WEATHER ANALYSIS REPORT USING MACHINE LEARNING

A PROJECT REPORT

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of

KAMLESH GUPTA
Project Carried Out At

Ardent Computech Pvt Ltd (An ISO 9001:2015 Certified)

CF-137, Sector - 1, Salt Lake City, Kolkata - 700 064

Submitted By

CHAYANIKA BHOWMICK

ORIJITA ADHIKARY

NILADRI MAITY

ANAMITRA CHOWDHURY

ANINDITA NARAYANI HALDER

FUTURE INSTITUTE OF TECHNOLOGY
BORAL, GARIA-700154, KOLKATA,WEST BENGAL,INDIA
JUNE – JULY 2018

ACKNOWLEDGEMENT

Success of any project depends largely on the encouragement and guidelines of many others. We take this sincere opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project work.

We would like to show our greatest appreciation to Mr. Kamlesh Gupta, Project Manager at Ardent, Kolkata. We always feel motivated and encouraged every time by his valuable advice and constant inspiration; without his encouragement and guidance this project would not have materialized.

Words are inadequate in offering our thanks to the other trainees, project assistants and other members at Ardent Computech Pvt. Ltd. for their encouragement and cooperation in carrying out this project work. The guidance and support received from all the members and who are contributing to this project, was vital for the success of this project.

INDEX

INTRODUCTION

India has a typical weather conditions consisting of various seasons and geographical conditions.Country has extreme high temperatures at Rajasthan desert, cold climate at Himalayas and heavy rainfall at Chirapunji. These extreme variations in temperatures make us to feel difficult in inferring / predictions of weather effectively. It requires higher scientific techniques / methods like machine learning algorithms applications for effective study and predictions of weather conditions. In this paper, we applied 3 different algorithms for tallying the different parameters of the weather report.

Weather prediction has been a challenging problem in meteorological department since years. Even after the technological and scientific advancement, the accuracy in prediction of weather has never been sufficient. Even in current date this domain remains as a research topic in which scientists and mathematicians are working to produce a model or an algorithm that will accurately predict weather. There have been immense improvements in the sensors that are responsible for recording the data from the environment and cancel the noise present in them; Indian Journal of Science and Technology, Vol 9(38), DOI: 10.17485/ijst/2016/v9i38/101962, October 2016 ISSN (Print) : 0974-6846 ISSN (Online) : 0974-5645 along with this new models have been proposed which include different attributes related to weather to make accurate prediction.

OBJECTIVE

To forecast weather, which is one of the greatest challenge in meteorological department. Weather prediction is necessary so as to inform people and prepare them in advance about the current and upcoming weather condition. This helps in reduction in loss of human life and loss of resources and minimizing the mitigation steps that are expected to be taken after a natural disaster occurs.

SCOPE

The project Weather Report Analysis is very useful in order to get the best weather prediction. The main purpose of this project is to provide better way to forecast the weather.

HARDWARE AND SOFTWARE REQUIREMENTS

HARDWARE REQUIREMENTS

- Computer that has a 1.6GHz or faster processor
- 1 GB (32 Bit) or 2 GB (64 Bit) RAM (Add 512 MB if running in a virtual machine)
- HDD 20 GB Hard Disk Space and Above Hardware Requirements 5400 RPM hard disk drive
- DirectX 9 capable video card running at 1024 x 768 or higher-resolution display

SOFTWARE REQUIREMENTS

- WINDOWS OS (XP/2000/200 Server/2003 Server/Vista or7)
- Internet Information Server 8.0 (IIS)
- .Net Framework 4.0
- SQL Server Express Edition
- Spyder3

LANGUAGE USED

PYTHON

Created by: Guido van Rossum

First released: 1991

Python is an interpreted high-level programming language for general-purpose programming.

Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales.

Syntax and Semantics: Python is meant to be an easily readable language. Its formatting is visually uncluttered, and it often uses English keywords where other languages use punctuation.

Indentation: Python uses whitespace indentation, rather than curly brackets or keywords, to delimit blocks.

Source code: The Python interpreter and extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

Features: Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

Python is Interactive: You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python is Object-Oriented: Python supports Object-Oriented style or technique of programming that encapsulates code within objects.

Databases: Python provides interfaces to all major commercial databases.

Scalable: Python provides a better structure and support for large programs than shell scripting.

ALGORITHMS USED

In our project we have used four different algorithms -

- Linear Regression
- SVR
- Kernel Ridge

REGRESSION

In statistical modelling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between dependent variable and one or more independent variables (or 'predictors').

More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Function: Regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile , or other location parameter of the conditional distribution of the dependent variable given the independent variables.

In all cases, a function of the independent variables called the **regression function** is to be estimated. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the prediction of the regression function using a probability distribution.

Methods: Many techniques for carrying out regression analysis have been developed.

Familiar methods such as linear regression and ordinary least squares regression are parametric , in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite dimensional.

Performance: The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process.

Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of casuality based on observational data, regression methods can give misleading results.

Uses:
(1) Determining the strength of predictors
(2) Forecasting an effect, and
(3) Trend forecasting

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, “how much additional sales income do I get for each additional $1000 spent on marketing?”

Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, “what will the price of gold be in 6 months?”

Types:

- **Simple linear regression**
  1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

- **Multiple linear regression**
  1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

- **Logistic regression**
  1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

- **Ordinal regression**
  1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

- **Multinominal regression**
  1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

LINEAR REGRESSION

In statistics, **linear regression** is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).
The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called **multiple linear regression**. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

- **Simple Linear Regression:** In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the criterion variable and is referred to as Y. The variable we are basing our predictions on is called the predictor variable and is referred to as X.

When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the predictions of Y when plotted as a function of X form a straight line.

Example: The example data in Table 1 are plotted in Figure 1. It can be seen that there is a positive relationship between X and Y. If Y is to be predicted from X, the higher the value of X, the higher is the prediction of Y.

Table 1: Example data

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

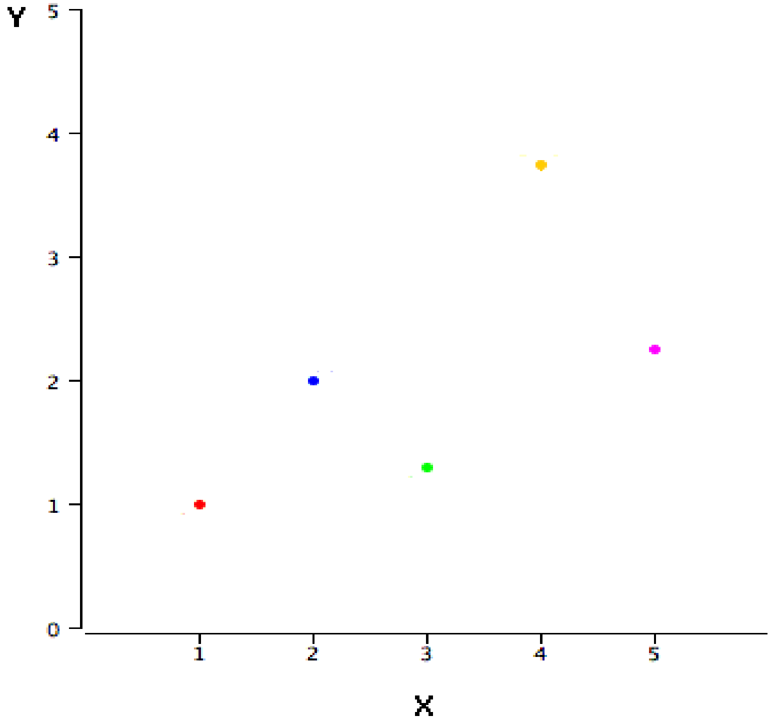


Figure 1: A scatter plot of the example data

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line.

The black diagonal line in Figure 2 is the regression line and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction. As it can be seen, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.

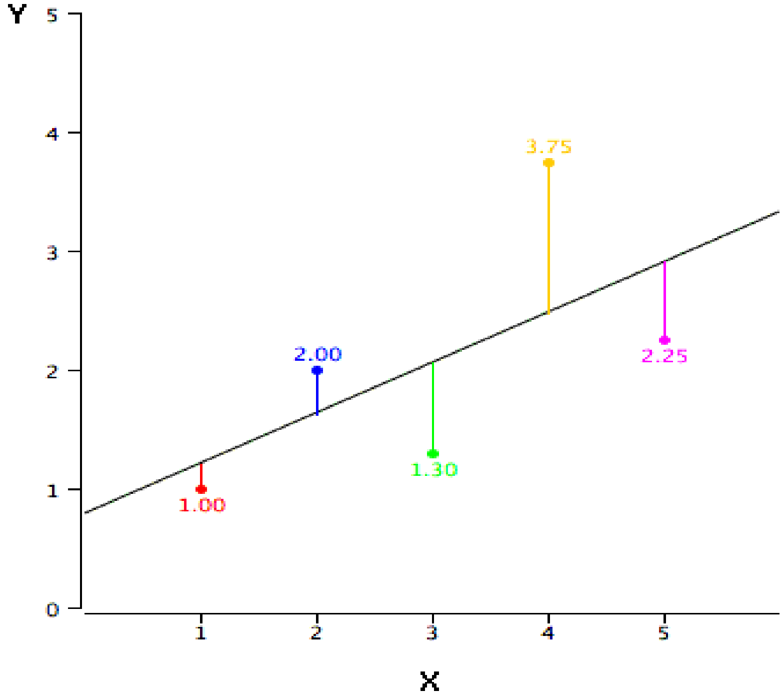


Figure 2: A scatter plot of the example data

The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

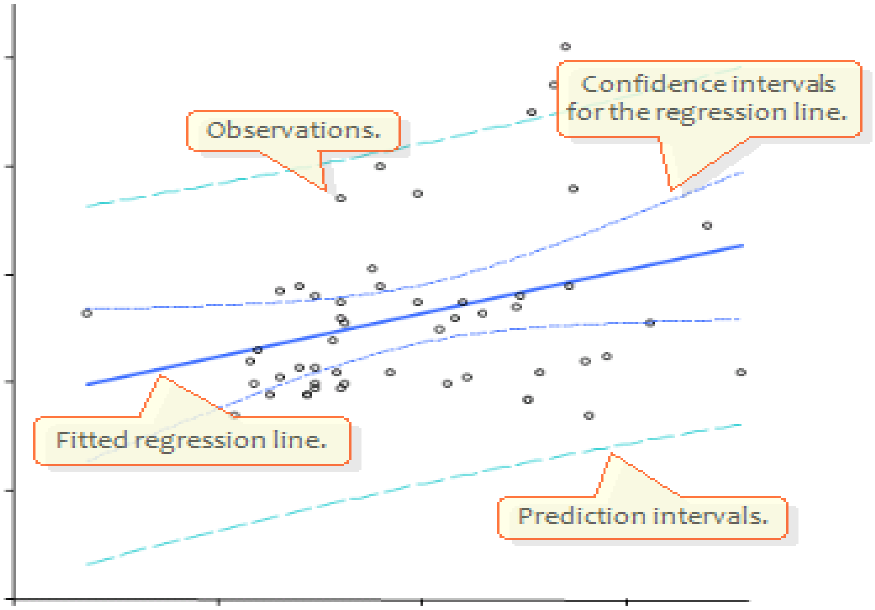
- **Multiple Linear Regression :** Multiple regression models describe how a single response variable Y depends linearly on a number of predictor variables.

  Examples:

  • The selling price of a house can depend on the desirability of the location, the number of bedrooms, the number of bathrooms, the year the house was built, the square footage of the lot and a number of other factors.

- The height of a child can depend on the height of the mother, the height of the father, nutrition, and environmental factors.

The least squares method is used to minimize the vertical distance between the response and the fitted linear line.



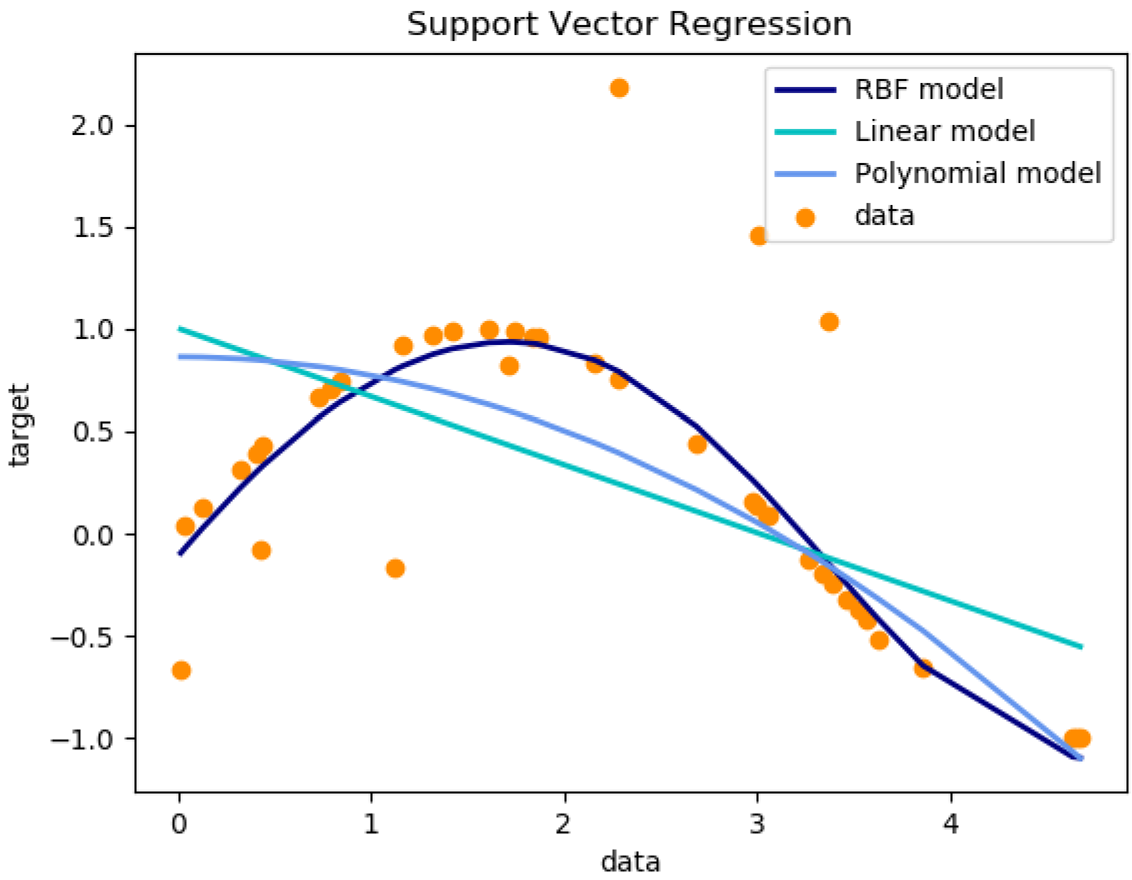
The scatter plot shows the fit, simultaneous confidence intervals and prediction intervals.

SVR (Support Vector Regression)

Support Vector Machine (SVM) which is, a supervised machine learning algorithm which can be used for both classification and regression problems. It follows a technique called the kernel trick to transform the data and based on these transformations, it finds an optimal boundary between the possible outputs.
In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (**SVR**) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem.

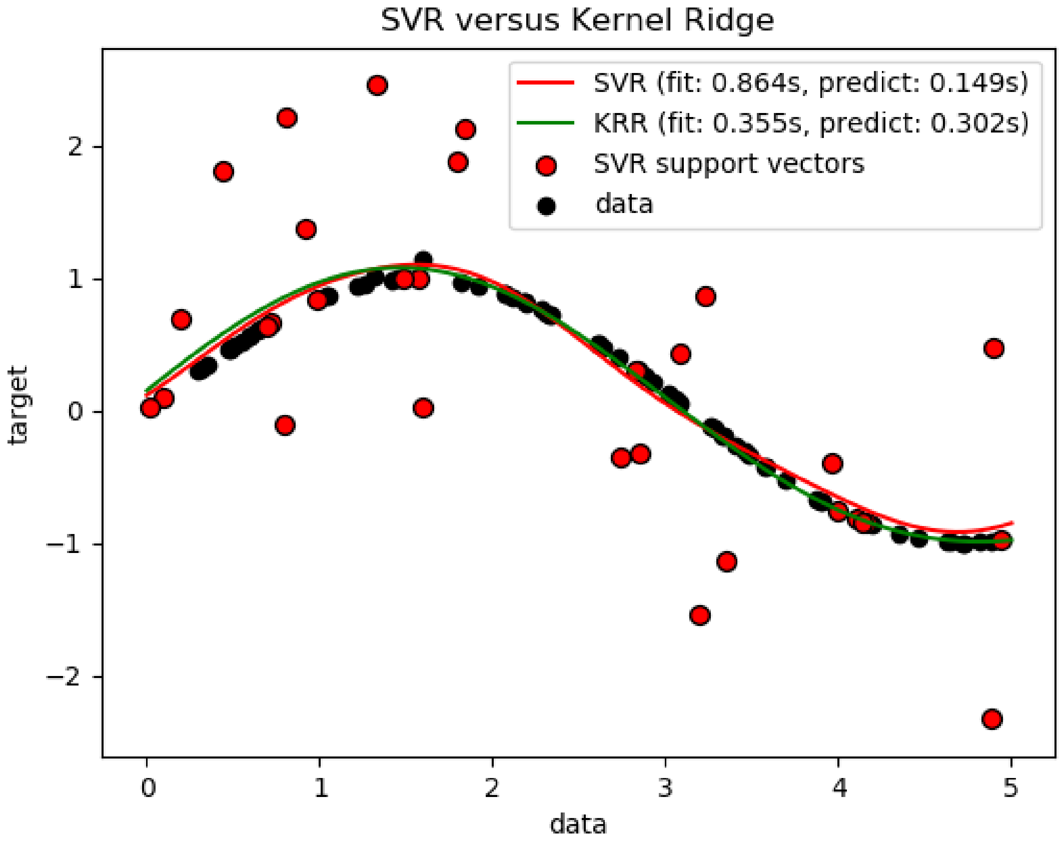
A radial basis function (**RBF**) is a real-valued function whose value depends only on the distance from the origin, so that ; or alternatively on the distance from some other point , called a center, so that . Any function that satisfies the property is a radial function. The norm is usually Euclidean distance, although other distance functions are also possible.



KERNEL RIDGE

Kernel ridge regression combines Ridge Regression with the kernel trick. It thus learns a linear function in the space induced by the respective kernel and the data. The form of the model learned by KernelRidge is identical to support vector regression (SVR). Kernel ridge regression (KRR) combines ridge regression (linear least squares with l2-norm regularization) with the kernel trick. It thus learns a linear function in the space induced by the respective kernel and the data. For non-linear kernels, this corresponds to a non-linear function in the original space.

The form of the model learned by KRR is identical to support vector regression (SVR). However, different loss functions are used: KRR uses squared error loss while support vector regression uses epsilon-insensitive loss, both combined with l2 regularization. In contrast to SVR, fitting a KRR model can be done in closed-form and is typically faster for medium-sized datasets. On the other hand, the learned model is non-sparse and thus slower than SVR, which learns a sparse model for epsilon > 0, at prediction-time. This estimator has built-in support for multi-variate regression (i.e., when y is a 2d-array of shape [n_samples, n_targets]).



LIBRARIES USED

- **NumPy** - NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. NumPy is open-source software and has many contributors. The core functionality of NumPy is its "ndarray", for n-dimensional array, data structure. These arrays are strided views on memory. In contrast to Python's built-in list data structure these arrays are homogeneously typed all elements of a single array must be of the same type.

  Example of NumPy -

  ```
  import numpy as np

  x = np.array([1, 2, 3])
  ```

  OUTPUT

  ```
  x

  array([1, 2, 3])
  ```

- **pandas** - pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Pandas, a convenient library that supports dataframes. Pandas is technically optional because Scikit-Learn can handle numerical matrices directly, but it'll make our lives easier.

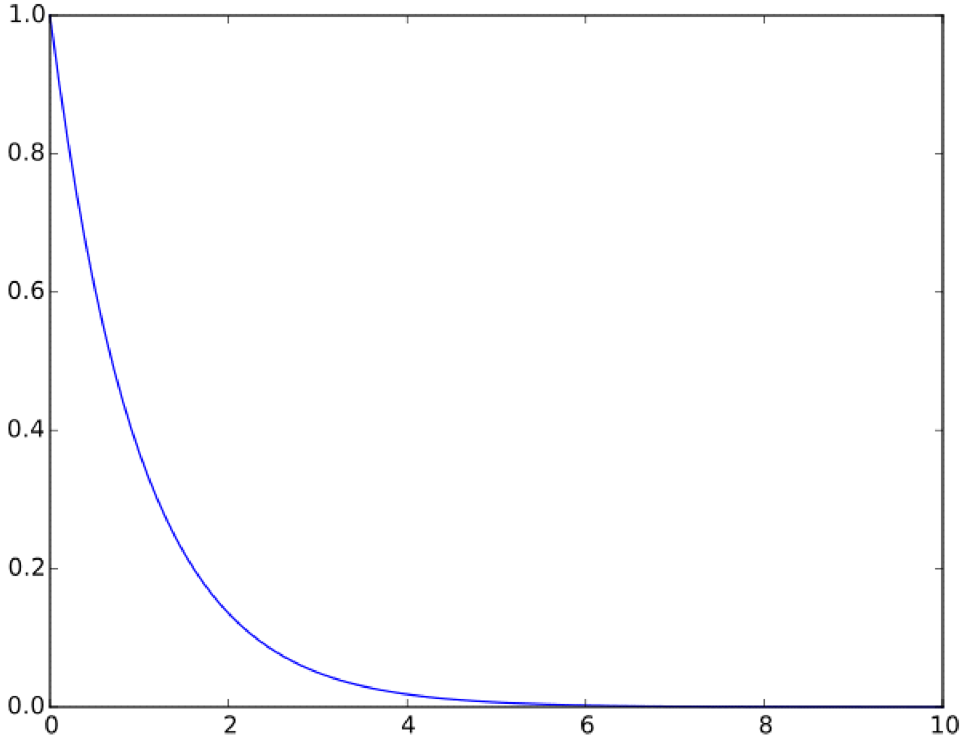
  Example-

  ```
  import pandas as pd

  DFnew=pd.DataFrame()
  ```

- **Matplotlib** - Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of matplotlib.

Example of Matlplotlib -

```
import matplotlib.pyplot as plt

import numpy as np

a = np.linspace(0, 10, 100)

b = np.exp(-a)

plt.plot(a, b)

plt.show()
```



- **SKLearn -** Scikit-learn (formerly scikits.learn) is a free software machine learning library for the Python programming language. It’s a fantastic library because it offers a high-level interface for many tasks (e.g. preprocessing data, cross-validation, etc.). This allows us to better practice the entire machine learning workflow and understand the big picture.It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

  Scikit-learn is largely written in Python, with some core algorithms written in Cython to achieve performance. Support vector machines are implemented by a Cython wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR.

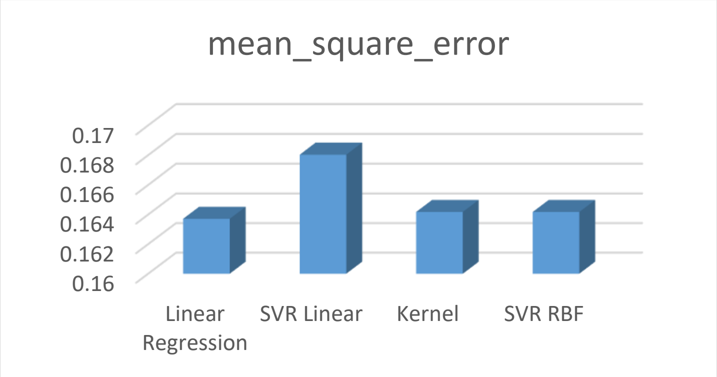
  Libraries of SKLearn used-

  - sklearn.metrics.mean_absolute_error(test_label,pred_y)
  - sklearn.metrics.median_absolute_error(test_label,pred_y)
  - sklearn.metrics.mean_squared_error(test_label,pred_y)
  - sklearn.metrics. r2_score(test_label,pred_y)

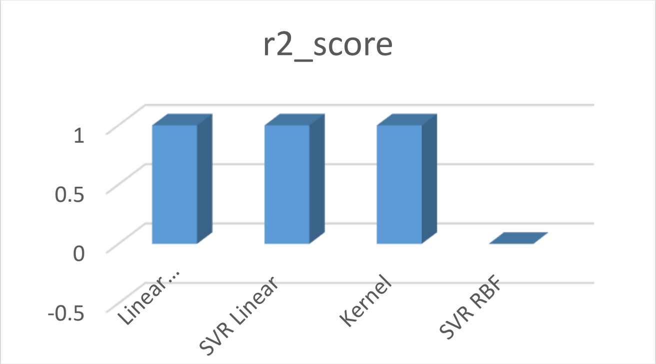
COMPARISON AMONG DIFFERENT ALGORITHMS

Studying the features of weather sets with respect to four different algorithms and finding out the best way out-

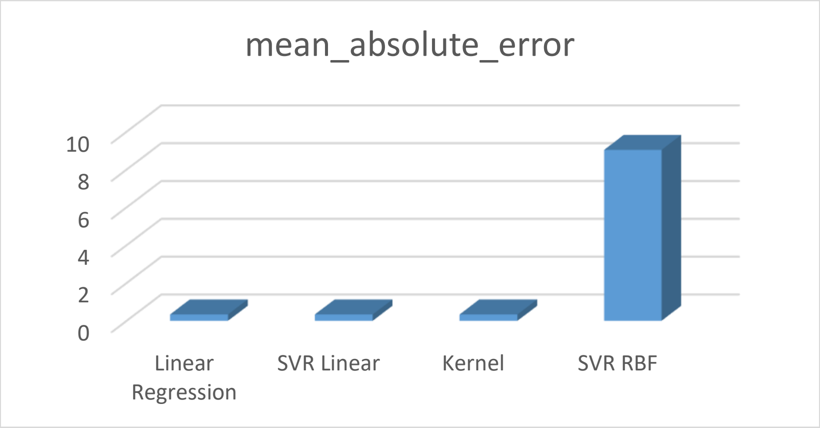
	Linear Regression	SVR Linear	Kernel	SVR RBF
mean_square_error	0.163713	0.168025	0.164174	0.164174
r2_score	0.998574	0.998537	0.998570	-0.00013
mean_absolute_error	0.335803	0.333127	0.334773	9.036589
median_absolute_error	0.279688	0.398637	0.282449	8.469625



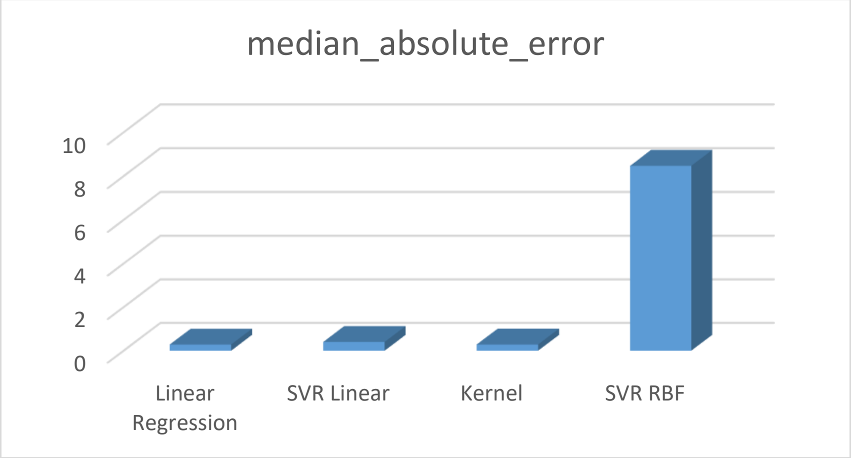
From the above graph we can observe that Linear Regression has got the least mean_squared_error.



From the above graph we can observe that Linear Regression has got the maximum r2_score.



From the above graph we can observe that SVR Linear has got the least mean_absolute_error.



From the above graph we can observe that Linear Regression has got the least median_absolute_error.

So, we can conclude that **Linear Regression** is the most relevant algorithm as it gives the best result in most of the cases.

CONCLUSION

This project has been appreciated by all the users in the organization.
Weather plays a major role in our daily life, and without the meteorologist and forecaster we would have difficulty planning our daily activities. As we can see, the weather is not a simple subject like we may have been thinking. The study of weather phenomenon requires the use of science, math, and different types of equipment and technology and data.
Even with all these equipment, data, and observation tools, the weather continues to be a topic to study because it is constantly changing. Meteorologist and forecasters predict the weather and its possible changes, but in reality, weather is still unpredictable.

BIBLIOGRAPHY

- www.kaggle.com
- http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
- http://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html
- https://sadanand-singh.github.io/posts/svmpython/

THANK YOU