**RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES**

**A PROJECT REPORT**

Submitted by

**S.KIRUTHIGA (810015205032)**

**K.NANTHINIMEENA (810015205049)**

In partialfulfillment for the award of the degree

Of

**BACHELOR OF TECHNOLOGY**

**IN**

**INFORMATION TECHNOLOGY**

****

**UNIVERSITY COLLEGE OF ENGINEERING,  
 BHARATHIDASAN INSTITUTE OF TECHNOLOGY CAMPUS, TIRUCHIRAPPALLI**

**ANNA UNIVERSITY: CHENNAI 600 025**

**APRIL 2019**

**UNIVERSITY COLLEGE OF ENGINEERING,**

**BHARATHIDASAN INSTITUTE OF TECHNOLOGY CAMPUS**

**TIRUCHIRAPPALLI-620 024**

**BONAFIDE CERTIFICATE**

Certificate that this project report **“RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES”** is the benefited work of **S.KIRUTHIGA (810015205032), K.NANTHINIMEENA (810015205049)** who carried out the project under my supervision.

**SIGNATURE SIGNATURE**

**Dr.D.VENKATESAN Mrs.R.JAYAMALA**

**HEAD OF THE DEPARTMENT SUPERVISOR**

Department of IT Assistant professor

University College of Engineering, Department of IT

BIT Campus, University College of Engineering,

Tiruchirappalli-620 024. BIT Campus,

Tiruchirappalli-620 024.

Submitted for the project Viva-Voce Examination held on\_\_\_\_\_\_\_\_\_\_\_\_

**INTERNAL EXAMINER EXTERNAL EXAMINER**

**DECLARATION**

I hereby that the work entitled” **RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES”** is submitted in partial fulfilment of the requirements for the award of the degree in B. Tech..Information Technology in University College of Engineering, Bharathidasan Institute of Technology (BIT), Tiruchirappalli, is a record of my own work carried out by me during the academic year 2018-2019 under the supervision and guidance of **Mrs.R.Jayamala**, Assistant Professor, Department of Information Technology, University College of Engineering, Bharathidasan Institute of Technology (BIT) Campus, Anna University, Tiruchirappalli. The extend and sources of information are derived from the existing literature and have been indicated through the dissertation at the appropriate places.

S.KIRUTHIGA (810015205032)

K.NANTHINIMEENA (810015205049)

I certify that the declaration made above by my candidate is true.

Signature of the Guide,

**Mrs.R.JAYAMALA**

Assistant professor

Department of Information Technology

University College of Engineering,

Anna University, BIT Campus,

Tiruchirappalli-620 024.

**ACKNOWLEDGEMENT**

We would like to thank our honorable Dean **Dr.T.SENTHIL KUMAR,** Professor for having provided us with all required facilities to complete our project without hurdles.

We would also like to express our sincere thanks to **Dr.D.VENKATESAN**, Head of the Department, Department of Information Technology, for his valuable guidance, suggestions and encouragement paved way for the successful completion of this project work.

We would like to thank our project coordinator **Mr. M. PRASANNA KUMAR,** Teaching fellow, Department of Information Technology for his kind support.

We would like to thank and express our deep sense of gratitude to our project guide **MRS.R.JAYAMALA**, Assistant Professor, Department of Information Technology, for her valuable guidance throughout the project. We also extend our thanks to all other teaching and non-teaching staff for their encouragement and support.

We thank our beloved parents, brothers and friends for their full support in the moral development of this project.

**ABSTRACT**

Rainfall Prediction is the application of science and technology to predict the state of the atmosphere. The purpose of predicting rainfall is to make use of water resources, crop productivity and pre-planning of water structures. Artificial neural network have been applied to compute the real-time rainfall prediction. The main problem with artificial neural network is, it takes much time to train the network. It does not specify the rule for determining structure. In proposed system, we used regression and classification algorithms under the supervised learning. Inspired by machine learning, a desired techniques viz., linear regression, multiple linear regression, logistic regression, support vector machine, naïve bayes algorithm are used in the proposed system. The parameters of the proposed system are mean square error (MSE), root mean square root(RMSE) and coefficient of determination(r2). The prediction of the proposed system is done by comparative analysis of these parameters. Based on the results, the best algorithm should be declared for rainfall prediction.

**TABLE OF CONTENTS**

**CHAPTER TITLE PAGE**

**NO. NO.**

**ABSTRACT v**

**LIST OF FIGURES vi**

**LIST OF ABBREVIATIONS vii**

**1 INTRODUCTION 01**

1.1 RAINFALL PREDICTION 01

1.2 RAINFALL FORECASTING METHODS 01

1.3 DATA MINING 02

1.4 MACHINE LEARNING TECHNIQUES 03

**2 LITERATURE SURVEY 05**

**3 SYSTEM ANALYSIS 14**

3.1 EXISTING SYSTEM 14

3.1.1 Advantages 15

3.1.2 Disadvantages 16

3.2 PROPOSED SYSTEM 17

**4 SYSTEM SPECIFICATION 18**

4.1 HARDWARE REQUIREMENTS 18

4.2 SOFTWARE REQUIREMENTS 18

4.3 ABOUT THE SOFTWARE 19

**5 SYSTEM DESIGN 20**

5.1 SYSTEM ARCHITECTURE 20

5.1.1 Methodology 21

5.1.2 Data flow diagram 21

5.2 MODULE DESCRIPTION 23

5.2.1 Data pre-processing 23

5.2.2 Implementation 24

5.2.3 Measures 27

5.3 EXPERIMENTAL ANALYSIS 28

**6 CONCLUSION AND FUTURE WORK 31**

6.1 CONCLUSION 31

6.2 FUTURE WORK 31

REFERENCES 31

APPENDICES 34

A1 SAMPLE SOURCE CODE 34

A2 SAMPLE SCREENSHOTS 38

**LIST OF FIGURES**

**3 3.1 Structure of ANN 15**

**5 5.1 System Architecture 20**

**5.1.2 Level 0 DFD 22**

**5.1.3 Level 1 DFD 22**

**5.3 Table 28**

**5.3.1 Graph 29**

**LIST OF ABBREVIATION**

|  |  |  |
| --- | --- | --- |
| **ABBREVIATION**  ANN  CRT | -  - | **EXPANSION**  Artificial Neural Network  Classification and Regression Tree |
| DB  DM  DT | -  -  - | Data Base  Data Mining  Decision Tree |
| GA | - | Genetic algorithm |
| HMM | - | Hidden Markov Model |
| KDD  KNN  MSE  NB  NWP  PRCP  RMSE  SVM  TMAX  TMIN | -  -  -  -  -  -  -  -  -  - | Knowledge Discovery  in Databases  K Nearest Neighbour  Mean Square Error  Naïve Bayes  Numerical Weather Prediction  Precipitation  Root Mean Square Error  Support Vector Machine  Temperature Maximum  Temperature Minimum |

**CHAPTER 1**

**1. INTRODUCTION**

**1.1 RAINFALL PREDICTION**

Rainfall prediction is the application of science and technology to predict the state of the atmosphere. It remains a serious concern and has attracted the attention of government, industries, risk management entities, as well as the scientific community. Rainfall is climatic factor that affects many human activities like agricultural production, construction, power generation, forestry, tourism and others. Rainfall prediction is essential since this variable is the one with the highest correlation with adverse natural events such as landslides, flooding, mass movements and avalanches. These incidents have affected society for years. Therefore, having an appropriate approach for rainfall prediction makes it possible to take preventive and mitigation measures for these natural phenomenon.

**1.2 RAINFALL FORECASTING METHODS**.

Weather forecasting is one of the most important and demanding operational responsibilities carried out by meterological services all over the world. It is complicated procedure that includes numerous specialized fields of know-how. The task is complicated because in the field of meteorology all decisions are to be taken in the visage of uncertainty.There are several rainfall forecasting methods,such as statistical methods and Numerical Weather Prediction models.Statistical methods cannot generate good results. Local climate situations also can affect the rainfall generation process a lot. NWP models cannot solve this local problem.

**1.3 DATA MINING:**

Data mining is used to turn raw data into useful information. It is a process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining is the analysis step of the “knowledge discovery in databases” (or) KDD. The difference between data analysis and data mining is that data analysis is to summarize the history such as analysing the effectiveness of a marketing campaing, in contrast, data mining focuses on using specific machine learning and statistical models to predict the future and discover the patterns among data. Data mining includes the following algorithm types viz., classification algorithms, regression algorithms, segmentation algorithms, association algorithms and sequence analysis algorithms. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

**DATA COLLECTION**

A dataset for proposed system has been taken from http://www.kaggle.com website. For the prediction model, we used weather data of Seattle, Washington from 1947 to 2017. The raw weather data collected consists of four measured attributes which are date, temperature (maximum , minimum) ,precipitation and rain. From the dataset we use date, temperature ( maximum , minimum), precipitation as an input and rain as a output. An input is a numerical value whereas output is a categorical value. We have ignored less relevant features in the dataset for better model computation and prediction.

**1.4 MACHINE LEARNING TECHNIQUES:**

Machine learning is a subset of artificial intelligence which provides machines the ability to learn automatically and improve from experience without being explicitly programmed. It is closely related to computational statistics, which focuseson making predictions using computers. It is also referred to as predictive analytics.In the proposed system , the machine learning techniques used are linear regression, multiple linear regression, logistic regression, support vector machine and naïve bayes. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering, and computer vision, where it is infeasible to develop an algorithm of specific instructions for performing the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as  predictive analytics.Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not spam", represented by the Boolean values true and false. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. 

**CHAPTER 2**

**LITERATURE SURVEY**

Aakash Parmar et al.,[1] states that heavy rainfall prediction is a major problem for meteorological department as it is closely associated with the economy and life of human. It is a cause for natural disasters like flood and drought which are encountered by people across the globe every year. Accuracy of rainfall forecasting has great importance for countries like India whose economy is largely dependent on agriculture. Due to dynamic nature of atmosphere, Statistical techniques fail to provide good accuracy for rainfall forecasting. Nonlinearity of rainfall data makes Artificial Neural Network a better technique. Review work and comparison of different approaches and algorithms used by researchers for rainfall prediction is shown in a tabular form. Intention of this paper is to give non-experts easy access to the techniques and approaches used in the field of rainfall prediction.

R. Usha Rani et al.,[2] states that interfacing through the continuously rising amounts of data in technical, medical, scientific, engineering, industrial and monetary fields and their renovation to logical form for the human user is one of the main requirements. To quickly discover and analyze complex patterns and requirements, we need the efficient techniques and need to learn from new data will be necessary for information-intensive applications. One of the solutions for this is that classification and clustering of largely available data. To partially fulfill the industry requirement, in this paper we proposed a two-level approach for clustering large data set for rain fall data prediction with Self Organized Maps (SOM) and Support Vector Machine (SVM) with ID3. In this paper, a novel approach to clustering of the SOM and SVM with ID3 are considered. In particular, the use of hierarchical agglomerative clustering and partitioned clustering with ID3 are investigated.The twostage procedure first using SOM to produce the prototypes and later it considers the SVM with ID3, that are then clustered in the second stage is found to perform well when compared with direct clustering of the data and to reduce the computation time.

Amruta A.Taksande et al.,[3] states that rainfall forecasting or Weather forecasting has been one of the most challenging problems around the world because it consists of multidimensional and nonlinear data such as in the field of agriculture to determine initial growing season. Recently, climate change causes much trouble in rainfall forecasting. This paper describes five data mining algorithms namely neural network (NN), random forest, classification and regression tree (CRT), support vector machine (SVM) and k-nearest neighbour. Generally these algorithms are used for the prediction. In this paper rainfall forecasting using Artificial Neural Network (ANN) and Genetic Algorithm ( GA) is made. In genetic algorithm we use Hidden Markov Model (HMM) for records the previous data. The data used within this research is taken from Yahoo Weather API is the type of Interface. Those data include temperature, air pressure, rainfall, relative humidity , and wind speed. Based on experiment result, it can be concluded that the combination of GA and HMM weather data can gives prediction graph with higher than 90% accuracy with several population size and crossover probability.

Chandrasegar Thirumalai et al.,[4] states that this paper is carried on the heuristic prediction of rainfall using machine learning techniques. As we know agriculture was the predominant of our country and economy. While a regular rain pattern is usually played vital for healthy agriculture but too much rainfall or too little rainfall can be harmful, even it led to devastating of crops. This paper discusses the rate of rainfall in previous years according to various crops seasons like rabi, Kharif, zaid and predicts the rainfall in future seasons. The paper also measures the different categories of data by linear regression method in metrics for effective understanding of agriculture in India. We have selected a real dataset which consists of past year’s rainfall rate according to various seasons. Results of this application help farmers to make a correct decision to harvest a particular crop accordingly to crops seasons. Linear regression helps to find.

S. Swain et al.,[5] states that estimation of precipitation is necessary for optimum utilization of water resources and their appropriate management.The economy of India being heavily dependent on agriculture becomes vulnerable due to lack of adequate irrigation facilities. In this paper, a multiplelinear regression model has been developed to reckon annual precipitationover Cuttack district, Odisha, India. The model forecasts precipitationfor a year consideringannual precipitation data of its three precedingyears. The model testing was performed over acentury-long dataset of annualprecipitation i.e. for 1904-2002. Assuming the intercept or constant of the multiple linear regressionmodel as zero, the equation developedthereby displayeda superbresult. The model predictions showedan excellentassociationwith theobserved data i.e. thecoefficient of determination (R2) and adjusted R2 value was obtained to be 0.974 and 0.963 respectively. This reconciliation justifies the application of the developed model over the study area to forecast rainfall, thereby aiding in proper planning and management.

Kavitha S et al.,[6] states that in business, consumers interest, behavior, product profits are the insights required to predict the future of business with the current data or historical data. These insights can be generated with the statistical techniques for the purpose of forecasting. The statistical techniques can be evaluated for the predictive model based on the requirements of the data. The prediction and forecasting are done widely with time series data. Most of the applications such as weather forecasting, finance and stock market combine historical data with the current streaming data for better accuracy. However the time series data is analyzed with regression models. In this paper, linear regression and support vector regression model is compared using the training data set in order to use the correct model for better prediction and accuracy.

N.Divya prabha et al.,[7] states that the extensive rainfall data series is acting as an most important position in all water related studies. Regularity and connection of rainfall data series are incredibly significant for obtaining some valuable or reliable results from such studies. Though, these rainfall data series frequently hold gaps or misplaced values due to different reasons like as the lack of observers, struggle with measuring devices, loss of information or records etc. The utilize of a rainfall data series with lost values may significantly authority the statistical power and correctness of a study. By estimating and extensive the nowhere to be found rainfall data, a series could be through longer to build the water related study additional dependable. Improved Multilayer Perception Neural Network is planned an intellectual tool for predicting Rainfall Time Series. This Rainfall Data series has been approved using the projected Multilayer Perceptron Neural Network. It seems that the presentation process such as MSE (Mean square error), and NMSE (Normalized mean square error) on testing as well as preparation of data set for small term forecast are found as best possible in assessment with other network like as Adanaive, AdaSVM.

Deepti Gupta et al.,[8] states that the India is an agricultural country which largely depends on monsoon for irrigation purpose. A large amount of water is consumed for industrial production, crop yield and domestic use. Rainfall forecasting is thus very important and necessary for growth of the country. Weather factors including mean temperature, dew point temperature, humidity, pressure of sea and speed of wind and have been used to forecasts the rainfall. The dataset of 2245 samples of New Delhi from June to September (rainfall period) from 1996 to 2014 has been collected from a website named Weather Underground. The training dataset is used to train the classifier using Classification and Regression Tree algorithm, Naive Bayes approach, K nearest Neighbour and 5-10-1 Pattern Recognition Neural Network and its accuracy is tested on a test dataset. Pattern Recognition networks has given 82.1% accurate results, KNN with 80.7% correct forecasts ranks second, Classification and Regression Tree(CART) gives 80.3% while Naive Bayes provides 78.9% correctly classified samples.

Nor SamsiahSani et al.,[9] states that data mining is a process that aims to extract useful knowledge from cluttered and unorganized information. Climate change is a discipline involved with analyzing the varying distribution of weather for a specific period of time. Specifically, rainfall forecasting analyzes specific features such as humidity and wind are used to predict rainfall in specific locations. Rainfall prediction has of recent been subjected to several machine learning techniques with different degree of short-term (daily) and long-term (monthly) prediction performance. Selecting an appropriate technique for specific rainfall duration is a challenging task. Several approaches have been proposed for rainfall forecasting using various machine techniques. This study aims to provide a comparative analysis of the multiple machine learning classifiers for rainfall prediction based on Malaysian data. Several classifiers were explored which are f Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Neural Network (NN) and Random Forest (RF). The analysis showed the most effective classifier to be the NN.

Razeef Mohd et al.,[10] states that prediction of rainfall is one of the most essential and demanding tasks for the weather forecasters since ages. Rainfall prediction plays an important role in the field of farming and industries. Precise rainfall prediction is vital for detecting the heavy rainfall and to provide the information of warnings regarding the natural calamities. Rainfall prediction involves recording the various parameters of weather like wind direction, wind speed, humidity, rainfall, temperature etc. From last few decades, it has been seen that data mining techniques have achieved good performance and accuracy in weather prediction than traditional statistical methods. This research work aims to compare the performance of few data mining algorithms for predicting rainfall using historical weather data of Srinagar, India, which is collected from http://www.wundergrounds.com website. From the collected weather data which comprises of 9 attributes, only 5 attributes which are most relevant to rainfall prediction are considered. Data mining process model is followed to obtain accurate and correct prediction results. In this paper, various data mining algorithms were explored which include decision tree based J48, Random forest, Naive Bayes, Bayes Net, Logistic Regression, IBk, PART and bagging. The experimental results show that J48 algorithm has good level of accuracy than other algorithms.

Jinglin Du et al.,[11] states that precipitation is a very important topic in weather forecasts. Weather forecasts, especially precipitation prediction, poses complex tasks because they depend on various parameters to predict the dependent variables like temperature, humidity, wind speed and direction, which are changing from time to time and weather calculation varies with the geographical location along with its atmospheric variables. To improve the prediction accuracy of precipitation, this context proposes a prediction model for rainfall forecast based on Support Vector Machine with Particle Swarm Optimization (PSO-SVM) to replace the linear threshold used in traditional precipitation. Parameter selection has a critical impact on the predictive accuracy of SVM, and PSO is proposed to ﬁnd the optimal parameters for SVM. The PSO-SVM algorithm was used for the training of a model by using the historical data for precipitation prediction, which can be useful information and used by people of all walks of life in making wise and intelligent decisions. The simulations demonstrate that prediction models indicate that the performance of the proposed algorithm has much better accuracy than the direct prediction model based on a set of experimental data if other things are equal.

Kumar Abhishek et al.,[12] states that the multilayered artificial neural network with learning by back-propagation algorithm configuration is the most common in use, due to of its ease in training. It is estimated that over 80% of all the neural network projects in development use back-propagation. In back-propagation algorithm, there are two phases in its learning cycle, one to propagate the input patterns through the network and other to adapt the output by changing the weights in the network. The back-propagation-feed forward neural network can be used in many applications such as character recognition, weather and financial prediction, face detection etc. The paper implements one of these applications by building training and testing data sets and finding the number of hidden neurons in these layers for the best performance. In the present research, possibility of predicting average rainfall over Udupi district of Karnataka has been analyzed through artificial neural network models. In formulating artificial neural network based predictive models three layered network has been constructed. The models under study are different in the number of hidden neurons.

Emilcy Hern´andez et al.,[13] states that the previous work has shown that the prediction of meteorological conditions through methods based on artiﬁcial intelligence can get satisfactory results. Forecasts of meteorological time series can help decision-making processes carried out by organizations responsible of disaster prevention. We introduce an architecture based on Deep Learning for the prediction of the accumulated daily precipitation for the next day. More speciﬁcally, it includes an autoencoder for reducing and capturing non-linear relationships between attributes, and a multilayer perceptron for the prediction task. This architecture is compared with other previous proposals and it demonstrates an improvement on the ability to predict the accumulated daily precipitation for the next day.

Mohini P. Darji et al.,[14] states that an accurate rainfall forecasting is very important for agriculture dependent countries like India. For analyzing the crop productivity, use of water resources and pre-planning of water resources, rainfall prediction is important. Statistical techniques for rainfall forecasting cannot perform well for longterm rainfall forecasting due to the dynamic nature of climate phenomena. Artificial Neural Networks (ANNs) have become very popular, and prediction using ANN is one of the most widely used techniques for rainfall forecasting. This paper provides a detailed survey and comparison of different neural network architectures used by researchers for rainfall forecasting. The paper also discusses the issues while applying different neural networks for yearly/monthly/daily rainfall forecasting. Moreover, the paper also presents different accuracy measures used by researchers for evaluating performance of ANN.

Mary N. Ahuna et al.,[15] states that a model for rainfall rate prediction 30 seconds ahead of time using an artificial neural network. The resultant predicted rainfall rate can then be used in determining an appropriate fade counter-measure, for instance, digital modulation scheme ahead of time, to keep the bit error rate (BER) on the link within acceptable levels to allow constant flow of data on the link during a rain event. The approach used in this paper is pattern recognition technique that considers historical rainfall rate patterns over Durban (29.8587°S, 31.0218°E). The resultant prediction model is found to predict an immediate future rain rate when given three adjacent historical rain rates. For our model validation, error analysis via root mean square (RMSE) technique on our prediction model results shown.

**CHAPTER 3**

**3. SYSTEM ANALYSIS**

**3.1 EXISTING SYSTEM**

**Artificial Neural Networks(ANNs)**

Artificial Neural Networks are constructed based on human brain by making the connections using neurons and dendrites . Each neuron is connected with Axons. Inputs from sensory organs are accepted by Dendrites. The series of electric impulses created by inputs travel along neural network. Then a neuron takes decision to transfer the signal to other neuron or block the signal to process forward. By this structure ANNs are constructed. There are 3 kinds of layers present in ANNs. They are Input Layer, Hidden Layer and Output Layer. Each layer consists fixed set of neurons depends on our application. There is no choice to decide how many neurons present in input and output layer because it depends on number of inputs and outputs of Data Set. But neuron selection by user is possible in hidden layer. ANN model allows us to select any number of hidden layers with any number of neurons in each hidden layer. Backpropagation(BP) is a method used in artificial neural networks to calculate a gradient that is needed in the calculation of the weights to be used. BP learning does not require normalization of input vectors, however normalization could improve performance. As the number of neurons increases in an ANN, the MSE decreases. Larger the amount of input data, lower the MSE after training. The input/output data should be normalized if they are of very high order. ANN is a bit time consuming.

**STRUCTURE OF ANN**

Input layer Hidden layer Output layer

I1 w11

I2 w22 y

I3 w33

Fig 3.1 Structure of ANN

**3.1.1 Advantages of Artificial Neural Networks ( ANN)**

* Storing information on the entire network.
* Ability to work with incomplete knowledge.
* Fault tolerance.
* Distributed memory.
* Gradual corruption.
* Ability to make machine learning.
* Parallel processing capability.

**3.1.2 Disadvantages of Artificial Neural Networks (ANN)**

* Hardware dependence**.**
* Unexplained behavior of the network**.**
* Determination of proper network structure.
* Difficulty of showing the problem to the network**.**
* The duration of the network is unknown.

**3.2 PROPOSED SYSTEM**

In the Proposed system, to overcome the flaws of existing system we use classification techniques under datamining. The classification techniques are linear regression, multiple linear regression, logistic regression, support vector machine and naïve bayes. These techniques are comes under the supervised learning. Supervised learning is the Data mining task of inferring a function from labelled training data. The training data consists of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a desired output. Linear Regression is used to estimate the relationship between target and one or more predictors. Multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. Logistic regression describe the data given and estimate the relationship between one or more independent variables and a dependent variable. The main purpose of using SVM is to classify the class from the given dataset. Naive Bayes assumes that attributes have no effect on each other that is they have independent distribution of values. The algorithm starts with computing prior probability, i for each class. Then for an input data instance q, compute posterior probabilities for each class.

**CHAPTER 4**

**4. SYSTEM SPECIFICATION**

**4.1 HARDWARE REQUIREMENTS**

Hardware is the physical components of the computer like microprocessor, hard disks, RAM and motherboard. Hardware devices are the executors of the commands provided by software applications. Computer hardware as the electronic, magnetic, and electric devices that carry out the computing functions.

* + - Platform: Windows10
    - Processor: INTEL Core i3
    - RAM Capacity: 8GB
    - Hard disk: 40GB

**4.2 SOFTWARE REQUIREMENTS**

Software includes all the various forms and roles that digitally stored data may have and play in a computer(or similar system), regardless of whether the data is used as code for a CPU, or other interpreter. Software thus encompasses a wide array of products that may be developed using different techniques such as ordinary programming languages, scripting languages and etc.

* + - Operating System : Windows10
    - Tool : R studio

**4.3 ABOUT THE SOFTWARE**

**R Studio**

RStudio is a set of integrated tools designed to help you be more productive with R. Includes a console, syntax-highlighting editor that supports direct code execution and a variety of robust tools for plotting, viewing history, debugging and managing your workspace. RStudio is an integrated development environment(IDE) that allows you to interact with R more readily. RStudio is similar to the standard RGui, but is considerably more user friendly. It has more drop-down menus, windows with multiple tabs and many customization options. When you open the RStudio , you will see three windows. A forth window is hidden by default, but can be opened by clicking the File drop-down menu, then New File and then R Script.

**CHAPTER 5**

**5 SYSTEM DESIGN**

**5.1 SYSTEM ARCHITECTURE**

A system architecture or systems architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system.A system architecture can consist of system components and the sub-systems developed, that will work together to implement the overall system. There have been efforts to formalize languages to describe system architecture, collectively these are called architecture description languages (ADLs).

Collection of Dataset

Result

Fig 5.1 System Architecture

**5.1.1 METHODOLOGY**

From the Fig 5.1,

Step1: Take the rainfall dataset for preceding algorithm.

Step2: Preprocess the dataset.

Step 3: Take away dissimilar attributes on both train and test dataset.

Step 4: Initialize the algorithms viz., regression models, support vector machine and naïve bayes.

Step 5: To apply the input to the algorithms and gather the computed output.

Step 6: To compute MSE, RMSE, Coefficient of determination.

Step 7: Apply the above steps for each training model until the desired output.

**5.1.2 DATA FLOW DIAGRAM**

A data flow diagram (DFD) is a graphical representation of the “flow” of data through an information system. It differs from the flowchart as it shows the data flow instead of the control flow of the program. A data flow diagram can also be used for the visualization of data processing. The DFD is designed to show how a system is divided into smaller portions and to highlight the flow of data between those parts.

**Level 0**

Result

Rainfall prediction Dataset

Fig 5.1.2 Level 0 DFD

In Fig 5.1.2, the process of level 0 DFD is used to analyse and classify historical and current data to predict the future data.

**Level 1**

Rainfall prediction Dataset

Result

Fig: 5.1.3 Level 1 DFD

In Fig: 5.1.3, Level 1 DFD defines the preprocessing of data and implementation of machine learning algorithms. Comparative analysis of various measures of these algorithms gives the desirable results.

**5.2 MODULE DESCRIPTION**

**5.2.1 DATA PREPROCESSING**

The main challenge in rainfall prediction is the poor data quality and selection. For this reason we try to preprocess data carefully to obtain accurate and correct prediction results. In this phase unwanted data or noise is removed from the collected data set which is done by removing the unwanted attributes and keeping the most relevant attributes that help in better prediction. Another major issue that is to be rectified is the missing values in the collected data set. Missing values in the data set is filled by using various techniques. In this work, the missing values for attributes in the dataset are replaced with the modes and means based on existing data. Adding the missing values provides a more complete dataset for the classifiers to be trained on. Data mining is the process of extracting the useful information from a large collection of data which was previously unknown . For extracting useful information we need to follow data mining process model that will give us clean valuable dataset for model computation and better prediction.Very rarely data are available in the form required by the data mining algorithms. Most of the data mining algorithms would require data to be structured in a tabular format with records in rows and attributes in columns. The methodological discovery of useful relationships and patterns in data is enabled by a set of iterative activities known as data mining process. Not all discovered patterns leads to knowledge. It is up to the practitioner to invalidate the irrelevant patterns and identify meaningful information.

**5.2.2 IMPLEMENTATION**

**LINEAR REGRESSION**

Linear Regression is one of the type of regression model which is used to study the relationship between dependent variable RAIN and independent variables TMAX and TMIN of the dataset. It is used to estimate the relationship between target and one or more predictors.

Y = a + bX (1)

where Y - dependent variable

X - independent variable

a - intercept of y

b - slope of the line

**MULTIPLE LINEAR REGRESSION**

Multiple Linear Regression is also known as multiple regression which is the common form of linear regression. It is used to explain the relationship between one continuous dependent variable RAIN and two or more independent variables TMAX, TMIN and PRCP.

Ỳ= b0 + b1 X1 + b2X2 (2)

Where Y – Dependent variable

X1, X2 – Risk factor

b0, b1, b2 – Estimated regression coefficient

**LOGISTIC REGRESSION**

Logistic Regression is used only when the dependent variable is binary. It also describe the data given and estimate the relationship between one or more independent variables TMAX, TMIN and PRCP and a dependent variable RAIN.

P = ea + Bx/1 + ea+bX (3)

where P – probability of 1

e – base of natural logarithm

a, b – parameters of the model

**SUPPORT VECTOR MACHINE**

Support vector machine is a supervised learning model. It analyses the data which is used for classification and regression. It performs well with a limited amount of data. The main purpose of using SVM is to classify the class from the given dataset. SVM is a supervised Machine Learning algorithm which can be used for both classification and regression challenges. However it is mostly used in classification problems. SVM are a subclass of supervised classifiers that attempt to partition a feature space into two or more groups. Then, we perform classification by finding the hyper-plane that differentiates the two classes.Support Vector Machines are particularly suited to handle such tasks. Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. In this algorithm, one hundred data are trained and fifty data are used for testing purposes.

**NAIVE BAYES**

Naive Bayes is a classification technique based on Bayes Theorem. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Bayesian approach for classification is a statistical and linear classifier which predicts class label for data instance on the basis of distribution of attribute values. This is a parametric classification where the size of classifier remains fixed. Distribution can be normal (Gaussian), kernel, multivariate or multi nominal. Assuming normal distribution for weather data, Bayesian classifiers use Bayes theorem to find posterior probabilities of occurrence of input data instance in all classes. Class label having maximum conditional probability is assigned to data instance. Naive Bayes assumes that attributes have no effect on each other that is they have independent distribution of values. The algorithm starts with computing prior probability, i for each class. Then for an input data instance q, compute posterior probabilities for each class.

P(c/x) = P(x/c) P(c)/P(x) (4)

Where P(c/x) – posterior probability of class

P(c) – prior probability of class

P(x/c) – probability of predictor

P(x) – prior probability of predictor

**5.2.3 MEASURES**

**MEAN SQUARED ERROR**

The mean squared error (MSE) or mean squared deviation (MSD) of an estimator  measures the average  of the squares of the errors that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.The MSE is a measure of the quality of an estimator it is always non-negative, and values closer to zero are better.

MSE = 1/n ∑ (yi – y)2 (5)

Where n – data points

Y – vector of observed values

(yi – y)2 - squares of the errors

**ROOT MEAN SQUARE ERROR**

The root mean square deviation (RMSD) or root mean square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed. The RMSD represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. These deviations are called residuals when the calculations are performed over the data sample that was used for estimation and are called errors when computed out of sample.

RMSE = √MSE (6)

Where MSE = Mean Square Error

**COEFFICIENT OF DETERMINATION**

The coefficient of determination (denoted by r2) is a key output of regression analysis.With linear regression, the coefficient ofdetermination is also equal to the square of the correlation between x and y scores. An r2 of 0 means that the dependent variable cannot be predicted from the independent variable.

r2 ≡ 1-SSres/SStot (7)

where SSres – sum of squares of residuals

SStot – total sum of squares.

**5.3 EXPERIMENTAL RESULTS**

|  |  |  |  |
| --- | --- | --- | --- |
| ALGORITHMS | MSE | RMSE | r2 |
| Linear regression | 69.5481 | 8.3395 | 0.0148 |
| Multiplelinear regression | 0.1756 | 0.4191 | 0.2920 |
| Logistic regression | 18.4009 | 4.2896 | 0.3622 |
| Support vector machine | 0.2030 | 0.4005 | 0.0319 |
| Naive Bayes | 0.2506 | 0.5006 | 0.0048 |

TABLE 5.3 EXPERIMENTAL RESULTS

**GRAPH** Fig:5.3.1 Experimental Results

The results of various machine learning algorithms are compared on the basis of mean square error, root mean square error and coefficient of determination. Performance measures of applied prediction models based on rainfall dataset is shown in TABLE 5.3 and is also graphically observed in fig. In this paper, the comparison of data mining tools is based on different types of data. The algorithms used for classification are linear regression, multiple linear regression, logistic regression, support vector machine and naive bayes. In linear regression, TMAX(maximum temperature) is input and the algorithm must predict the output RAIN. By using these input and output, the values of MSE, RMSE and Coefficient of determination respectively 69.5481, 8.3395 and 0.0148.In multiple linear regression,TMIN(minimum temperature) and PRCP(precipitation) are input and the algorithm must predict the output RAIN. By using these input and output, the values of MSE,RMSE and Coefficient of determination are predicted respectively 0.1756, 0.4191 and 0.2920 . In logistic regression, TMIN(minimum temperature) and PRCP(precipitation) are input object and the algorithm must predict the output RAIN. By using these input and output, the values of MSE, RMSE and Coefficient of determination respectively 18.4009, 4.2896 and 0.3622 are predicted. In support vector machine,TMIN(minimum temperature),TMAX(maximum temperature) and PRCP(precipitation) are input object and the algorithm must predict the output RAIN. By using these input and output, the values of MSE, RMSE and Coefficient of determination respectively 0.2030, 0.4005 and 0.0319 are predicted. In Naïve bayes, a dataset is partitioned in the range of probability(0.8,0.2). Calculation of confusion matrix for both the test and trained data was done. The values of MSE, RMSE and Coefficient of determination are predicted respectively 0.2506, 0.5006 and 0.0048. Compared to other algorithms SVM exhibits lowest value of root mean squared error.

**CHAPTER 6**

**6. CONCLUSION AND FUTURE WORK**

**6.1 CONCLUSION**

It is very important to estimate rainfall properly for an improved water resources planning, development and management. A support vector machine model was developed to estimate the annual rainfall for Seattle, Washington from 1947 to 2017. The model is able to produce very good result and delivered an excellent matching with the actual data thereby obtaining a very low value of root mean square error equal to 0.4005. Such a low RMSE value is enough to justify the capability of the model to estimate annual rainfall over the area, that ma

**6.2 FUTURE WORK**

We have worked with one hundred and fifty historical data of rainfall. Having a very few data may sometimes lead to the problem of undecidable results. Similarly if we provide millions of records to a machine, the above mentioned problem can be solved. Working with day – to –day records is left as a future scope of this research. SVM has time complexity issues to overcome this flaw it may be compared with some other algorithms for desirable results.

**REFERENCES**

[1] Aakash Parmar, Mithila Sompura and Kinjal Mistree,“Machine Learning Techniques For Rainfall Prediction: A Review”,International Conference on Innovations in information Embedded and Communication Systems (ICIIECS),pp.401-408,2017.

[2] R.Usha Rani, R. Kiran Kumar Reddy and T.K.Rama Krishna Rao,“An Efficient Machine Learning Regression Model for Rainfall Prediction”, International Journal of Computer Applications, Vol.116,No.23,pp.25-30,2015.

[3] Amruta A.Taksande,Dr.S.P.Khandait and Prof.Manish Katkar, “Rainfall Forecasting Using Artificial Neural Network: A Data Mining Approach”, International journal of engineering sciences&research technology,pp.1-3,2014.

[4] Chandrasegar Thirumalai, K Chaitanya Krishna, K Sri Harsha and M Lakshmi Deepak, “Heuristic Prediction of Rainfall Using Machine Learning Techniques”,International Conference on Trends in Electronics and Informatics,pp.1114-1117,2017.

[5] S.Swain,P.Patel and S.Nandi,“A Multiple Linear Regression Model for Precipitation Forecasting over Cuttack District, Odisha, India”,International Conference for Convergence in Technology,pp.355-357,2017.

[6]Kavitha S, Ramya R and Varuna S,“A Comparative Analysis on Linear Regression and Support Vector Regression”,International Conference on Green Engineering and Technologies(IC-GET),pp.978-983,2016.

[7] N.Divya prabha,P.Radha,“prediction of weather and rainfall forecasting using classification techniques”,International Research Journal of Engineering and Technology (IRJET), Vol.06,pp.802-806,2019.

[8] Deepti Gupta and Udayan Ghose,“A Comparative Study of Classification Algorithms for Forecasting Rainfall”,International Journal of Computer Applications,pp.176-181,2015.

[9] Nor SamsiahSani,Israa Shlash,Mohammed Hassan,Abdul Hadi and Mohd Aliff,“Enhancing Malaysia Rainfall Prediction Using Classification Techniques”, Journal of Applied Environmental and Biological Sciences,pp.20-29,2017.

[10] Razeef Mohd,Muheet Ahmed Butt and MajidZaman Baba,“Comparative Study of Rainfall Prediction Modeling Techniques”,Journal of Computer Science and Technology,Vol.7,No.3,pp.13-19,2018.

[11] Jinglin Du,Yayun Liu,Yanan Yu and Weilan Yan,“A Prediction of Precipitation Data Based on Support Vector Machine and Particle Swarm Optimization (PSO-SVM) Algorithms”,Article, pp.1-15,2017.

[12] Kumar Abhishek,Abhay Kumar,Rajeev Ranjan and Sarthak Kumar,“A Rainfall Prediction Model using Artificial Neural Network”,IEEE Control and System Graduate Research Colloquium, pp.82-87,2012.

[13] Emilcy Hern´andez,Victor Sanchez-Anguix,Vicente Julian, Javier Palanca and N´estor Duque,“Rainfall prediction: A Deep Learning approach”,IEEE Conference Paper,pp.1-13,2016.

[14] Mohini P. Darji,Vipul K. Dabhi and Harshadkumar B.Prajapati, “Rainfall Forecasting Using Neural Network: A Survey”,International Conference on Advances in Computer Engineering and Applications (ICACEA),pp.706-713,2015.

[15] Mary N. Ahuna,Thomas J. Afullo and Akintunde A.Alonge,“Rainfall Rate Prediction based on Artificial Neural Networks For Rain Fade Mitigation over EarthSatellite link”,IEEE Africon, pp.579-584,2017.

**APPENDICES**

**A1 SAMPLE SOURCE CODE**

**LINEAR REGRESSION**

regressiondata = read.csv(file.choose(), header=TRUE)

regressiondata

model = lm (TMAX~RAIN, data = regressiondata)

summary(model)

model$residuals

mean(model$residuals^2)

sqrt(mean(model$residuals^2))

plot(model)

**MULTIPLE LINEAR REGRESSION**

regressiondata = read.csv(file.choose(), header=TRUE)

regressiondata

multi.fit =lm (RAIN ~ PRCP + TMAX + TMIN , data = regressiondata)

summary(multi.fit)

multi.fit$residuals

mean(multi.fit$residuals^2)

sqrt(mean(multi.fit$residuals^2))

plot(multi.fit)

**LOGISTIC REGRESSION**

input <- RD[,c("DATE","PRCP","TMAX","TMIN")]

print(head(input))

input <- RD[,c("DATE","PRCP","TMAX","TMIN")]

DATE.data = glm(formula = DATE ~ PRCP + TMAX + TMIN, data = input, family = binomial)

print(summary(DATE.data))

DATE.data$residuals

mean(DATE.data$residuals^2)

sqrt(mean(DATE.data$residuals^2))

**SUPPORT VECTOR MACHINE**

library(e1071)

plot(RD)

plot(RD$DATE, RD$PRCP, col=RD$RAIN)

plot(RD$TMAX, RD$TMIN, col=RD$RAIN)

RD

s<-sample(150,100)

col<- c("TMAX", "TMIN", "RAIN")

RD\_train <- RD[s,col]

RD\_test <- RD[-s,col]

svmfit <- svm(RAIN ~., data = RD\_train, kernal = "linear", cost = .1, scale = FALSE)

print(svmfit)

plot(svmfit, RD\_train[,col])

tuned <- tune(svm, RAIN~., data = RD\_train, kernel = "linear", ranges=

list(cost=c(0.001,0.01,.1,.1,10,100)))

summary(tuned)

p<-predict(svmfit, RD\_test[,col], type="class")

plot(p)

table(p,RD\_test[,3])

mean(p== RD\_test[,3])

svmfit$residuals

mean(svmfit$residuals^2)

sqrt(mean(svmfit$residuals^2))

chisq.test(svmfit$coefs^2)

**NAIVE BAYES**

library(naivebayes)

library(dbplyr)

library(ggplot2)

library(psych)

data <- read.csv(file.choose(), header = T)

str(data)

xtabs(~TMAX+RAIN, data = data)

data$RAIN <- as.factor(data$RAIN)

data$TMAX <- as.factor(data$TMAX)

#Data Partition

set.seed(1234)

ind <- sample(2,nrow(data), replace = T, prob = c(0.8,0.2))

train <- data[ind == 1,]

test <- data[ind == 2,]

#Naive Bayes Model

model <- naive\_bayes(RAIN ~., data = train)

model

plot(model)

#Predict

p <- predict(model, train, type = 'prob')

head(cbind(p, train))

model$prior

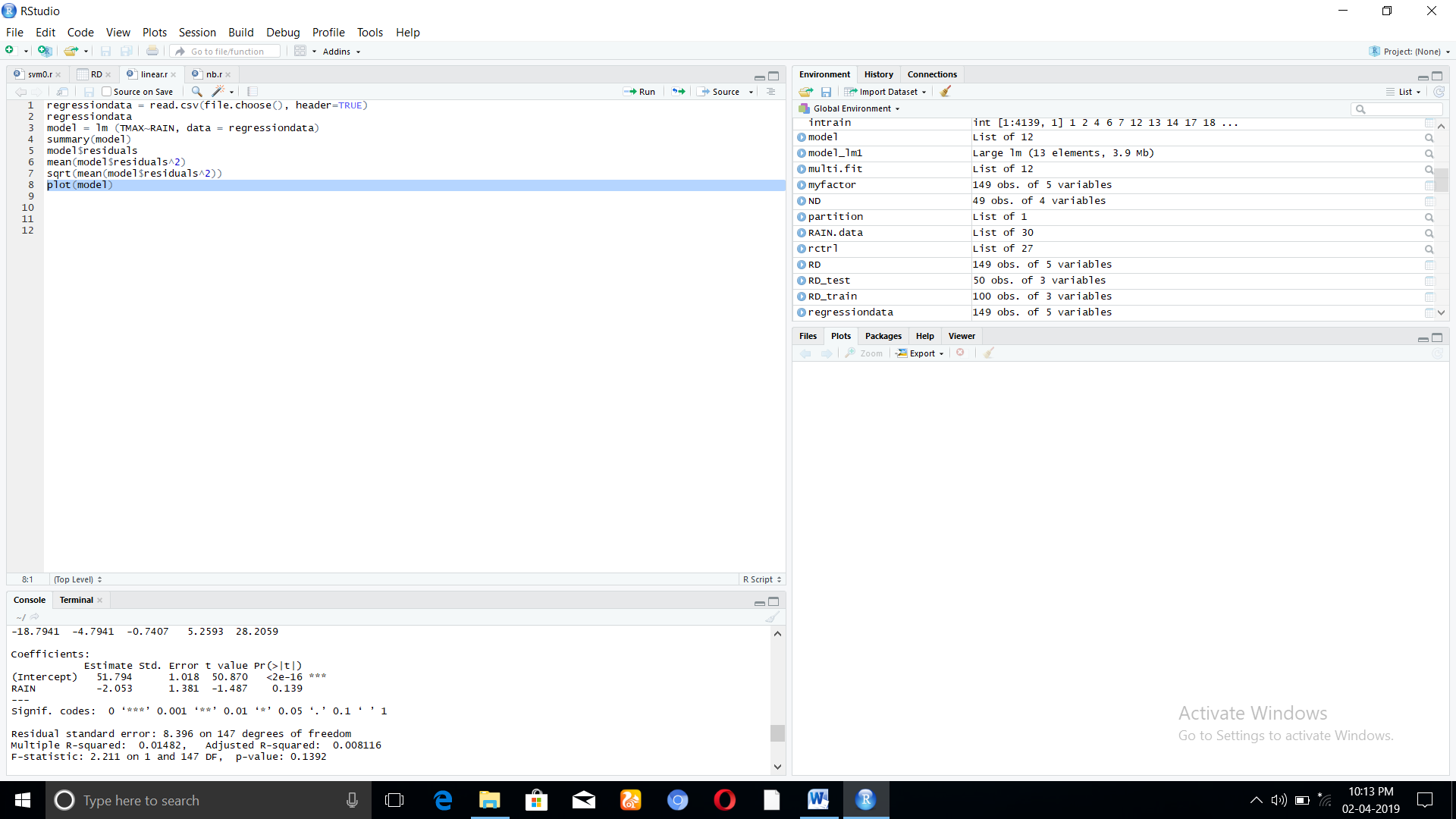
mean(model$prior^2)

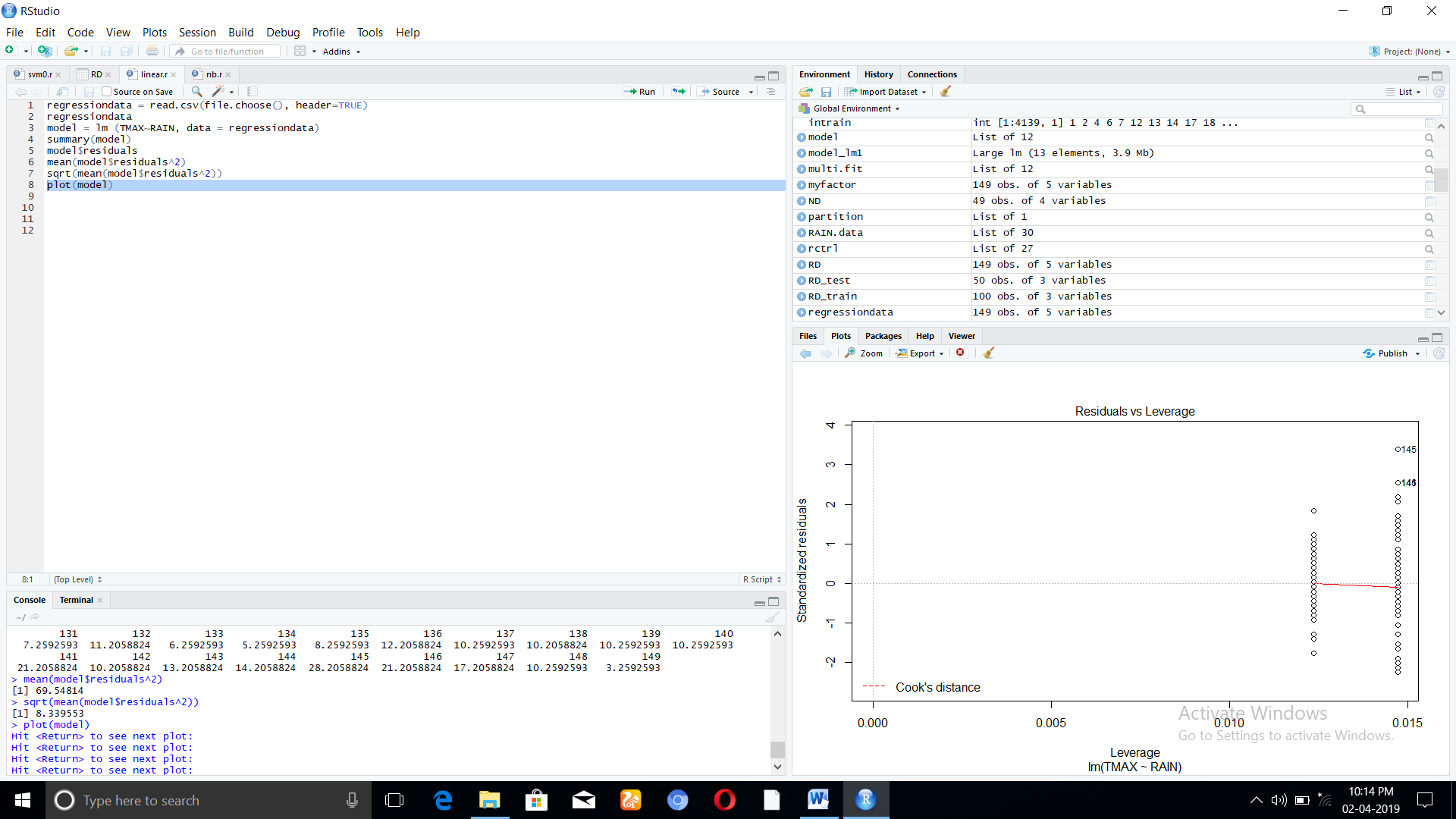
sqrt(mean(model$prior^2))

chisq.test(model$prior^2)

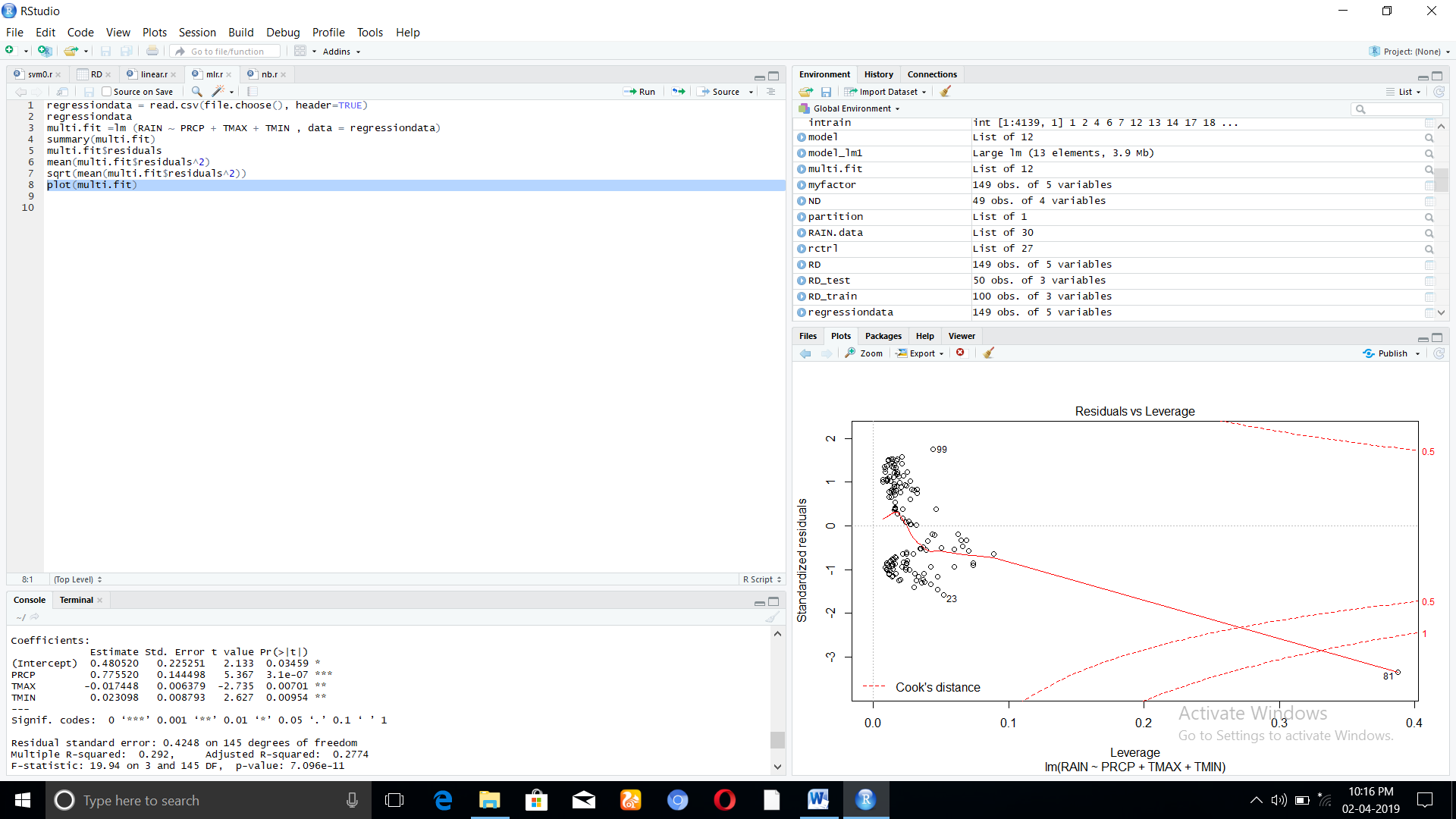
**A2 SAMPLE SCREENSHOTS**

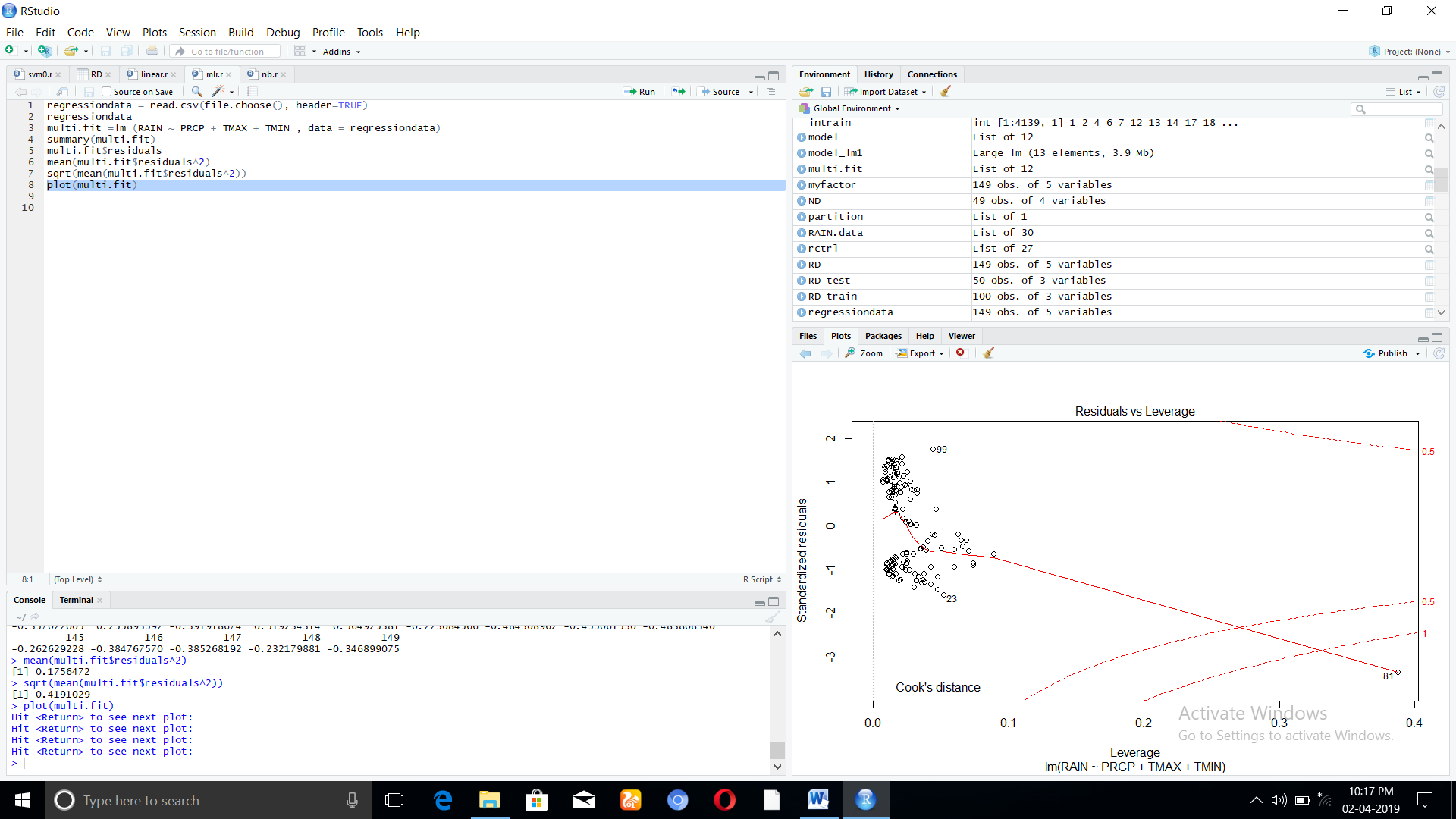
**LINEAR REGRESSION**



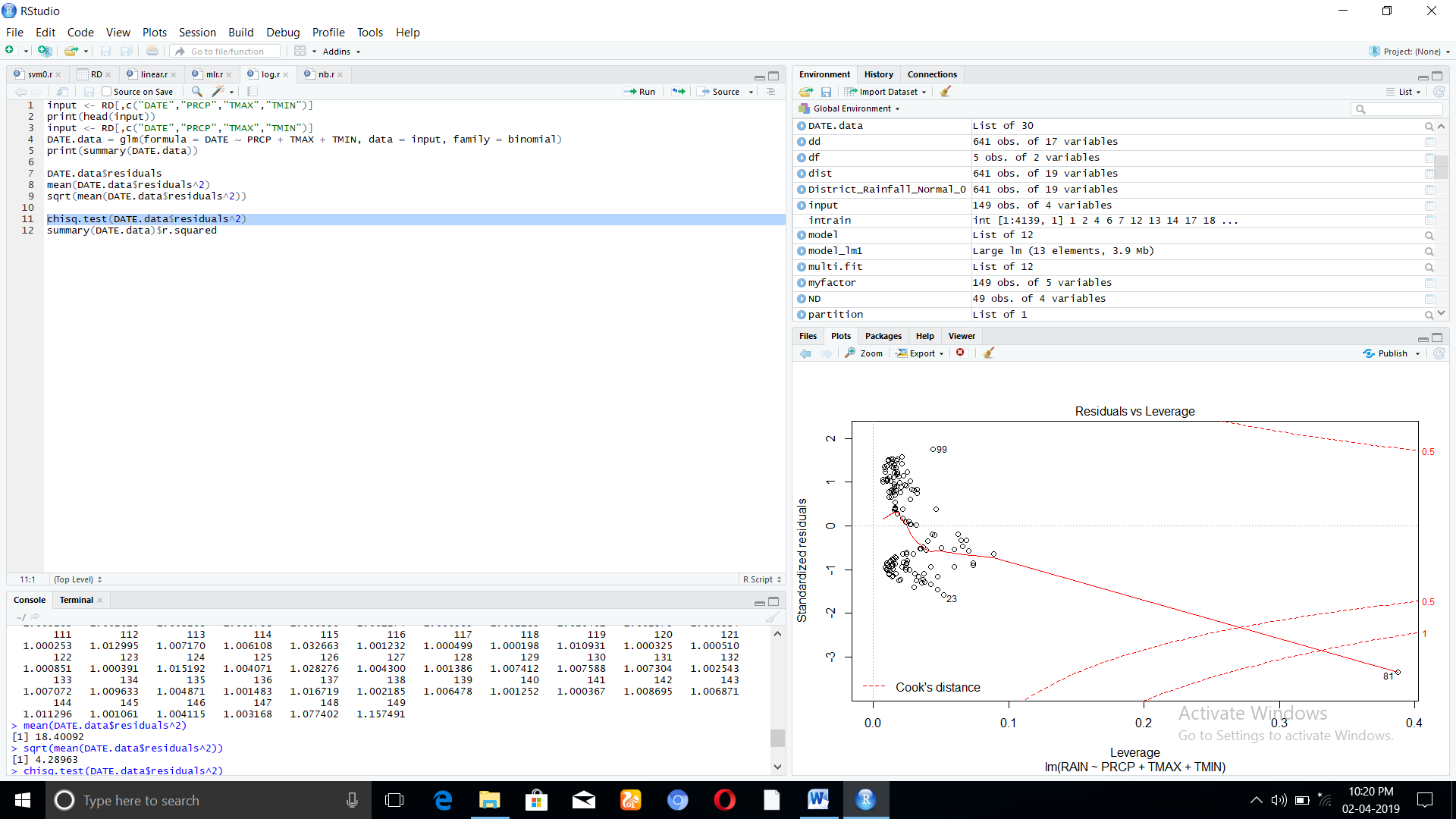


**MULTIPLE LINEAR REGRESSION**

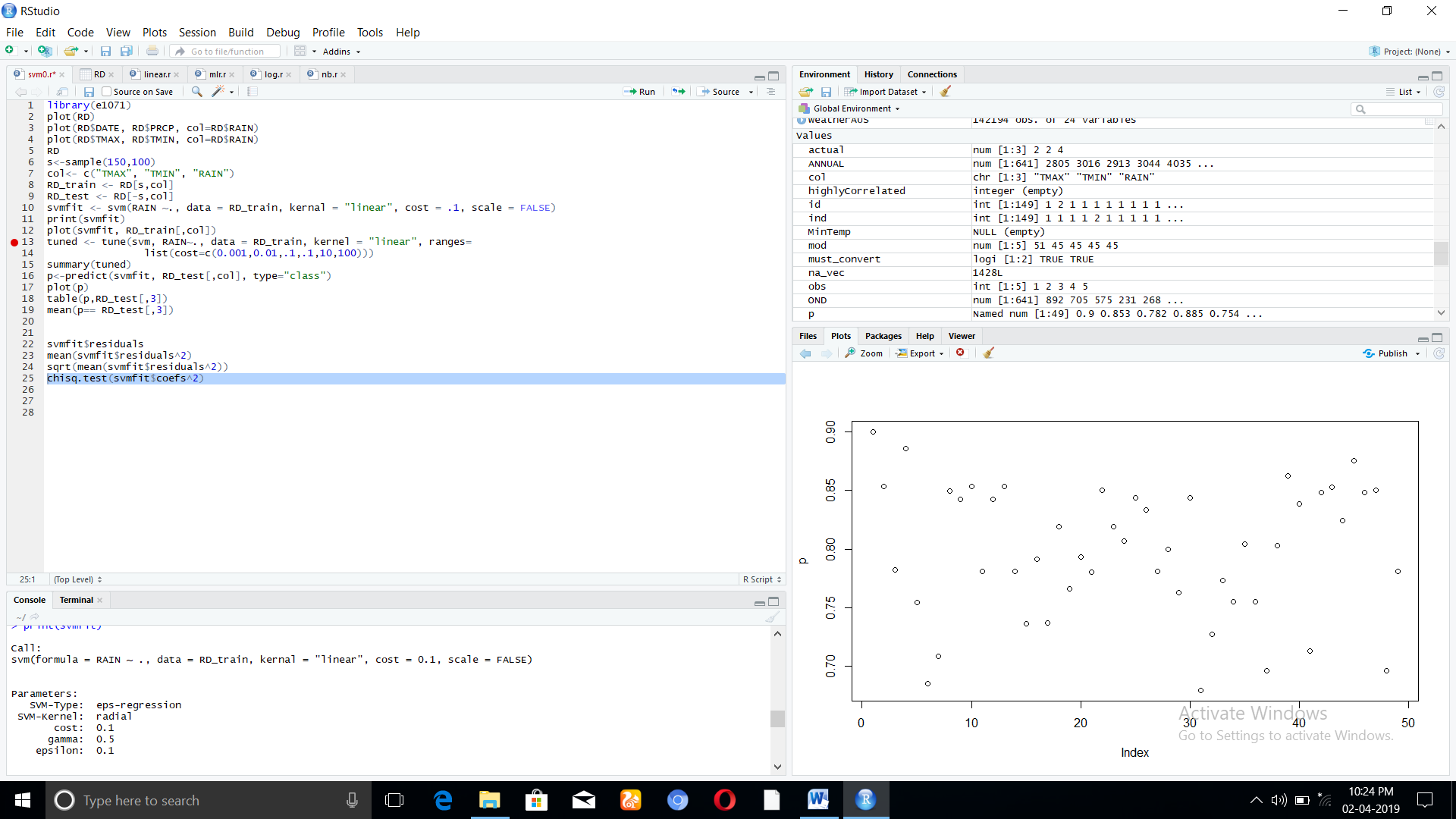


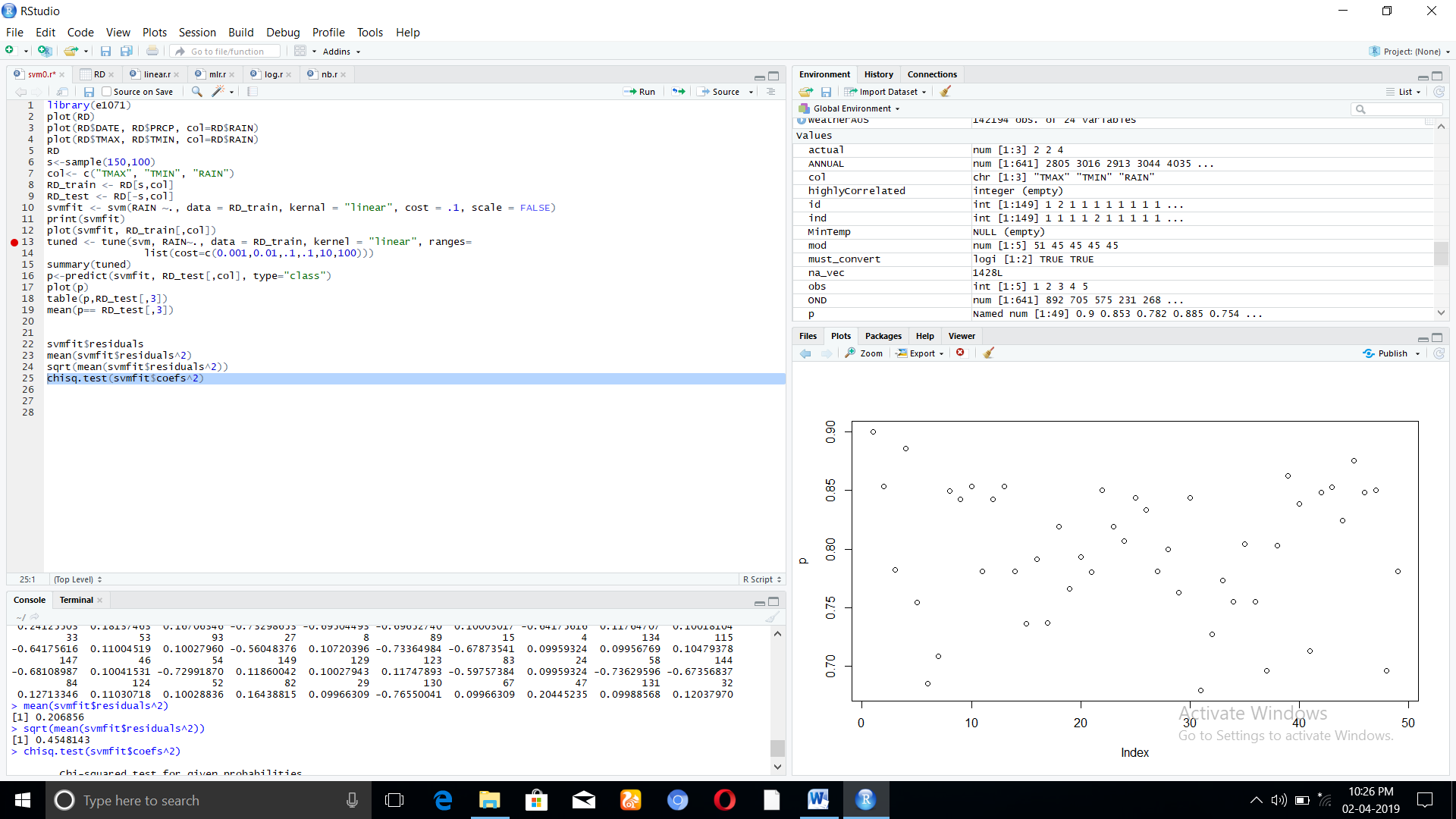


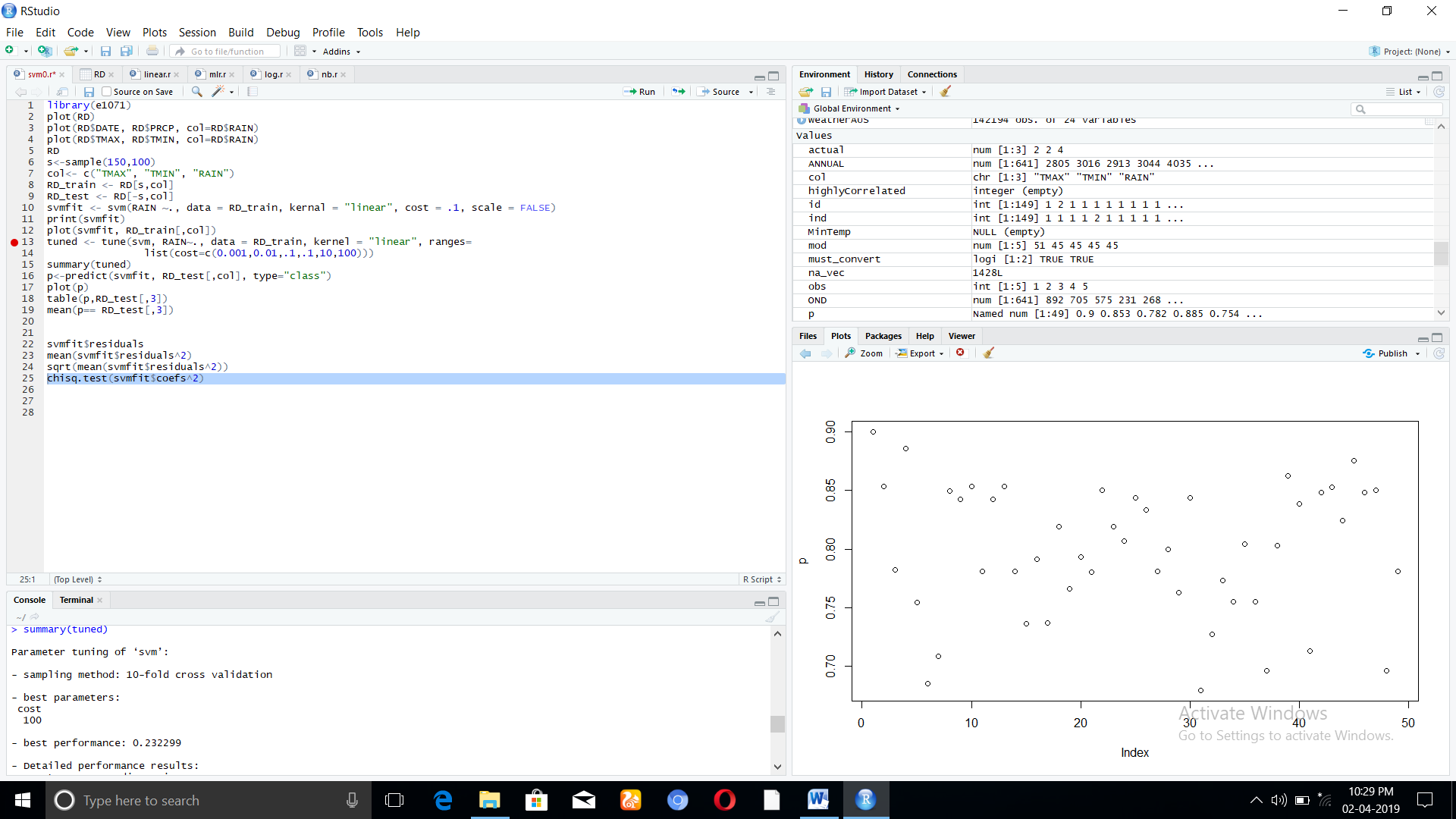
**LOGISTIC REGRESSION**

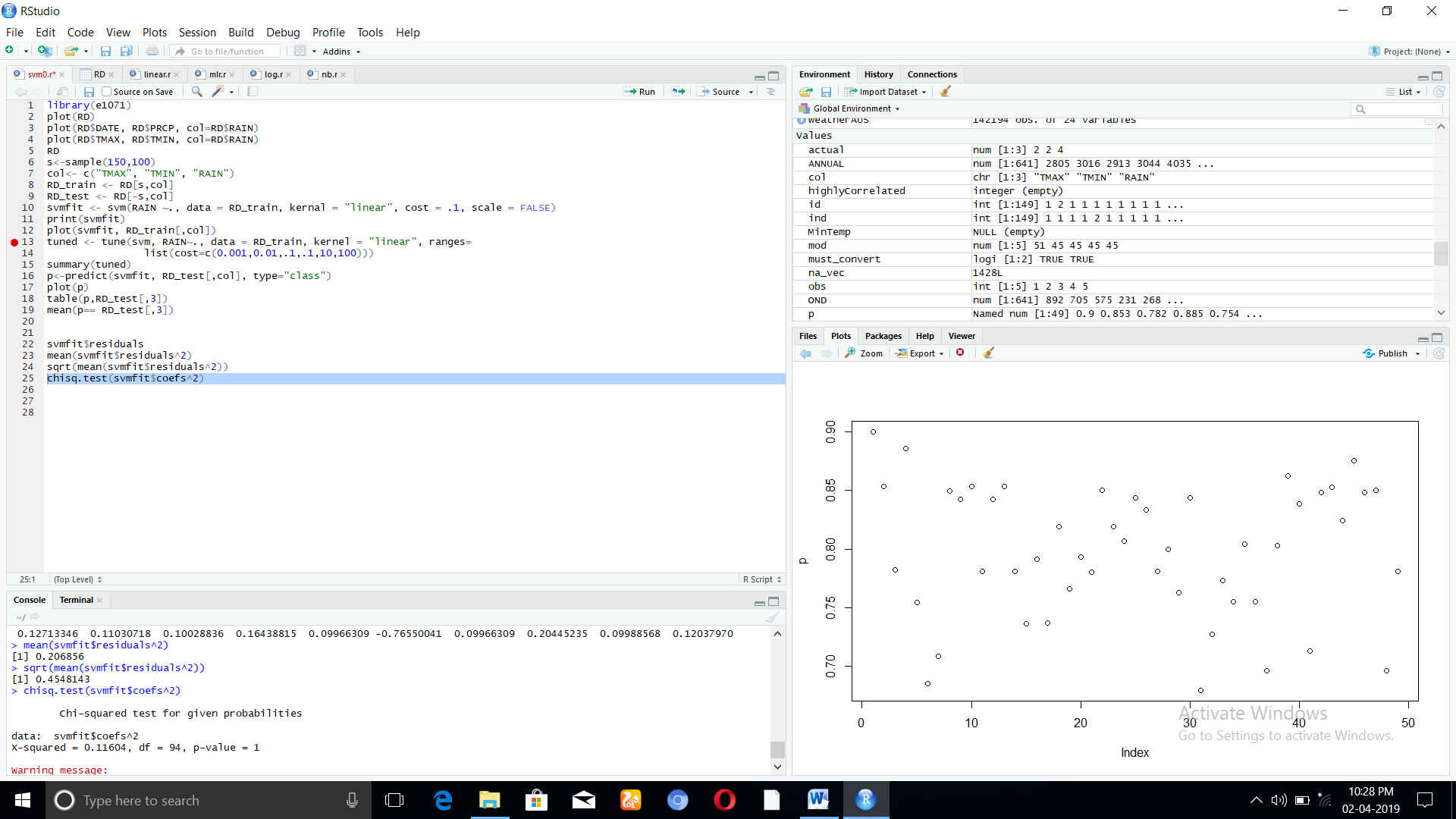


**SUPPORT VECTOR MACHINE**









**NAIVE BAYES**

