IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 6, June 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106034 |

Sales Prediction Using Machine Learning Techniques and Measuring Parameters

Rajesh Kumar Vishwakarma, Dr. Jitendra Sheetlani

Ph.D. Research Scholar, SSSUTMS, Sehore, India

Associate Professor, SSSUTMS, Sehore, India

ABSTRACT*:* A name sales which is acquiring every industry in today's era. Every service and product need a good Advertising and Selling team to properly advertise and sell their service or product. Today the use of data science can be seen in any of the fields such as education, health, sales, entertainment, and many more. Data Science/ Machine Learning/ Deep learning these terms are playing a very important role in analysing and predicting sales according to the particulars of product, service, area, etc. Data Science is playing a vital role in predicting customer behaviour over products and services. Data Science is transforming every aspect of products and services on every step of the same. From predicting and analysing the needs of the customer to sales analysis and prediction data science is used everywhere in the industry. That's the reason why data science has become a huge contributor to the wellness of customers and industries. Data science helped companies to overcome the common issues with their traditional approach of predicting sales and analysing sales over different particulars. This helped them a lot in getting good customers as well as in planning their future in the industry, side by side this helped in coping up their competitors in the market. Mostly the effect can be seen in the way of approaching one product on different dependent particulars. So, all over this concludes that analysing and predicting the sales with good accuracy is very important. So, aim for this is we are building a system for predicting sales with good accuracy and a unique approach to predicting sales. Using different models and it is carry out the comparative analysis and pick and best one out, using feature engineering to increase the efficiency and accuracy of the model.

KEYWORDS: - Prediction, Sales, Machine Learning, Deep Learning, Forecasting

I. INTRODUCTION

Now a day there are large amount of reviews available online on numerous e-commerce sites. Besides offering products to the customers these websites also take feedback from these customers to know if the product is liked or disliked by the customer, by the help of these feedback (Reviews) a customer needing to buy a specific product can know the insight of all the products present on an e-commerce website. In the beginning only, some of the websites showed these feedback to the customer but as the time passed customers start checking the reviews of a product and dependent on its rating and comments decide whether to buy a product or not, In one of the surveys, it's discussed that how negative reviews effects the decision of customer to not buy a product, many initiatives are taken by the brand to keep their customers happy and likewise the reviews positive [1].

In their process of buying/comparing a product, they need to find useful information as quickly as possible, but Amazon made it easy adding a field rating of every single review by its side which is measured on the scale of five. How much a customer likes/recommends a product to anyone else. According to the marks given out of 5 if a product gets 4.0 or above means it's a very good product and the customer likes that product. In addition, if some other customer is searching for reviews and found any of the reviews which is helpful for him/her and he clicks on helpful button. A different field is populated i.e. how many people finds a review helpful. However, product reviews (from other that amazon) are not accompanied by 5-star rating and helpful to people system [2].

All these reviews are given by a genuine customer as only the customer who bought products from Amazon can give a review on that product, but no other person can which is also specified in every review. However, it is tough to predict the sales in coming month or week of that product without considering the reviews of these people, if there is problem in even one model then people may stop buying other models of cell phones too. Therefore, models able to predict the sales using ratings and helpful to people is very important. Getting all supporting columns which may cause variance in sales of a product are to be considered [3].

Nevertheless, this predictive analysis come with some challenges as the sales are not just dependent on the reviews but many more aspects which may come from the people perspective or may be some changes in a specific brand. As all users have different perspective and does not see a product in the same way. In this research a single year's data is used

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106034 |

which is of 9 different cell phone models of OnePlus and likewise the sales of each model are different fromeach other (price, specifications, camera) and with the launch of every new product having high end finishing and new features increases its sales and decreases the sales of old model. This research is not model specific but brand specific as OnePlus sell cell phones and no other product so only the cell phone review's data is considered.
The question that arises is how this data can be predicted using reviews. One of the solutions is to rely on machine learning algorithms which will help in predicting the total sales of these products considering other values of the same review and how efficiently that helps in predicting the sales. For an instance, depending on the launch date (month) of a new product the sales may increase also if the rating given by many customers is above 4.0 out of 5. While on the other hand rating is 2.0 or below tend to decrease in sale, but that totally depends on the sum of rating in that month given by customers. If average rating is good, then sales always tends to increase. This dissertation uses data science (polynomial regression) technique for the sales prediction of amazing mart and the analysis of this technique is done in Python programming language and Spyder tool of data science is used. Data Science is the mixture of various algorithms, tools, data inference, various machine learning principle and technology in order to solve different analytics complex problems. [4]

- The main aim of data science is to find the useful insights from the data and the main thing is to mine the raw data and discover the various trends and complex behavior of data.
- Data science is just like taking out insights from the data that can help the various companies to make better business-related decisions that can give the company a very nice boost.
- Data Science involves using automated methods to analyse massive amounts of data and to extract knowledge from them.

Statement of the Problem
In this work, looking at the various Stores Sales around the world are tasked with predicting their daily sales in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of sales records we will be predicting sales based on their unique circumstances, the accuracy of results can be quite varied, The task is to forecast the "Sales" column.

II. RELATED WORK

Kumari Punam et al., (2018) Prediction ofsales of a product from a particular outlet is performed via atwo-level approach that produces better predictiveperformance compared to any of the popular single modelpredictive learning algorithms. The approach is performed onBig Mart Sales data of the year 2013. Data exploration, datatransformation and feature engineering play a vital role inpredicting accurate results. The result demonstrated that thetwo-level statistical approach performed better than a singlemodel approach as the former provided more information thatleads to better prediction [5].

Yanrong Ni, Feiya Fan (2011)The difficulty with fashion retail forecasting is due to a number of factors such as the season, region andfashion effect and causes a nonlinear change in the original sales rules. To improve the accuracy of fashionretail forecasting, a two-stage dynamic forecasting model is proposed, which is combined with bothlong-term and short-term predictions. The model introduces the improved adjustment methods, themain adjustment model and error forecasting model in the adjustment system collaborated with eachother. The real-time data are demonstrated by applying the model in wireless mobile environment.The experiment shows that the model provides good results for fashion retail forecasting [6].

Kilimci et al. (2019) an intelligent demand forecasting systemis developed. This improved model is based on the analysis and interpretation of the historical data by using different forecastingmethods which include time series analysis techniques, support vector regression algorithm, and deep learningmodels. To the bestof our knowledge, this is the first study to blend the deep learning methodology, support vector regression algorithm, and differenttime series analysis models by a novel decision integration strategy for demand forecasting approach.The other novelty of this workis the adaptation of boosting ensemble strategy to demand forecasting system by implementing a novel decision integration model. The developed system is applied and tested on real life data obtained from SOK Market in Turkey which operates as a fast-growing company with 6700 stores, 1500 products, and 23 distribution centers. A wide range of comparative and extensive experiments demonstrate that the proposed demand forecasting system exhibits noteworthy results compared to the state-of-art studies. Unlike the state-of-art studies, inclusion of support vector

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106034 |

regression, deep learning model, and a novel integration strategy to the proposed forecasting system ensures significant accuracy improvement [7].

Mohit Gurnani et al. (2017) evaluates andcompares various machine learning models, namely, ARIMA, Auto Regressive Neural Network(ARNN), XGBoost, SVM, Hy-bridModels like Hybrid ARIMA-ARNN, Hybrid ARIMA-XGBoost,Hybrid ARIMA-SVM and STL Decomposition (using ARIMA,Snaive, XGBoost) to forecast sales of a drug store company called Rossmann. Training data set contains past sales and supplemental information about drug stores. Accuracy of these models ismeasured by metrics such as MAE and RMSE. Initially, linearmodel such as ARIMA has been applied to forecast sales. ARIMAwas not able to capture nonlinear patterns precisely, hencenonlinear models such as Neural Network, XGBoost and SVM wereused. Nonlinear models performed better than ARIMA and gave lowRMSE. Then, to further optimize the performance, composite models were designed using hybrid technique and decomposition technique. Hybrid ARIMA-ARNN, Hybrid ARIMA-XGBoost,Hybrid ARIMA-SVM were used and all of them performed betterthan their respective individual models. Then, the composite modelwas designed using STL Decomposition where the decomposed components namely seasonal, trend and remainder componentswere forecasted by Snaive, ARIMA and XGBoost. STL gave better results than individual and hybrid models. This paper evaluates andanalyzes why composite models give better results than an individual model and state that decomposition technique is betterthan the hybrid technique for this application [8].

Sahar and Ramaswamy (2019), used a combination of institutional, academic, demographic, psychological and economic factors to predict students' performances using a multi-layered neural network (NN) to classify students' degrees into either a good or basic degree class. To our knowledge, the usage of such an extended profile is novel. A feed-forward network with 100 nodes in the hidden layer trained using Levenberg-Marquardt learning algorithm was able to achieve the best performance with an average classification accuracy of 83.7%, sensitivity of 77.37%, specificity of 85.16%, Positive Predictive Value of 94.04%, and Negative Predictive Value of 50.93%. The NN model was also compared against other classifiers specifically k-Nearest Neighbour, Decision Tree and Support Vector Machine on the same dataset using the same features. The results indicate that the NN outperforms all other classifiers in terms of overall classification accuracy and shows promise for the method to be used in Student Success ventures in the universities in an automatic manner. [9]

Sekeroglu et al. (2019), for a productive and a good life,education is a necessity and it improves individuals' life with value and excellence. Also, education is considered a vital need for motivating self-assurance as well as providing the things are needed to partake in today's World. Throughout the years, education faced a number of challenges. Different methods of teaching and learning are suggested to increase the learning quality. In today's world, computers and portable devices are employed in every phase of daily life and many materials are available online anytime, anywhere. Technologies like Artificial Intelligence had a surprising evolution in many fields especially in educational teaching and learning processes. Higher education institutions have started to adopt the use of technology into their traditional teaching mechanisms for enhancing learning and teaching. In this paper, two datasets have been considered for the prediction and classification of student performance respectively using five machine learning algorithms. Eighteen experiments have been performed and preliminary results suggest that performances of students might be predictable and classification of these performances can be increased by applying pre-processing to the raw data before implementing machine learning algorithms. [10]

III. PROPOSED METHDOLOGY

3.1 Polynomial Regression

In polynomial regression, the relationship between the independent variable x and the dependent variable y is described as an nth degree polynomial in x. Polynomial regression, abbreviated E(y |x), describes the fitting of a nonlinear relationship between the value of x and the conditional mean of y. It usually corresponded to the least-squares method. According to the Gauss Markov Theorem, the least square approach minimizes the variance of the coefficients. This is a type of Linear Regression in which the dependent and independent variables have a curvilinear relationship and the polynomial equation is fitted to the data; we'll go over that in more detail later in the article. Machine learning is also referred to as a subset of Multiple Linear Regression. Because we convert the Multiple Linear Regression equation into a Polynomial Regression equation by including more polynomial elements. Here we use various types of regression such as XGB Regressor, Gradient Boosting Regressor, Ada Boost Regressor, K Neighbors Regressor, Random Forest Regressor, Decision Tree Regressor, Linear Regression, Polynomial Regression (Degree 2), Polynomial Regression (Degree 3), Polynomial Regression (Degree 4), Polynomial Regression (Degree 5). We have chosen the regression

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106034 |

model because it reduces the overfitting together with bias and variance. This machine learning is also easy in implementation which produce better results as compared to other regression model. The best approximation of the connection between the dependent and independent variables is a polynomial. It can accommodate a wide range of functions. Polynomial is a type of curve that can accommodate a wide variety of curvatures.

3.2 Layout of Proposed Architecture

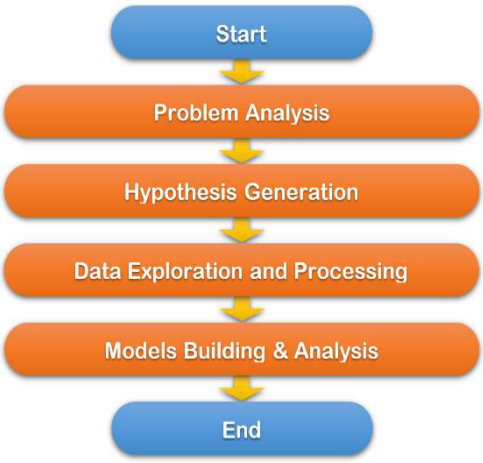

Fig.1 Layout of Proposed methodology

The whole task is divided into 4 different categories which are further divided into sub-tasks.
Stages of complete Task

A. Problem Analysis
In this step, we analyse the problems statement and carry hypothesis analysis. Usually, problems analysis includes articulation of problems and understanding the problem. This helps in getting the answers to questions like, what are the requirements of the problems? What is the problem? What is the desired output? and so on.

B. Hypothesis Generation
According to hypothesis analysis, we analyse the problems and finalize the needs so that while solving the complete problems, what are the things which can help us in generating good efficiency in last or what are the factors which are directly or indirectly impacting the output. So that if we are aware of such factors at the start, we can plan things better and can follow a proper systematic flow.

C. Data Exploration and Processing
In this step, we look for the most suitable database for our problem. then we collect the database so this is the time where we need to filter or clean the dataset. Cleaning and pre-processing the data is the first task in ML & DS workflow. Without this, we will face many issues in exploring the required terms. As cleaning just not means cleaning the data. It exactly means filtering and modifying your data such that it is easy to explore, understand, and can be processed further to models.

D. Models Building & Analysis
Basically, in this step, we have used the training partition of the dataset to train different models and then tested them with the help of the test partition of the dataset. As in the implementation part, all the models trained in the previous step will be compared based on different factors. The comparison is easy when it is carried out in pictorial representation. So, we use the graphical representation of all those five factors to compare all the models. In the last, we pick one as the best model.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106034 |

IV. RESULTS

4.1 Experimental Sup

The experiment for sales prediction of amazing mart, the dataset is chosen from official website of Big mart. For the simulation of proposed methodology following setup is performed such as desktop machine, running Intel Quad Cores i7-3770 @ 3.4 GHz, with 12 GB RAM, a hard disk with 7200 rpm storage seeking speed, and the Windows 10 operating system. For simulation of the experiment, python programming language were used and analysis is performed using Spyder simulation toolbox.

Here for the analysis of proposed methodology we have chosen the dataset from the official website of Big mart which is available online. This dataset contains 8523 rows and 12 column. Dataset is of Mart of the total of 1559 products of 10 different cities. The dataset was designed in 2013 this is the official dataset of big mart.

4.2 Result Analysis

Score

This section presents the comparison of score parameter to show the predictive analysis of amazing mart and their comparative analysis of this parameter is done among different XGB Regressor,. Gradient Boosting Regressor. Ada Boost Regressor, K Neighbors Regressor, Random Forest Regressor, Decision Tree Regressor, Linear Regression, Polynomial Regression (Degree 2), Polynomial Regression (Degree 3), Polynomial Regression (Degree 4), Polynomial Regression (Degree 5). The simulation results of our proposed method and existing method is shown in table 5.1 and the score value of XGB Regressor is 47%,. Gradient Boosting Regressor is 58%. Ada Boost Regressor is 45%, K Neighbors Regressor is 54%, Random Forest Regressor is 54%, Decision Tree Regressor is 14%, Linear Regression is 50%, Polynomial Regression (Degree 1) is 50%, Polynomial Regression (Degree 2) is 59%, Polynomial Regression (Degree 3) is 61%, Polynomial Regression (Degree 4) is 63%, Polynomial Regression (Degree 5) is 68%. The analysis is done using the comparison graph shown in figure 1 and it is found that our proposed method has higher value than the others existing method.

Table 1. Score comparison table of proposed and existing techniques

S. No.	Models	Score
1	XGB Regressor	0.47
2	Gradient Boosting Regressor	0.58
3	Ada Boost Regressor	0.45
4	K Neighbors Regressor	0.54
5	Random Forest Regressor	0.54
6	Decision Tree Regressor	0.14
7	Linear Regressor	0.5
8	Polynomial Regressor 1	0.5
9	Polynomial Regressor 2	0.59
10	Polynomial Regressor 3	0.61
11	Polynomial Regressor 4	0.63
12	Polynomial Regressor 5	0.68

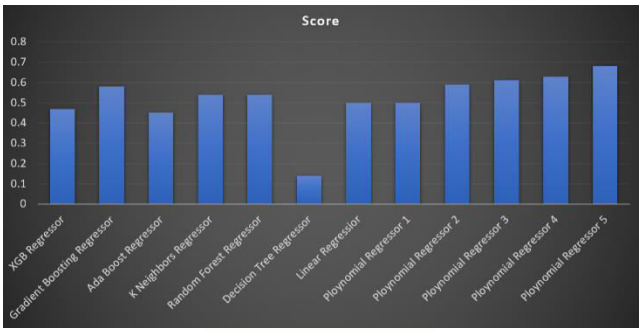

Fig. 1: Score analysis

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106034 |

MAE Analysis

This section presents the comparison of MAE parameter to show the predictive analysis of amazing mart and their comparative analysis of this parameter is done among different XGB Regressor, Gradient Boosting Regressor. Ada Boost Regressor, K Neighbors Regressor, Random Forest Regressor, Decision Tree Regressor, Linear Regression, Polynomial Regression (Degree 2), Polynomial Regression (Degree 3), Polynomial Regression (Degree 4), Polynomial Regression (Degree 5). The simulation results of our proposed method, existing method is shown in table 5.2, and the MAE value of XGB Regressor is 851.0382. Gradient Boosting Regressor is 761.8774, Ada Boost Regressor is 986.4784, K Neighbors Regressor is 813.8725, Random Forest Regressor is 798.895, Decision Tree Regressor is 1091.1664, Linear Regression is 898.8358, Polynomial Regression (Degree 1) is 898.8358, Polynomial Regression (Degree 2) is 778.1174, Polynomial Regression (Degree 3) is 741.9602, Polynomial Regression (Degree 4) is 732.6907, Polynomial Regression (Degree 5) is 694.7307. The analysis is done using the comparison graph shown in figure 2 and it is found that our proposed method has higher value than the others existing method.

Table 2 MAE comparison table of proposed and existing techniques

S. No.	Models	Mean Absolute Error
1	XGB Regressor	851.0382
2	Gradient Boosting Regressor	761.8774
3	Ada Boost Regressor	986.4784
4	K Neighbors Regressor	813.8725
5	Random Forest Regressor	798.895
6	Decision Tree Regressor	1091.1664
7	Linear Regressor	898.8358
8	Polynomial Regressor 1	898.8358
9	Polynomial Regressor 2	778.1174
10	Polynomial Regressor 3	741.9602
11	Polynomial Regressor 4	732.6907
12	Polynomial Regressor 5	694.7307

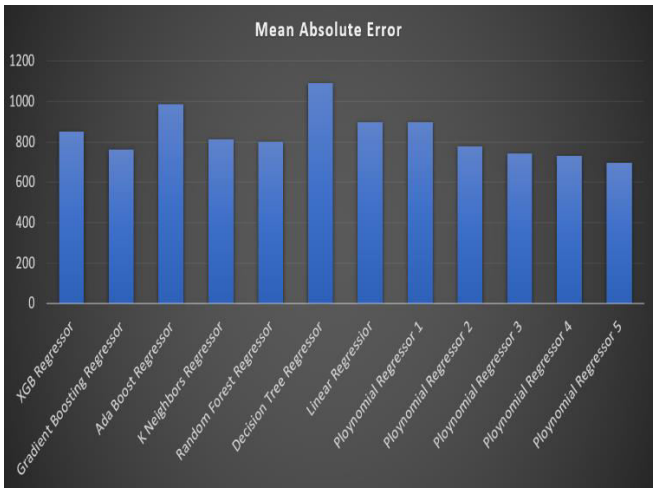

Fig. 2 Comparative analysis of MAE parameter between proposed and existing method

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106034 |

V. CONCLUSION

We made different systems of predicting the sales of the products and also made a comparative analysis of the systems and models used. The first step of the process is to start with an analysis of the problem and what should be a good dataset for the same. After extracting and basic cleaning of the dataset we carried out the hypothesis analysis and pre-processed the dataset according to the rules and filters we came up with. Made some of the label encodings as a part of dataset pre-processing. Moving on to the splitting of the dataset into training and testing datasets & then training models over the training dataset and testing all the models with the testing segment dataset. We used the following models for training and testing purposes, XGB Regressor, Gradient Boosting Regressor, Ada Boost Regressor, K Neighbours Regressor, Random Forest Regressor, Decision Tree Regressor, Linear Regression, and Polynomial Regression of various degrees. Then carried out the analysis of the systems developed and got that the polynomial regressor gives the best values. So finally, we came up with a system with a very good score for predicting the sales of all 10 products. Overall analysis of proposed work is the it give better predictive score value and also generate less error than the existing methodology of Amazing mart analysis.

REFERENCES

[1] "ReviewTrackers Online Reviews Survey: Statistics and Trends", [online] Available at: https://www.reviewtrackers.com/online-reviews-survey/ [Accessed 2 Aug. 2018].
[2] Neda Khalil Zadeh, Mohammad Mehdi Sepehri, and Hamid Farvaresh "Intelligent Sales Prediction for Pharmaceutical Distribution Companies: A Data Mining Based Approach", Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2014, Article ID 420310, 15 pages.
[3] https://www.quora.com/What-is-Data-Science-and-Data-Analytics.
[4] Vaseem Naiyer, Jitendra Sheetlani, Harsh Pratap Singh, " Software Quality Prediction Using Machine Learning Application", Smart Intelligent Computing and Applications, 2020, Springer.
[5] Kumari Punam et al., "A Two-Level Statistical Model for Big Mart SalesPrediction", International Conference on Computing, Power and Communication Technologies (GUCON)Galgotias University, Greater Noida, UP, India. Sep 28-29, 2018.
[6] Yanrong and Fan "A two-stage dynamic sales forecasting model for the fashion retail", Expert Systems with Applications 38 (2011) 1529–1536.
[7] Kilimci et al., "An Improved Demand Forecasting Model UsingDeep Learning Approach and Proposed DecisionIntegration Strategy for Supply Chain", HindawiComplexity, Volume 2019, Article ID 9067367, 15 pages,https://doi.org/10.1155/2019/9067367.
[8] Mohit Gurnani et al., "Forecasting of sales by using fusion ofMachine Learning techniques", International Conference on Data Management, Analytics and Innovation (ICDMAI)Zeal Education Society, Pune, India, Feb 24-26, 2017.
[9] Marquez-Vera, C., A. Cano, C. Romero, A.Y.M. Noaman and H. Mousa Fardoun et al., 2016. Early dropout prediction using data mining: A case study with high school students. Expert Syst., 33: 107-124. DOI: 10.1111/exsy.12135.
[10] Zhang, W., S. Zhang and S. Zhang, 2015. Predicting the graduation thesis grade using SVM. Int. J. Intell. Inform. Process., 5: 60-68.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details