated qualities gives us a degree for the estimate precision.

iv. It can happen that a less stunning model is definitely not hard to complete when stood out from an all around really tremendous model as it just depends upon our objective.

v. Regression regularization systems (Lasso, Ridge and ElasticNet) works outstandingly if there ought to emerge an event of high dimensionality and multi colinearity among an impressive parcel of the components in the educational record.

In our project, we have used Linear, SVR, Lasso and Ridge Regression.

Why SVR over Linear Regression?

In simple regression we try to minimize the error rate. While in SVR we try to fit the error within a certain threshold. Support Vector Machine, has all the features like SVR, the main features through which the algorithms (maximal margin) are characterize.SVR works on same principle as SVM works with only a few differences. In the same way as with classification approach the main idea is to seek and optimize the generalization bounds mentioned for regression.

In this project we are using the following regression techniques namely-

Linear Regression, Support Vector Regression, Ridge Regression and Lasso Regression.

CHAPTER 4

IMPLEMENTATION OF PREDICTIVE ANALYSIS

The main aim of our project is to find the predictive model for absenteeism. We have used different steps consisting data collection, preprocessing, feature engineering, feature selection, implementation of regression for implementation the predictive analysis of Absenteeism shown in Figure 4.1. All the simulation was carried out in Jupyter Notebook using Python language.

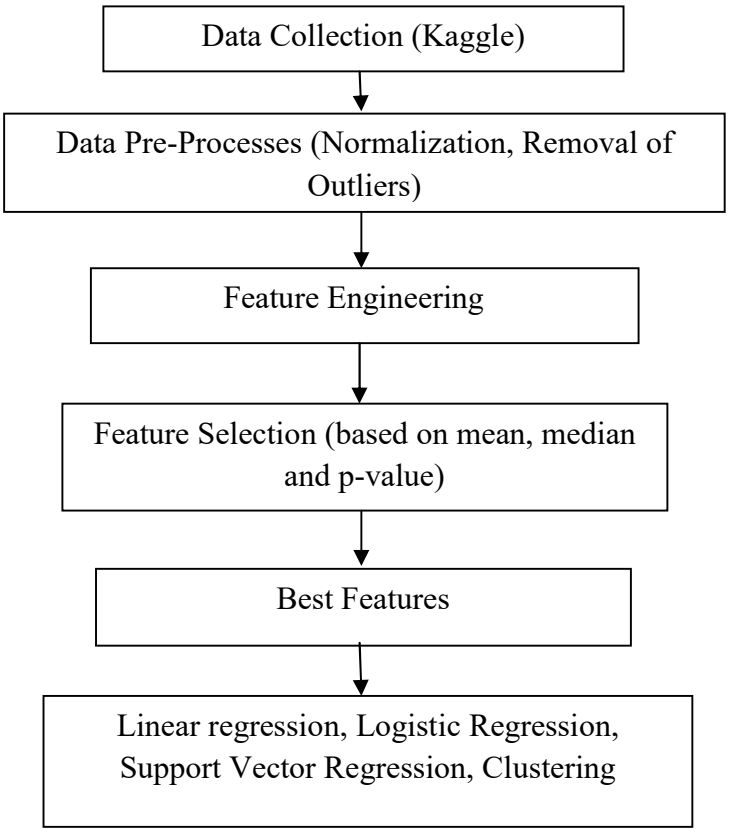


Figure4.1: Predictive Analysis Mode

4.1 Data Collection:

In data collection process we collect the dataset on which we have to apply predictive analysis and regression techniques .Here we have collected the data set from kaggle.com, the dataset includes 22 factors on which absenteeism in MNCs depend and 786 observations.

absent – absent (0 = No; 1 = Yes), e class – Employee Class (1 = 1st; 2 = 2nd; 3 = 3rd), name – Name, sex – Sex , age – Age.

4.2 Data Pre-processing:

Data pre-preprocessing is an information mining approach that includes adjusting crude measurements into an intelligible design as clarified in Figure 4.2. Certifiable information is regularly inadequate, conflicting, and additionally missing in positive practices or qualities, and is probably going to incorporate numerous mistakes. Information pre-handling is a method of settling such inconveniences.[5]

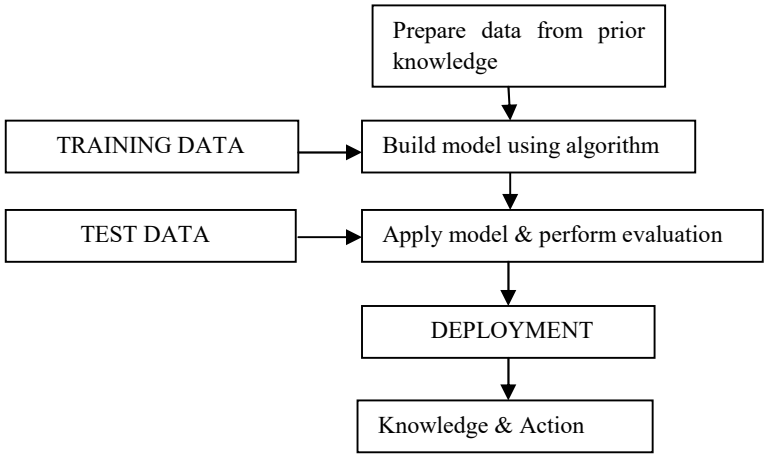


Figure4. 2**: Pre-processing using Data mining process**

Figure 4.2 elaborates the data mining process that we have incorporated in our project. It explains how we started off by applying regression techniques to our data set in order to build a model. We split the data set into training data set and testing data set in the ratio of 70:30 and using training data set, we came up with a

model to be deployed. To this model, we applied test data set and came up with a predictive model. [6]

In real world, data are generally incomplete, noisy and inconsistent.

ALGORITHM 1

Input: Input from dataset

Output: Pre processed Image

BEGIN

Step 1: Import the libraries.

Step 2: Import the informational indexes.

Step 3: Check out the missing qualities**.**

Step 4: Check for all out qualities.

Step 5: Splitting the informational collections into preparing informational collection and test informational index.

Step 6: Apply feature scaling

END

4.3 Feature Engineering

Since the data can have missing fields, incomplete fields, or fields containing unknown information, a fundamental step in building any prediction system is Feature Engineering. For instance, the fields Age, Fare, in the training and test data, had missing values that had to be packed in. The field "Name" while being of no uses itself, contained employee's Title (Mr., Mrs., etc.), we also used

employee's class and age to figure out various reasons for absenteeism. Most of the times, our data set contains features rather differing in sizes, units and range.[7] In any case, taking into account that, limit of the AI calculations use Eucledian separate between two variables of their calculations, this is an issue.

Since name is unique for each employee, it is not useful for our prediction system. However, an employee's title can be extracted from his or her name. We found 3 titles:

Table 1: Title from Name The field Name

Index	Title	No of Occurrence
1.	MR	757
2.	MRS	198
3.	MS	2

.Title indicates employee's sex (Mr. And Mrs.) and age (Ms. And Mrs.)

4.4 Feature Selection / Data Exploration

On this phase we're going to discover the dataset. This is an critical step inside the machine learning procedure as first off we want to recognize more information approximately the information we're the usage of and secondly we need to make a few changes to the data itself. Feature choice is being performed on three elements specifically p value, median and mean of information set.

p value- *p*-value encourages you decide the noteworthiness of your outcomes. The *p*-value is a number somewhere in the range of 0 and 1 and translated in the accompanying manner: A little *p*-value (normally ≤ 0.05) demonstrates solid proof against the invalid speculation, so you dismiss the invalid theory.

Median- The median is one of many measures of central tendency.

Mean- The mean is the sum-total average of the dataset.

On the basis of feature selection we have found that only five features out of 21 are used further analysis for our data set.

We are dropping ‘Employee Id’and ‘Name’. The reasons for these are

- Employee Id – This is a completely unique incrementing quantity, it must not influence the ‘risk’ isn't any use.
- Name – It is not for our concern whether a person named Daniel is more likely to be present as compared to Mike.

4.5 Regression Analysis:

Regression analysis is a measurable procedure for evaluating the connections between factors. All the more explicitly, relapse examination empowers one secure how the standard estimation of the organized variable (or 'rule variable') adjustments while any of the fair factors is various, while the contrary free factors are consistent.

4.5.1: Linear Regression:

Linear regression is a strategy for showing the association between a scalar ward variable y and in any event one useful elements (or autonomous components) demonstrated as X. The case of one consistent variable is called straightforward linear regression. For more than one useful variable, the methodology is called multiple linear regression.

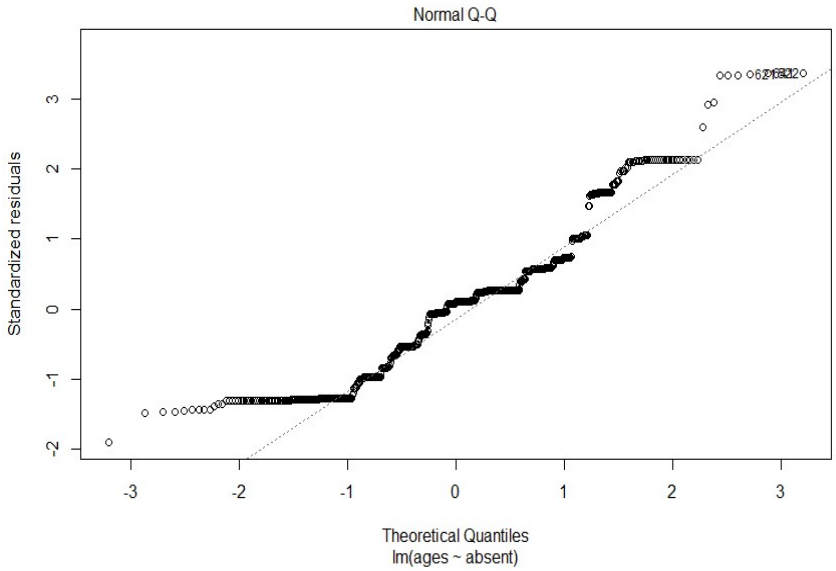


Figure4.3: Result of performing Linear Regression on our data set**.**

In Figure 4.3 , it is depicting Linear Regression with **age** of the employee being out predictor variable and **number of days our employee is absent** as our response variable. We performed linear regression of all the parameters given to us, and found out that absenteeism is linearly varying with age parameter only.

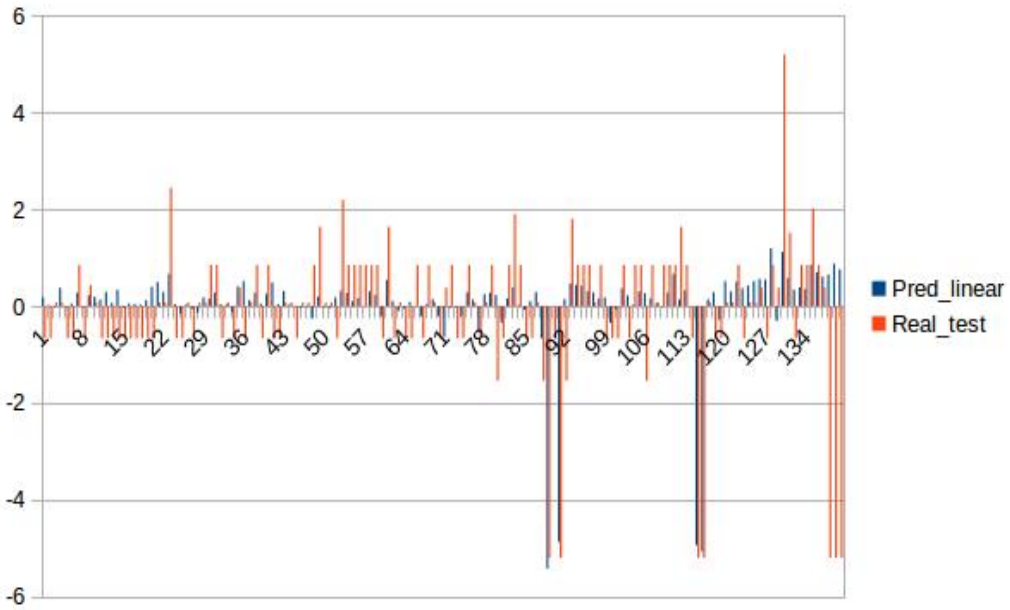


Figure 4.4 : Graph between predicted and real values using linear regression

DISADVANTAGE OF LINEAR REGRESSION

Linear regression is the best model connections among reliant and autonomous factors which may be direct. It accept there is a straight-line connection between them which isn't right sometimes. Linear regression could be sensitive to the inconsistencies inside the data (or exceptions). Consequently we moved to SVR.

4.5.2: SVR

The main idea of SVR is to seek and optimize the generalization bounds mentioned for regression. The main idea is to rely on defining the loss function that ignores errors, which are located within the certain range or distance of the true value.

We have used SVR on our datasets. For SVR we have considered 3 parameters - Age, seasons of the year and days of the week.

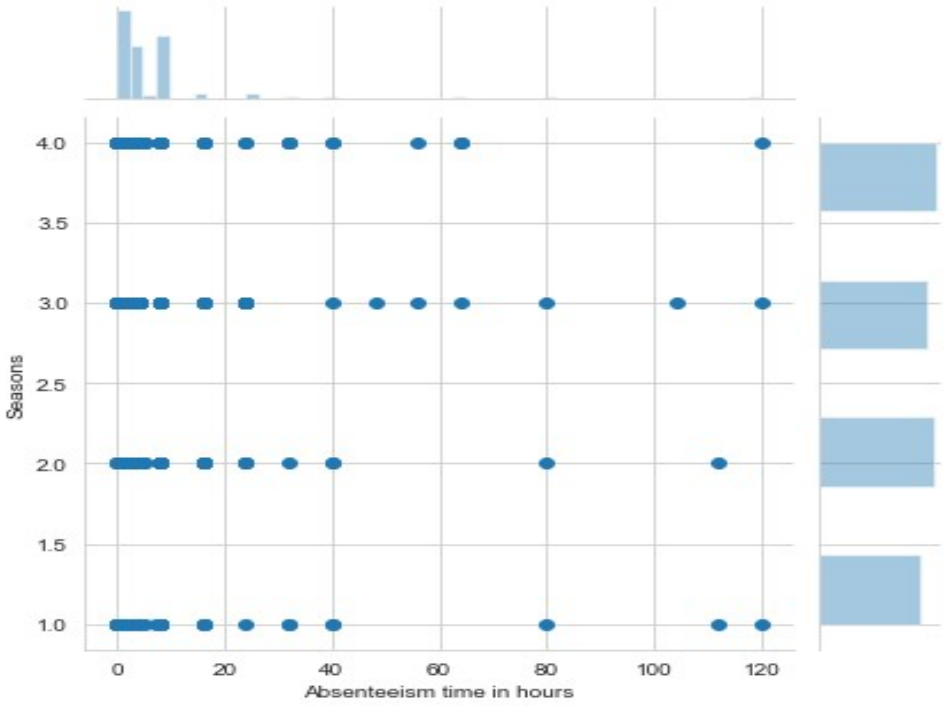


Figure4.5: Graph between seasons of the year Vs absenteeism using SVR

In Figure 4.5, it is depicting that we have divided the year into 4 major seasons with a interval of 0.5 but as shown in above graph we can infer that absenteeism rate does not get much affected by a particular season of the year. The absenteeism rate remains unaffected by this factor as the absenteeism rate is almost equal in all seasons.

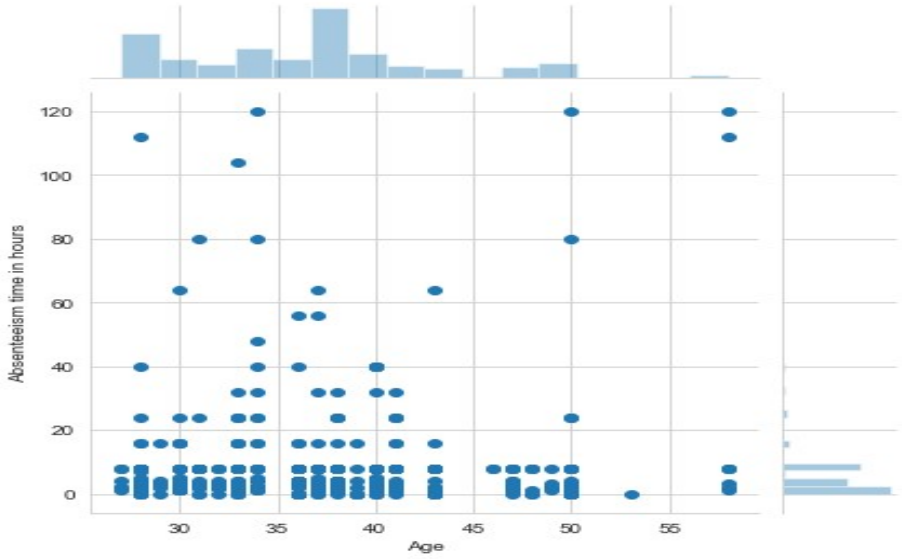


Figure4.6: Graph between Age Vs Absenteeism using SVR

In Figure 4.6, it is depicting that the datasets are normalized and SVR is implemented the absenteeism is not linearly related to age. As we can infer from the graph the age group of 35-40 years have the highest absenteeism rate as compared to the age group of 50-60.

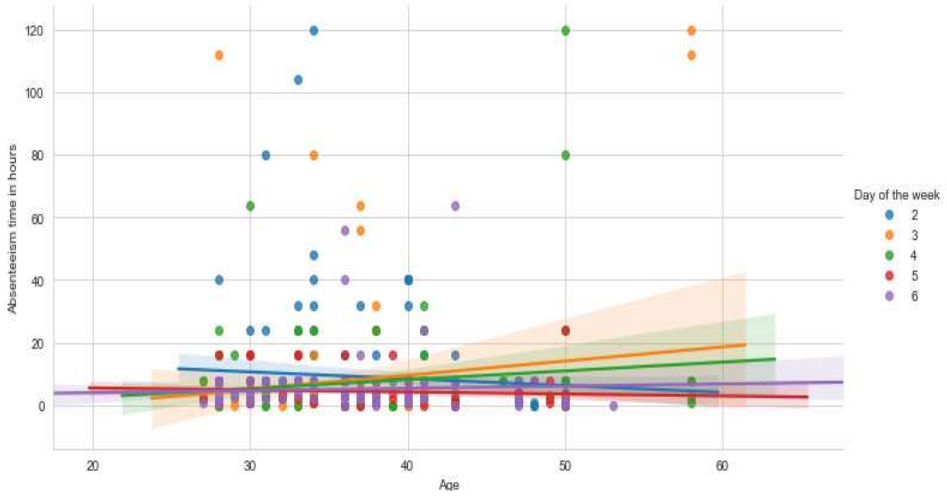


Figure4.7: Graph between age and absenteeism in hours taking consideration of the days of the week.

In Figure 4.7, is depicting with Age parameter, days of the week with hours is also considered.

Graph depicts the absenteeism on different days of the week, Here each day starting from Monday to Saturday is assigned a different color and the gradual decrease in absenteeism on 2nd day of the week is observed as the age increases and the absenteeism on the 6th day i.e Saturday is fairly high than any other day, reason being the end of the week. In fig, 2 defines 2nd day, 3 defines the 3rd day of the week.

The most significant issue with SVMs is the high algorithmic intricacy and broad memory prerequisites of the required quadratic programming in enormous scale errands. The choice of part is one more issue that happens in SVR.

4.5.3: Lasso regression

Lasso (Least Absolute Shrinkage and Selection Operator) likewise punishes indisputably the measure of the relapse coefficients. Furthermore, it is fit for improving the precision of direct relapse models.

Figure. 4.8 show the predicted results from training data and real results from test data and comparison between them when lasso regression is applied on dataset.

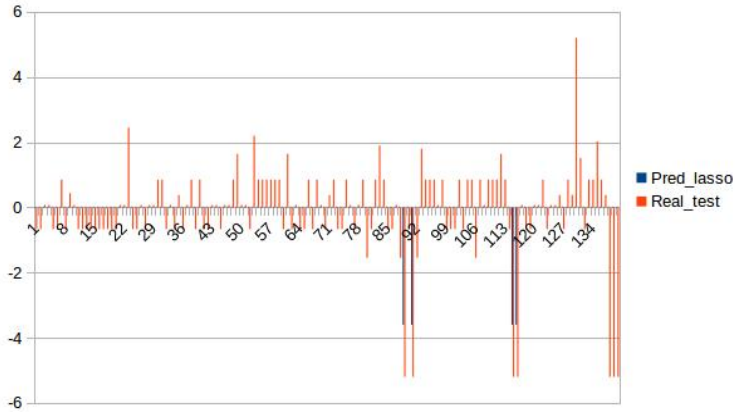


Figure 4. 8: Graph between predicted and real values using lasso regression

4.5.4: Ridge regression

Ridge Regression is a technique used when the data encounters multi-co linearity (self-sufficient variables are exceedingly related). In multi-co linearity, regardless of the way that the least squares checks (OLS) are unbiased, their progressions are colossal which veers off the viewed an impetus far from the certified regard.

Figure 4.9 shows the predicted results from training data and real results from test data and comparison between them when lasso regression is applied on dataset.

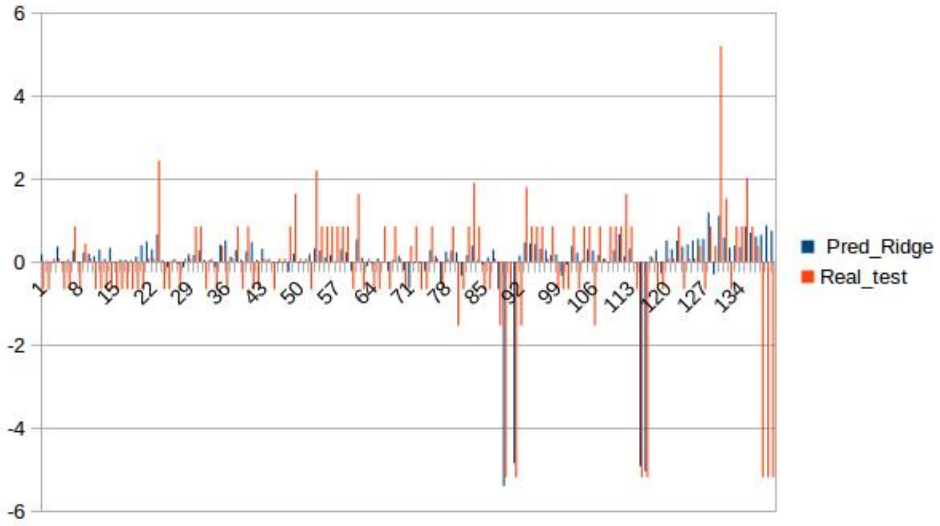


Figure4.9: Graph between predicted and real values using ridge regression

Table 2: Results considering different parameters

	Evaluation of error function for linear regression	**Evaluation of error function for ridge regression**	**Evaluation of error function for lasso regression**
Mean squared error	1.407	1.405	1.491
Mean absolute error	0.75437	0.75438	0.83806
R squared value	0.3528	**0.3534**	0.3138
Median absolute error	0.608	0.613	0.688

From comparison of linear regression, lasso regression and ridge regression on different factors like mean squared error, mean absolute error, r square value and median absolute error, we find that as r square value of ridge is more efficient than linear or lasso regression. Therefore we came to conclusion that ridge regression will yield the best predictive result in this case.

CONCLUSION AND FUTURE WORK

The motivation of this study developed from a desire to learn, understand, and apply Linear and Logistic algorithm. *Predictive Analysis of Absenteeism in MNCs*, served as a framework for introductory predictive analytic methods. The problem posed in this project was to build a predictive model to predict the absenteeism and avoid it by studying a data set from Kaggle. A significant amount of time and effort was spent in cleaning, organizing, and redefining variables in the data. We successfully implemented the Data Cleaning as well as advancements of this methodology such as calculation of ROC. We used the prediction accuracy, provided by Kaggle, to assess the efficacy of our model in correctly classifying the absenteeism in MNCs.

We obtained two sets of results coming from two different approaches for handling the data. We consider the first analysis (linear) to be the one which required minimal adjustments to the data provided by Kaggle.

From the experiments we come to know that Age parameter of the age group of 35-40 linearly related to absenteeism and the maximum absenteeism is on the 6th day of the week and minimum absenteeism is on the 2nd day of the week. There is no much affect of season on absenteeism.

We hypothesize that this phenomenon could be attributed to the structure of the data. That is, we believe there is more to the data than what we have discovered and the data may require additional restructuring to obtain further improvement when using the methods that have been covered.

It is apparent that the effort put forth when working on the *Absenteeism in MNCs* problem has achieved our aims and goals of this study. In attempts to provide a superior solution and further our knowledge of predictive analytics, future work that can be applied to this problem involves learning how to apply methods such as feature engineering.

PAPER PUBLISHED

Krittika , Shriya Vandita, Shruti Jain, " Predictive Analysis of Absenteeism in MNCs using Machine Learning Algorithm", 2nd International Conference on Recent Innovations in Computing (ICRIC-2019), March 8 – 9, 2019, Central University of Jammu, J & K.

Science and Engineering Research Board (SERB)
Department of Science and Technology (DST)
Govt. of India

ICRIC-2019

Springer
Lecture Notes in Electrical Engineering (LNEE)

Second International Conference on Recent Innovations in Computing

ICRIC-2019
March 08-09, 2019
(Sponsored by SERB, DST, Govt. of India)
Organized by
Department of Computer Science & Information Technology
CENTRAL UNIVERSITY OF JAMMU, J&K, INDIA

Certificate

This is to certify that Prof./Dr./Mr./Ms Krittika, Shriya Vandita & Shruti Jain of Jaypee University of Information Technology, Solan participated/presented a paper entitled Predictive Analysis of Absenteeism in MNCs using Machine Learning Algorithm during 2nd International Conference on Recent Innovations in Computing organised by Department of Computer Science & Information Technology, Central University of Jammu, Rahya Suchani (Bagla), Samba, J & K, India.

Dr. Yashwant Singh
General Chair

Prof. Devanand
Co-Patron

REFERENCES

[1] Finlay "Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods″, 1st edition, New York: Palgrave Macmillan, 2014.

[2] Abbott, "Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst″, 1st edition, Indianapolis: John Wiley & Sons, Inc, 2014.

[3] "Spark Lighting fast computing cluster computing". http://spark.apache.org

[4] "GraphLab Create: A Python-based machine learning platform". http://graphlab.org

[5] "Spark Programming Guide". https://spark.apache.org/docs/latest/ programming-guide.html

[6] Y. Ma, "Remote sensing big data computing: Challenges and opportunities". *Future Generation Computer. System.* vol. 51, pp. 47–60, 2015.

[7] Z. Sun, F. Chen, M. Chi, and Y. Zhu, "A spark-based big data platform for massive remote sensing data processing," in *Data Science*, New York, NY, USA: Springer, pp. 120–126 2015.

[8] Zhang, L., Tan, J., Han, D. and Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery; *Drug discovery today*, 22(11), 1680-1685.

[9] Delen, D., Zaim, H., Kuzey, C. and Zaim. S. (2013). A comparative analysis of machine learning systems for measuring the impact of knowledge management Practices; *Decision Support Systems,* 54(2), 1150-1160.