

**REPUBLIC OF TURKEY
FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND
APPLIED SCIENCES**



**THE ESTIMATION OF CLIMATE PARAMETERS
USING DATA MINING TECHNIQUES**

SATTAR NABEE RASOOL

**Master Thesis
Department: Software Engineering
Supervisor: Prof. Dr. Ahmet KOCA**

August - 2017

REPUBLIC OF TURKEY
FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**THE ESTIMATION OF CLIMATE PARAMETERS USING DATA MINING
TECHNIQUES**

MASTER THESIS
SATTAR NABEE RASOOL
(151137105)

Date of Application: 20 June 2017

Thesis Defense Date: 9 August 2017

Thesis Supervisor : Prof. Dr. Ahmet KOCA

Other Juries Members: Assoc. Prof. Dr. Cihan VAROL

: Assoc. Prof. Dr. Murat KARABATAK

August - 2017

ACKNOWLEDGEMENT

I dedicate this work to my family, who have supported and encouraged me in this work, and to my beloved mother and my wife who have always given me compassion and love. I would also like to express my gratitude to my supervisor, Prof.Dr. Ahmet Koca, for his patient guidance and invaluable advice, for numerous discussions and encouragement throughout the course of the research. I would also like to thank Him that conveys the light of His knowledge to others, who guides perplexed questioners to the right answers, who has stood by me every step of the way to completion of this work, every appreciation and respect to Him

I would like to thank all in the Software Engineering department at Firat University, particularly Prof. Dr. Asaf Varol, the head of department, for his help, support, advice and encouragement.

Special thanks also go Dr. Murat Karbatak for his never-ending support, love and feedback which helped me face the immense pressure during my masters and made my stay in Elazig comfortable. I appreciate his vast knowledge and skills in the parallel systems and network area. Thank you for your tutelage, advice and guidance during my first year at Firat University. Working with you has already helped me so much and will continue to inspire me. Your devotion and vineyard with your time in helping me with other side-projects are greatly appreciated also. It is my pleasure to send you this very sincere thank you.

You certainly are a great mentor for me and you have been very generous in sharing your rich and worthy knowledge in your field!

I also want to thank my all friends, who have bolstered me all through the whole process, both by keeping me congruous and helping me assembling pieces. I will be eternally grateful. Last but not least, want to thank my family: my folks, my brothers and sister for supporting me profoundly all through composing this proposal and my life overall.

Sincerely

SATTAR NABEE RASOOL

LIST OF CONTENT

	Page No
ACKNOWLEDGEMENT	II
LIST OF CONTENT	III
ABSTRACT	V
ÖZET	VI
LIST OF FIGURES	VII
LIST OF TABLE	VIII
ABBREVIATION	IX
1. INTRODUCTION	1
1.1. Need for Estimating Climate Parameters	1
1.2. Motivation	1
1.3. Problem Definition	2
1.4. Aims and Objectives	2
1.4.1. Aims	2
1.4.2. Objectives.....	2
1.5. Scope of the Research	3
1.6. Structure of the Thesis.....	3
2. LITERATURE REVIEW	5
2.1. Insights on Data Mining Techniques	5
2.2. Previous Research Works in the Area.....	6
2.3. Summary	13
3. SYSTEM ANALYSIS	14
3.1. Existing System.....	14
3.2. Proposed System	14
3.3. Functional Requirements.....	15
3.4. Non-Functional Requirements	15
3.5. Software Requirements	16
3.5.1. Weka.....	16
3.6. Methodology	19
3.6.1. Neural Networks	20
3.6.2. Support Vector Machine (SVM).....	23

3.6.3. Linear Regression.....	25
3.7. Summary	27
4. IMPLEMENTATION.....	28
4.1. Case Study.....	30
4.2. Architecture Diagram	31
4.2.1. Two-Tier Architecture.....	32
4.3. Modules (Weka software)	33
4.3.1. Load Dataset.....	33
4.3.2. Linear Regression.....	33
4.4.3. Neural Networks	33
4.3.4. Support Vector Machine	33
5. RESULTS DISCUSSION AND EVALUATIONS	35
5.1. Applied Linear regression and SVM and ANN	35
5.2. Evaluation Results.....	39
5.2.1. Obtained Results for Wind Perdition	39
5.2.2. Obtained results for three Statically Matrices	49
5.3. Summary	52
6. CONCLUSION.....	53
6.1. Summary of Findings	53
6.2. Conclusions and Future Work.....	53
REFERENCES	55
APPENDICES	59
Appendix A: Climate Dataset	59
CIRRUCULUM VITAE	69

ABSTRACT

The Estimation of Climate Parameters Using Data Mining Techniques

Nowadays, estimation of climate parameters is a crucial phenomenon. The rationale behind this phenomenon is that important decisions in many applications are based on predicting the weather. Data mining is widely used by enterprises to discover knowledge from large databases and data mining techniques are essential for estimation of climate parameters. The data of these parameters are huge and need prediction mechanisms. In this thesis, a methodology for wind velocity prediction is proposed. Predictive data mining algorithms neural networks, linear regression and Support Vector Machine (SVM), are used to estimate wind velocity. A prototype application is built to demonstrate proof of the concept. The prototype exploits Weka data mining API provided using Java programming language. The climate dataset used in the experiments has a large amount of data from intelligent stations, for every day of the month from several different regions. It has attributes such as station number, month, day, pressure, humidity and temperature. Wind velocity is the parameter for which prediction is made by using the three different algorithms. Wind velocity estimation has its utility in weather forecasting. The relation between velocity and weather forecasting is the start point of the thesis. Based on the existing data available on climate parameters, the chosen data mining algorithms perform their logic in order to estimate the wind velocity. After completing computations, the observations are presented and the results are compared with actual wind velocity. The error rate is also considered to evaluate the performance of the three algorithms. From the empirical study, it is understood that the prediction performance of Linear Regression is higher than the other two data mining algorithms.

Keywords: Data mining techniques, climate parameter, renewable energy, weka

ÖZET

Veri Madenciliği Tekniklerinin Kullanarak İklimlendirme Parametrelerinin Tahmini

İklim parametrelerinin tahmini, günümüz dünyasında çok önemli bir olgudur. Bu olgunun arkasındaki mantık, pek çok uygulamada önemli kararların hava durumu bilgisi tahmine dayalı olmasıdır. Veri madenciliği, büyük veritabanlarından bilgi keşfetmek için işletmeler tarafından yaygın şekilde kullanılır. Bu tez çalışmasında rüzgar hızı tahmini için bir metodoloji önerilmektedir. Rüzgar hızını tahmin etmek için Sinir Ağları, Doğrusal Regresyon ve Destek Vektör Makinesi (DVM) gibi öngörülü veri madenciliği algoritmaları kullanılmıştır. Bir prototip uygulaması, konseptin kanıtını göstermek için oluşturulmuştur. Prototip, Java programlama dili kullanılarak sağlanan Weka veri madenciliği uygulama programlama arabirimi ile çalıştırılmıştır. Deneylerde kullanılan iklim veri seti akıllı istasyonlardan ve ayın her günü birkaç farklı bölgeden gelen çok miktarda veriye sahiptir. İstasyon numarası, ay, gün, basınç, nem ve sıcaklık gibi özellikleri vardır. Rüzgar hızı, üç farklı algoritma kullanılarak tahmin edilen parametredir. Rüzgar hızı tahmininin hava durumu tahmininde faydası vardır. Hız ile hava tahmini arasındaki ilişki tezin başlangıç noktasıdır. İklim parametreleri üzerinde bulunan mevcut verilere dayanarak, seçilen veri madenciliği algoritmaları, rüzgar hızını tahmin etmek için kullanılmıştır. Hesaplamaları tamamladıktan sonra, gözlemler sunulmuş ve sonuçlar gerçek rüzgar hızı ile karşılaştırılmıştır. Hata oranı da üç algoritmanın performansını değerlendirmek için düşünülmüştür. Ampirik çalışmadan Doğrusal Regresyonun öngörme performansının diğer iki veri madenciliği algoritmalarından daha yüksek olduğu anlaşılmaktadır.

Anahtar Kelimeler: Veri Madenciliği teknikleri, İklim Parametreleri, Yenilenebilir Enerji, Weka

LIST OF FIGURES

	Page No
Figure 3.1. Weka start-up screen.....	17
Figure 3.2. Weka Explorer	18
Figure 3.3. Proposed methodology	20
Figure 3.4. A sample of simple Neural Network of Canonical Activation Function.	21
Figure 3.5. Neural Networks diagram	22
Figure 3.6. Three-layer-ANN-architecture	23
Figure 3.7. The solution One-dimensional regression with epsilon intensive band ..	24
Figure 3.8. Shown Pre-Specified Value	24
Figure 3.9. Support Vector Machine diagram.....	25
Figure 3.10. Linear Regression model	27
Figure 4.1. Two tier architecture used in the proposed system.....	32
Figure 5.1. Apply linear regression in weka software	35
Figure 5.2. Apply support vector machine in weka software	36
Figure 5.3. Apply Multilayer perceptron	38
Figure 5.4. Actual and predict wind velocity in city INEBOLU in month June.....	40
Figure 5.5. Actual and predict wind velocity in city INEBOLU in month July.....	40
Figure 5.6. Actual and predict wind velocity in city INEBOLU in month August....	41
Figure 5.7. Actual and predict wind velocity in city SINOP in month June.....	42
Figure 5.8. Actual and predict wind velocity in city SINOP in month July.....	42
Figure 5.9. Actual and predict wind velocity in city SINOP in month August.....	43
Figure 5.10. Actual and predict wind velocity in city AMASRA in month June	44
Figure 5.11. Actual and predict wind velocity in city AMASRA in month July.....	44
Figure 5.12. Actual and predict wind velocity in city AMASRA in month August..	45
Figure 5.13. Actual and predict wind velocity in city SILE in month June.....	46
Figure 5.14. Actual and predict wind velocity in city SILE in month July	46
Figure 5.15. Actual and predict wind velocity in city SILE in month August	47
Figure 5.16. Actual and predict wind velocity in city AKCAKOCA in month June	48
Figure 5.17. Actual and predict wind velocity in city AKCAKOCA in month July .	48
Figure 5.18. Actual and predict wind velocity in city AKCAKOCA in month July .	49

LIST OF TABLE

	Page No
Table 4.1. Sample datasets usage for getting results.....	31
Table 4.2. Geographical position	31
Table 5.1. Performance comparison in terms of wind prediction value	39
Table 5.2. Results for three algorithms by months	50
Table 5.3. Results for three algorithms by Stations	52



ABBREVIATION

ANN	: Artificial Neural Network
API	: Application Program Interface
ARIMA	: Autoregressive integrated moving average
DM	: Data mining
MC	: Markov Chain
EP	: Evolutionary Programming
GUI	: Graphical user interface
GRNN	: General Regression Neural Network
IDE	: Integrated Development Environment
JDK	: Java Development Kit
JDBC	: Java Database Connectivity
LR	: Linear Regression
MLP	: Multilayer Perception
RP	: Resilient Propagation
RPE	: Recursive Prediction Error
SVM	: Support Vector Machine
SVR	: Support Vector Regression
RPE	: Recursive Prediction Error
SCG	: Scaled Conjugate Gradient
TSMS	: Turkish State Meteorological Service
TLS	: Total Least Squares
UI	: User interface
PC	: personal computer
RMSE	: Root Mean Square Error

1. INTRODUCTION

Estimation of climate data in real time is very essential as it can provide valuable information to people of different domains, such as agriculture, aviation and tourism to mention but a few. However, since climate data is growing exponentially it is difficult to analyse manually.

Therefore, machine-learning techniques, such as unsupervised and supervised learning methods, are used to mine extensive data and discover valuable knowledge. Predictive modeling in data mining is required to estimate climate parameters. In this thesis, we propose a framework that exploits data mining techniques such as neural networks, SVM and linear regression. The framework takes climate dataset as input, completes the training phase and creates different models using data mining algorithms. Finally, it exploits linear regression, which models the relationship between a dependent variable and an exploratory variable. The framework results in estimating wind velocity. We built a prototype application based on Weka, which is used to demonstrate proof of the concept. The empirical results revealed that the proposed framework is useful to have a predictive model with respect to the estimation of climate parameters.

1.1. Need for Estimating Climate Parameters

Estimating climate parameters is very useful to people nowadays because it estimates weather information. This study used three different data mining techniques to estimate the wind speed using the existing climate parameters, using six parameters, input station number, month, day, pressure, humidity and temperature. Based on these input parameters, wind velocity is computed using different classification and regression algorithms

1.2. Motivation

The estimation of climate parameters using data mining techniques is useful for calculating or estimating the weather, based on the data captured in different locations and environments, considering all sensitivity climate parameters that directly impact the wind speed. Then, input parameters are calculated and wind velocity predicted using the

different classification and regression algorithms, likewise the consideration of different times and months. Here, the classification algorithms SVM, neural networks and the regression algorithm of linear regression are used in this thesis. When the weather parameter, such a wind speed, is estimated, it can help all stakeholders in making important decisions. People with different agendas all require weather information. Players need it, soldiers need it and fishermen need it. This is the motivation behind taking up this project.

1.3. Problem Definition

Climate data are huge and needs to be processed in order to obtain the important information required to decision- making. Making expert decisions is a main problem if information on the weather is lacking. Manual analysis of data is time-consuming and error prone; therefore, it is essential to have an automated process which can analyse the data and provide valuable information. The reason being that enterprises and people of all lifestyles need information on the weather. Thus, automatic estimation of weather is the problem to be addressed. In this project, data mining techniques are explored with the help of the Weka tool in order to estimate weather information.

1.4. Aims and Objectives

1.4.1. Aims

The aim of the thesis is to analyse weather data and provide an estimation of weather conditions, such as wind velocity. This information has utility in the real world.

1.4.2. Objectives

To fulfill the aim of the thesis, the following are the identified objectives.

1. To investigate the present state-of-the-art of estimation of climate parameters, including techniques used, advantages and limitations
2. To propose a methodology that helps in the estimation of wind velocity by analyzing weather data.

3. To implement the proposed system by using Weka software with the help of data mining techniques.
4. To evaluate the work done.
5. To implement the proposed system by using the Weka tool for different algorithms. Weka is a tool used in machine learning for solving data mining tasks to evaluate the work done.
6. To use the weka tool for classifying the dataset
7. Use different algorithms from regressions under different categories.

1.5. Scope of the Research

The scope of the project is limited to developing an application for estimating the climate parameters using data mining techniques. The proposed system provides the techniques that can help individuals and organizations to obtain useful information on climate conditions. The proposed solution explores different data mining techniques, such as neural networks, Support Vector Machine (SVM) and linear regression. It takes climate dataset as input and performs a series of mining activities in order to provide an estimation of climate parameters.

1.6. Structure of the Thesis

The remainder of the thesis is structured as follows. A brief overview of each chapter is provided here to help the reader to understand the essence of the chapters

Chapter 2

This reviews the literature on the present state-of-art of the estimation of climate parameters using data mining techniques. It provides the different methodologies and different techniques used.

Chapter 3

This chapter covers system design. It includes the modelling of the proposed system for estimation of climate parameters using data mining techniques.

Chapter 4

This chapter covers the implementation details of the proposed system. It talks about the environment in which the application was implemented and the technologies used to implement. The business logic used to implement the functionalities of the system is also covered.

Chapter 5

This chapter provides the results of the implementation. It provides a graphical view of the proposed system implementation. It presents the images that indicate the functionalities of the data mining techniques used for estimating the climate parameters.

Chapter 6

This chapter provides conclusions and recommendations. The conclusions are drawn from the experience of the application development while the recommendations provide possible future work to be carried out.

.

2. LITERATURE REVIEW

This chapter provides a review of the literature on different data mining techniques. It also throws light on the techniques that are used for the different purposes of weather estimation. There are many data mining techniques and these are used to discover hidden knowledge from databases. The business intelligence discovered through data mining is used in expert decision-making. The data mining techniques particularly reviewed in this chapter include neural networks, regression and SVM. The SVM algorithm is explored as a binary classifier and the classifier, which can predict multiple class labels with the help of a kernel mechanism. This chapter also throws light on the data mining techniques that cater to the needs of different domains. It reviews the approaches used by different researchers on the usage of data mining techniques.

2.1. Insights on Data Mining Techniques

Langone et al. [1] proposed a machine learning method known as Least Squares Support Vector Machine (LS-SVM) for fault detection online with industrial machine data. It is a supervised learning method, which helps to know early detection and classification of faults. Claesen et al. [2] proposed a supervised learning model using SVM. The model makes use of positives and unlabelled objects. It actually makes use of an ensemble of SVM models in order to have an intuitive approach for an appropriate learning model. It has features such as supervised learning, positive and unlabelled learning, and learning with false positives. Frandi et al. [3] propose another SVM classification approach along with an algorithm for large-scale benchmark data classification. Drumetz et al. [4] focused on dimensionality estimation algorithms for finding intrinsic dimensionality in datasets. Understanding dimensionality can help in gaining prior knowledge for making decisions. When the kernel is used, it is possible to have dynamism and it supports multiple class labels. Thus, it is possible to overcome the limitations of binary classifier SVM.

Schmidhuber [5] made a review of data mining techniques that are used for deep supervised and unsupervised learning. In particular, they focused on neural networks. Neural networks are the technology that mimics neurons in the human brain. As the human brain and neurons are so complex, the neural networks also contain complexity with

different layers, such as input and output, and are widely used to solve complex problems. Aziz and Yusof [6] proposed a classification model to analyse graduates' employment. They employed different algorithms, such as J48, k-Nearest Neighbor (KNN), multilayer perception, logistic regression and Naive Bayes. They found that logistic regression resulted in the highest accuracy, 95.2%. Sanghavi et al. [7] explored three kinds of algorithms related to feature selection. They are known as Filter, Wrapper and Embedded. They built a logistic regression algorithm in order to identify diseases using medical data mining.

In [8], the Weka data mining tool is used with different algorithms and datasets. They explored many datasets with various classification algorithms. These come under supervised learning and are machine learning techniques that are widely used to label unlabelled classes. Generally, they are used to predict objects with the right labels assigned. In [9], classification algorithms such as SVM and KNN are combined and used. Thus, the hybrid solution was able to cater to the needs of the application. In particular, it was able to obtain the benefits of both worlds.

2.2. Previous Research Works in the Area

Barbounis and Theocharis [10] proposed a method for wind speed prediction that can be employed in wind farms. They took spatial information pertaining to weather from remote stations. As the data contained temporal information, they determined to make use of local current neural networks coupled with spatial correlation. Thus, they provided a methodology for wind speed prediction. They employed online learning algorithms for training based on an approach named Recursive Prediction Error (RPE). For a better prediction model, they used a Global RPE (GRPE) that considers a concurrent update of all weights associated with the prediction. Then, they tried an enhanced RPE, known as Decoupled RPE (DRPE), which works with less computational costs. An ad-joint model approach was used in order to have the partial derivatives needed by online learning algorithms. They considered a real world wind form problem and evaluated the efficiency of the solution they proposed. When compared with other models, such as recurrent forest models and gradient descent algorithms, the DRPE showed superior performance improvement.

Kani and Ardehali [11] proposed a hybrid approach for short-term wind speed prediction. They used Markov Chain (MC) and Artificial Neural Network (ANN) for synergic effect in the performance. They considered one hour for the short term and a few minutes for very short-term experimentations. The ANN approach was used for short patterns in wind speed prediction while long-term patterns are considered in the MC model. The rationale behind the usage of MC for the long term is that it is capable of memorising long-term behavioural aspects. When these two models are combined, the method is termed ANN-MC, which leverages the prediction performance. Thus, the hybrid model can handle both short-term and long-term patterns in wind speed prediction. Another advantage of the proposed combined approach is to reduce error rates. Prediction error was minimised. At the same time, the resultant model ensured a reduction in computation time and uncertainty in formal predictions of wind speed.

ANN is used in many different applications. For instance, Kok et al. [12] used it for thermal mixing phenomena. In other words, they used ANN for estimating the efficiency of thermal mixing in the channel. The parameters used in the experiments of thermal mixing include temperature difference between cold and hot jet inlets diameters, flow rate ratio of inlet fluid and angle of the channel. These parameters were used while experimenting with thermal mixing and the thermal index was computed-based on the temperatures measured. The ANN model was employed in order to reduce computational time. With respect to a forward model associated with ANN, a very limited number of measurements is used. Thus, the ANN could predict results or output parameters without actually doing an experiment. The mixing index, MI, is used to measure the closeness of temperature profile with regards to mean temperature observed.

Senpinar and Karabatak [13] also employed ANN for a different purpose in the experiments carried out in Turkey. They employed ANN for estimation of something related to sustainable energy, incomplete bacteriological and solar radiation. Solar radiation in specified places or cities was considered for experimentation. For unforeseen reasons, it was found that the data collected from different cities were incomplete. The incomplete values were completed by employing the ANN for better prediction. In order to provide training for ANN, it is essential to have data pertaining to bacteriological and geological factors. The data used by ANN include temperature, latitude, longitude, pressure, relative humidity, year, month, altitude, cloudy day, sunny day, etc. These parameters are effectively used by ANN for prediction of incomplete data and to complete

them. They used an ANN back propagation algorithm with two variations, known as Scaled Conjugate Gradient (SCG) and Levenberg-Margardt. They found that ANN is feasible for such prediction problems.

Karabatak and Senpinar [14] also employed ANN for prediction of monthly average daily global solar radiation based on the trends in the historical data. They also made a prediction model for the temperature of the future month in order to have a useful prediction of weather. The experiments were performed at six stations geographically located in Turkey. The geographical and bacteriological data were used for prediction purposes. The ANN input layer takes historical data and produces desired prediction values into output layer. The measured data, which is the ground truth, and the details obtained from the ANN method were compared for performance evaluation. The average accuracy of solar radiation was found at 99.35% and the average temperature found to be 99.57%. These results reveal the significance of the ANN in prediction models. In fact, ANN is also useful in different applications, such as wind velocity prediction.

Pinson et al. [15] focused on a wind power application as pertaining to wind velocity. They employed an adaptive orthogonal fitting along with local linear regression and observed that it is possible to convert meteorological data, such as wind velocity, into a power production model. This model is meant for short-term forecasting of wind generation. In this case, the resultant power curve is non-linear. For power curve estimation, there is a nonparametric approach known as local linear regression. Moreover, the recursive Least Square (LS) method are used for model building and tracking. However, there is an assumption needed, this being the presence of noise component. Such assumption may lead to results with inaccuracy. Therefore, there needs to be an efficient approach for relaxing this assumption. To this end, they proposed an adaptive orthogonal fitting with local linear regression in order to overcome the issue. With this, the criterion used is known as weighted Total Least Squares (TLS). Thus, it is made non-stationary and efficient for estimating wind power with short-term forecasting.

Bilgili et al. [16] studied wind speed prediction using data collected from real world weather stations. They used a technique known as Artificial Neural Networks (ANN) for the monthly wind speed prediction. The data were collected for hourly wind speed from the Turkish State Meteorological Service (TSMS), which was collected from eight different measuring stations across the nation. Using ANN, they applied an algorithm known as the Resilient Propagation (RP) learning algorithm. The hidden layer possesses a

logistic sigmoid transfer function while the output layer gives linear transform function. A model was finally built using ANN which could be updated with new data from time to time. The model produces wind speed predicted values that are compared with ground truth values for estimating the performance of their proposed system. The error rate was at 14.13% and, at a selected station, was just 4.49%.

Mohandas et al. [17] employed Support Vector Machine (SVM), which is one of the most used data mining algorithms for classification. They used it for prediction of wind speed. The performance of this algorithm was compared with another algorithm named Multilayer Perception (MLP) neural networks. The data were taken from the mean daily wind speed data from Madina City in Saudi Arabia. Root mean square errors were used to compare the two algorithms. This experiment was done to have renewable energy resources in the country. Mean square error was computed and compared for performance evaluation. Normalised daily wind speed exhibited the dynamics of wind speed in Madina City. The SVM was used with a Gaussian kernel. The rationale behind this is that SVM is, by default, a binary classifier. To make it support multiple classes, the concept of the kernel is employed in order to have a better application of SVM for wind prediction.

Salcedo-Sanz et al. [18] employed evolutionary algorithms based on SVM for short-term wind speed prediction. Regression SVM was used for parameter estimation while other algorithms were used for performance evaluation, namely Particle Swarm Optimization (PSO) and Evolutionary Programming (EP). It was dining in Spanish wind form with regression SVM for wind speed prediction. PSO is a population-based approach used for the purpose. The procedure followed for the experiments included global forecasting model, mesoscale model (MM), regressor and, finally, wind speed prediction on each turbine. Physical downscaling and statistical downscaling are two approaches followed in order to have a hybrid approach for short-term wind speed forecasting. Two turbines were used for experiments for evaluation of the proposed method.

Liu et al. [19] shed light on the wind speed prediction problem and compared a hybrid approach with other such algorithms. The hybrid was made up of two method combinations, known as Autoregressive Integrated Moving Average- Artificial Neural Network (ARIMA-ANN) and Autoregressive Integrated Moving Average-Kalman (ARIMA-Kalman). The two hybrid methods were employed and theory performance compared. With ARIMA, the two variants used were Kalman filter and ANN. The structure of ANN was used to determine ARIMA while the ARIMA was used to have

initialisation of the Kalman filter in the latter hybrid model. Both cases proved that they can be used for non-stationary wind speed prediction. These algorithms are best used with real world wind power systems. The proposed framework uses original wind speed series data which contain historical values. ARMA is a common phase for both hybrid algorithms. Then, the ANN model was used in one approach while the other approach followed the Kalman Filter model. Both hybrid algorithms involve a step known as multi-step prediction, which results in a comparative study of wind speed prediction performance of the two algorithms.

Cheng and Guo [20] proposed a method for wind speed prediction for short-term. They employed fuzzy logic-based SVM. SVM is used as a regression prediction method. Fuzzy information granulation is the underlying logic used along with SVM. The result of this method is information granulation. Then, the information granulation is exploited by SVM for better prediction of wind speed. Once the prediction is completed, the predicted values are compared with the original values. Actual experiments were meant for wind power prediction, which has an important function in generating renewable energy sources. Prediction error was used to have a comparison among different methods. In particular, the number of the granulated windows was used to compare the original and predicted values with respect to error and granulated value. The statistical learning method SVM showed better performance with fuzzy information granulation.

Lee and He [21] used General Regression Neural Network (GRNN) for prediction of wind speed. The GRNN has different layers to achieve this. There are different kinds of units associated with layers, these being input units, pattern units, summation units and output units. CKS International airport data related to weather were used for the empirical study. Around 120 hours of data for three consecutive years from 2006 to 2008 was used for experiments. They compared the result with traditional models based on time-series. The regression neural networks model showed superior performance. The error rate of the GRNN was the least when compared with other models. The GRNN was used to have a model for a prediction that could be updated and used further. It showed that there is a possibility of predicting wind speed more accurately with GRNN.

Zhao et al. [22] studied Support Vector Regression (SVR) for wind speed prediction. The result of the prediction was compared with neural network models. They made important observations and found that both are useful in predicting wind speed. They also observed that the performance of SVR was better than that of neural network models. In

particular, neural networks could not perform well when there are fluctuations in the wind speed. The SVG takes training data and generates a model that is used for wind speed prediction. The testing data is subjected to the SVG mechanism in high-dimensional feature space for effective estimation of wind speed. The base line algorithm SVR and BPNN were compared to know the performance differences. SVR showed more accurate wind speed when compared to other methods.

Kulkarni et al. [23] explored two techniques for wind speed prediction: neural network and statistical regression. Some variants of these two were explored. The actual algorithms used for experiments include Artificial Neural Networks (ANN), Auto Regressive Integrated Moving Average Model (ARIMA) and curve fitting. They found that all these algorithms can be used to obtain wind speed prediction. However, the results revealed the effectiveness of ensemble methods where two or more methods are combined. Actual wind speed and predicted wind speed were compared to all algorithms. The difference between them was considered as an error for finding the performance of the algorithms. The metric used for comparison was Root Mean Square Error (RMSE). They found that regression technique with neural networks can be used for better prediction accuracy. Another observation made by them is that prior information of wind speed can help in the prediction of wind speed with results that are more accurate.

Pinson et al. [24] studied adaptive orthogonal fitting along with local linear regression for wind power applications. They used it for short-term forecasting. Local linear regression approach was the nonparametric approach used for power curve estimation. Noise component and its presence was the assumption used in the approach. Subsequently, they relaxed the assumption to have better prediction performance. Total Least Squares is used as a criterion that paves the way for better results. Moreover, it lowers computational cost incurred in the prediction process. They also proposed an optimisation method for improving the robustness of the method. The results were encouraging, as the regression approach proved to have accurate results when compared with ground truth-values.

Douak et al. [25] explored algebra to have a model for wind speed prediction. The technique they employed was general regression neural network. To this end, they designed a multi-block regression approach for wind speed estimation. The multi-block general regression neural network is known as MBGRNN. They used a large number of training samples effectively for training purposes. They were also able to reduce the time

for the training sample collection. Multi-block general regression is used for effective training, which makes the resultant model suitable for prediction of wind speed. The model thus built was used for the actual prediction process. The sample selection method while choosing the training data also played an important role in the performance of MBGRNN. Human experts were also involved in the labelling and usage of them in training. This improved the chances of accuracy in prediction.

Lahouar et al. [26] used two complementary approaches for wind speed prediction: Support Vector Regression (SVR) and direction prediction. They followed nonlinear wind evolution in order to obtain wind speed prediction. The approach was used for short-term wind speed prediction. They also studied the role of wind direction in a prediction of wind speed. They found that there is a relationship between wind speed and wind direction. In particular, the wind direction can influence the power generation process. They evaluated their method by using data collected from the Sidi Daoud wind farm in Tunisia. They compared actual wind velocity with predicted values in order to estimate the performance of the methods.

Botha and Vault [27] studied wind speed prediction for different purposes. They considered the problem of carbon emissions into the atmosphere as a cause of global warming. As global warming has an influence on the eco-system, they used wind velocity experiments and linked that with the emissions. In particular, their intention was to promote renewable energy resources for power generation and other purposes. They used data collected from South Africa and used different regression algorithms for prediction. They compared RMSE values for different approaches, such as persistence, ordinary least squares, Bayesian ridge regressor and SVR. Out of them, SVR showed better performance.

Finamore et al. [28] studied the notion of one-hour ahead wind speed prediction for generating wind power to provide renewable energy resources. They used the ANN method for this. The Feed Forwarded NN model was used, which uses different parameters, such as maximum epochs, error criterion function, transfer functions for hidden layer and an output layer. There are a number of output layers, a number of hidden layers and a number of input layers. The forecasting model has an input layer in which weather data is taken from multiple sources, the hidden layer makes use of weights, and the output layer produces a forecast. The forecast accuracy is compared with real wind speed so as to estimate the performance of the proposed method and MSE is used for performance evaluation.

Koka et al. [29] investigated the problem of estimating solar radiation, which is used for many applications, including renewable energy sources. They carried out experiments in different Turkish cities with various input parameters. Artificial neural network (ANN) technique was used with multiple input layers and the outcome was used in understanding the effectiveness of solar devices. Turkey's solar radiation intensities were explored in the experiments in order to obtain good understanding and insights on solar radiation in the Mediterranean region of Anatolia in Turkey. They used the solar radiation data of 2006 for training and 2005, 2007 and 2008 data for estimation of solar radiation parameters. Solar radiation is the output layer considered in the ANN method. Modelling solar radiation was done and a model built. The model was used for further exploration of the dynamics of solar radiation in the country. The parameters used for experiments included longitude, latitude, altitude, month, average temperature, average wind velocity, average cloudiness and sunshine duration.

Olaiya and Adeyemo [30] investigated the utility of data mining techniques for predicting weather information such as wind speed, evaporation, rainfall and maximum temperature. They used two algorithms, namely, decision tree and artificial neural network, and employed data mining techniques for weather forecasting and study of climate changes.

2.3. Summary

This chapter has covered a review of the literature on data mining techniques that are used to predict weather information and other purposes. It is understood that data mining techniques are widely used for predictive and descriptive purposes. There are algorithms that come under both supervised learning and unsupervised learning, many of which have been reviewed in this chapter. These algorithms are known as machine learning algorithms wherein some sort of learning is involved. In this thesis, many data mining algorithms, such as SVM, are used for predicting climate parameters from a real world weather dataset.

3. SYSTEM ANALYSIS

This chapter covers system requirements, functional requirements, and non-functional requirements. It also shows the difference between the existing system and the proposed system. It sheds light on the advantages of the proposed system. In addition to this, this chapter provides a detailed methodology that helps in understanding the procedure by which the proposed system is built. The aim of the project is in calculating the wind velocity using the input parameters of the climate dataset.

3.1. Existing System

The existing system for prediction of weather is found in the literature. According to the literature reviewed in Chapter 2, there has been considerable research on the prediction of weather. Most of the current research uses either a single technique or a hybrid one as solutions for weather prediction, they do not use multiple algorithms and make a comparative analysis to see the accuracy and the performance of each technique. Since Weka is supporting different data mining algorithms, it is important to use them to obtain predictions with multiple methods. The proposed system shown in the next section throws light on the possible use of different data mining techniques and gives error prediction as well as prediction of wind velocity.

3.2. Proposed System

In the proposed system, hybrid approaches are used to estimate climate parameters using data mining techniques. In a hybrid approach, different algorithms are used. For estimating the climate parameters using data mining techniques, a hybrid method is used. The project is to estimate weather conditions using data mining techniques. Here we used different algorithms for estimation of climate parameters. In order to perform the prediction of climate parameters, such as wind velocity by using a given weather dataset, different algorithms, i.e. SVM, neural networks and linear regression, are used. These algorithms are used according to the methodology presented in this chapter.

3.3. Functional Requirements

There are many functional requirements of the proposed system and it has two important roles, user and system and the user uses the different classification and regression algorithms to calculate the wind velocity.

1. It is essential to use an input dataset known as a weather dataset. The dataset contains data related to weather.
2. Once the inputs are taken from the dataset, the data is loaded using the Weka tool support in the form of an .arff file and various algorithms are performed on the given dataset.
3. Data mining algorithms, such as neural networks, SVM and linear regression, are used to predict wind velocity in addition to comparing them.
4. The prediction error is computed and compared in order to differentiate performance among the algorithms.
5. A weka tool application is built in order to analyse weather data using data mining techniques. The application demonstrates proof of the concept.
6. The visualisation of the results is done in order to have a better understanding of the performance difference among the algorithms.

3.4. Non-Functional Requirements

Non-functional requirements are important characteristics, but not compared to functional parameters. For example, attributes such as performance, usability, security, compatibility are not features of the system, but are required characteristics.

Each requirement must be objective and countable, which means there must be some way to measure. The following are some examples of non-functional requirements:

Performance Requirements: Requirements about resources are response time, transaction rates, throughput and benchmark specifications.

Operating Constraints: System Resources, people and required needed software come under this category.

Platform Constraints: User requirements are generally under this category. If the user does not care, there are still platform constraints

Accuracy and Precision: Requirements relating to the accuracy and precision.

Modifiability: Requirements about the ability to make changes in the software.

Portability: The effort required to easily move the software from one place to another place.

Reliability: The reasons for failure of the software.

Security: Requirements about protection of the system and data. This measurement is expressed in the various ways to break into the system.

Usability: Requirements about the difficulty to operate the system or to learn the process.

Legal: Legal issues involving privacy of data, property rights, export of restricted technologies, etc.

3.5. Software Requirements

3.5.1. Weka

This is one of the available data mining tools and is widely used in academics and research for data mining purposes. Weka stands for Waikato environment for knowledge analysis. It provides a set of data-mining algorithms of different categories, such as the clustering and classification needed for data mining. The algorithms provided by Weka tool can be used either directly or called from Java applications. Weka is based on Java programming language. It has other algorithms for data preprocessing, regression, association rule mining and visualization and also helps in building new machine learning algorithms and is included as part of the tool. It is an open source tool that can be used and improved by developers.

Weka is an accumulation of machine learning algorithms for information mining errands. The algorithms can either be connected specifically to a dataset or called from your own Java code. Weka highlights incorporate classification, regression, clustering, association rules, attribute selection, experiments, workflow and visualization. Weka is composed in Java and was created at the University of Waikato, New Zealand. The greater part of Weka's systems are predicated on the presumption that the information is accessible as a solitary level record or connection, whereby every information point is depicted by a settled number of credits Weka gives access to SQL databases utilizing Java Database

Connectivity and can process the outcome returned by a database inquiry. It is not fit for multi-social information mining in Figure 3.1 as shown Weka start-up screen.

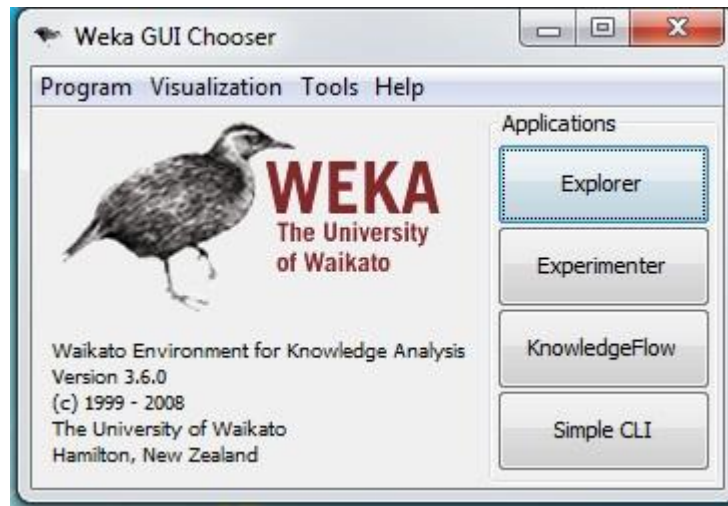


Figure 3.1. Weka start-up screen

As shown in Figure 3.2, Wake's principle UI is the Explorer, a similar usefulness additionally can be gained through the part-based Knowledge Flow interface and from the order line. There is, likewise, the Experimenter, which permits the efficient correlation of the prescient execution of Weka's machine learning algorithms on an accumulation of datasets.

The Explorer interface includes a few boards giving access to the principal parts of the workbench, for example, the preprocess board which is a platform for bringing in information, the order board which empowers the client to apply classification and regression algorithms, relate board which gives access to association lead students, and a visualise panel showing a scatter plot matrix.

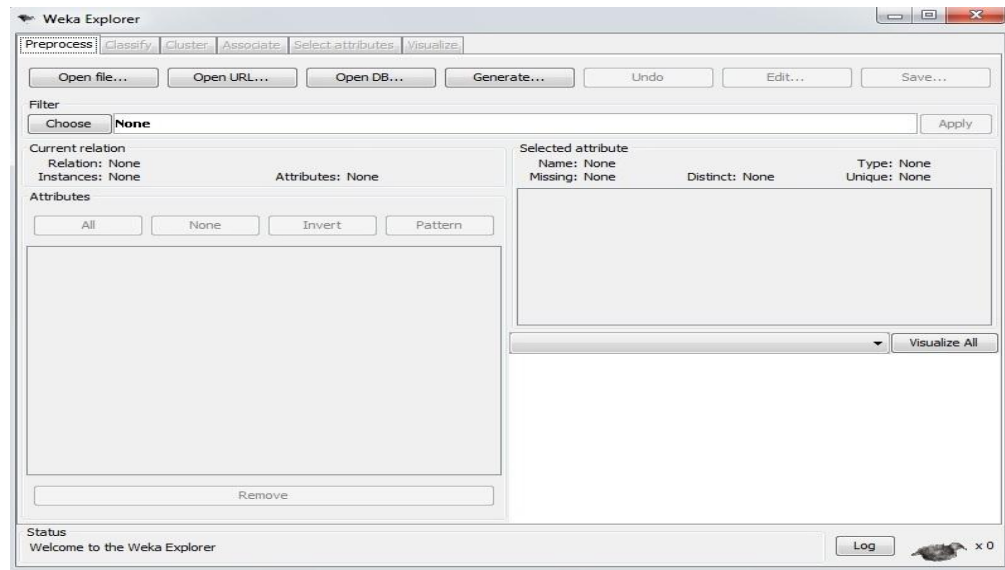


Figure 3.2. Weka Explorer

Weka gives a far-reaching set of data pre-handling apparatuses, learning algorithms and assessment strategies, graphical UIs and a domain for contrasting learning algorithms. The data can be foreign made from a document in different organizations, for example, ARFF, CSV and parallel. Information can likewise be perused from a URL or from an SQL database (utilising JDBC). Pre-preparing devices in WEKA are called “filters” and there are filters available for discretisation, normalisation, resembling, attribute selection, transforming and combining attributes.

The executed learning plans are choice trees and records, and case-based, while the implemented learning schemes are decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptions, logistic regression, and locally weighted learning.

Weka is a collection of tool for classification: to anticipate ostensible or numeric amounts, we have classifiers in Weka. Accessible learning plans are choice trees and records, support vector machines, example-based classifiers, and logistic regression nets. Once the data have been stacked, all of the tabs are empowered. In view of the prerequisites and by experimentation, can discovering the most appropriate calculation to create an effortlessly justifiable portrayal of information.

Before running any order algorithm, having to set test alternatives. Accessible test choices are recorded below. Before running any order algorithm, having to set test alternatives. Accessible test choices are recorded underneath.

Utilize preparing set: Evaluation depends on how well it can foresee the class of the examples it was prepared on.

Provided preparing set: Evaluation depends on how well it can anticipate the class of an arrangement of occasions stacked from a record.

Cross-approval: Evaluation depends on cross-approval by utilizing the quantity of folds entered in the "Folds" content field.

3.6. Methodology

In this thesis, we propose a methodology for the exploration of data mining algorithms for prediction of the wind velocity parameter of a climate dataset. In other words, the methodology throws light on building various models using different data mining techniques. The methodology takes climate dataset as input and performs training phase so as to obtain training data ready for an exploration of different classification algorithms. The data mining algorithms used in the exploration of the models include neural networks, SVM and linear regression. The methodology is illustrated in Figure 3.3.

As shown in Figure 3.3, it is evident that the proposed methodology provides different algorithms of data mining. These algorithms are used in order to have a prediction of weather data. In other words, the proposed methodology is for predicting weather information. The following subsections provide more detail on the proposed framework.

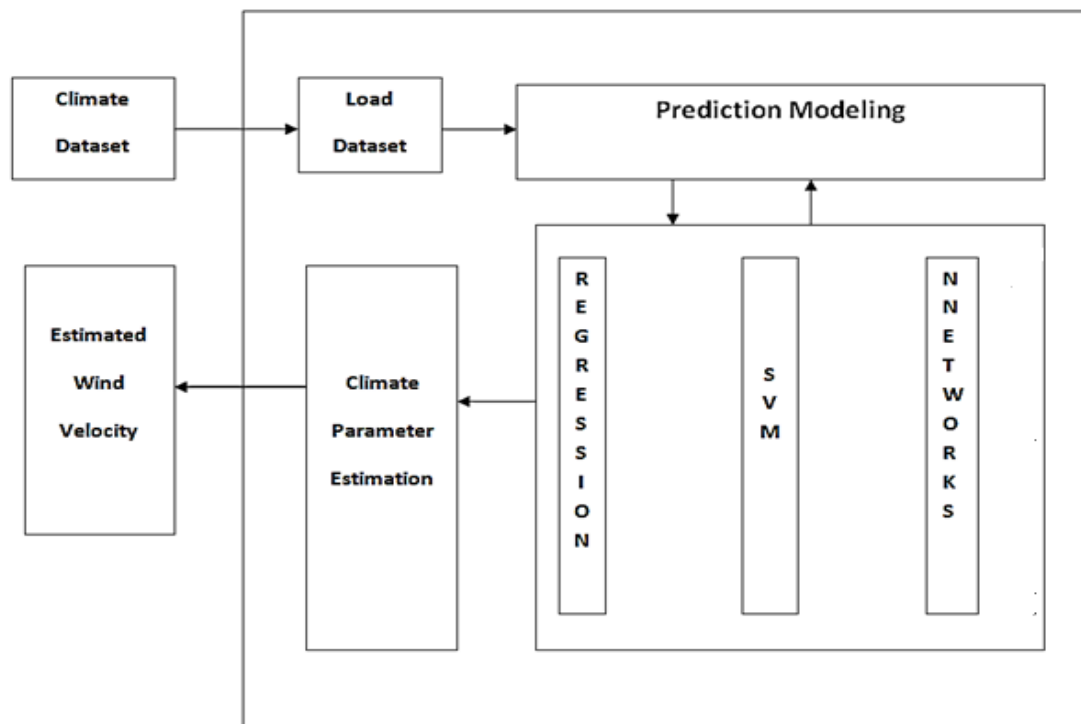


Figure 3.3. Proposed methodology

3.6.1. Neural Networks

Neural networks are the networks used in computer science that model the human brain and nervous system. It is a computational approach, which is based on a collection of artificial neurons that mimic how biological neurons work. Artificial Neural Network (ANN) is a paradigm for information processing and which has a novel structure for processing information. It is widely used as part of machine learning and artificial intelligence.

Artificial Neural Networks (ANN) is one of the machine learning techniques. Unlike other techniques, it is based on the neurons in the human brain and biological operations that resemble human brain operations. A neuron is the important component of ANN and which is modelled after biological neurons, which produce activities as outputs based on a certain amount of activation. ANN is made up of several neurons and they form a network in the form of a weighted graph. A node in ANN is nothing but an artificial equivalent of the neuron. The node takes a collection of weighted inputs and produces outputs after processing in addition to passing on the outputs to the nodes further down in the graph. It is

based on activation function and its aggregation in the process. The input to activations can be represented as follows:

$$(\sum_i^n (w_i a_i)) = \phi(\mathbf{w}^T \mathbf{a}) \quad (3.1)$$

The above representation of input activations can be represented in the form of a node that takes inputs and produces outputs. The node is nothing but a neuron here. The node representation is as in Figure 3.4.

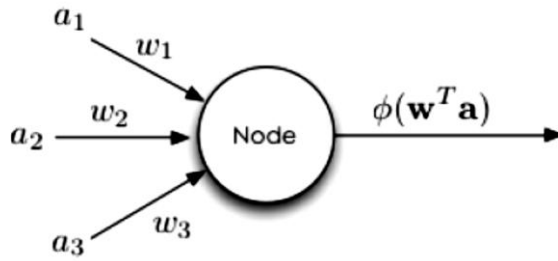


Figure 3.4. A sample of simple Neural Network of Canonical Activation Function

Canonical activation functions are available. One such function is known as linear activation, which is represented as follows:

$$\phi(\mathbf{w}^T \mathbf{a}) = \mathbf{w}^T \mathbf{a} \quad (3.2)$$

The above representation is also called as identity activation. There is another example known as sigmoid activation function, which is represented as follows:

$$\phi(\mathbf{w}^T \mathbf{a}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{a})} \quad (3.3)$$

exp=2.718

Each activation function can be imagined as a node. When all nodes are chained together, it forms a network of neurons that play a vital role in the ANN technique. This process is done in the form of layers. One then takes input and produces an output of the

next layer. The purpose of this is to train the network with certain labelled data, feeding it as a collection of inputs and expected outputs. Correct edge weights are considered in the process of training in order to produce outputs correctly. With a trained network that is ANN, it is possible to solve prediction problems in the real world.

$$a_i = f^{(1)}\left(b_j^{(1)} + \sum_{j=1}^N W_{ij}X_j\right) \quad (3.4)$$

Here, $f(1)$ refers to an activation function to the weighted inputs. Its biased term is denoted as $b_i^{(1)}$. Weighted matrix is represented by W_{ij} . X_j refers to input signals.

In Figure 3.5, $(X_1 \dots X_n)$ is an input signal, $(W_1 \dots W_n)$ is the weight which represents the strength of the synapse, the connecting neuron is the activation function and b is the threshold value.

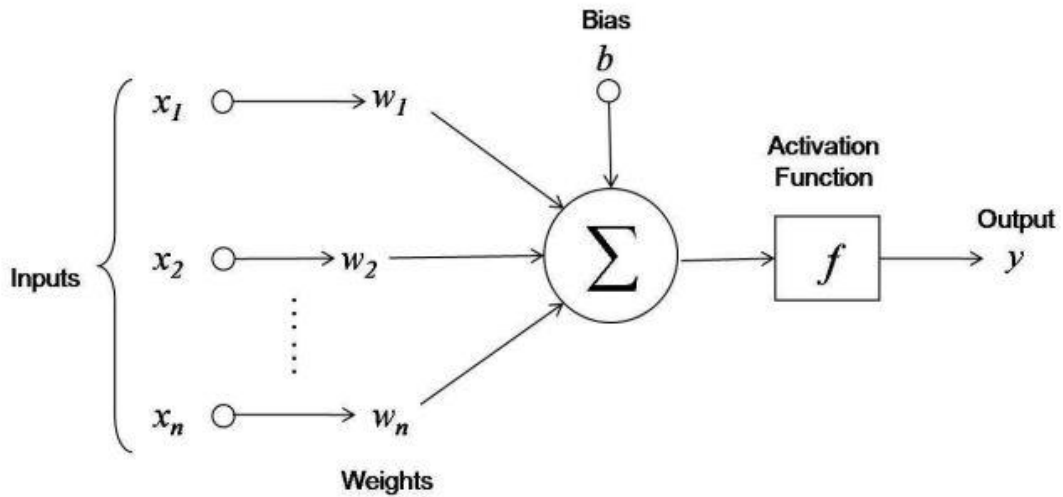


Figure 3.5. Neural Networks diagram [31]

In Figure 3.6 below, F is the transfer function, b the Bias value and w the weight value and illustrates a sample of three layers of artificial neural network which distinctly represent all layers and shows how the data will be inputted then forwarded and processed to the output layer.

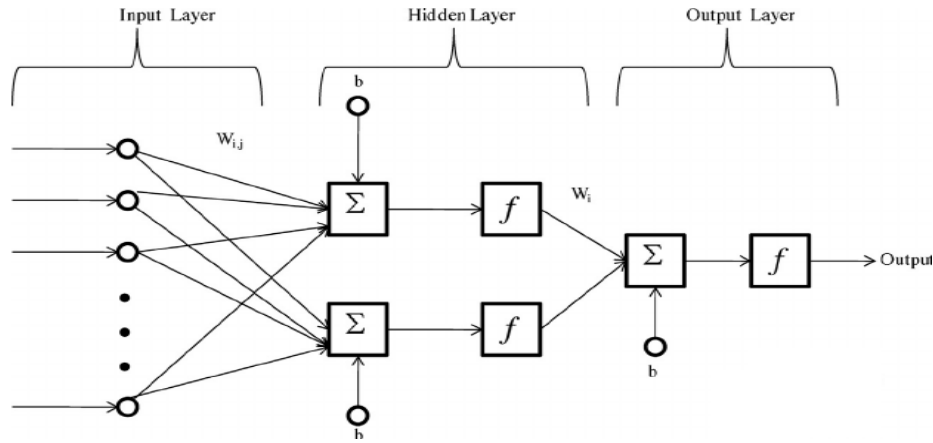


Figure 3.6. Three-layer-ANN-architecture[32]

3.6.2. Support Vector Machine (SVM)

This is a supervised learning model which can have associated learning algorithms for both regression and classification tasks. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples. Support Vector Machine is a frontier which best segregates the two classes.

Support Vector Machine (SVM) is another machine learning technique, one which is supervised learning in nature. It can be used in many real world classification problems. It is also used as a regression technique and has features of a maximum margin algorithm. It is controlled by parameters that are not dependent on the dynamics of dimensionality in the feature space. In a similar way, classification can be made with optimization and generations for regression. The loss function is defined as ignoring errors in order to have better optimization. Such loss function is known as epsilon intensive.

As shown in the below Figure 3.7, it is evident that there is an example for one-dimensional linear regression with aforementioned Epsilon intensive loss function. The cost of errors is measured with respect to training points. Actually, the cost of errors is zero for all points.

$$\text{Minimize } \frac{1}{2} \mathbf{w} \cdot \mathbf{w}$$

$$y = \mathbf{w} \cdot \mathbf{x} + b, \text{ Constraints: } y_i - \mathbf{w}x_i - b \leq \varepsilon, \mathbf{w}x_i + b - y_i \leq \varepsilon$$

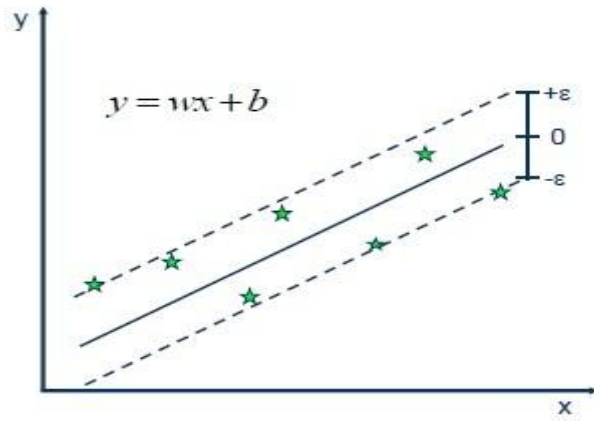


Figure 3.7. The solution One-dimensional regression with epsilon intensive band.[43]

Figure 3.8 below shows C as pre-specified value. Upper and lower constraints are denoted by slack variables and outputs of the system. C is a given parameter value, and Lagrange relaxation can be obtained. The variable b denotes a bias value, ξ_i is slack variables or is the error item of the wrong classification, is absolute variables. The support vector regression is represented as follows:

$$\text{Minimize } \frac{1}{2} \mathbf{w} \cdot \mathbf{w} + C \sum_{k=1}^R \varepsilon_k$$

Constraints:

$$y_i - \mathbf{w}x_i - b \leq \varepsilon + \xi_i$$

$$\mathbf{w}x_i + b - y_i \leq \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

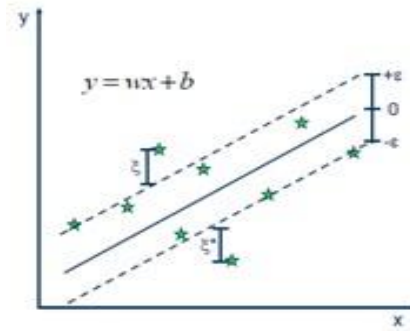


Figure 3.8. Shown Pre-Specified Value[43]

$$y = F(X) \sum_{k=1}^m \bar{a}_k \cdot K(X, X(k)) + b \quad (3.5)$$

SVM is a binary classifier. The function operates on the training set 1 through m . The k th training example is represented as $x_{(k)}$, $y_{(k)}$. The input vector is denoted as $x_{(k)}$ while the class label is denoted as $y_{(k)}$. K is known as kernel function while x is the input vector which contains data to be classified. A scalar value used by SVM is denoted as b .

In Figure 3.9, X is input vector, $(X_1, \dots, X_n)$ is support vector machine, $(a_1, \dots, a_m)$ represents weight and is known as kernel function and output b is the threshold value.

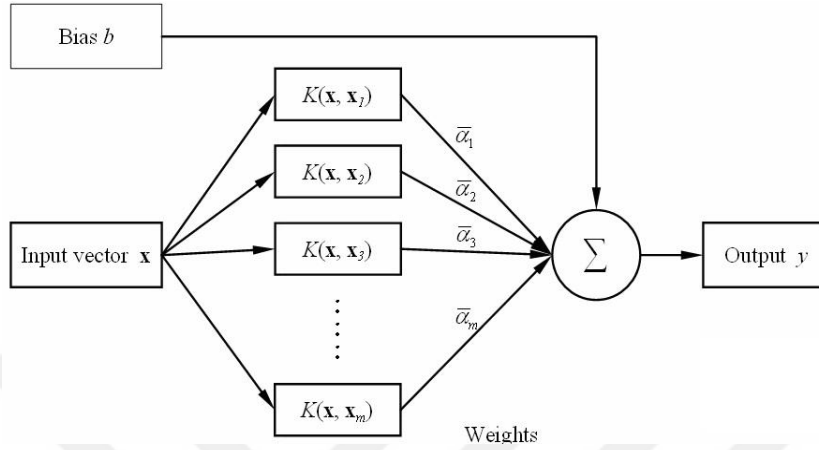


Figure 3.9. Support Vector Machine diagram[33]

There are a number of kernels that can be used in Support Vector Machines models. These include linear. There exists a way to compute inner product in feature space as a function of original input points' kernel functions. The kernel function is expressed as K , where, x_i and x_j are the input vectors, as the function:

$$\text{Kernel Linear: } K(X_i, X_j) = X_i^T \cdot X_j \quad (3.6)$$

3.6.3. Linear Regression

Linear regression analysis is an efficient method that can be used for predicting the numeric value of the unknown value of a variable and relies on the known value of another variable.

The main equation of two linear regressions, Y on X and X on Y , is calculated as $Y = a + bX$ where a and b are defined as constants recognized as intercept and slope of the equation. This technique is utilized to predict the Y variable of unknown value when X variable value is known:

$$Y = a + bX:$$

Moreover, the X line of regression that is on Y is computed as $X = c + dY$, whilst this can be used to make a prediction of the unknown value of X variable depending on the value of the known Y variable. Frequently, only one of these lines makes sense.

In linear regression, the model is formed with just one dependent variable and one independent variable. Conventionally, the dependent variable is that variable whose value is to be predicted and the independent variable is known as the value used for prediction.

In detail, any of these will be convenient for the analysis in hand and will rely on the labeling of dependent and independent variable in the problem which is to be analysed.

The regression coefficient of Y on X is defined by any coefficient of X in the line of regression of Y. It is defined as the change in the value of dependent (Y) variable corresponding to a unit change in the value of the independent (X) variable. In equation (3.7) y is dependable value, x in dependable value, an intercept, b slope of the line.

$$Y = a + bX \quad (3.7)$$

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \quad (3.8)$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \quad (3.9)$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \quad (3.10)$$

In Figure 3.10, X_1 , X_2 , X_3 are independent variables, β_1 , β_2 , β_3 propitiation Y intercepts parameters representing the contributing of the independent variables, Y dependent variable. β is called the "parameter" or "weight" of each feature. Each β is multiplied with corresponding feature x and all values are added to obtain predictions of the target y.

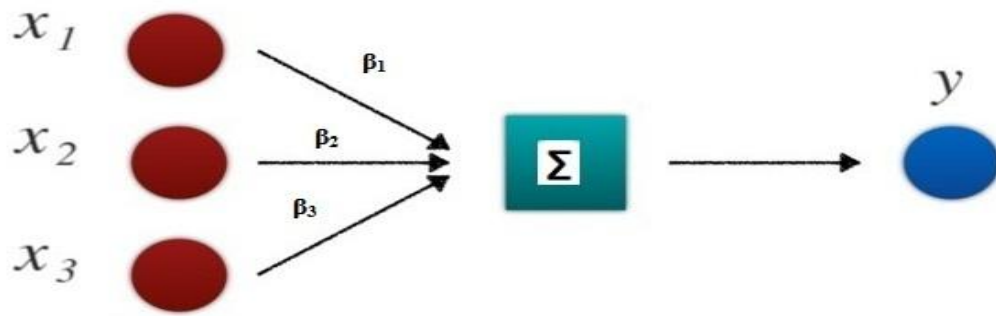


Figure 3.10. Linear Regression model [34]

3.7. Summary

This chapter has provided the system requirements specification with details such as functional requirements, non-functional requirements and system requirements such as hardware and software requirements. Apart from this, it shed light on the methodology that covers the approach in building a web portal, which is hybrid in nature, as well as the activities needed for language acquisition and the non-functional requirements that improve the user experience. The methodology presented in this chapter is used to complete the empirical study. The results of the empirical study and the design of the proposed system are presented in the subsequent chapters. The next chapter provides details of the system design, which is the basis for the implementation of the proposed system.

4. IMPLEMENTATION

This chapter provides implementation details for the project of the estimation of climate parameters using Data Mining Techniques. It covers the Architecture Diagram which is used for the implementation. The architecture diagram reflects the aspects of the project. This project is based on two-tier architecture. There are many data mining algorithms involved in the implementation. The data mining tool Weka is used in order to implement the functionality of the proposed system. Weka is one of the data mining tools widely used in the real world application development. Here we use climate dataset taking as input. Here the data set contains parameters such as station number, month, day, pressure, humidity and temperature. Station number is a code for the area. Month is for taking the values from January to December, day is for taking the day number in the month and pressure is an important parameter in monitoring the climate. Humidity takes a major role in the impact of temperature and temperature means the intensity of heat present in the substance.

Weka is a collection of algorithms related to machine learning for solving data mining tasks. The algorithms that are applied by directly using a dataset with GUI and also called through from Java code. Weka contains tools for pre-processing, classification, regression, clustering, association rules and visualization. It is also used for developing new machine learning techniques. It is an open source software under the license of GNU, a general public license. The Weka Knowledge Explorer is an easy to use graphical user interface. Each of the major Weka packages, filters, clusters, classifiers, associations and attribute selection, are represented in the Explorer along with a visualization tool which allows the datasets and the predictions of classifiers and clusters to be visualized. Weka contains the following panels:

Preprocess panel: This is the starting point of the Weka Explorer. Based on this panel, we can load datasets, browse the characteristics of attributes and apply any combination of unsupervised filters to the data.

Classifier Panel: This panel allows to configure and execute any weak classifiers in the browsing dataset. We can choose to perform a cross-validation or test on a separate dataset. Classification errors appear in the form of pop-up data. If the classifier produces a decision tree, it can be displayed on the popup tree visualize.

Cluster Panel: From the cluster panel we can configure and execute any Weka clusters in the browsing datasets. Clusters can be visualized in a popup visualization tool.

Associate Panel: From the associated panel, we can apply the association rules by using Weka associates in the dataset.

Select Attributes Panel: This panel allows configuring and applying any combination of Weka attribute evaluator and searching method to select the most pertinent attributes in the dataset. If an attribute selection scheme transforms the data, then the transformed data can be appear in the pop-up visualization tool.

Visualize Panel: This panel displays a scatter plot matrix for the current data set. The size of the individual cells and the size of the points can be displayed and adjusted using the slider controls at the bottom of the panel. The number of the cells in the matrix can be changed by pressing the select attributes button and then choosing those attributes to display.

Weka tool is used for data mining algorithms to produce the outputs in the format of three measurements, which are:

Correlation Coefficient: This explains how well the predictions are correlated or change the actual output value. A value of 0 is the worst and a value of 1 is a perfectly correlated set of predictions.

$$r_{xy} = \frac{n \sum_{i=1}^n xy - \sum_{i=1}^n x \sum_{i=1}^n y}{\sqrt{\left[n \sum_{i=1}^n x^2 - \left(\sum_{i=1}^n x \right)^2 \right] \left[n \sum_{i=1}^n y^2 - \left(\sum_{i=1}^n y \right)^2 \right]}} \quad (4.1)$$

Root Mean Squared Error: This is the average amount of error occurring in the test set in the units of the output value. This measure helps provide an idea of the degree to which a given prediction value may be wrong on average.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x_i - x'_i)^2}{n}} \quad (4.2)$$

Mean Absolute Error: This is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. This measures the average magnitude of the errors in a set of forecasts. It measures the accuracy of the Continuous variable

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - x'_i| \quad (4.3)$$

4.1. Case Study

A dataset that contains weather data with attributes is taken as a case study for this research work. The first column of the file is the station number and every station represents a Turkish city, which is five stations. These cities were chosen by their geographical position and include INEBOLU, SINOP, AMASRA, SILE and AKCAKOCA. Every station place is near the sea; therefore, their wind potential is higher than other places. All data are obtained from the Turkish Government Meteorology Services for summer months of 2013.

The wind potential of a place is related to the geographical position, humidity, atmospheric pressure and the ambient temperature. All of these are datasets parameters. In addition, it has 456 instances covering data of five stations for summer months. An excerpt of the dataset is shown in Table 4.1. Weka is the environment used to explore a climate dataset, which is very useful, and the mining of such data provides useful information to various domains in the world. The Weka tool is used to exploit all existing methods. Weka software with different kinds of data mining techniques is used to perform mining on a given dataset. The dataset is related to climate data that were captured in one of the regions in Turkey within the first fifteen days of June. The attribute of pressure, humidity, temperature and actual wind speed are presented in the datasets. To predict the wind velocity using all data mining techniques, the data are inputted into the proposed model. In the supervised simulation, the predefined classes are achieved through the training phases of the existing datasets. In the NN phase, the backward propagation used is based on an excerpt of datasets, as presented in Table 4.1 (except the last column, which is the predicted wind velocity using linear regression), which is used as input to the proposed framework.

As shown in Table 4.1, it is evident that the input dataset contains different attributes, such as station number, month, day, pressure, humidity and temperature. The proposed framework is applied to have training samples and different models. All these experiments are done using Weka, which is one of the widely used data mining tools.

Table 4.1. Sample datasets usage for getting results.

Station Number	Month Number	Day Of month	Pressure(hPa)	Humidity (%)	Temperature(°C)	Wind Velocity (m/s)
1	6	1	1008.7	67.7	19.7	1.7
1	6	2	1005	80.3	19.6	1.3
1	6	3	1010.1	69.3	17.3	2.5
1	6	4	1011.1	72	19.4	1.4
1	6	5	1013.4	82.7	17.4	1.3
1	6	6	1014.6	80.7	18.8	1.8
1	6	7	1013.6	72	19.5	1.5
1	6	8	1010.1	77.7	20.4	1
1	6	9	1012.3	81.3	19.2	2.4
1	6	10	1012.9	77.7	19.2	1.7

Getting datasets for five cities in Turkish country, every city has a geographical position, as shown in Table 4.2.

Table 4.2. Geographical position

Station number	City name	Latitude (°N)	Longitude (°E)	Altitude (m)
1	INEBOLU	41.97	33.75	120
2	SINOP	42.02	35.15	59
3	AMASRA	41.74	32.38	14
4	SILE	41.17	29.61	22.45
5	AKCAKOCA	41.08	31.12	59

4.2. Architecture Diagram

The architectural diagram for this program's application, Weka software, is shown in Figure 4.1. It contains different components involved in the system. It sheds light on different aspects of wind velocity estimation. The estimation of climate parameters is done empirically with the help of data mining techniques supported by the Weka tool. The methodology presented in Chapter 3 is used to implement the proposed application. As shown in Figure 4.1, it is evident that the prediction modelling makes use of different data

mining techniques using the Weka tool. Climate dataset is given as input while the output of the proposed system is estimated wind velocity.

4.2.1. Two-Tier Architecture

The Weka tool application is technically based on two-tier architecture. This architecture contains two tiers. The first tier is known as client tier, while the second tier is the data tier. The two tiers contain three layers of the application. The first layer is known as the presentation layer or application layer. It provides a user interface. The second layer is known as a business logic layer, which is responsible for performing activities on data as per business logic. The third layer is responsible for managing data. These three layers are fit into two tiers. Therefore, it is said to have two-tier architecture.

As shown in Figure 4.1, it is evident that the client application is interacting with a data source in order to obtain data mining operations. The data present in the data source are taken by the client application. Then, the client application performs different operations in order to achieve the goal of the system. This kind of architecture is simple and effective.

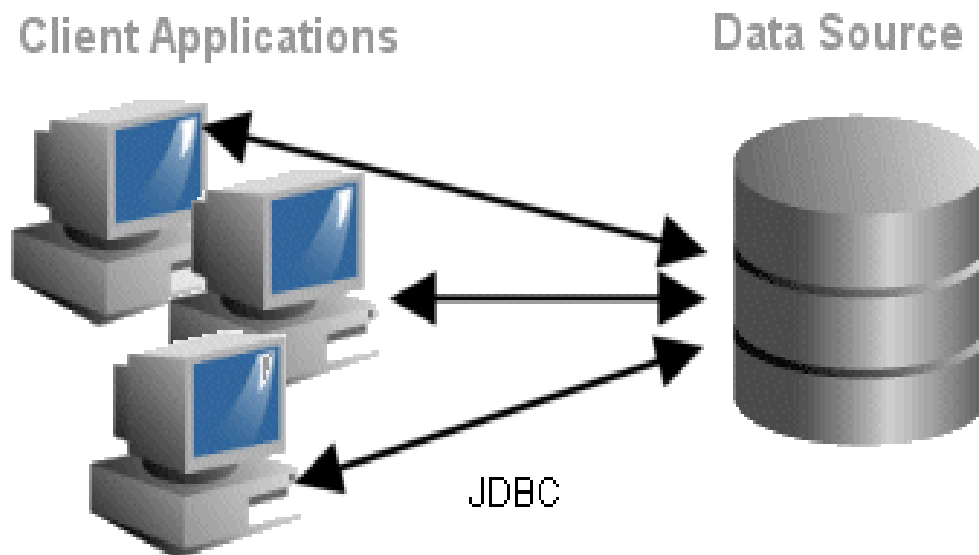


Figure 4.1. Two tier architecture used in the proposed system

4.3. Modules (Weka software)

Using a system with Weka software has many functional requirements. Those are logically divided into a number of modules. The modules present in the system are loading a dataset, regression, SVM and neural networks.

4.3.1. Load Dataset

Load Dataset takes the input file from the selected location. For that, we use the select browse option to select the input dataset. Here we write the code for selecting the input file from that location. The dataset used in this thesis is climate dataset

4.3.2. Linear Regression

Regression is a data mining approach to model the relationship between a dependent variable and one or more independent variables. Simple linear regression makes use of one independent variable. A linear model is built using a linear regression approach, which models the relationship between the model parameters' linear predictor functions. In this thesis, the linear regression model is used as a prediction model. The goal of the analysis is the prediction of wind velocity, which is accomplished by using linear regression.

4.4.3. Neural Networks

Neural networks are the networks used in computer science that model the human brain and nervous system. It is a computational approach based on a collection of artificial neurons that mimic how biological neurons work. Artificial Neural Network (ANN) is a paradigm for information processing which has a novel structure for processing information. It is widely used as part of machine learning and artificial intelligence.

4.3.4. Support Vector Machine

This is a supervised learning model which can have associated learning algorithms for both regression and classification tasks. It needs training examples that contained

labeled data. It uses a training algorithm to build a model and then uses the model for classification of objects. As new examples are added, the model gets updated and the performance of the SVM will get improved with more training samples.



5. RESULTS DISCUSSION AND EVALUATIONS

This chapter presents the results of the estimation of climate parameters using data mining techniques in this thesis. It is very helpful to estimate the climate parameters in all seasons. The following section provides the usage in a user-friendly manner. This chapter throws light on the different aspects of the system and how it is useful.

5.1. Applied linear regression and SVM and ANN

To obtain results for linear regression in Weka software, after loading datasets in classifications, then test options are selected (use training set), after in which further options are checked for output predictions, in the Weka Classification form there is a command button called Choose which has several functions for selecting linear regressions.

As shown in Figure 5.1, the Attribute selection method is linear regression. First set the method used to select attributes. Available methods are no attribute selection, attribute selection using M5's method (step through the attributes removing the one with the lowest standardized coefficient until no improvement is observed in the estimate of the error given by the Akaike information criterion), and a greedy selection using the Akaike information metric.

Debug is false outputs debug information to the outputs debug information to the console. Eliminate Collinear Attributes are true Eliminate collinear attributes, ridge $1.0E-8$ is the value of the Ridge parameter.

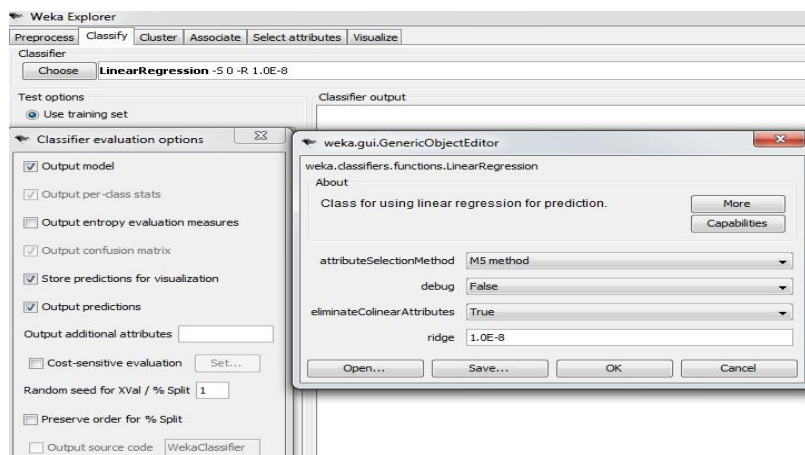


Figure 5.1. Apply linear regression in weka software

Figure 5.2 shows the implement Alex Smola and Bernhard Scholkopf's sequential minimal optimization algorithm for training a support vector regression model. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. (Note that the coefficients in the output are based on the normalized/standardized data, not the original data.), C(1.0). The complexity parameter C. checks Turned Off(true), Turn time-consuming checks off, use with caution. Debug is false. If set to true, the classifier may output additional info to the console.eps (1.0E-12). The epsilon for round off error (shouldn't be changed). Epsilon value is 0.001, the degree up to which deviations are tolerated. Watch out, the value of epsilon is used with the (normalized/standardized) data. In filter, Type choice normalizing training data is used for determining how/if the data are transformed.

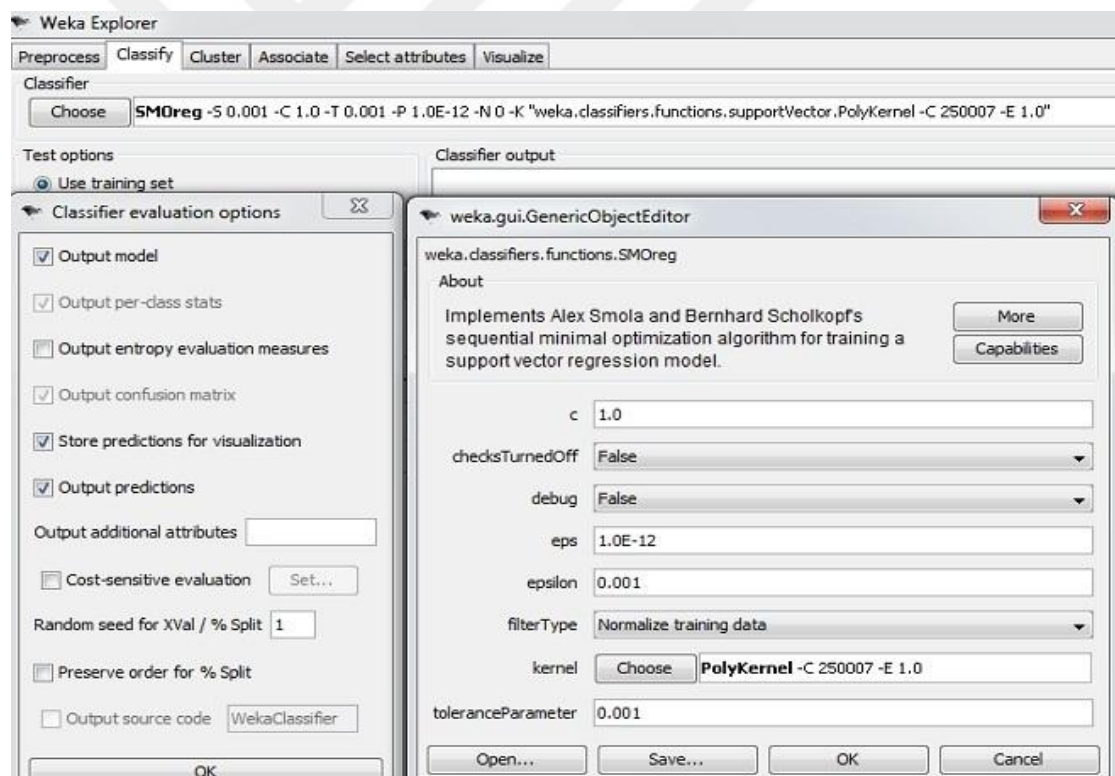


Figure 5.2. Apply support vector machine in weka software

Figure 5.3 shows the implement function multilayer perceptron, a classifier that uses back propagation to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time.

The nodes in this network are all sigmoid (except for when the class is numeric, in which case the output nodes become un-thresholded linear units).

GUI (false): brings up a GUI interface. This will allow the pausing and altering of the neural network during training.

Auto build (true): adds and connects up hidden layers in the network.

Debug (False): if set to true, the classifier may output additional info to the console.

Decay (False): this will cause the learning rate to decrease. This will divide the starting learning rate by the epoch number, to determine what the current learning rate should be. This may help to stop the network from diverging from the target output, as well as improve general performance. Note that the decaying learning rate will not be shown in the GUI, only the original learning rate. If the learning rate is changed in the GUI, this is treated as the starting learning rate.

Hidden layers (a): this defines the hidden layers of the neural network. This is a list of positive whole numbers. One for each hidden layer. Comma separated. To have no hidden layers put a single 0 here. This will only be used if auto-build is set. There are also wildcard values 'a' = (attrib + classes) / 2, 'i' = attribute, 'o' = classes, 't' = attrib + classes.

Learning rate (0.3): the degree to which the weights are updated.

Momentum (0.2): momentum applied to the weights during updating.

Normalise attributes (true): this will normalise the attributes. This could help improve the performance of the network. This is not reliant on the class being numeric. This will also normalise nominal attributes as well (after they have been run through the nominal to binary filter if that is in use) so that the nominal values are between -1 and 1

Normalise numeric class (true): this will normalise the class if it is numeric. This could help improve the performance of the network; it normalises the class to be between -1 and 1. Note that this is only internally; the output will be scaled back to the original range.

RandomSeed (0): seed is used to initialise the random number generator. Random numbers are used for setting the initial weights of the connections between nodes, and also for shuffling the training data.

Reset (0): this will allow the network to reset with a lower learning rate. If the network diverges from the answer, this will automatically reset the network with a lower learning rate and begin training again. This option is only available if the GUI is not set.

Note that if the network diverges, but isn't allowed to reset, it will fail the training process and return an error message

Training time (500): the number of epochs to train through. If the validation set is non-zero then it can terminate the network early

Validation set size (0): the percentage size of the validation set. (The training will continue until it is observed that the error on the validation set has been consistently getting worse, or if the training time is reached). If this is set to zero, no validation set will be used and, instead, the network will train for the specified number of epochs.

Validation threshold (20): used to terminate validation testing. The value here dictates how many times in a row the validation set error can get worse before training is terminated.

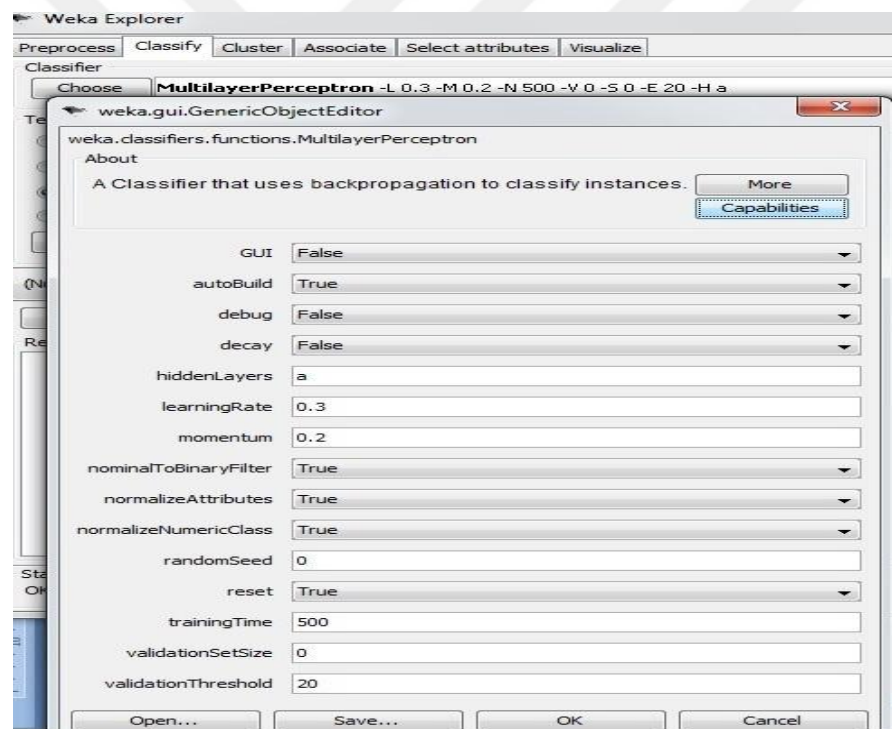


Figure 5.3. Apply Multilayer perceptron

5.2. Evaluation Results

5.2.1. Obtained Results for Wind Prediction

After loading the datasets, the results in the Weka software are shown. The obtained results of wind prediction for SVM, neural network and linear regression are shown in Table 5.1 below.

Table 5.1. Performance comparison in terms of wind prediction value

Station Number	Month	Day	Pressure(hPa)	Humidity(%)	Temperature(°C)	Wind Velocity (m/s)	Prediction		
						Actual	ANN	SVM	LR
1	6	1	1008.7	67.7	19.7	1.7	1.73	1.65	1.65
1	6	2	1005	80.3	19.6	1.3	1.6	1.34	1.41
1	6	3	1010.1	69.3	17.3	2.5	1.92	1.9	2.04
1	6	4	1011.1	72	19.4	1.4	1.05	1.74	1.70
1	6	5	1013.4	82.7	17.4	1.3	2.88	1.80	1.79
1	6	6	1014.6	80.7	18.8	1.8	2.07	1.63	1.66
1	6	7	1013.6	72	19.5	1.5	1.69	1.78	1.79
1	6	8	1010.1	77.7	20.4	1	1.44	1.44	1.56
1	6	9	1012.3	81.3	19.2	2.4	1.58	1.74	1.69
1	6	10	1012.9	77.7	19.2	1.7	1.70	1.65	1.81

Figure 5.4 below shows wind speed data covering a period of a month during summer time in INEBOLU city. As illustrated, on the graph the x axis represents the number of days that the data are captured, while Y dimension shows the wind speed in miles per second. The obtained results on the testing data indicate that the LR and SVM system outperforms the ANN model, as indicated by the predication graph and by the distance between the actual wind and the predictions of the three method. However, in more than one place ANN fluctuated with the actual wind. The research is ongoing to validate the performance of each technique based on three different statistical tools.

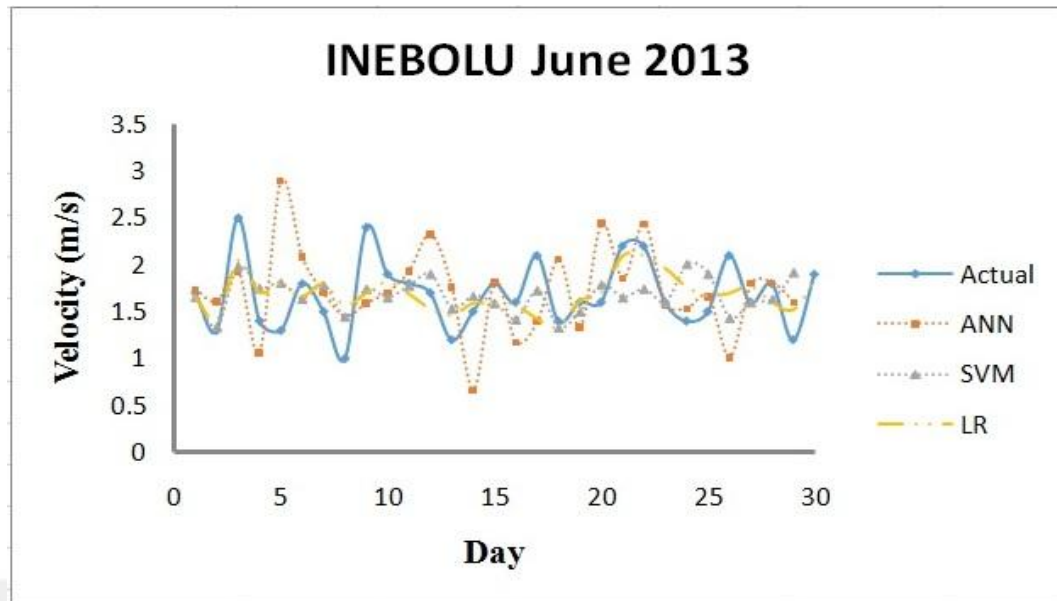


Figure 5.4. Actual and predict wind velocity in city INEBOLU in month June

Figure 5.5 shows the line curve prediction of wind speed obtained by the used methods and compares with the actual wind in INEBOLU during July. As illustrated by the graph, it can clearly be seen that, based on the distance with the actual wind line curve, that linear regression and support vector machine achieved better and closer results over ANN. In spite of that, ANN in Figure 5.5 is getting much closer to the actual wind between day 8 to 30.

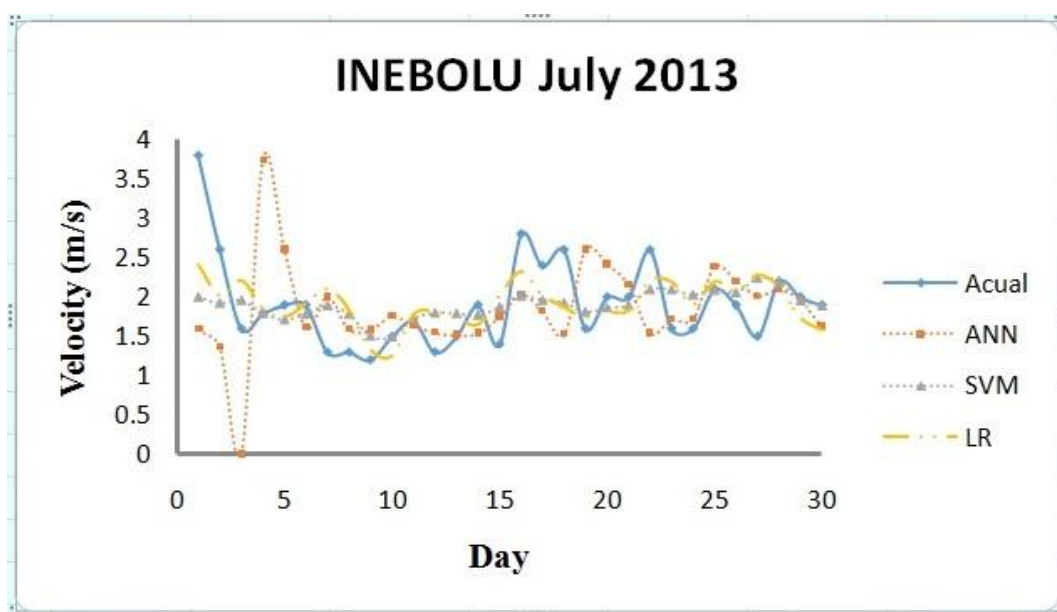


Figure 5.5. Actual and predict wind velocity in city INEBOLU in month July

Figure 5.6 shows the line curve prediction of wind speed obtained by the used methods and compares with the actual wind in INEBOLU during August. As illustrated by the graph it can clearly be seen, based on the distance with the actual wind line curve, that support vector machine achieved better and closer results over ANN and linear regression. In spite of that, ANN in Figure 5.6 is getting much closer to the actual wind between day 1 to 20, whereas the LR just draws a straight line curve along the specified periods.

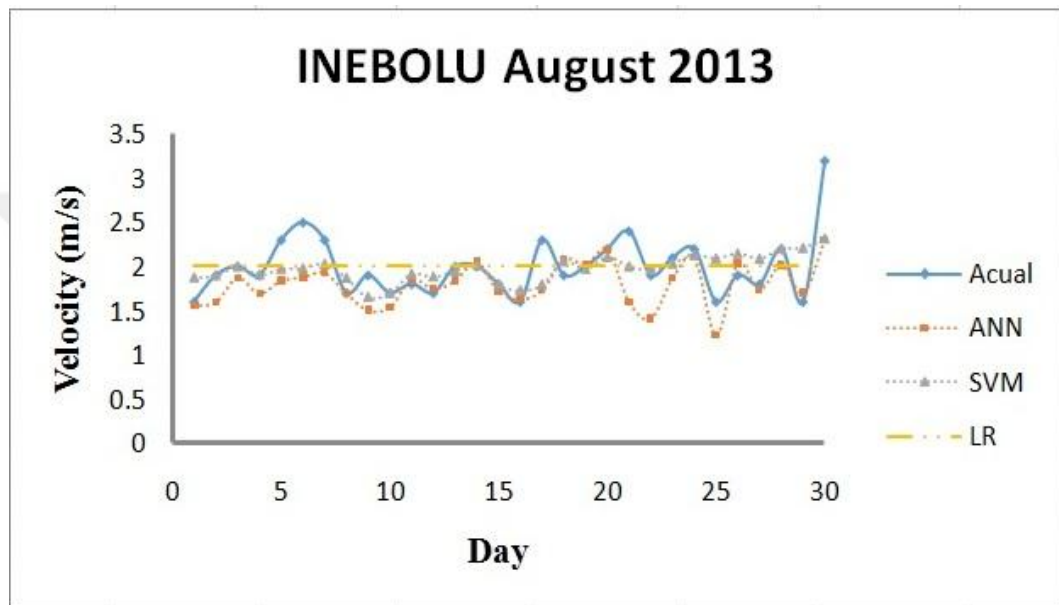


Figure 5.6. Actual and predict wind velocity in city INEBOLU in month August

According the below graph in Figure 5.7, which presents the SINOP wind predictions, it shows that the actual wind curve fluctuated and gives a variation of results in different days. In some points, its points are close to LR while, in some other sharps, it converges with ANN and in other places the curves show closer to SVM. Generally, to get the right decision regarding the closest right estimation, in the section of validations, three different statistical metrics were used for the purpose of demonstrating the performance for each used method.

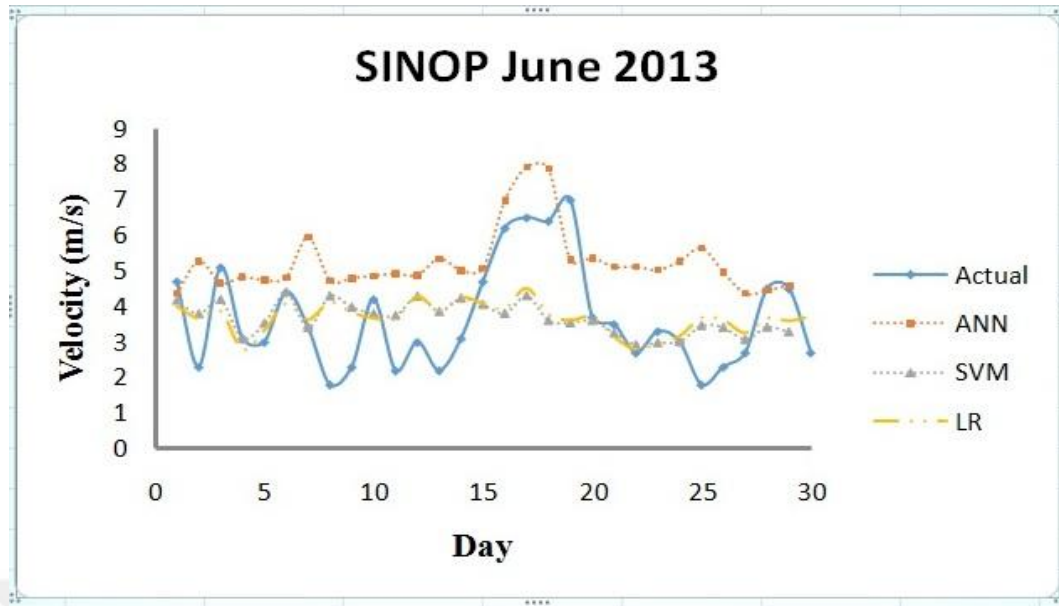


Figure 5.7. Actual and predict wind velocity in city SINOP in month June

Figure 5.8 shows, as was highlighted from the prior explanations of the graphs, that these figures only demonstrate the estimation distance of actual wind and the wind velocity predictions calculated by SVM, LR and ANN. Particularly, in Figure 5.8 the foreseeable red line curve of ANN obtained a better prediction result relying on scaling the gaps and oscillating with the actual wind, which is shown in blue. For SVM and LR from day 10 to day 30, the results are good.

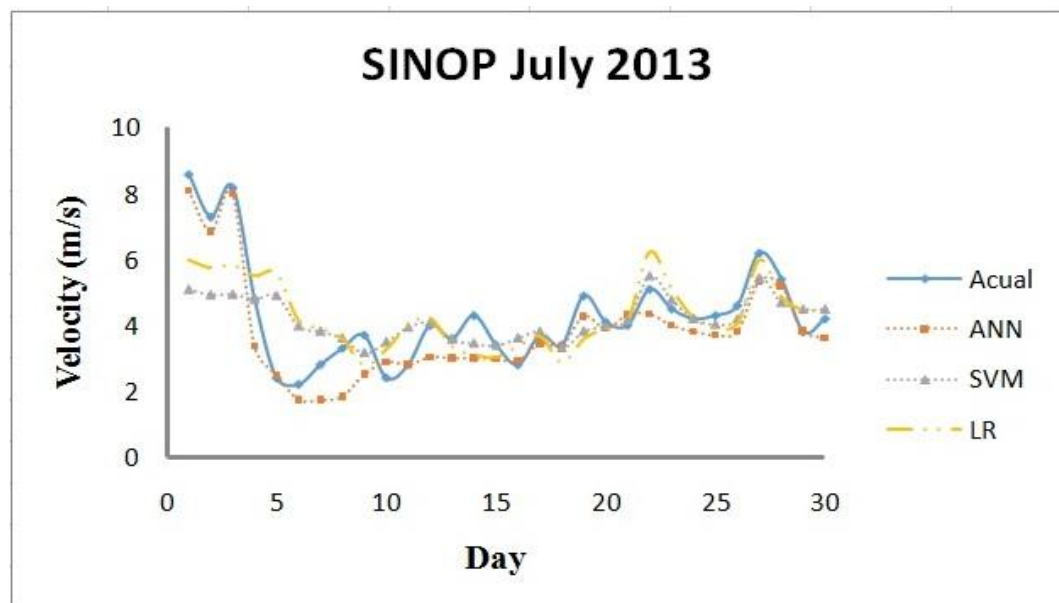


Figure 5.8. Actual and predict wind velocity in city SINOP in month July

Figure 5.9 shows the SINOP wind predictions, depicting that the actual wind curve fluctuates and gives variation results in different days. The actual wind in August is getting closer to the purple line of SVM, but the overall curve shows that SVM performs better than ANN and LR. In other places, the curves show closer to LR. Generally, to get the right decision regarding the most correct estimation, in the section of validations, three different statistical metrics were used for demonstrating the performance for each used method.

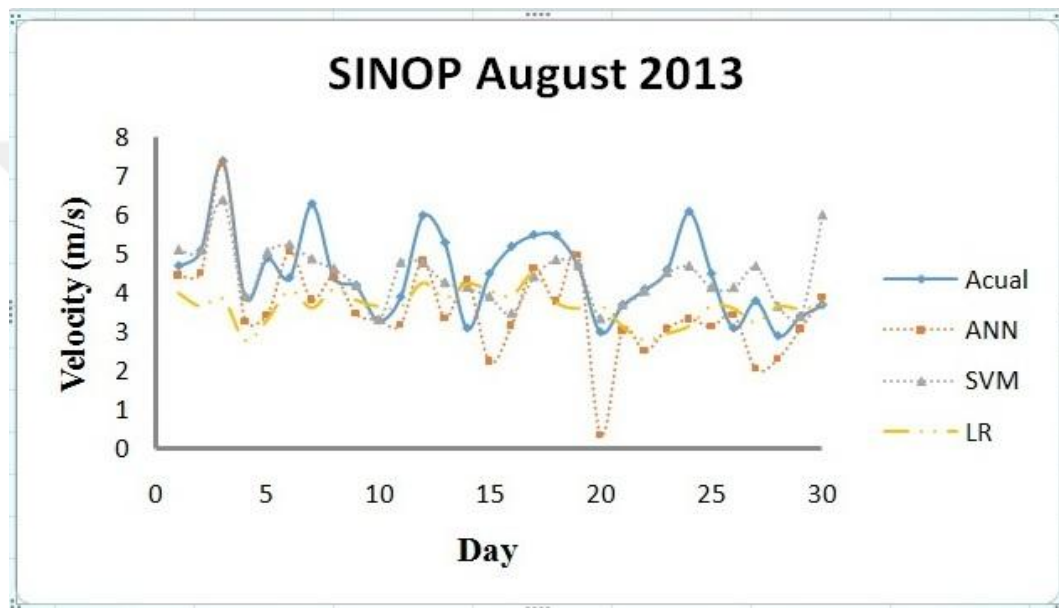


Figure 5.9. Actual and predict wind velocity in city SINOP in month August

Regarding station 3 shown in Figure 5.10, which is AMASRA, as illustrated in the above figure, the actual wind falls and rises with ANN in most of the periods. Nevertheless, in some parts of the actual wind, the graph shows that the curve gaps become reduced with LR and SVM. Particularly, more than once it converges with SVM and LR, but, comparatively, the graph indicates that all data points of the actual wind congregated with ANN in more points than LR and SVM. Once more, the statistical measurements will conclude the most accurate prediction for each method.

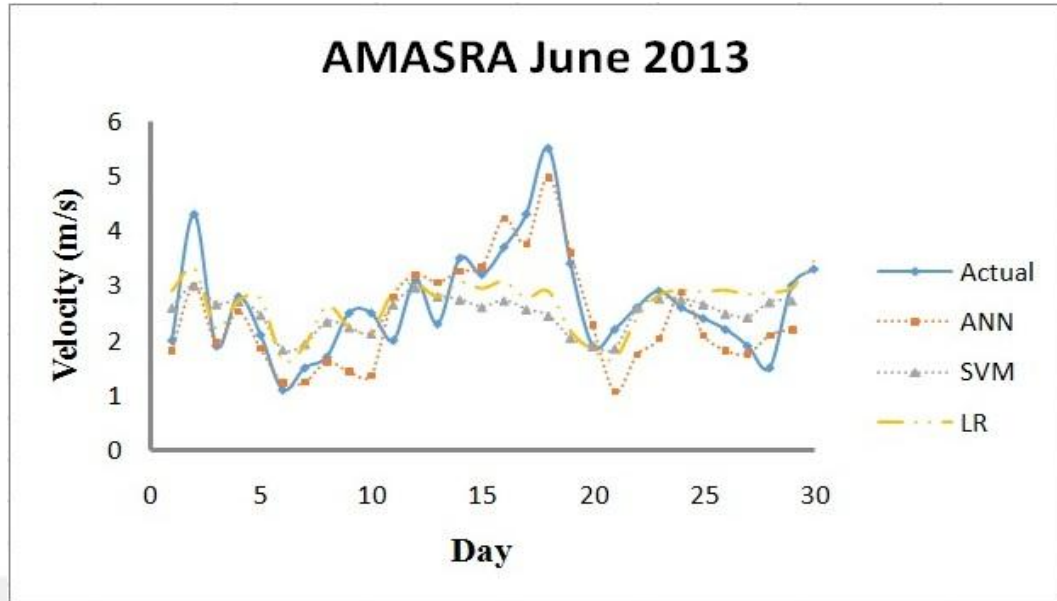


Figure 5.10. Actual and predict wind velocity in city AMASRA in month June

For the month of July regarding AMASRA city in Figure 5.11, it presents the AMASRA wind predictions as depicting that the actual wind curve fluctuated and gives variation of results on different days. ANN results are better than SVM and LR because of the distance values between ANN and the actual value is less than SVM and LR, while, in other places the SVM and LR curves are close with the actual wind curves.

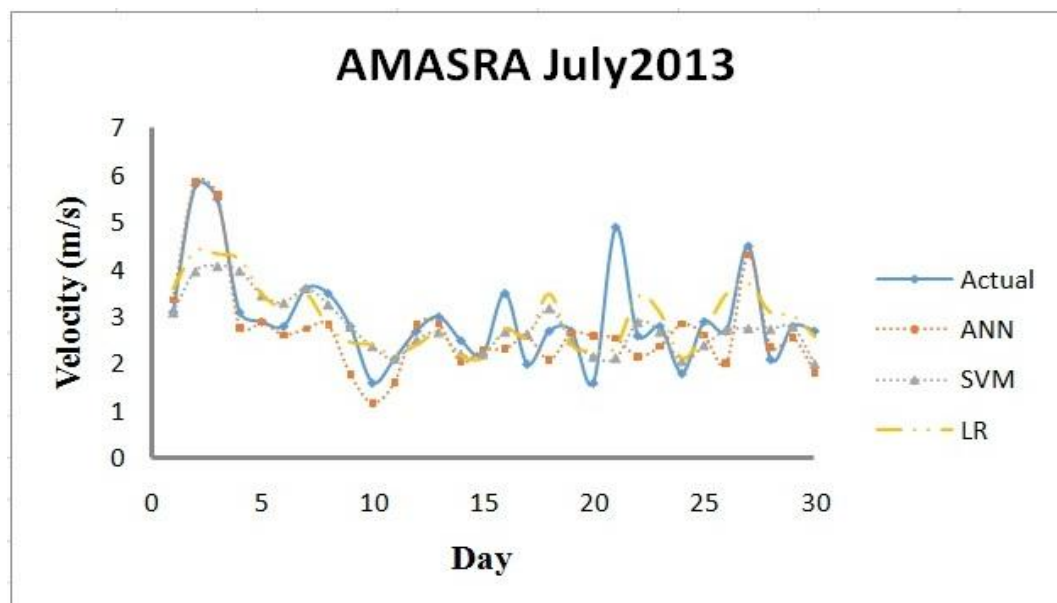


Figure 5.11. Actual and predict wind velocity in city AMASRA in month July

For July regarding AMASRA city in Figure 5.12 AMASRA wind predictions depict that the actual wind curve fluctuated and gives variation of results on different days. It is difficult to select which algorithm is better, but generally SVM and LR results are better than ANN because, on some days, the distance between actual values and SVM and LR are closer than ANN, while on some days the ANN curves are close to the actual wind curves.

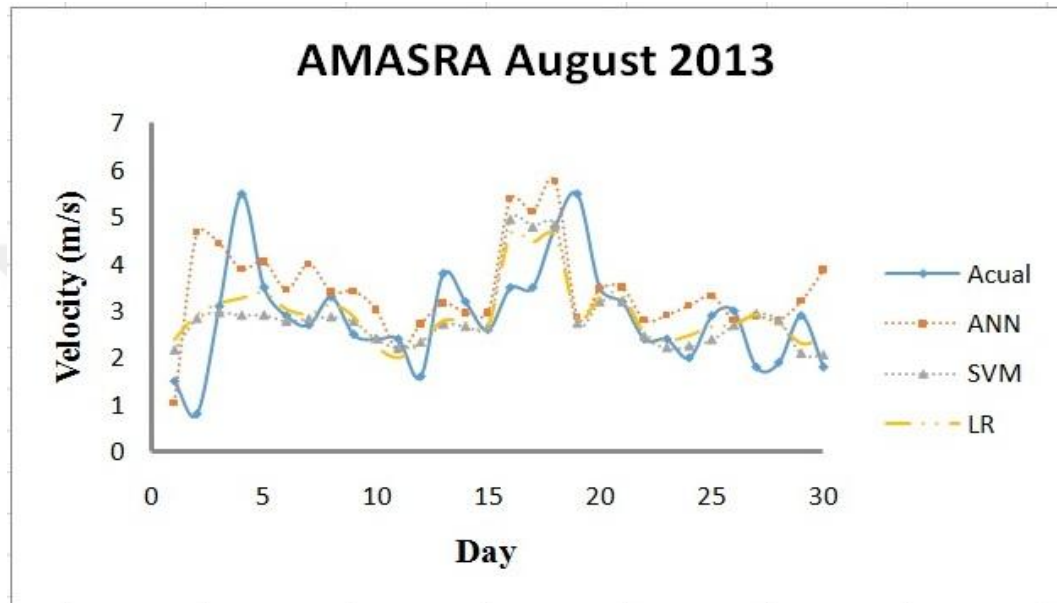


Figure 5.12. Actual and predict wind velocity in city AMASRA in month August

In this study, as shown in Figure 5.13 and as highlighted in the earlier sections, the daily wind speed data for five different stations were taken during the months of June, July and August to see the comparison. As seen in Figure 10, the prediction of ANN for the first five days integrated with the actual winds and mostly followed the variation of actual wind data exactly. In the case of SILE station, in the actual wind there are some sharp points with a level of 15m/s, whereas the lowest level of wind speed is with a level below 2m/s. The prediction of ANN agrees better with the actual value rather than SVM and LR. The worst case prediction is LR where it only gives the bottom line during the entire period.

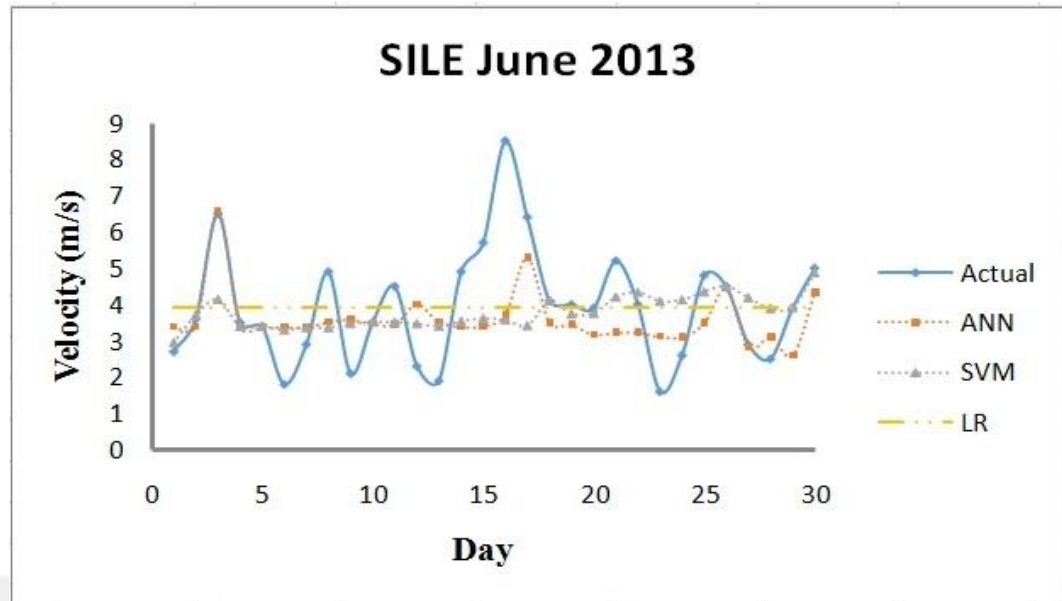


Figure 5.13. Actual and predict wind velocity in city SILE in month June

The experimental results for SILE city are depicted in Figure 5.14. ANN, in most of the fluctuations, follows the actual wind. It is difficult to choose which value prediction obtained for the three algorithms is better because all results have little distance from the actual wind value, but, generally ANN is better because, on some days, the ANN values are closer to the actual value compared with LR and SVM.

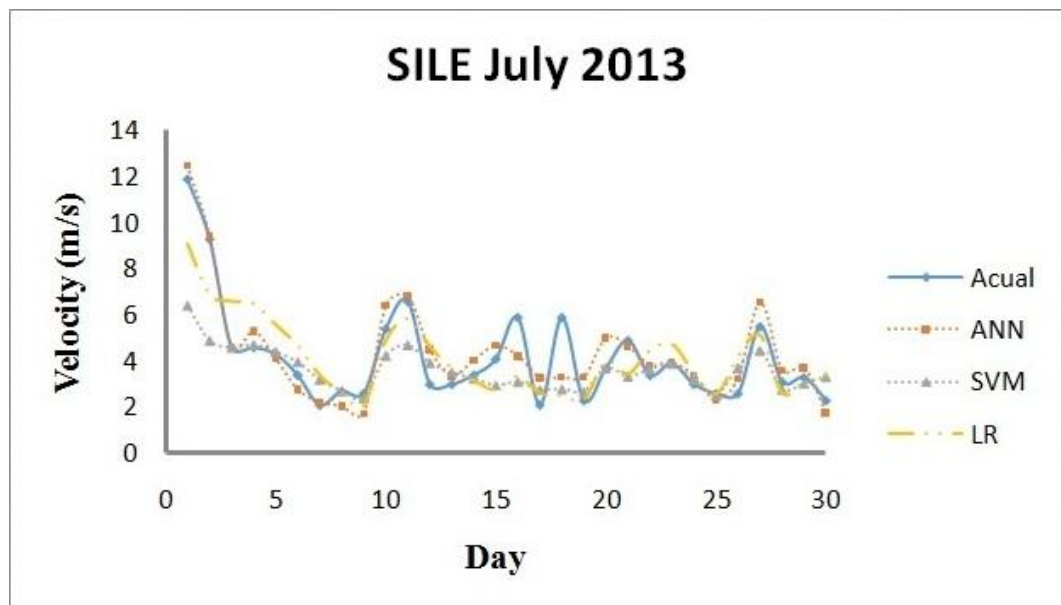


Figure 5.14. Actual and predict wind velocity in city SILE in month July

For August in this study, Figure 5.15 shows the SILE wind predictions, depicting that the actual wind curve fluctuated and gives variation of results on different days. In selecting the better algorithm to obtain the results for August, generally this is ANN, because the ANN prediction value is closer to the actual wind value compared with SVM and LR, while, on other days, SVM and LR were closer in value to the actual value.

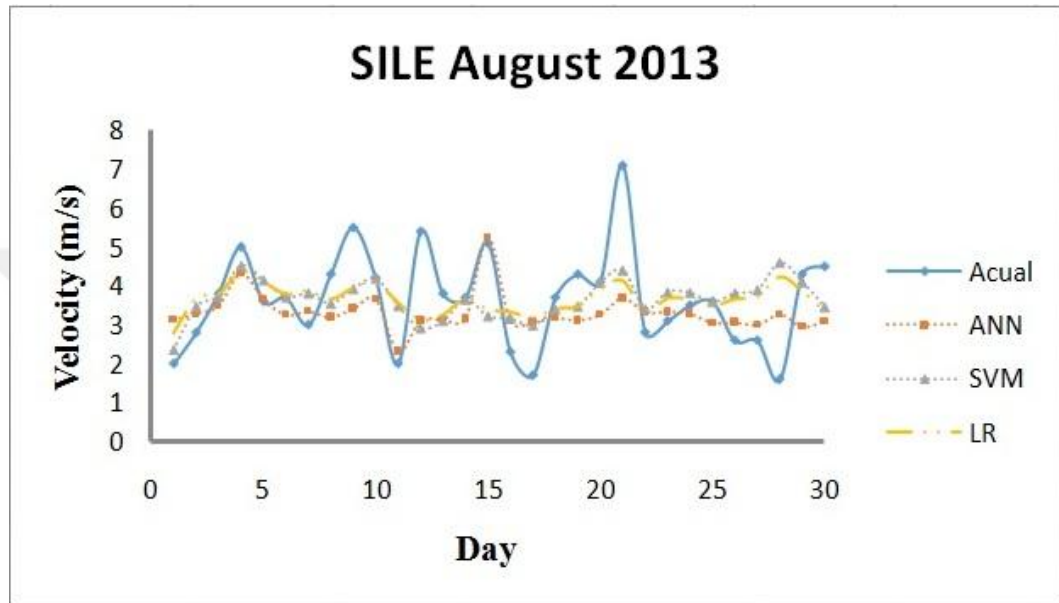


Figure 5.15. Actual and predict wind velocity in city SILE in month August

Once again, as shown in Figure 5.16, in AKCAKOCA station the actual wind oscillates. Results obtained in the case study for station 5 for June in Figure 5.16 show s ANN to be better than SVM and LR because the value obtained from the ANN algorithm is closer to with the actual value when comparing with the values obtained by SVM and LR. Therefore, SVM and LR have good results on more days.

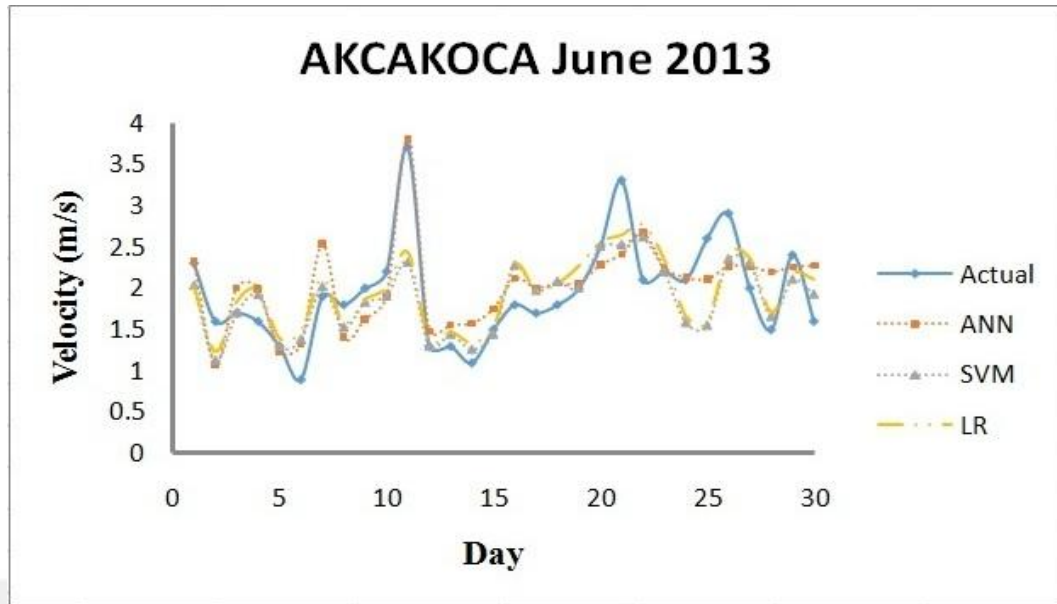


Figure 5.16. Actual and predict wind velocity in city AKCAKOCA in month June

The experiment results in Figure 5.17 relate to the wind speed prediction in AKCAKOCA for July. This presents an ambiguous outcome in indicating the superior attainment. It is difficult to choose which algorithm's result is better because all algorithm results in value with greater distance compared with actual value. However, on some days, the values of ANN are better than RL and SVM.

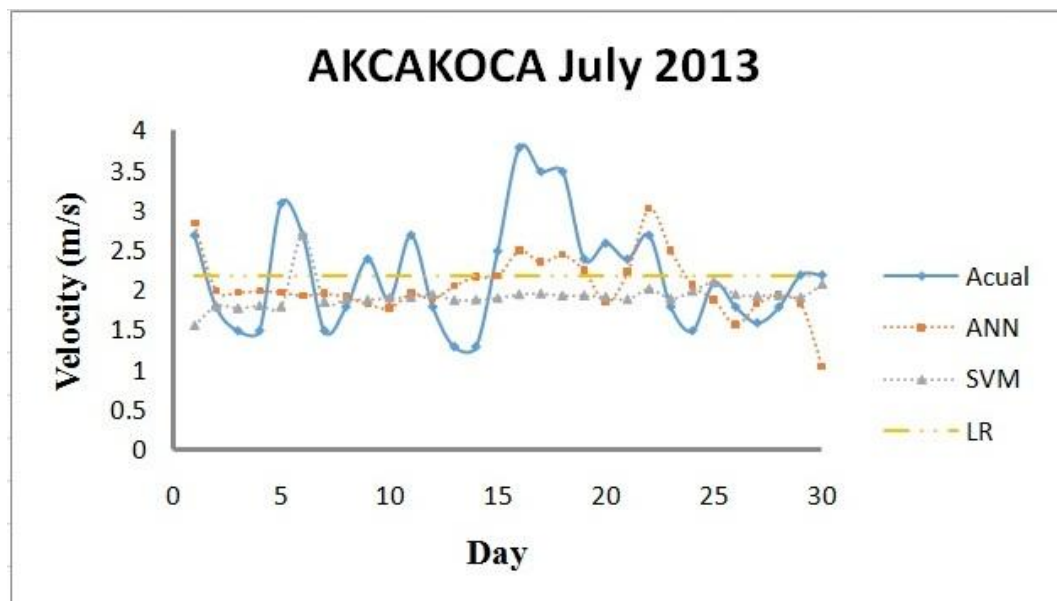


Figure 5.17. Actual and predict wind velocity in city AKCAKOCA in month July

Figure 5.18 shows the results of wind prediction for August in station 5 and illustrates that all results obtained for wind prediction are good for SVM, ANN and LR. However, for an estimated period of days (4 to 7) SVM and LR values are better than ANN because, in this period, SVM and LR values are closer than the actual curve comparing with the ANN curve.

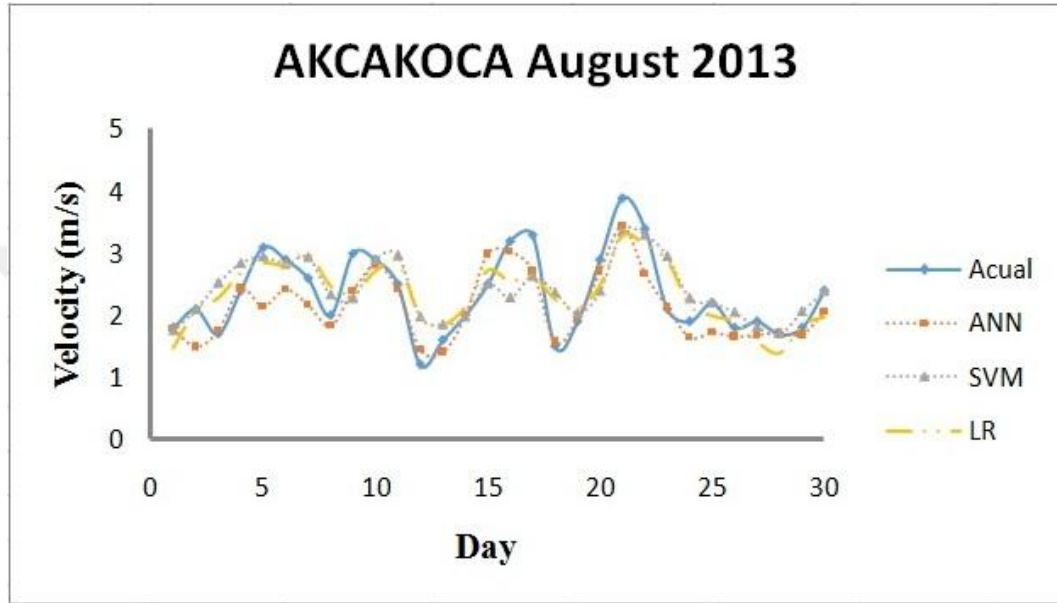


Figure 5.18. Actual and predict wind velocity in city AKCAKOCA in month July

5.2.2. Obtained results for three Statically Matrices

After loading datasets to obtain results for three algorithms (SVM, Neural Network, Linear Regression), the Weka tool is used for finding correlation coefficient, mean absolute errors and root mean squared error for the linear regression algorithm. Here we first browse the input dataset. Aside from showing correlation coefficient, mean absolute errors and root mean squared error, Weka shows various metrics. In this case, the correlation coefficient results are between 1 and -1.

In Table 5.2 values for measurements (correlation coefficient, mean absolute error, root mean squared error) by using SVM and linear regression and neural network for the June, July and August are shown.

The results reveal the climate analysis of the time taken for processing and different kinds of errors and correlation coefficient for five stations in Turkey. Given an

arrangement of named records, manufacturing a model, for example, a choice tree, and envisioning names for the future is not labeled records

In Table 5.2, or the five stations, each station contains a dataset of three months. In station 1, the neural network for measurement of correlation coefficient obtained the highest value, as required. For SVM for measurement of mean absolute error, the value is the lowest. Linear regression for measurement of root means squared error, has the lowest value. In station 2, neural network for measurement of correlation coefficient has the highest value compared to the other two values. Neural network for measurements of root and root mean squared error have the lowest value. In station 3, neural network for measurement of correlation coefficient has the highest value. In station 4, neural network for measurement of correlation coefficient has a higher value than the other values. SVM for mean absolute error and measurement of root mean squared error have the lowest value. In station 5, neural network for measurement of correlation coefficient has the hest value and SVM for mean absolute error and measurement of root mean squared error have the lowest values.

Table 5.2. Results for three algorithms by months

Station	Month	Linear regression			SVM			Neural Network		
		Correlation coefficient	Mean absolute error	Root mean squared error	Correlation coefficient	Mean absolute error	Root mean squared error	Correlation coefficient	Mean absolute error	Root mean squared error
1	6	0.52	0.24	0.30	0.51	0.23	0.30	0.70	0.19	0.25
1	7	0.48	0.37	0.47	0.42	0.27	0.35	0.83	0.23	0.29
1	8	0.41	0.24	0.30	0.44	0.20	0.29	0.59	0.24	0.33
2	6	0.28	1.09	1.37	0.21	1.07	1.40	0.51	1.73	1.98
2	7	0.65	0.85	1.14	0.63	0.80	1.22	0.94	0.55	0.68
2	8	0.60	0.71	0.85	0.57	0.65	0.88	0.60	1.08	1.35
3	6	0.49	0.61	0.81	0.68	0.52	0.59	0.80	0.51	0.63
3	7	0.64	0.61	0.78	0.61	0.54	0.82	0.80	0.50	0.69
3	8	0.63	0.61	0.79	0.61	0.60	0.82	0.75	0.72	0.85
4	6	0	1.92	1.53	0.19	1.05	1.52	0.48	1.02	1.38
4	7	0.74	1.11	1.40	0.74	1.03	1.65	0.92	0.67	0.86
4	8	0.28	0.89	1.16	0.27	0.87	1.18	0.46	0.86	1.12
5	6	0.68	0.34	0.44	0.6	0.32	0.44	0.75	0.34	0.40
5	7	0	0.56	0.67	0.15	0.52	0.71	0.44	0.51	0.62
5	8	0.74	0.36	0.42	0.72	0.33	0.45	0.88	0.27	0.36

The results for the five stations over all three months are shown in Table 5.3 below, and were obtained by loading datasets in Weka software for the algorithms

Table 5.3 shows three different metrics to calculate the performance of all used methods for wind velocity prediction. The correlation coefficient (r) is one of the metrics and ranges between 1 to -1, so the value is close to 1 or -1 it is considered as perfectly aligned. Conversely, data sets with values of r close to zero show little to no straight-line relationship. In this experiment, according to the obtained results and taking consideration of r , ANN achieved a better result, 0.57, compared to LR and SVM, which attained 0.36 and 0.37, respectively. Both MAE and RMSE demonstrate the average error of the prediction model in units of the variable of interest. The two measurements can run from 0 to ∞ and are not interested towards fault. They are adversely situated scores, which implies bring down values are better. Once more, looking at the above table and taking the value of MAE and RMSE, it distinctly shows that SVM has a lower error value (MAE=0.7 and RMSE=0.71) compared to the error values of LR (MAE=0.72 and RMSE=0.9732) and ANN (MAE= 0.78732 and RMSE=1.02).

SVM in the best case for measurement of means absolute error in station 1 for month 8, station 2 for month 6 and month 8, station 3 for month 6 and month 8, and station 5 for month 6 because SVM value obtained is the lower than ANN and linear regression.

The linear regression algorithm in the best case for measurement of root means squared error in station 1 for month 6, station 2 for month 8, station 3 for month 8, and obtained fewer than ANN and SVM.

The artificial neural network algorithm in the best case for measurement of root mean squared error and root mean squared error in station 2 and station 3 for month 7, in station 4 for every month, and in station 5 for month 7 and month 8 because the value obtained is lower than SVM and linear regression.

Table 5.3. Results for three algorithms by Stations

Station	Liner regression			SVM			Neural Network		
Number	Correlation coefficient	Mean absolute error	Root mean squared error	Correlation coefficient	Mean absolute error	Root mean squared error	Correlation coefficient	Mean absolute error	Root mean squared error
1	0.50	0.28	0.38	0.49	0.27	0.38	0.50	0.29	0.38
2	0.35	0.98	1.30	0.33	0.97	1.3	0.53	0.93	1.22
3	0.21	0.72	0.98	0.27	0.69	0.96	0.63	0.62	0.864
4	0.28	1.17	1.61	0.29	1.13	0.29	0.56	1.58	1.99
5	0.48	0.46	0.58	0.49	0.45	0.58	0.62	0.50	0.64
Average	0.36	0.72	0.97	0.37	0.70	0.71	0.57	0.78	1.02s

5.3. Summary

This chapter has provided sufficient evidence of the results of the research and implementation. This project is helpful to estimate the climate parameters. Estimating the climate parameter is today very essential for humans. Here, in this project, we estimate the climate parameters using data mining techniques. In this thesis, we used regression algorithms for estimating the climate parameters

In this chapter, three data mining techniques have been tested in predicting the wind velocity for five different stations in three different months during the summer time in Turkey. The datasets used in these experiments included various climate features such as temperature, pressure and humidity. The results have revealed that the statistical metrics based on the correlation coefficient, mean absolute error and root mean square error achieved comparable results between all existing methods in the thesis framework. Thus, the value of MAE and RMSE clearly shows that SVM has a lower error value compared to the error values of LR and ANN, whereas in regards to the value which represents the correlation coefficient, the ANN technique aligned the result much more perfectly than SVM and LR.

6. CONCLUSION

This section provides conclusions on the research work carried out on the estimation of climate parameters using data mining techniques. It also throws light on a summary of research and useful insights on the work done in this thesis

In this thesis, the wind velocity was found from the input parameters of station number, month, day, temperature and humidity. Here, we use different algorithms, linear regression, neural networks and SVM. In comparing the predicted wind velocity to actual wind velocity, the linear regression matches near values to actual compared to the other algorithms.

In this thesis, the Weka tool was used to find the correlation coefficient, mean absolute error and root mean squared error for all three algorithms.

In conclusion, the evaluation results revealed that SVM earned better error values than LR and ANN, whereas ANN obtained a better coefficient result than SVM and LR.

6.1. Summary of Findings

In the review of the literature, it is understood that there are many techniques used in data mining. However, little information is found for finding the wind velocity using such techniques. We found this is to be a research gap and motivation to work on this thesis. The literature review insights resulted in a research hypothesis that investigated the need for having an intuitive interface. This research hypothesis has been tested with the proposed system, which proved to be very user-friendly.

6.2. Conclusions and Future Work

1. This thesis focused on data mining techniques used to build different models using training samples.
2. The thesis explored different data mining techniques using Weka environment for prediction of climate parameters from a climate dataset.
3. The techniques employed include simple algorithms, neural networks, SVM and linear regression

4. The results revealed that different algorithms have different performance as well as having a correlation coefficient, mean absolute error and root mean squared error in prediction capabilities.
5. This work can be improved to develop the used method or integrated with some additional methods into a hybrid model to overcome the limitations of each existing method.
6. Generally, results obtained the wind predictions exactly. Artificial neural network and support vector machine compared with actual wind value are perfect.
7. According to the comparison of different algorithms based on the results of the correlation coefficient, neural network is better than SVM and linear regression because its results are the highest value.
8. According to the comparison of different algorithms based on the results of the mean absolute error, support vector machine is better than linear regression and artificial neural network.
9. Estimation climate parameters for five stations in Turkru used three algorithms, linear regression, SVM and neural network. Datasets were collected by month for every station and then a value for three measurements was obtained. For every station and all months, the neural network algorithm for correlation coefficient was the best case, because the value for all situations and months was greater than the other algorithms. Generally, the SVM algorithm for errors is better than the other algorithms.
10. The work can be improved by finding the best geographical position for taking the wind velocity.

REFERENCES

- [1] **Langone, R., Alzate, C., De Ketelaere, B., Vlasselaer, J., Meert, W. and Suykens, J.A.**, 2015. LS-SVM based spectral clustering and regression for predicting maintenance of industrial machines. *Engineering Applications of Artificial Intelligence*, 37, pp.268-278.
- [2] **Claesen, M., De Smet, F., Suykens, J.A. and De Moor, B.**, 2015. A robust ensemble approach to learn from positive and unlabeled data using SVM based models. *Neurocomputing*, 160, pp.73-84.
- [3] **Frandi, E., Nanculef, R. and Suykens, J.A.**, 2015, July. A PARTAN-accelerated Frank-Wolfe algorithm for large-scale SVM classification. In *Neural Networks (IJCNN), 2015 International Joint Conference on* (pp. 1-8). IEEE.
- [4] **Drumetz, L., Veganzones, M.A., Gómez, R.M., Tochon, G., Dalla Mura, M., Licciardi, G.A., Jutten, C. and Chanussot, J.**, 2016. Hyperspectral local intrinsic dimensionality. *IEEE Transactions on Geoscience and Remote Sensing*, 54(7), pp.4063-4078.
- [5] **Schmidhuber, J.**, 2015. Deep learning in neural networks: An overview. *Neural networks*, 61, pp.85-117.
- [6] **Aziz, M.T.R.A. and Yusof, Y.**, 2016, August. Graduates employment classification using data mining approach. In *AIP Conference Proceedings* (Vol. 1761, No. 1, p. 020002). AIP Publishing.
- [7] **Sanghavi, D., Patel, H., and Nair, S.**, 2014. Logistic regression in data mining and its application in identification of disease. *International Journal of Current Engineering and Technology* , vol.4, (1-6), pp. 3837-3839, 2014.
- [8] **Arora, R.**, 2012. Comparative analysis of classification algorithms on different datasets using WEKA. *International Journal of Computer Applications*, 54(13).
- [9] **Zhang, H., Berg, A.C., Maire, M. and Malik, J.**, 2006. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (Vol. 2, pp. 2126-2136). IEEE.

- [10] **Barbounis, T.G. and Theocharis, J.B.,** 2007. Locally recurrent neural networks for wind speed prediction using spatial correlation. *Information Sciences*, 177(24), pp.5775-5797.
- [11] **Kani, S.P. and Ardehali, M.M.,** 2011. Very short-term wind speed prediction: a new artificial neural network–Markov chain model. *Energy Conversion and Management*, 52(1), pp.738-745.
- [12] **Kok, B., Uyar, M., Varol, Y., Koca, A. and Oztop, H.F.,** 2013. Analyzing of thermal mixing phenomena in a rectangular channel with twin jets by using the artificial neural network. *Nuclear Engineering and Design*, 265, pp.554-565.
- [13] **Senpinar, A. and Karabatak, M.,** 2012. Estimation of incomplete meteorological and solar radiation data in turkey using the artificial neural network. *11th International Conference on Sustainable Energy Technologies.* , on (pp.1476-1483), Vancouver, Canada.
- [14] **Karabatak, M. and Senpinar, A.,** 2012. Prediction of monthly average daily global solar radiation and an average temperature of next month using ANN based on previous months' data of some cities in turkey. *11th International Conference on Sustainable Energy Technologies.*, on (pp.1476-1483), Vancouver, Canada.
- [15] **Pinson, P., Nielsen, H.A., Madsen, H. and Nielsen, T.S.,** 2008. Local linear regression with adaptive orthogonal fitting for the wind power application. *Statistics and Computing*, 18(1), pp.59-71.
- [16] **Bilgili, M., Sahin, B. and Yasar, A.,** 2007. Application of artificial neural networks for the wind speed prediction of target station using reference stations data. *Renewable Energy*, 32(14), pp.2350-2360.
- [17] **Mohandes, M.A., Halawani, T.O., Rehman, S. and Hussain, A.A.,** 2004. Support vector machines for wind speed prediction. *Renewable Energy*, 29(6), pp.939-947.
- [18] **Salcedo-Sanz, S., Ortiz-Garcı, E.G., Pérez-Bellido, Á.M., Portilla-Figueras, A. and Prieto, L.,** 2011. Short term wind speed prediction based on evolutionary support vector regression algorithms. *Expert Systems with Applications*, 38(4), pp.4052-4057.

- [19] **Liu, H., Tian, H.Q. and Li, Y.F.**, 2012. Comparison of two new ARIMA-ANN and ARIMA-Kalman hybrid methods for wind speed prediction. *Applied Energy*, 98, pp.415-424.
- [20] **Cheng, X. and Guo, P.**, 2013, May. Short-term wind speed prediction based on support vector machine of fuzzy information granulation. In *Control and Decision Conference (CCDC), 2013 25th Chinese* (pp. 1918-1923). IEEE.
- [21] **Lee, C.Y. and He, Y.L.**, 2012, January. Wind prediction based on general regression neural network. In *Intelligent System Design and Engineering Application (ISDEA), 2012 Second International Conference on* (pp. 617-620). IEEE.
- [22] **Zhao, P., Xia, J., Dai, Y. and He, J.**, 2010, June. Wind speed prediction using support vector regression. In *Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on* (pp. 882-886). IEEE.
- [23] **Kulkarni, M.A., Patil, S., Rama, G.V. and Sen, P.N.**, 2008. Wind speed prediction using statistical regression and neural network. *Journal of Earth System Science*, 117(4), pp.457-463.
- [24] **Pinson, P., Nielsen, H.A., Madsen, H. and Nielsen, T.S.**, 2008. Local linear regression with adaptive orthogonal fitting for the wind power application. *Statistics and Computing*, 18(1), pp.59-71.
- [25] **Douak, F., Benoudjit, N. and Melgani, F.**, 2013, May. Design of a multiblock general regression neural network for wind speed prediction in Algeria. In *Systems, Signal Processing and their Applications (WoSSPA), 2013 8th International Workshop on* (pp. 390-395). IEEE.
- [26] **Lahouar, A. and Slama, J.B.H.**, 2014, March. Wind speed and direction prediction for wind farms using support vector regression. In *Renewable Energy Congress (IREC), 2014 5th International* (pp. 1-6). IEEE.
- [27] **Van der Walt, C.M. and Botha, N.**, 2016, November. A comparison of regression algorithms for wind speed forecasting at Alexander Bay. In *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 2016* (pp. 1-5). IEEE.
- [28] **Finamore, A.R., Galdi, V., Calderaro, V., Piccolo, A., Conio, G. and Grasso, S.**, 2016. Artificial neural network application in wind forecasting: An one-hour-ahead wind speed prediction.

- [29] **Koca, A., Oztop, H.F., Varol, Y. and Koca, G.O.,** 2011. Estimation of solar radiation using artificial neural networks with different input parameters for Mediterranean region of Anatolia in Turkey. *Expert Systems with Applications*, 38(7), pp.8756-8762.
- [30] **Olaiya, F. and Adeyemo, A.B.,** 2012. Application of data mining techniques in weather prediction and climate change studies. *International Journal of Information Engineering and Electronic Business*, 4(1), p.51.
- [31] **Kang, J. S.** Modeling of biological signal transduction pathway using artificial neural networks. *APEC Youth Scientist Journal Vol. 5*, pp. 72–79.
- [32] **Vignan, D. B.** and Technology, I. (2017) ‘Use of Artificial Neural Network in Pattern Recognition’, (August)
- [33] **Sarıkaya, B.,** 2009. *Lung mass classification using wavelets and support vector machines* (Master dissertation, Fen Bilimleri Enstitüsü).
- [34] **Paisitkriangkrai, P.,** 2012. Linear regression and support vector regression. *University Lecture*.
- [35] **Platt, J.C.,** 1998. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, 1st ed.; Schölkopf, B., Burges, C.J.C., Smola, A.J., Eds.; MIT Press: Cambridge, UK,

APPENDICES

Appendix A: Climate Dataset

Station Number	Month	Day	Pressure (hPa)	Humidity (%)	Temperature (°C)	Wind Velocity (m/s)
1	6	1	1008.7	67.7	19.7	1.7
1	6	2	1005	80.3	19.6	1.3
1	6	3	1010.1	69.3	17.3	2.5
1	6	4	1011.1	72	19.4	1.4
1	6	5	1013.4	82.7	17.4	1.3
1	6	6	1014.6	80.7	18.8	1.8
1	6	7	1013.6	72	19.5	1.5
1	6	8	1010.1	77.7	20.4	1
1	6	9	1012.3	81.3	19.2	2.4
1	6	10	1012.9	77.7	19.2	1.9
1	6	11	1008.4	75.7	20.6	1.8
1	6	12	1005.3	80.3	21.2	1.7
1	6	13	1006.5	80.3	21.9	1.2
1	6	14	1009.8	88.7	20	1.5
1	6	15	1010.7	85.3	20.8	1.8
1	6	16	1011.8	80.7	21.8	1.6
1	6	17	1013	80.3	23.1	2.1
1	6	18	1014.5	79.7	24	1.4
1	6	19	1015.1	79	22.5	1.6
1	6	20	1016.4	72.3	23.2	1.6
1	6	21	1015.4	71.7	20.9	2.2
1	6	22	1013.2	70.3	21.3	2.2
1	6	23	1008.4	78.3	21.2	1.6
1	6	24	1008.7	84	21.9	1.4
1	6	25	1011.4	82	23.1	1.5
1	6	26	1011.3	83	23	2.1
1	6	27	1009.5	81.3	23	1.6
1	6	28	1008.3	84.3	23.9	1.8
1	6	29	1004.1	80.3	25.3	1.2
1	6	30	1010.1	85.3	22.7	1.9
1	7	1	1015.4	73.3	21.2	3.8
1	7	2	1014.2	80.3	19.2	2.6
1	7	3	1012.7	70.3	19.4	1.6
1	7	4	1011.7	77	21	1.8
1	7	5	1011.5	79	23	1.9
1	7	6	1013.5	80.7	24.1	1.9
1	7	7	1015.3	81.3	23.9	1.3
1	7	8	1012.9	81	23.6	1.3
1	7	9	1007.9	79.3	24.9	1.2
1	7	10	1007.5	80.3	24.5	1.5
1	7	11	1010.7	75.7	24.7	1.7

1	7	12	1011.7	77.7	23.6	1.3
1	7	13	1011.4	79.3	23	1.5
1	7	14	1011.3	81	23.2	1.9
1	7	15	1012.5	75	24.8	1.4
1	7	16	1013.7	70.3	24.4	2.8
1	7	17	1014.2	80.3	23.6	2.4
1	7	18	1014.1	83.7	23.5	2.6
1	7	19	1010.5	76	24.2	1.6
1	7	20	1011.6	77.7	24.1	2
1	7	21	1012.3	79.3	24.1	2
1	7	22	1011.5	66.7	20.9	2.6
1	7	23	1012.2	69.3	21.7	1.6
1	7	24	1012.1	76.7	21.2	1.6
1	7	25	1011.5	67	22.2	2.1
1	7	26	1011.7	71	22.5	1.9
1	7	27	1016.1	79	21.6	1.5
1	7	28	1015.5	81.3	21.6	2.2
1	7	29	1009.4	73	21.9	2
1	7	30	1008.3	73.7	22.1	1.9
1	7	31	1011.4	76.3	23.4	2.1
1	8	1	1014.2	85.3	22.6	1.6
1	8	2	1014.5	84.3	22.7	1.9
1	8	3	1014.9	76	22.7	2
1	8	4	1015.5	67.7	24.4	1.9
1	8	5	1013.4	68.3	23.3	2.3
1	8	6	1013.8	76.7	22.9	2.5
1	8	7	1015.2	80	22.8	2.3
1	8	8	1011.8	80.3	23.2	1.7
1	8	9	1006.8	79	23.7	1.9
1	8	10	1006.9	80	23.4	1.7
1	8	11	1011.9	78	23.4	1.8
1	8	12	1012.3	82.7	23.6	1.7
1	8	13	1011.3	82.3	22.9	2
1	8	14	1012.5	78.3	23.1	2
1	8	15	1012.7	82	24.9	1.8
1	8	16	1012.4	83.3	25.4	1.6
1	8	17	1013.2	83.7	25.2	2.3
1	8	18	1013.5	74.7	23.9	1.9
1	8	19	1012.7	77.3	24.3	2
1	8	20	1012.3	77	23.2	2.2
1	8	21	1010.7	73.3	24	2.4
1	8	22	1010.5	75.7	24.2	1.9
1	8	23	1011.2	78	23.8	2.1
1	8	24	1010.7	79	22.9	2.2
1	8	25	1010.7	75.7	23.6	1.6
1	8	26	1010.8	81.3	22.9	1.9

1	8	27	1009.8	84.3	23	1.8
1	8	28	1010.2	83.7	22.4	2.2
1	8	29	1011.6	81.7	23	1.6
1	8	30	1014.1	84.7	22.8	3.2
2	6	1	1003	80.3	18.4	4.7
2	6	2	997.9	74	19.5	2.3
2	6	3	1001	77.7	16.9	5.1
2	6	4	1004.9	57.7	19.2	3.1
2	6	5	1007.1	67.7	19.3	3
2	6	6	1008.3	81.7	16.5	4.4
2	6	7		73.3	18.3	3.4
2	6	8	1004.9	82.3	17.8	1.8
2	6	9	1005.7	77	18.7	2.3
2	6	10	1007	74	19.2	4.2
2	6	11	1003.2	73.7	19.6	2.2
2	6	12	999.1	85.3	19.2	3
2	6	13	1000	78	20.3	2.2
2	6	14	1002.5	85	19.6	3.1
2	6	15	1002.8	81.7	19.9	4.7
2	6	16	1002.6	79	21.3	6.2
2	6	17	1004.2	89.7	21	6.5
2	6	18	1005.8	76	22	6.4
2	6	19	1008.4	73	21.2	7
2	6	20	1010.3	74.7	21.3	3.7
2	6	21	1010.1	64.3	20.1	3.5
2	6	22	1007.8	57.7	20.5	2.7
2	6	23	1002.5	61	21.4	3.3
2	6	24	1003.5	64.7	22.8	3
2	6	25	1005.4	74.3	22.5	1.8
2	6	26	1005.6	73	22.4	2.3
2	6	27	1003	66.3	22.8	2.7
2	6	28	1001.8	74	22.7	4.5
2	6	29	996.8	72.7	23.4	4.5
2	6	30	998.7	75	23.8	2.7
2	7	1	1005.2	80.7	20.4	8.6
2	7	2	1006.2	74.7	20.7	7.3
2	7	3	1006.1	64.3	20.6	8.2
2	7	4	1005.6	71.3	21	4.8
2	7	5	1005.8	84	20.9	2.4
2	7	6	1007.3	75.3	22.7	2.2
2	7	7	1009	73	23	2.8
2	7	8	1007	62.7	23.4	3.3
2	7	9	1002.4	61.7	24.4	3.7
2	7	10	1002	65.3	23.8	2.4
2	7	11	1004.4	68	22.9	2.8
2	7	12	1005.2	64.7	22.6	4

2	7	13	1004.8	69.7	23.7	3.6
2	7	14	1005	73	24	4.3
2	7	15	1006.4	70.3	24.1	3.4
2	7	16	1007.2	74	23.7	2.8
2	7	17	1007.3	71.3	23.3	3.7
2	7	18	1007.8	69.3	24.3	3.4
2	7	19	1004.4	72.3	23.4	4.9
2	7	20	1005	60	23	4.1
2	7	21	1005.7	68.3	22.7	4
2	7	22	1005.4	76.3	20.1	5.1
2	7	23	1005.7	65	21.5	4.5
2	7	24	1006	55.7	22.6	4.2
2	7	25	1004.3	55.3	23	4.3
2	7	26	1003.9	64.3	22.8	4.6
2	7	27	1008.8	87.7	20.4	6.2
2	7	28	1009.6	75.3	21.8	5.4
2	7	29	1003.9	69.7	22.3	3.8
2	7	30	1001.9	60.7	22.3	4.2
2	7	31	1004.5	59	23.2	4.2
2	8	1	1007.3	70.3	23.5	4.7
2	8	2	1007.4	70.3	23.5	5.1
2	8	3	1008.3	74.3	22	7.4
2	8	4	1009.5	58.7	23.7	3.9
2	8	5	1007.3	65.3	22.3	4.9
2	8	6	1007.2	68.7	22.6	4.4
2	8	7	1009.1	66.7	23.3	6.3
2	8	8	1007	70.7	24.1	4.4
2	8	9	1001.8	68	23	4.2
2	8	10	1000.3	67	24.1	3.3
2	8	11	1004.4	69	22.6	3.9
2	8	12	1005.5	69.7	23	6
2	8	13	1005.9	66.3	23.3	5.3
2	8	14	1006.2	68.7	24.1	3.1
2	8	15	1006.1	64.7	23.6	4.5
2	8	16	1005.4	65	24.3	5.2
2	8	17	1006.4	73	24.4	5.5
2	8	18	1007.7	70.7	23.3	5.5
2	8	19	1007.6	73.3	24.1	4.7
2	8	20	1007.4	62	24.2	3
2	8	21	1004.5	64.7	23.3	3.7
2	8	22	1003.9	67.3	23	4.1
2	8	23	1004.5	73	23.4	4.6
2	8	24	1005.1	70	22.5	6.1
2	8	25	1004	64.3	22	4.5
2	8	26	1005.2	69	23.3	3.1
2	8	27	1003.8	73	22.7	3.8

2	8	28	1003.1	65.3	22.8	2.9
2	8	29	1005.1	67.3	24.2	3.4
2	8	30	1007.4	77.3	21.8	3.7
3	6	1	1007.2	93	19.1	2
3	6	2	1003.8	91	19.5	4.3
3	6	3	1004.6	96.7	17.1	1.9
3	6	4	1009.1	67.7	19.7	2.8
3	6	5	1011.3	73	20.5	2.1
3	6	6	1012.7	99	18.2	1.1
3	6	7	1012.4	96.7	19	1.5
3	6	8	1009.5	95	20.3	1.7
3	6	9	1009.4	99	19.5	2.5
3	6	10	1010.6	98.7	19.8	2.5
3	6	11	1007.7	87	21.1	2
3	6	12	1002.9	99	20.5	3.1
3	6	13	1004.2	99	20.4	2.3
3	6	14	1005.6	98	21.6	3.5
3	6	15	1006.8	99	21.8	3.2
3	6	16	1006.8	91	22.3	3.7
3	6	17	1007	99	21.8	4.3
3	6	18	1008.7	98	22.7	5.5
3	6	19	1011.7	99	21.8	3.4
3	6	20	1013.8	91.3	21.6	1.9
3	6	21	1014.6	87	21.6	2.2
3	6	22	1012.2	54.7	23.1	2.6
3	6	23	1008.3	75	23.4	2.9
3	6	24	1006.7	89	23.4	2.6
3	6	25	1008.2	86	23.9	2.4
3	6	26	1010	87.3	24.6	2.2
3	6	27	1010	92	24.6	1.9
3	6	28	1006.8	93	24.1	1.5
3	6	29	1005.9	99	24.3	3
3	6	30	1001.2	93.3	24.6	3.3
3	7	1	1003.3	94.7	22.3	3.1
3	7	2	1007.6	99	20.9	5.8
3	7	3	1009.4	79.7	21	5.5
3	7	4	1009.6	95	21.2	3.1
3	7	5	1008.7		22.5	2.9
3	7	6	1009.3		23	2.8
3	7	7	1011		22.5	3.6
3	7	8	1012.5		23.7	3.5
3	7	9	1010.9	95.3	24.3	2.8
3	7	10	1007	97.3	24.4	1.6
3	7	11	1005.9	84.3	24.8	2.1
3	7	12	1008.3		24.4	2.7
3	7	13	1008.8		24	3

3	7	14	1008.2	99	24.8	2.5
3	7	15	1008.8	95.3	24.9	2.2
3	7	16	1009.8	91.7	23.8	3.5
3	7	17	1010.9	87.7	24.1	2
3	7	18	1010.7	97.3	22.5	2.7
3	7	19	1011.3	96.3	24.4	2.7
3	7	20	1008.5	83.7	24.6	1.6
3	7	21	1008.1	84.3	24.5	4.9
3	7	22	1008.9	74.7	22.6	2.6
3	7	23	1009.4	74.7	23.2	2.8
3	7	24	1009.8	61.7	24.9	1.8
3	7	25	1009.8	73.3	23.9	2.9
3	7	26	1007.9	76	22.5	2.7
3	7	27	1007.1	95.7	22.1	4.5
3	7	28	1011.8	91	23.2	2.1
3	7	29	1013.2	94.7	23.3	2.8
3	7	30	1007.9	80	24.1	2.7
3	7	31	1005.9	83.7	24.9	1.5
3	8	1	1008.2	99	23.4	0.8
3	8	2	1010.8	94.7	24.5	3.1
3	8	3	1010.7	82.3	24	5.5
3	8	4	1011.7	85	23.6	3.5
3	8	5	1013	90	23.5	2.9
3	8	6	1011.2	92.3	24.1	2.7
3	8	7	1010.7	93.3	24.7	3.3
3	8	8	1012.6	99	24.5	2.5
3	8	9	1011.1	99	24.9	2.4
3	8	10	1006.7	92.7	25	2.4
3	8	11	1004.4	86	24.8	1.6
3	8	12	1007.8	98.3	24.8	3.8
3	8	13	1009.2	87.7	24.8	3.2
3	8	14	1009.9	95.3	25	2.6
3	8	15	1009.8	99	25.4	3.5
3	8	16	1009.4		25.9	3.5
3	8	17	1008.4		25.8	4.8
3	8	18	1009.3		25.6	5.5
3	8	19	1010.9	99	25.4	3.5
3	8	20	1011.4	81.7	25.5	3.2
3	8	21	1011.1	77	25.3	2.4
3	8	22	1008.3	93.3	25.2	2.4
3	8	23	1007.4	93.3	24.7	2
3	8	24	1008.1	93.7	24.6	2.9
3	8	25	1009.2	93.7	24.7	3
3	8	26	1009.1	91.3	25.8	1.8
3	8	27	1009.7	85	26	1.9
3	8	28	1008.9	87.7	26.1	2.9

3	8	29	1006.9	92	24.7	1.8
3	8	30	1008.4	98.7	24.5	3
4	6	1	1001.7	73	19.6	2.7
4	6	2		72.7	19.3	3.6
4	6	3		75.7	17.2	6.5
4	6	4	1003.7	66	19.1	3.5
4	6	5	1006.1	61.7	19.9	3.4
4	6	6	1007.2	73	19.7	1.8
4	6	7	1006.4	71.3	19.9	2.9
4	6	8	1002.4	79.7	19.5	4.9
4	6	9	1004.6	80.7	19.2	2.1
4	6	10	1005.8	78	19.6	3.5
4	6	11	1001.3	77.7	20	4.5
4	6	12	997.9	84.3	20	2.3
4	6	13	999.2	80	21.1	1.9
4	6	14	1001.8	76.7	21.1	4.9
4	6	15	1002.6	77.7	21	5.7
4	6	16	1002.8	82.3	21.1	8.5
4	6	17	1004.4	88.7	21.7	6.4
4	6	18		84.3	22.1	4.1
4	6	19	1007.8	80	21.6	4
4	6	20	1009.1	74.7	22.4	3.9
4	6	21	1008.3	58	22.1	5.2
4	6	22	1005.9	65.3	21.1	4
4	6	23	1002.9	75	21.8	1.6
4	6	24	1001.1	77	21.8	2.6
4	6	25	1001.4	80.7	20.7	4.8
4	6	26	1003.7	84.7	19.9	4.5
4	6	27	1003.8	78.3	22.5	2.9
4	6	28	1002.4	84	23.8	2.5
4	6	29	1001	83	24.2	3.9
4	6	30		81	23	5
4	7	1	1000.1	81.3	20.1	11.9
4	7	2	1006.6	81.7	19.5	9.3
4	7	3	1006.4	71.7	19.9	4.6
4	7	4	1005.3	76.3	20.5	4.6
4	7	5	1004.3	81	21.9	4.3
4	7	6	1004.2	80	22.9	3.4
4	7	7	1005.9	80	23.5	2.1
4	7	8	1007.7	79.7	23.5	2.7
4	7	9	1005.6	68.7	25	2.6
4	7	10	1000.1	71	24.5	5.4
4	7	11	999.9	68.7	23.6	6.6
4	7	12	1003.2	71	23.4	3
4	7	13	1004.3	77.7	23.9	3
4	7	14	1004	76	24.6	3.4

4	7	15	1004	69.7	25	4.1
4	7	16	1005.1	65.3	23.9	5.9
4	7	17	1006.2	70.3	24.2	2.1
4	7	18	1006.4	74	24.2	5.9
4	7	19	1006.8	73	24.2	2.3
4	7	20	1003.1	68.7	24.1	3.7
4	7	21	1004.2	67.7	24.2	4.9
4	7	22	1004.9	57.3	22.8	3.4
4	7	23	1004.3	60.7	22.8	3.9
4	7	24	1004.8	59.7	23.9	3
4	7	25	1004.9	47	24.8	2.6
4	7	26	1004	63	23.5	2.6
4	7	27	1003.6	75	22.7	5.5
4	7	28	1008.4	69.7	23.2	3.1
4	7	29	1008.4	69.3	22.7	3.3
4	7	30	1002.5	54.3	25	2.3
4	7	31	1000.9	56	25.3	2.3
4	8	1	1003.8	84.3	22.2	2
4	8	2	1006.7	71	24.7	2.8
4	8	3	1007.3	69.3	23.1	3.8
4	8	4	1007.5	60.7	23.6	5
4	8	5	1008.4	64.7	23.8	3.6
4	8	6	1006.2	69.3	24.4	3.7
4	8	7	1006.5	68	24.9	3
4	8	8	1008	71.3	24.5	4.3
4	8	9	1004.8	67	25.1	5.5
4	8	10	999.4	65	23.9	4.2
4	8	11		72.7	23.1	2
4	8	12	1004	79.3	23.6	5.4
4	8	13	1004.9	77.3	23.8	3.8
4	8	14	1004.2	71	24.7	3.7
4	8	15		75.3	24.6	5.1
4	8	16	1005.2	76.3	25.2	2.3
4	8	17	1005.1	78.3	25.3	1.7
4	8	18	1005.8	75	23.7	3.7
4	8	19	1006.5	74	24.3	4.3
4	8	20	1005.5	67.3	24.7	4.1
4	8	21	1005.1	64.3	24.3	7.1
4	8	22	1003.3	76	22.9	2.8
4	8	23	1003.1	71	23.8	3.1
4	8	24	1003.9	71.3	23.9	3.5
4	8	25	1003.6	73.7	24.5	3.6
4	8	26	1003	71.3	24.7	2.6
4	8	27	1003.7	70.3	25.1	2.6
4	8	28	1002.4	62.7	25.4	1.6
4	8	29	1002.4	68.3	25.4	4.3

4	8	30	1004.1	76.3	23.8	4.5
5	6	1	999.9	73.7	20.4	2.3
5	6	2	996.5	87.3	17.2	1.6
5	6	3	1001.4	77.7	17.8	1.7
5	6	4	1002.5	75.7	19.6	1.6
5	6	5	1004.7	86	18.3	1.3
5	6	6	1006.3	85	18.5	0.9
5	6	7	1005	74.7	19.5	1.9
5	6	8	1002	84.3	20	1.8
5	6	9	1004.1	80	21	2
5	6	10	1004.2	78	20.1	2.2
5	6	11	999.8	71.3	20.5	3.7
5	6	12	997.6	89.7	20.2	1.3
5	6	13	998.7	88.3	21.2	1.3
5	6	14	1001.4	91	20.3	1.1
5	6	15	1002.5	88.3	20.7	1.5
5	6	16	1003.7	76	23.3	1.8
5	6	17	1005.3	81.7	23.2	1.7
5	6	18	1006.7	80.7	23.9	1.8
5	6	19		77.7	23.5	2
5	6	20	1008.2	73.3	23.7	2.5
5	6	21	1007.5	72.3	22.8	3.3
5	6	22	1005.1	70.7	22.6	2.1
5	6	23	1001.9	78.3	22.4	2.2
5	6	24	1000	90	22.7	2.1
5	6	25	1000.7	91.7	23.5	2.6
5	6	26	1003.6	78.3	24.4	2.9
5	6	27	1003.5	79.7	24.7	2
5	6	28	1001.8	91	24	1.5
5	6	29	1000.3	83.7	24.6	2.4
5	6	30	995.8	85.3	22.7	1.6
5	7	1	1002.5	82.7	19.5	2.7
5	7	2	1007.4	64.7	21.5	1.8
5	7	3	1005.9	65.3	21.1	1.5
5	7	4	1004.3	70.7	22.1	1.5
5	7	5	1003.5	80	22.8	3.1
5	7	6		83.3	23.5	2.7
5	7	7	1005.5	86.7	24.3	1.5
5	7	8	1007.2	86.3	24.8	1.8
5	7	9	1005.1	88.7	24.9	2.4
5	7	10	1000.1	86.7	25	1.9
5	7	11	999.7	82.7	24.7	2.7
5	7	12	1002.7	80	25	1.8
5	7	13	1003.6	81.7	24.1	1.3
5	7	14	1003.5	77	23.7	1.3
5	7	15	1003.3	77	24	2.5

5	7	16	1004.5	69.3	24	3.8
5	7	17	1006.1	71	24.3	3.5
5	7	18	1006.7	70	23.8	3.5
5	7	19	1006.1	72.7	24.1	2.4
5	7	20	1002.4	78.3	24.3	2.6
5	7	21	1003.8	75.7	23.6	2.4
5	7	22	1004.4	54	23.5	2.7
5	7	23	1003.1	68.3	23	1.8
5	7	24	1003.8	66.3	24.2	1.5
5	7	25	1003.6	52	24.5	2.1
5	7	26	1003.3	73	24.2	1.8
5	7	27	1003.8	72.3	23.9	1.6
5	7	28	1007.7	69.7	23.8	1.8
5	7	29	1006.8	73.3	23.7	2.2
5	7	30	1001.1	60.3	24.9	2.2
5	7	31	1000.5	80.7	23.9	1.1
5	8	1	1003.6	83.3	23.4	1.8
5	8	2	1006	79	24.6	2.1
5	8	3	1006	71	23.9	1.7
5	8	4	1006.7	67	24.5	2.4
5	8	5	1007.2	65.7	24.8	3.1
5	8	6	1005.3	70	25.4	2.9
5	8	7	1005.7	69	25.8	2.6
5	8	8	1006.9	77	25.5	2
5	8	9	1004	80	25.6	3
5	8	10	999	73.7	25.7	2.9
5	8	11	999.3	73.3	26.1	2.5
5	8	12	1003.7	83.7	25	1.2
5	8	13	1004	85.3	24.9	1.6
5	8	14	1003.2	86	25.9	2
5	8	15	1004.5	78.3	26.5	2.5
5	8	16	1004.6	81.7	26.5	3.2
5	8	17	1004.9	74	25.5	3.3
5	8	18	1005.2	76.3	24.7	1.5
5	8	19	1005.4	82.7	25.3	1.9
5	8	20	1004.7	78.7	25.8	2.9
5	8	21	1004.6	64	25.9	3.9
5	8	22	1002.8	65.7	25.7	3.4
5	8	23	1002.6	71.7	25.9	2.1
5	8	24	1003.1	80	25	1.9
5	8	25	1002.5	80.7	24.6	2.2
5	8	26	1002.6	83.3	24.7	1.8
5	8	27	1002.6	85.7	24.1	1.9
5	8	28	1001.9	87.3	23.8	1.7
5	8	29	1002.2	83	24.5	1.8
5	8	30	1003.9	75	23.6	2.4

CIRRUCULUM VITAE

Sattar Nabee Rasool was born in 1981 at Erbil-Iraq. His Completed bachelor's degree in Computer Science at College of Science from Salahaddin University in 2004 to 2005, He has been working at General Directorate of Education, He has been a Master Student at Firat University Faculty of Technology Software Engineering Department.

