Machine Learning Approach to Forecast Average Weather Temperature of Bangladesh

Ashfaq Ali Shafin[1]

[1] Stamford University Bangladesh

Received: 14 December 2018 Accepted: 2 January 2019 Published: 15 January 2019

Abstract

Weather prediction is gaining popularity very rapidly in the current era of Artificial Intelligence and Technologies. It is essential to predict the temperature of the weather for some time. In this research paper, we tried to find out the pattern of the average temperature of Bangladesh per year as well as the average temperature per season. We used different machine learning algorithms to predict the future temperature of the Bangladesh region. In the experiment, we used machine learning algorithms, such as Linear Regression, Polynomial Regression, Isotonic Regression, and Support Vector Regressor. Isotonic Regression algorithm predicts the training dataset most accurately, but Polynomial Regressor and Support Vector Regressor predicts the future average temperature most accurately.

Index terms— machine learning, linear regression, isotonic regression, support vector regressor, polynomial regression, temperature prediction.

1 I. Introduction

rediction for the future using the correct algorithm is viral nowadays. This prediction is applicable for the weather prediction as well. We can use machine learning to know whether it will rain tomorrow or what will be the temperature tomorrow. Machine learning algorithms can correctly forecast weather features like humidity, temperature, outlook, and airflow speed and direction. This sector is immensely dependent on previous data and artificial intelligence. Predicting future weather also helps us to make decisions in agriculture, sports and many aspects of our lives.

We aimed to predict the average temperature of Bangladesh in this research paper. As a subtropical country, Bangladesh has very different weather from other countries due to periodic disparities of rainfall, sophisticated temperatures, and humidity. Mainly three distinct seasons are present in Bangladesh, and those are Summer, Rainy, and Winter [1]. The summer season consists from March to June, while Rainy season lasts June to October and the Winter is from October to March. Even though Bangladesh is known as the sixseasoned country, mainly three seasons can be observed in this current time.

The dataset used in this paper contains the average temperature from the year 1901 to 2018 on a once-a-month basis. We calculated the sum of the values of temperature of the twelve months and then divided by 12 to get the average temperature of that particular year. Then we used different machine learning algorithms to extrapolate our findings and the generalize the output result.

After the modeling, also known as training or fitting in machine learning, we have forecasted the average temperature for Bangladesh in upcoming days using the machine learning prediction. Future weather forecast can use the predicted result.

2 II.

3 Literature Review

Mizanur et al. used a model, produced for predicting mean temperature that adjusted with groundbased watched information in Bangladesh during the time of 1979-2006. For the comprehension of the model execution, they have utilized the Climate Research Unit (CRU) information. Better implementation of MRI-AGCM got through approval procedure expanded trust in using it later temperature projection for Bangladesh [2].

An assessment of air temperature and precipitation conduct is significant for momentary arranging and the forecast of future atmospheric conditions. Patterns in precipitation and temperature at yearly, regular and month to month time scales for the times of 1981-2008 have been dissected utilizing BMD information and MPI-ESM-LR (CMIP5) model information. Likewise, the outcomes thus structure a decent premise of future examinations on temperature changeability. Thinking about all seasons (winter, prestorm, rainstorm and post-storm), most extreme temperature has expanded altogether in all seasons except winter which is immaterial over the entire investigation zone for BMD information however for MPI-ESM-LR (CMIP5) model information highest temperature is on increment in the area. Heat over the whole area expanded by 0.29?celcius and 5.3?celcisus every century individually for BMD information and MPI-ESM-LR (CMIP5) model information [3].

Holmstrom et al. recommended a method to determine the highest and lowest temperature of the subsequent seven days, given the data of the past couple days [7]. They employed a linear regression model and a variation of a functional linear regression model. Expert weather forecasting services for the prediction outperformed the two models. As a classification problem, Radhika et al. used support vector machines for climate forecast [8]. Krasnopolsky For seasonal average temperature, the following table (**??**) is used to calculate the average temperature. We have added the average temperature for those months respectively and then divided it by 4 for the seasonal average temperature. and Rabinivitz offer a crossbreed model that employed neural networks to model weather forecasting [9]. A predictive model based on data mining was presented in [10] to establish fluctuating weather patterns

4 III. Methodology a) Dataset

We collected the dataset from the website www.kaggle.com/yakinrubaiat/bangladeshweather-dataset. This dataset contains the monthly average value of Bangladesh temperature and rain from 1901 to 2015. Then we manually, added the data from the year 2016 to 2018 from the Bangladesh Meteorological Department which is the official weather forecasting department of Bangladesh government.

5 b) Pre-processing

For yearly average temperature, we have added all the monthly average temperature of a particular year and then divided it by 12 to get the annual average temperature. A mathematical equation presented for the average temperature of a year in equation (1). Table (2) describes the overview of the yearly and seasonal average temperature data statistics is. Standard Deviation of the annual average temperature is 0.42 while the standard deviation for the summer, rainy, and winter is 0.41, 0.29, and 0.56 respectively.

6 d) Estimator Selection

In this paper, we have used several machine learning algorithms described in table (3) to train our data and to predict future average temperature.

7 Estimator

Parameter Linear Regression Default Isotonic Regression Default Polynomial Regression Degree = 2 and 3 Non-linear Support Vector Regressor (SVR) Polynomial regression is a structure of regression analysis in which the connection between the independent variable ?? and the dependent variable ?? displayed as n-th degree polynomial in x. Polynomial relapse fits a nonlinear relationship between the worth of x and the corresponding conditional mean of y. In this paper, we have used polynomial regression of 2nd-degree, and 3rd-degree equations represented as follows: Degree = 3?? = ?? 0 + ?? 1 ?? + ?? 2 ?? 2 + ??**(4)**?? = ?? 0 + ?? 1 ?? + ?? 2 ?? 2 + ?? 3 ?? 3 + ??(?? = ?? 0 + ?? 1 ?? + ?? (2)

Here, ?? 0 and ?? 1 are the weight vectors, and ?? is the error term.

Isotonic regression is the method of fitting a freestyle line to a succession of perceptions under the accompanying requirements: the provided freestyle line needs to be non-decreasing all over, and it needs to lie as near the opinion as would be prudent.

The isotonic regression optimization is expressed by:??????????????? ? ?? ?? ?? (?? ?? ? ?????????? ??) 2**(3)**

Here ?? ?? is actual output, ?????????? ?? is a prediction, and ?? ?? are strictly positive weights (default to 1.0).

Linear regression is a direct technique of demonstrating the connection between a scalar reaction, also known as the dependent variable and one or more explanatory variables or independent factors. The instance of one logical variable is called univariate linear regression. For more than one explanatory variable, the procedure is called multiple linear regression [4]. This term is unmistakable from multivariate direct relapse, where numerous associated ward factors are anticipated, as opposed to a single scalar variable. [5] For one variable feature ??, year in our case and target value ??, the average output temperature the linear regression equation is:SD* indicates Standard Deviation.

For Support Vector Regressor (SVR), the model delivered by help vector arrangement depends just on a subset of the preparation information because the cost capacity for structure the model does not think about preparing focuses that lie past the edge. Comparably, the model created by SVR depends just on a subset of the preparation information, because the cost capacity for structure the model overlooks any preparation information near the model forecast. The equation for non-linear SVR represented as:?? = ? ?? ?? ?? ?? + ?? ?? ??=0 (6)

Here, ?? is the weight vector, ?? is the input vector of years, ?? is being outputted vector of average temperature and b is the bias term, and n is the degree of the equation. In our case, we have used n=3 for the experiment. Here, ?? 0 , ?? 1 , ?? 2 , ?? 3 are weight vectors, ?? is the independent input variable of year, ?? is the output variable average temperature, and e is the error term. From figure (11), (**??**2), (**??**3) we can observe that, a. For the rainy season dataset, Isotonic Regression fits the dataset accurately than other estimators. b. Both Polynomial Regression of 3 rd degree and Polynomial Regression of 2 nd degree fits quite similarly with a minimal margin line c. SVR of 3 rd degree performs better than Polynomial and Linear Regressor. We used the figures (**??**4), (15), and (**??**6) to plot different estimators result for the winter season average temperature dataset. In diagram 14, we applied Linear and Isotonic Regression for the winter season. Figure hazard work, relating to the conventional estimation of the squared mistake misfortune. The way that MSE is quite often carefully positive (and not zero) is a direct result of haphazardness or because the estimator does not represent data that could create a progressively precise estimate [6]. If a vector of n likelihoods created from a sample of n data points on all variables, and Y is the vector of experimental values of the variable being forecast, then the within-sample MSE of the predictor is computed as:

8 IV. Result Analysis

?????? = 1 ?? ? (?? ?? ? ?? ?? ?) 2 ?? ??=1**(7)**

Mean Absolute Error (MAE) is a proportion of the contrast between two consistent factors. Accept X and Y are factors of combined perceptions that express a similar of n focuses, where the i number point have organized (?? ?? , ?? ??). Mean Absolute Error (MAE) is the normal vertical separation between each position and the character line. MAE is likewise the normal flat separation between each point and the personality line. If a vector of n likelihoods created from a sample of n data points on all variables, and Y is the vector of experimental values of the variable forecast, then the within-sample MSE of the marvel. Instances of Y versus X incorporate correlations of anticipated versus watched, ensuing time versus starting time, and one method of estimation versus an elective strategy of evaluation. Consider a dissipate plot predictor computed as:= 1 ?? ? | ?? ?? ? ?? ?? ? | 2 ?? ?? =1**(8)**

The Median Absolute Error (MedAE) is especially fascinating because it is secure to exceptions. The misfortune determined by taking the middle of every apparent distinction between the objective and the forecast. If ?? ?? ? is the predicted value of then the Median Absolute Error estimated over n samples defined as:?????????? = ???????????? (| ?? 1 ? ?? 1 ? |, ? ? , | ?? ?? ? ?? ?? ??? |)**(9)**

R2_Score represents the extent of fluctuation that has been clarified by the autonomous factors in the model. It gives a sign of decency of fit and subsequently a proportion of how well inconspicuous examples are probably going to be anticipated by the model, though the extent of clarified change. R2_Score will be in the range of 0.0 to 1.0. A higher value of R2_Score means better performance of the estimator. If a vector of n likelihoods created from a sample of n data points on all variables, and Y is the vector of experimental values of the variable being forecast, then the R2_Score of the predictor is computed as: 4), (**??**), (6), and (7) represent the estimation of the estimators. We can observe that for training data, Isotonic Regression outperforms all the other estimators. As MSE, MAE, and MedAE are the lower and R2_Score of higher value means better performance for the estimator. We can notice that the boldfaced benefits of Isotonic Regressor perform better than other estimators. R2_Score for the yearly average temperature dataset is 0.835687, which is the highest value among all the datasets and estimators.??2 ?????????? = 1 ? ? (?? ?? ? ?? ?? ?) 2 ?? ??=1 ? ??? ?? ? ?? ? ? 2 ?? ??=1 (10) ????????? ?? ? = 1 ?? ? ?? ?? ?? ?? =1

Polynomial Regressor of degree 3 and Support Vector Regressor of 3 rd degree performs quite similar. All the estimator's values are the same as the other. Both Regressors perform better than Linear Regression and second-degree Polynomial Regression.

After training the dataset, we experimented with predicting the future yearly average temperature and seasonal average temperature. Table (8), (9), (10), and (11) denote the future average temperature for Bangladesh from 2019 to 2040. Extrapolated from the tables that, Isotonic Regression prediction is a constant value. As Isotonic Regressor cannot predict future projection for the average temperature, we should not rely on this estimator. The forecast for SVR or Polynomial of degree 3, can be considered as the future average temperature values. We can conclude our paper by extrapolating that, even though Isotonic Regression has performed better on the training dataset, for testing data, it performs very poorly. So, we cannot recommend this estimator for the prediction of upcoming annual or seasonal temperature for Bangladesh. We recommend using Polynomial Regression or SVR of higher degrees to predict the temperature for upcoming years.

Average temperature let alone will not be very useful for the weather forecast. That is why in the future, we want to forecast weather attributes like outlook prediction, rain prediction, and rainfall amount for the imminent future. Maximum temperature and minimum temperature prediction will also be sufficient for weather estimation.

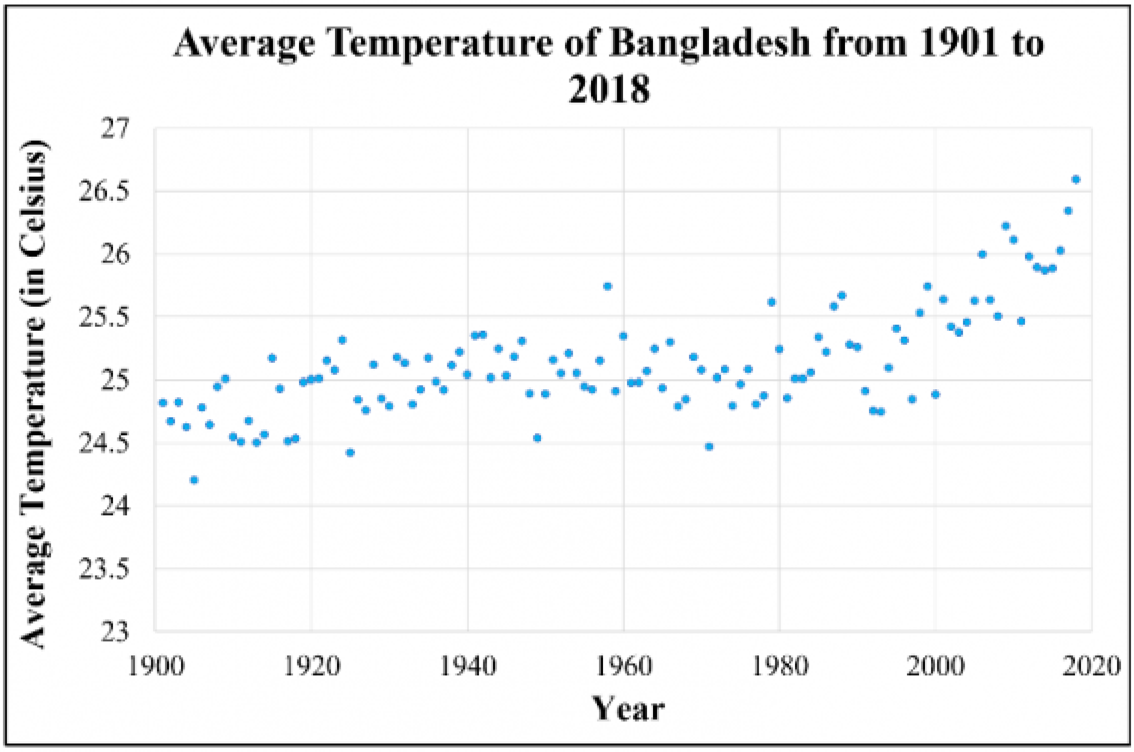


Figure 1:

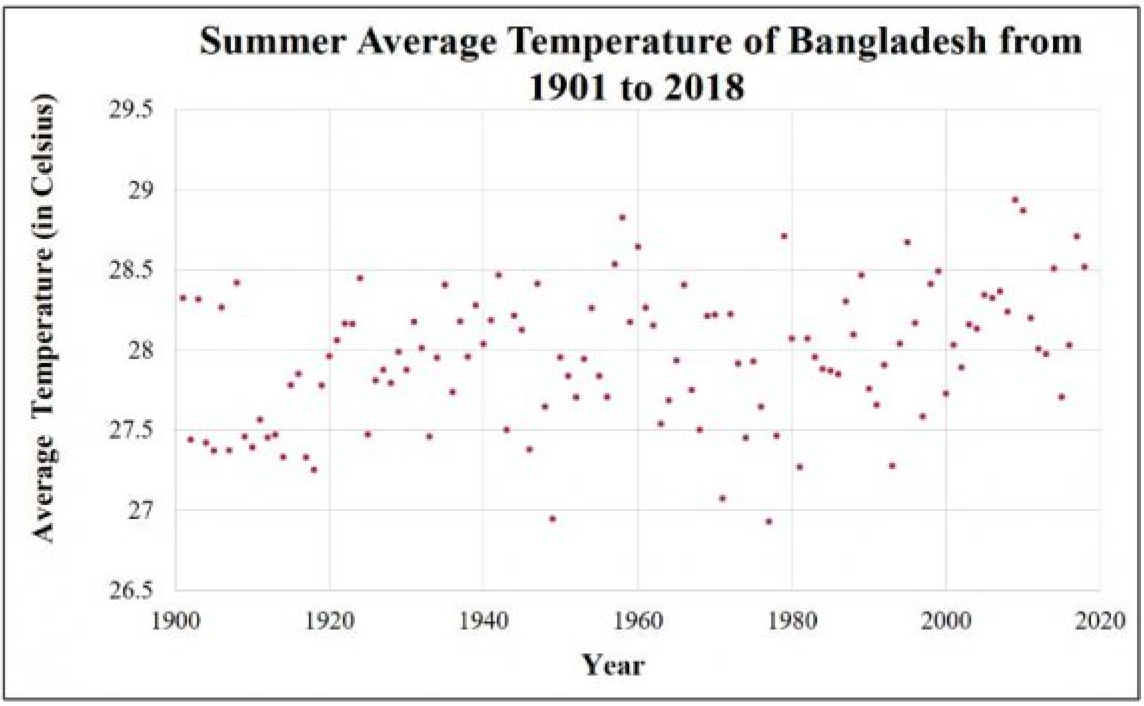
1 

Figure 2: Figure 1 :

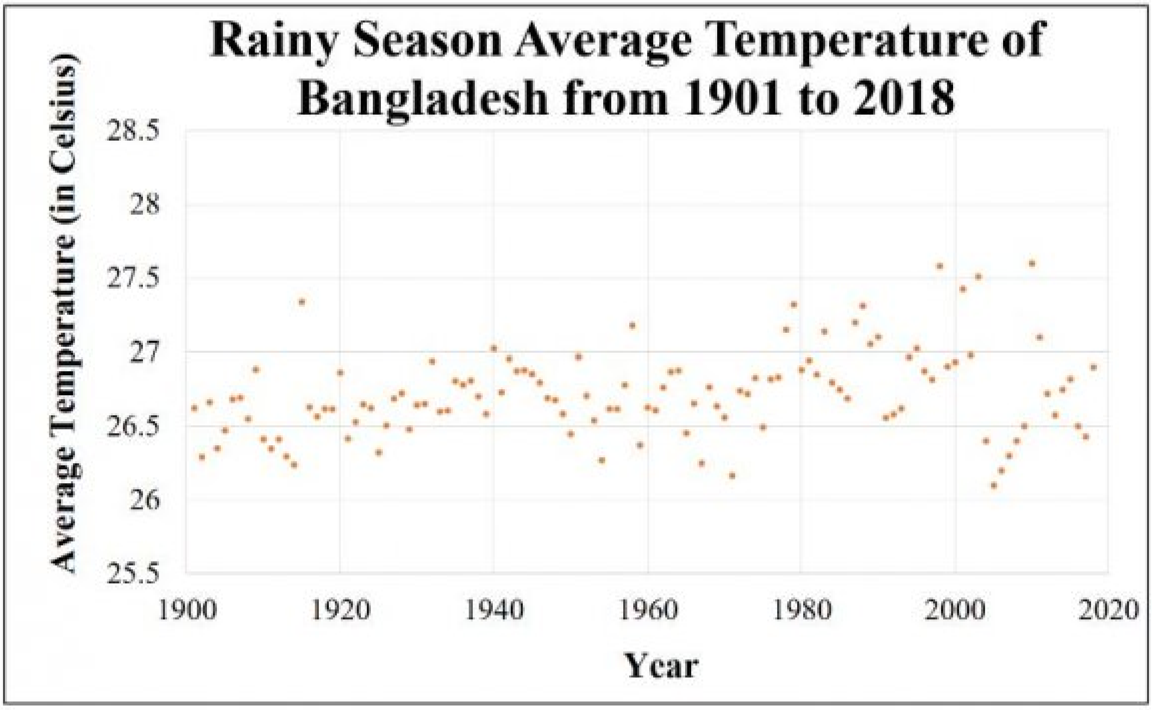
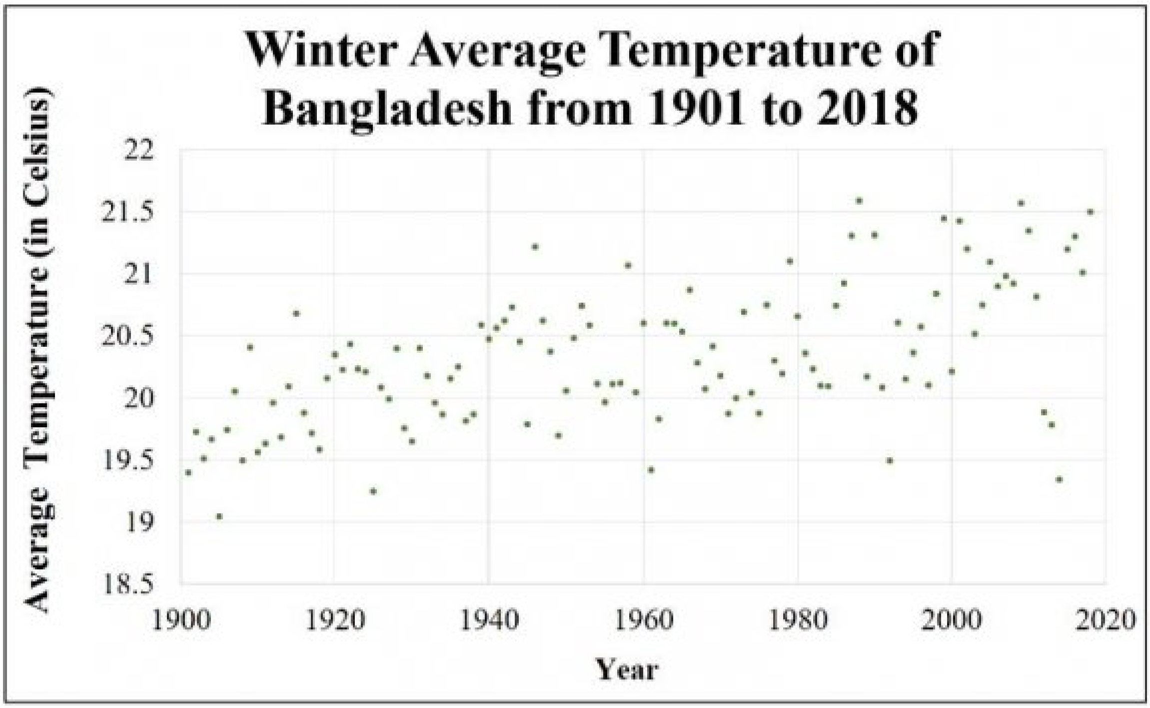
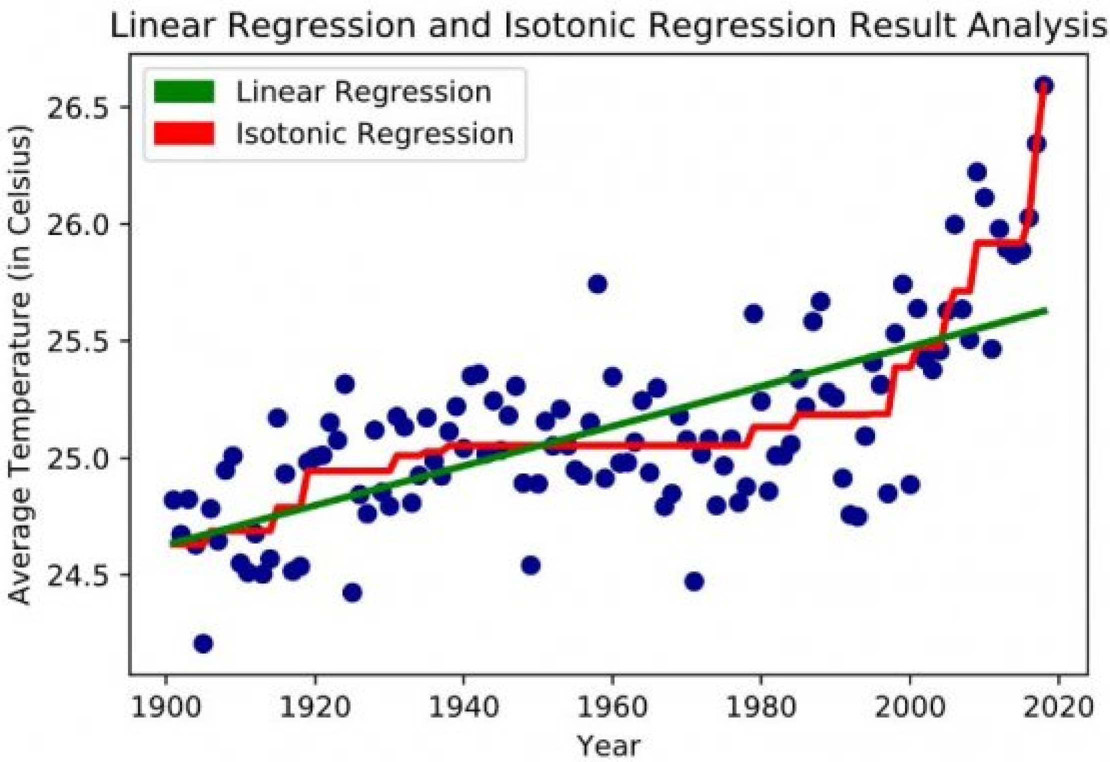
2 

Figure 3: Figure 2 :



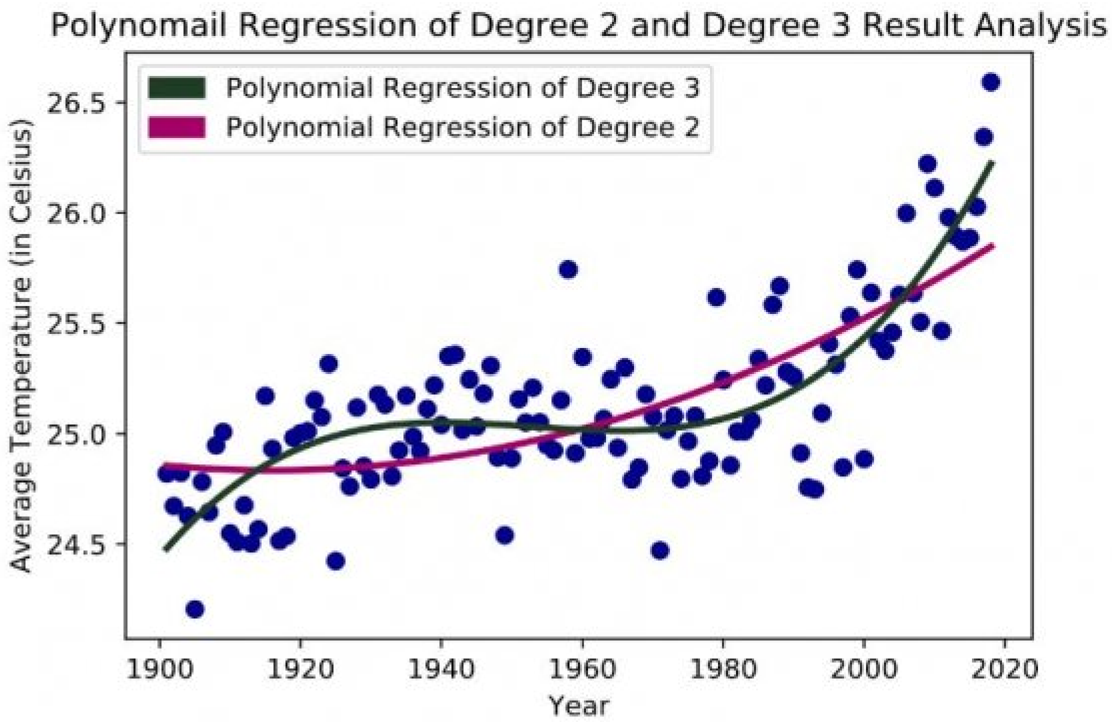
3

Figure 4: Figure 3 :



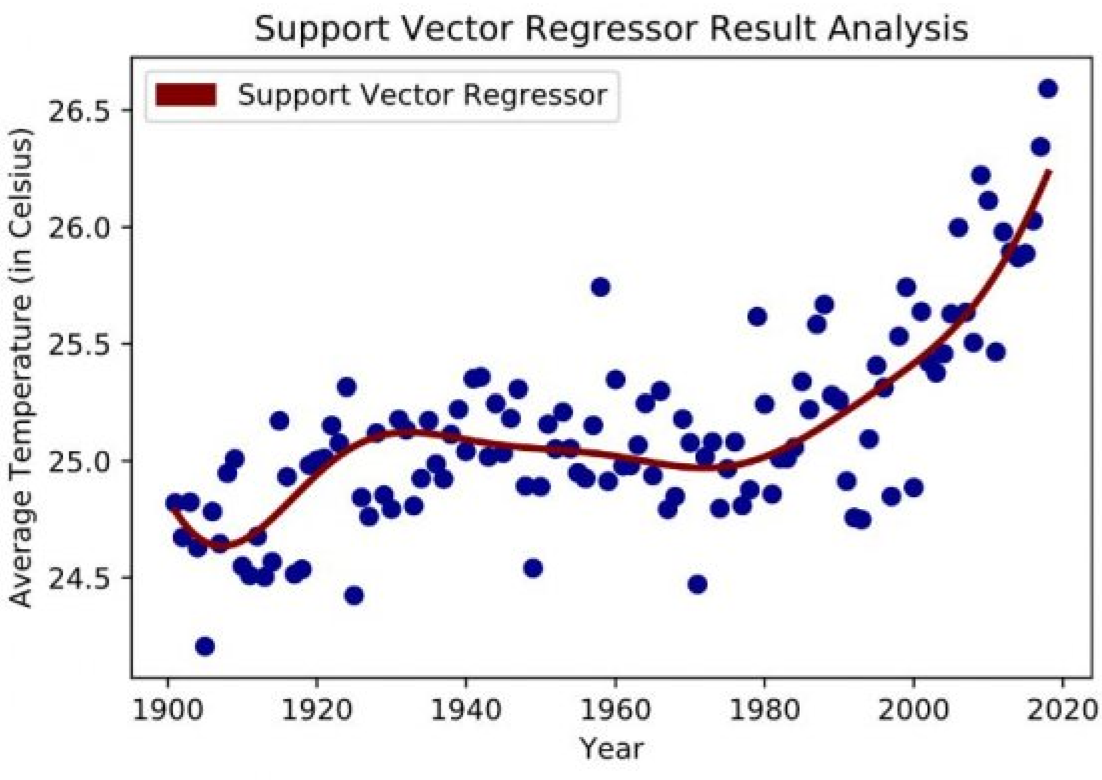
14

Figure 5: 1)Figure 4 :



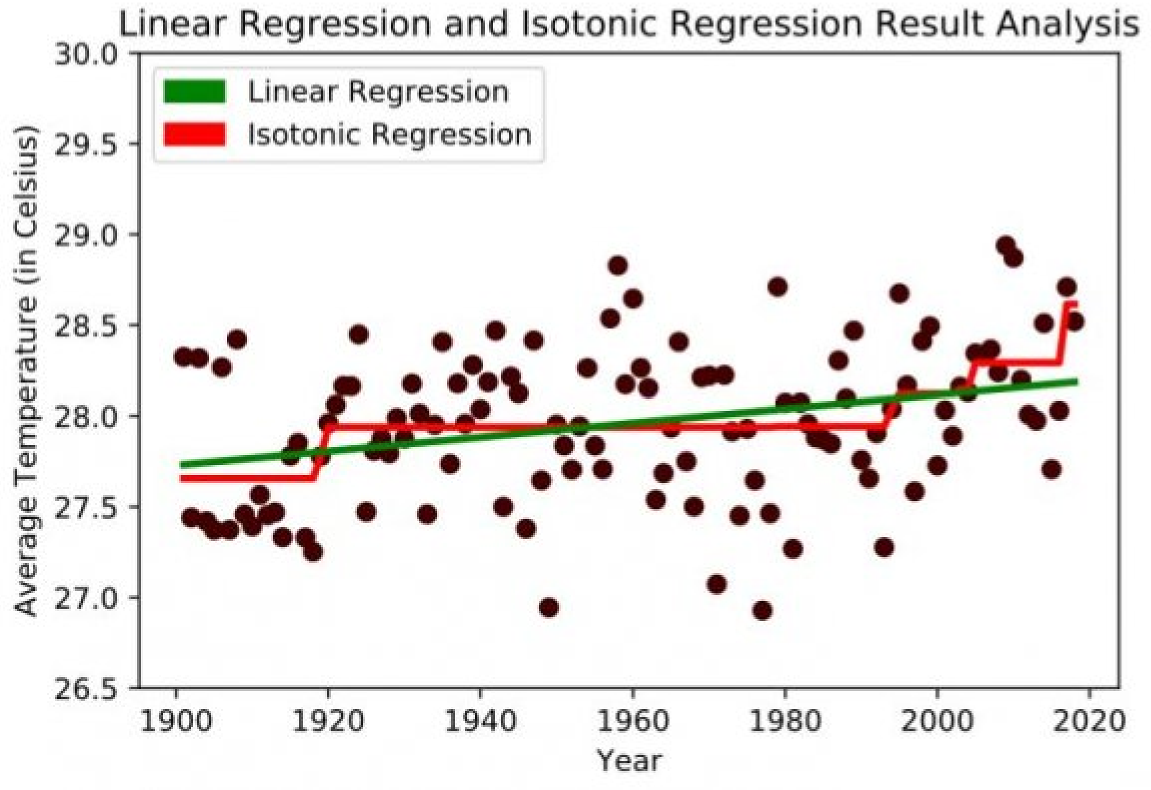
5

Figure 6: 5)



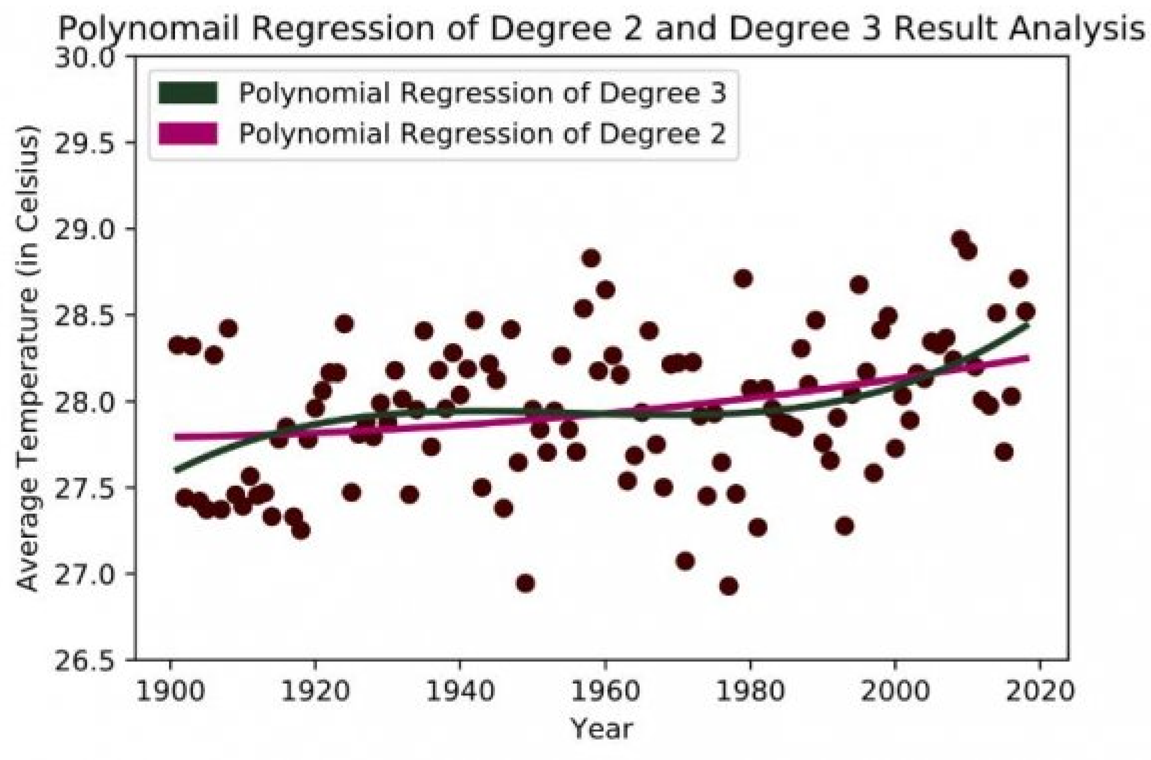
5

Figure 7: Figure (5)



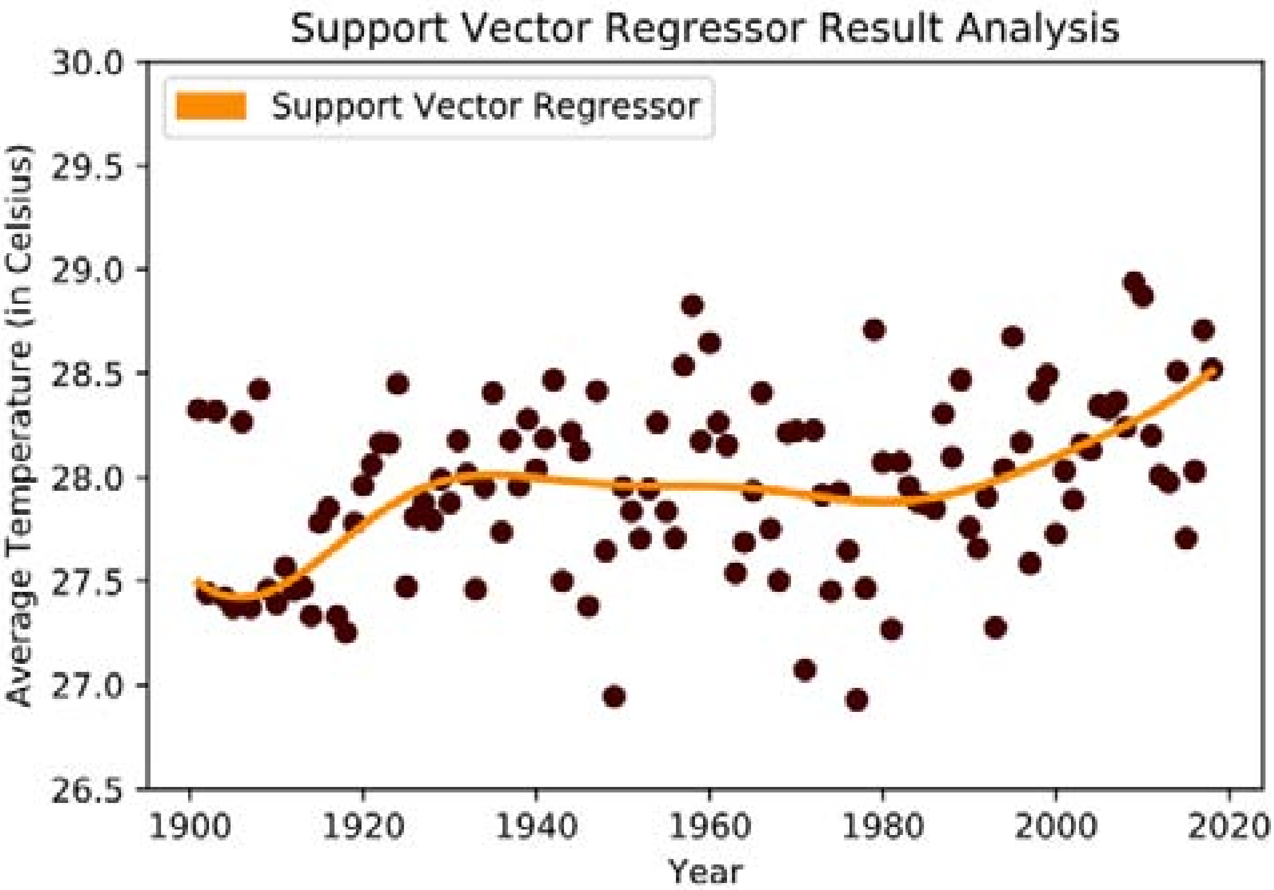
8

Figure 8: Figure (8)



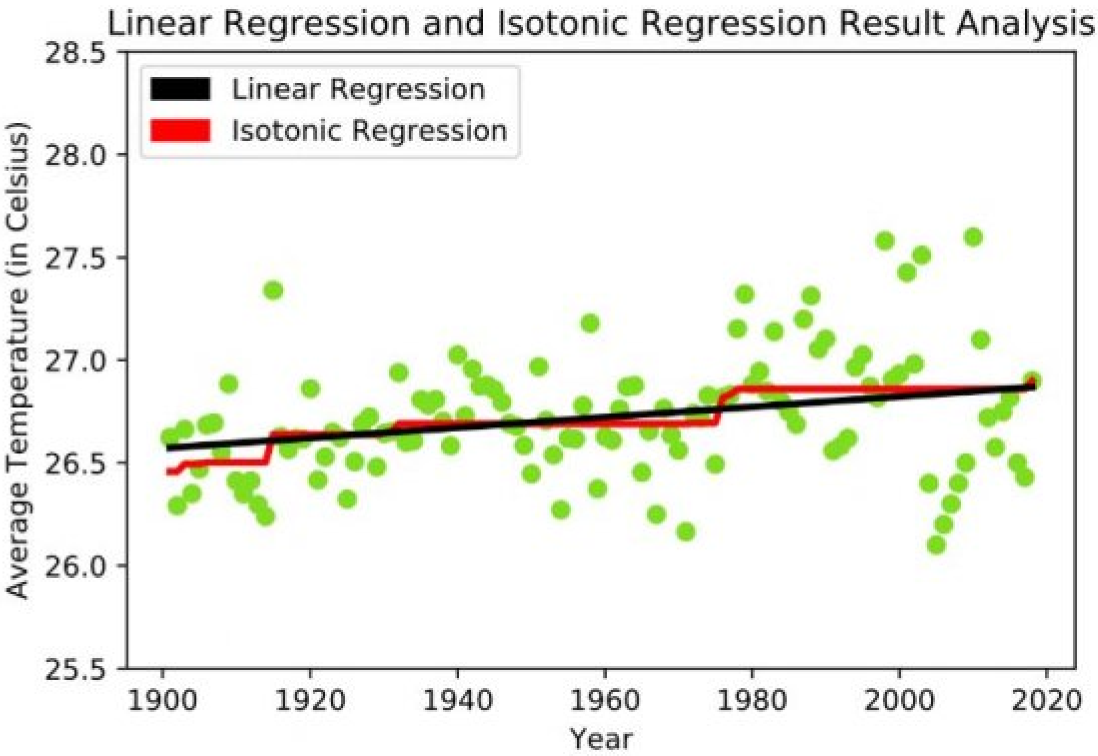
9

Figure 9: Figure (9)



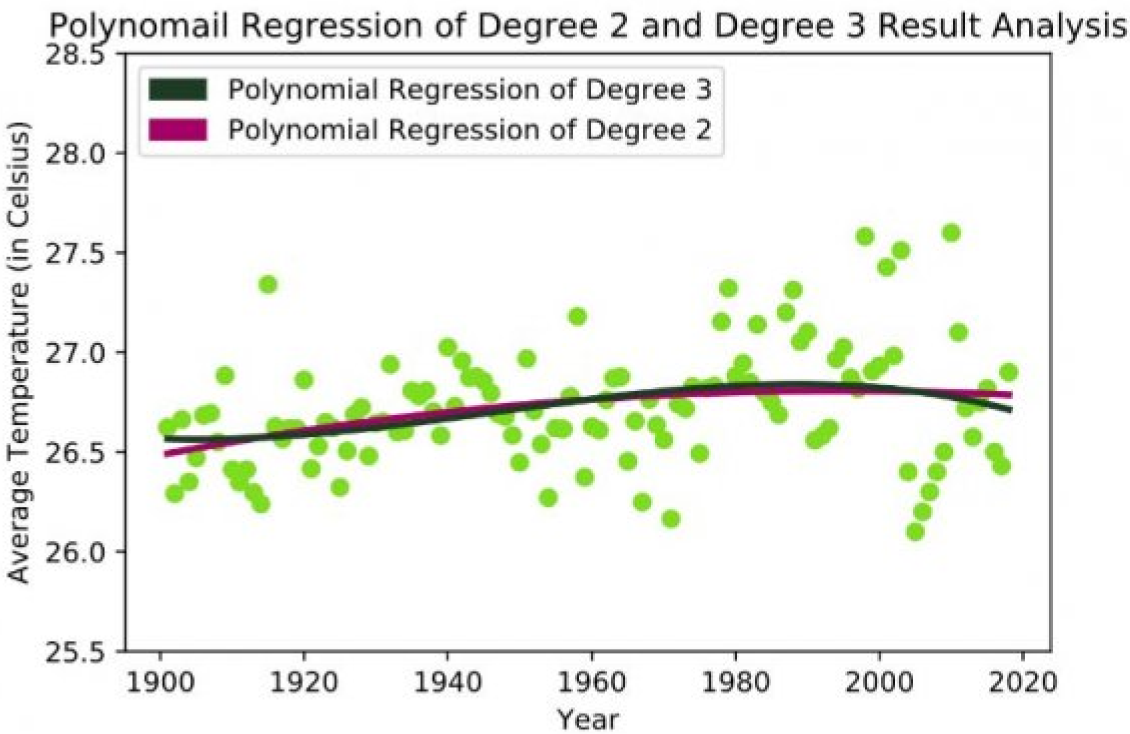
567

Figure 10: Figure 5 :Figure 6 :Figure 7 :



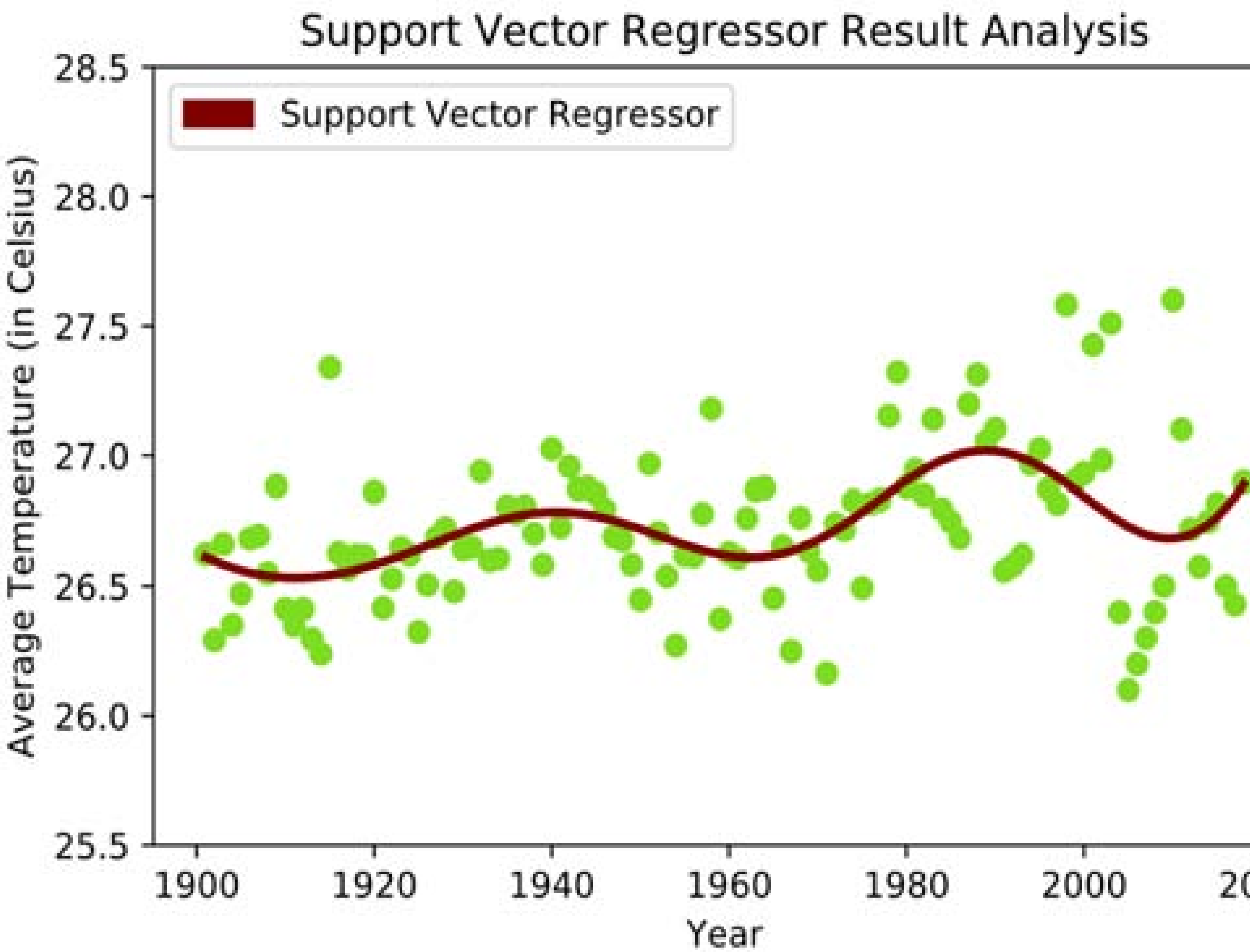
11

Figure 11: Figure 11 ,



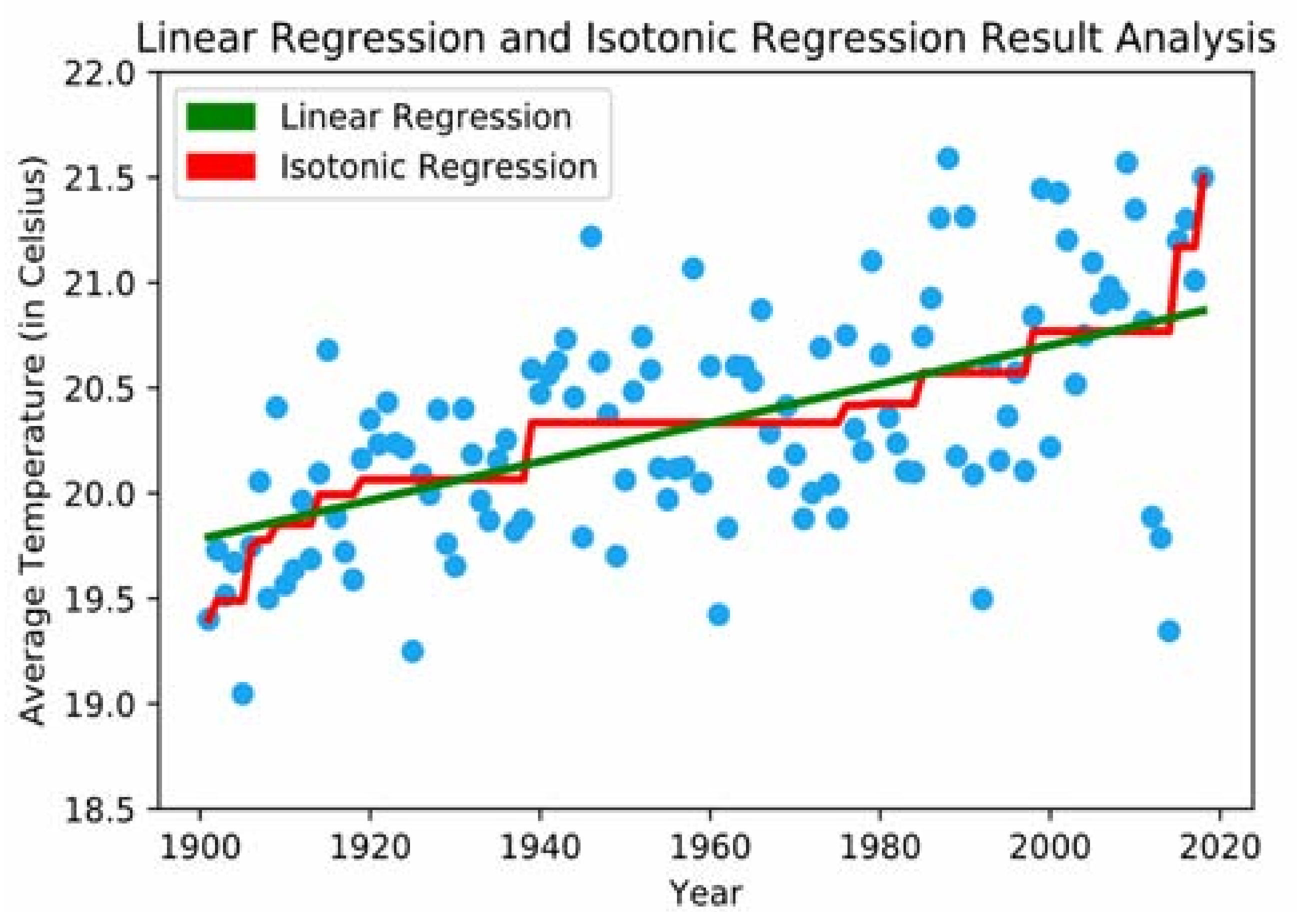
8

Figure 12: Figure 8 :



9101112

Figure 13: Figure 9 :Figure 10 :Figure 11 :Figure 12 :



141516

Figure 14: Figure 14 :Figure 15 :Figure 16 :

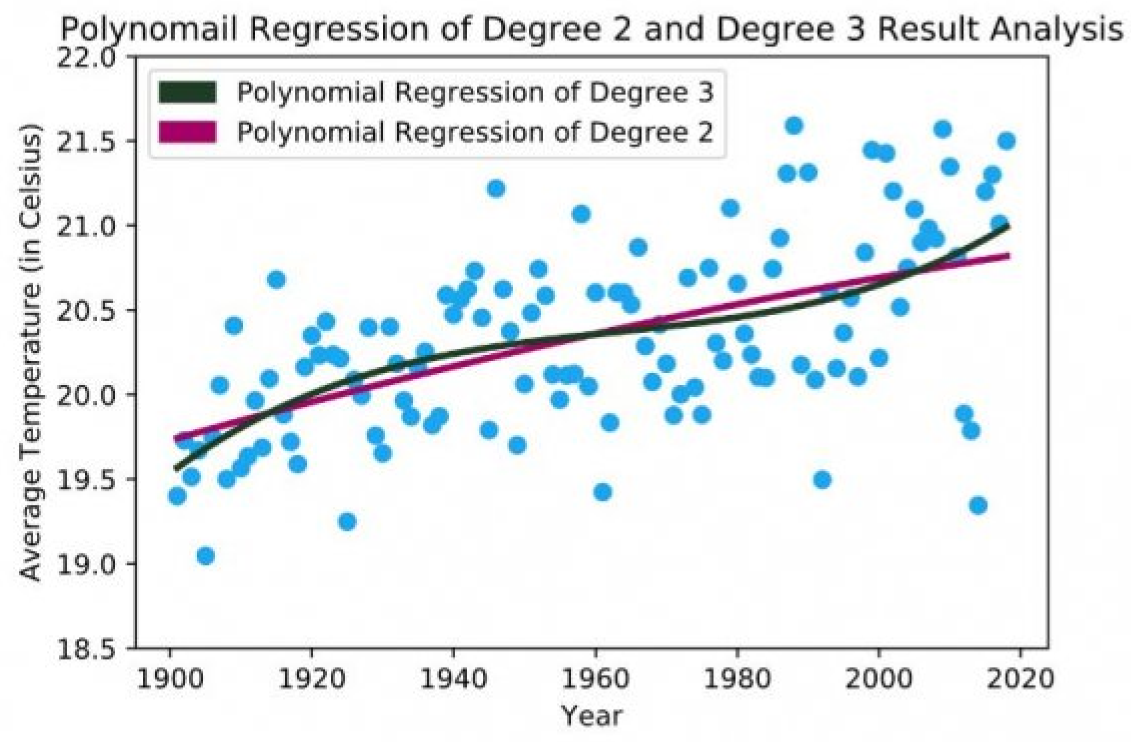


Figure 15:

1

Season	Months
Summer	March, April, May, June,
Rainy	July, August, September, October
Winter	November, December, January, February

Figure 16: Table 1 :

2

Attributes	Year	Yearly Average Temperature (in Celsius)	Summer Season Average Temperature (in Celsius)	Rainy Season Average Temperature (in Celsius)	Winter Season Average Temperature (in Celsius)
Count	118	118	118	118	118
Mean	1959.5	25.13	27.96	26.72	20.33
SD*	34.2077	0.42	0.41	0.29	0.56
Minimum	1901	24.21	26.93	26.10	19.05
25%	1930.25	24.86	27.69	26.55	19.96
50%	1959.5	25.06	27.96	26.69	20.24
75%	1988.75	25.31	28.24	26.87	20.67
100%	2018	26.59	28.94	27.60	21.59

Figure 17: Table 2 :

3

Figure 18: Table 3 :

4

	Mean Squared Error	Mean Absolute Error	Median Absolute Error	R2_score
Linear	0.093863	0.242148	0.205785	0.539885
Isotonic	0.050341	0.172628	0.139777	0.835687
Polynomial 2	0.083724	0.230284	0.19991	0.617148
Polynomial 3	0.061494	0.197245	0.159327	0.722697
SVR	0.061455	0.189903	0.140335	0.712918

Figure 19: Table 4 :

5

	Mean Squared Error	Mean Absolute Error	Median Absolute Error	R2_score
Linear	0.155614	0.318922	0.297744	0.3500776
Isotonic	0.107546	0.228602	0.236929	0.695183
Polynomial 2	0.154835	0.29976	0.272486	0.555278
Polynomial 3	0.14919	0.27711	0.257635	0.657899
SVR	0.151835	0.27976	0.262486	0.645278

Figure 20: Table 5 :

6

Machine Learning Approach to Forecast Average Weather Temperature of Bangladesh

??????

the k-th sample and ?? ?? is the corresponding actual value,

Year 2 019
45
Volume XIX Issue III Version I
( ) D
Global Journal of Computer Science and Technology

	Mean Squared Error	Mean Absolute Error	Median Absolute Error	R2_score
Linear	0.077955	0.206872	0.146128	0.3800776
Isotonic	0.030075	0.134821	0.11163	0.715183
Polynomial 2	0.066564	0.174726	0.139304	0.535278
Polynomial 3	0.0557	0.153632	0.124492	0.687899
SVR	0.056564	0.151726	0.121304	0.695278

Figure 21: Table 6 :

7

	Mean Squared Error	Mean Absolute Error	Median Absolute Error	R2_score
Linear	0.206298	0.370612	0.305485	0.472405
Isotonic	0.114431	0.265708	0.266803	0.697074
Polynomial 2	0.155757	0.307127	0.31012	0.594085
Polynomial 3	0.140985	0.305768	0.277769	0.609855
SVR	0.145712	0.307321	0.31021	0.614085

Figure 22: Table 7 :

8

Year	Linear (in Celsius)	Isotonic (in Celsius)	Polynomial 2 (in Celsius)	Polynomial 3 (in Celsius)	SVR (in Celsius)
2019					

Figure 23: Table 8 :

(

Figure 24: Table (

11

Figure 25: Table 11 :

[Krasnopolsky and Fox-Rabinovitz ()] , V Krasnopolsky , M Fox-Rabinovitz . 2006.

[Weather Online ()] , Weather Online . `https://www.weatheronline.co.uk/reports/climate/Bangladesh.html` 2019.

[Bhuyan et al. ()] 'A Trend Analysis of Temperature and Rainfall to Predict Climate Change for Northwestern Region of Bangladesh'. M Bhuyan , M Islam , M Bhuiyan . 10.4236/ajcc.2018.72009. *American Journal of Climate Change* 2018. (02) p. .

[Kumar and Khatri ()] 'A Weather Forecasting Model using the Data Mining Technique'. R Kumar , R Khatri . *International Journal of Computer Applications* 2016. (14) p. 139.

[Radhika and Shashi ()] 'Atmospheric Temperature Prediction using Support Vector Machines'. Y Radhika , M Shashi . 10.7763/ijcte.2009.v1.9. *International Journal of Computer Theory and Engineering* 2009. p. .

[Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather 'Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction'. 10.1016/j.neunet.2006.01.002. *Neural Networks* 19 (2) p. .

[Freedman ()] David A Freedman . *Statistical Models: Theory and Practice*, 2009. Cambridge University Press. p. 26.

[Holmstrom et al. ()] M Holmstrom , D Liu , C Vo . *Machine Learning Applied to Weather Forecasting*, 2016.

[Lehmann and Casella ()] E L Lehmann , George Casella . *Theory of Point Estimation*, (New York) 1998. Springer. (2nd ed.)

[Rencher and Christensen ()] 'Multivariate regression -Section 10.1, Introduction, Methods of Multivariate Analysis'. Alvin C Rencher , William F Christensen . *Wiley Series in Probability and Statistics* 2012. John Wiley & Sons. 10 p. 19. (3rd ed.)

[Mizanur Rahman et al. ()] 'Rainfall and temperature scenario for Bangladesh using 20 km mesh AGCM'. M Mizanur Rahman , N Ferdousi , Y Sato , S Kusunoki , A Kitoh . 10.1108/17568691211200227. *International Journal of Climate Change Strategies and Management* 2012. 4 (1) p. .

[References Références Referencias] *References Références Referencias*,