WEATHER PREDICTION USING MACHINE LEARNING ALGORITHMS

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

by

PREETHAM KASARLA	2103A52185
MAHANTH.P	2103A52190
RADHA KRISHNA.J	2103A52192

Under the guidance of

Mr. D. Ramesh Kumar

Assistant Professor, Department of CSE.

Department of Computer Science and Artificial Intelligence

Department of Computer Science and Artificial Intelligence

CERTIFICATE

This is to certify that the Project entitled **"WEATHER PREDICTION USING MACHINE LEARNING ALGORITHMS"** is the bonafied work carried out by **B. PREETHAM KASARLA, MAHANTH POTHUGANTI AND RADHA KRISHNA** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ARTIFICAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2022-2023 under our guidance and Supervision.

.

Mr. D Ramesh Kumar
Asst. Professor,
SR University,
Anantha Sagar, Warangal.

Dr. M. Sheshikala
Assoc. Prof. & HOD (CSE),
SR University,
Anantha Sagar, Warangal.

ABSTRACT

Weather Project application is a web-based application where you will be able to access all reports related to weather forecasts for any locations. Its location detected by your browser setting and server configuration will automatically identify the location and be able to present its weather information such as temperature, wind direction, rainfall, humidity etc. To change location, you will have to select the options provided below to get its details. Its new avatar and feed burner will also allow its users to receive weather reports directly from their mail, where they have not been able to access this particular domain even if the server is down. People have been trying to predict the weather informally for thousands of years and since the 19th century. Weather forecasts are made by collecting information about the current state of the atmosphere in a particular area and then using the weather to predict how the atmosphere will change. Individual input is still required to select the best predictive model to establish the prediction. Predictability inaccuracies are due to the prevailing weather conditions, the high calculation power required to solve atmospheric calculations, the error involved in estimating the initial conditions, and an incomplete understanding of atmospheric processes. Therefore, the predictions are less accurate as the difference between the current time and the time the forecast is made (the range of the forecast) increases. The use of ensembles and a harmonious model helps to minimize error and select the possible outcome.

ACKNOWLEDGEMENT

We express our thanks to Course co-coordinator **Mr. D. Ramesh Kumar, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude to Head of the department CS&AI, **Dr. M. Shashikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

Table of Contents

CHAPTER 1
INTRODUCTION

The Machine Learning Model that we are working on is WEATHER PREDICTION. **The remarkable improvement in the quality of weather forecasts** is one of the great successes of environmental science in the 20th century, which continues at a sustained pace at the beginning of the 21st century (see Figure 1 and Bauer et al, 2015). This is due to the progress of numerical prediction systems and the increasing **number** and variety of **observations of** the state of the atmosphere and related media (ocean, soils, vegetation, cryosphere), including observations from Earth observation satellites. The **rapid development of supercomputers** has been one of the keys to this success, which has also required significant scientific work.
Each country in the world has a National Meteorological Service (NMS), whose mission is to make regular observations of the atmosphere and to issue forecasts for government, industry and the public. But **only the most advanced countries have Numerical Weather Prediction (NWP) centres**, whose products are also distributed to other countries, in exchange for their observations, within the framework of the World Meteorological Organization.

Among the main NWP centers outside Europe are those in the United States, Canada, Japan, Korea, China, Russia, Australia, India, Morocco, South Africa and Brazil. In Europe, only France, the United Kingdom and Germany make numerical forecasts for the entire globe, while the other countries have NWP center's covering only regional areas. The European countries have also come together in a "super-center", which is responsible for providing them with medium-range numerical forecasts.

The "**weather prediction dataset**" is a **novel tabular dataset that was specifically created for teaching machine learning and deep learning to an academic audience**. The **dataset** contains intuitively accessible **weather** observations from 18 locations in Europe. It was designed to be suitable for a large variety of different training goals, many of which are not easily giving way to unrealistically high **prediction** accuracy.

PROBLEM STATEMENT

The problem statement is to recognize the weather condition in the given picture by using extracted useful features from the image.

For some image recognition tasks or computer vision studies, intuition and illumination kinds of difficulties may cause unexpected results. It can be tricky to define weather conditions from a given scene. Even human eyes and brain can be insufficient to recognize weather patterns. That's where an image recognition model can be helpful.

Existing system

The Existing weather system **predicts the weather and the sudden change in the forecast with some delay**. In order to overcome these limitations, we need Sensors, Microcontrollers and Software Applications to make a Smart Weather Reporting System Using IoT. Advantages of Weather Reporting System using IoT.

Proposed system

Weather Forecasting is crucial since it helps to determine future climate changes. With the use of latitude, we can determine the probability of snow and hail reaching the surface. We are able to identify the thermal energy from the sun that is exposed to a region. Climatology is the scientific study of climates, which in simple words mean weather conditions over a period. A bunch of studies within atmospheric sciences also takes the help of the variables and averages of short-term and long-term weather conditions accumulated. Climatology is different from meteorology and can be divided into further areas of study. Different approaches to this segment can be taken. Currently, our primary research goal is to motivate and help the development of efficient and effective measures of Environmental activities.

Objectives***

The objective of this project is to predict weather. Seasons and nature play a major role in agriculture and farming. When it comes to the farming of various fruits, vegetables, and pulses, temperature is extremely important. Farmers didn't have a better understanding of weather forecasts before, so they had to rely on estimates to do their jobs. They do, however, sometimes suffer losses as a result of inaccurate weather forecasts. Farmers will now get all of their forecasts on their smartphones, thanks to advances in technology and the use of unique weather forecasting mechanisms. Of course, education in this area is critical, but the majority of the farmer community at this point understands the fundamentals, making it simple for them to use the features.
It helps us optimize the machine learning models and report on expected accuracy by applying few methodologies such as KNN and logistic regression.

Architecture

The architecture of this machine learning model is “SUPERVISED LEARNING” and the process involved is data acquisition, data processing, data modelling and execution (parameter tuning and making predictions). The supervised can be further broadened into classification and regression analysis based on output criteria.

CHAPTER 2
LITERATURE SURVEY

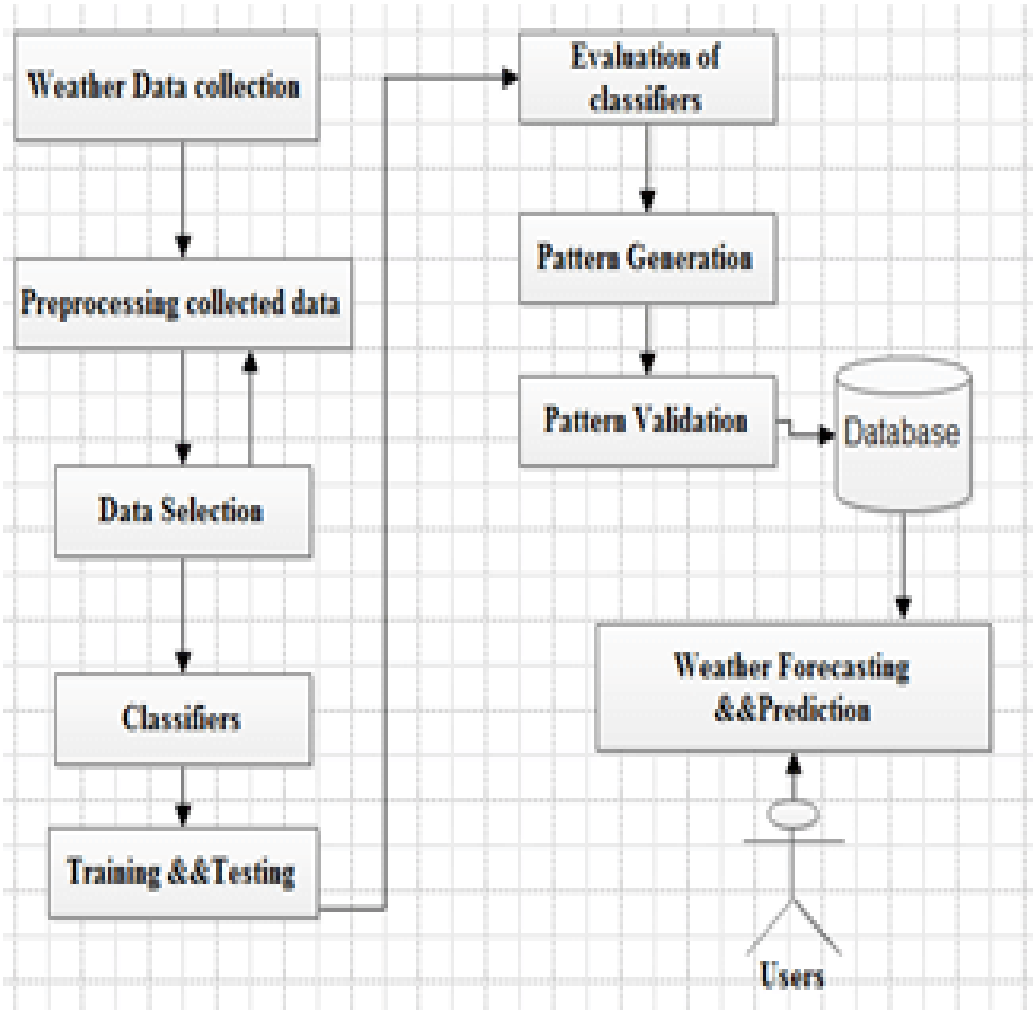
Document the survey done by you****

Numerous data-mining and machine-learning models and algorithms have been applied to security, stability prediction/monitoring, management, and control of the Weather prediction. Weather forecasting plays a fundamental role in the early warning of weather impacts on various aspects of human livelihood. For instance, weather forecasting provides decision making support for autonomous vehicles to reduce traffic accidents and congestions, which completely depend on the sensing and predicting of external environmental factors such as rainfall, air visibility and so on. Accurate and timely weather prediction has always been the goal of meteorological scientists. However, the conventional theory-driven numerical weather prediction (NWP) methods face many challenges, such as incomplete understanding of physical mechanisms, difficulties in obtaining useful knowledge from the deluge of observation data, and the requirement of powerful computing resources. With the successful application of data-driven deep learning method in various fields, such as computer vision, speech recognition, and time series prediction, it has been proven that deep learning method can effectively mine the temporal and spatial features from the patio-temporal data. Meteorological data is a typical big geospatial data.

CHAPTER 3
DATA PRE-PROCESSING

This is an analysis of weather prediction problem using different machine learning models. It predicts the weather by given data.

FLOW CHART:



DATASET:

For this project, we have obtained my dataset from Kaggle. This dataset contains 1463 rows of data and 6 columns(features)that we could focus onto build our prediction model i.e., we use 14 attributes to predict the weather.

DATA SET

1	date	precipitatio	temp_max	temp_min	wind	weather
2	01-01-2012	0	12.8	5	4.7	drizzle
3	02-01-2012	10.9	10.6	2.8	4.5	rain
4	03-01-2012	0.8	11.7	7.2	2.3	rain
5	04-01-2012	20.3	12.2	5.6	4.7	rain
6	05-01-2012	1.3	8.9	2.8	6.1	rain
7	06-01-2012	2.5	4.4	2.2	2.2	rain
8	07-01-2012	0	7.2	2.8	2.3	rain
9	08-01-2012	0	10	2.8	2	sun
10	09-01-2012	4.3	9.4	5	3.4	rain
11	10-01-2012	1	6.1	0.6	3.4	rain
12	11-01-2012	0	6.1	-1.1	5.1	sun
13	12-01-2012	0	6.1	-1.7	1.9	sun
14	13-01-2012	0	5	-2.8	1.3	sun
15	14-01-2012	4.1	4.4	0.6	5.3	snow
16	15-01-2012	5.3	1.1	-3.3	3.2	snow
17	16-01-2012	2.5	1.7	-2.8	5	snow
18	17-01-2012	8.1	3.3	0	5.6	snow
19	18-01-2012	19.8	0	-2.8	5	snow
20	19-01-2012	15.2	-1.1	-2.8	1.6	snow
21	20-01-2012	13.5	7.2	-1.1	2.3	snow
22	21-01-2012	3	8.3	3.3	8.2	rain
23	22-01-2012	6.1	6.7	2.2	4.8	rain
24	23-01-2012	0	8.3	1.1	3.6	rain
25	24-01-2012	8.6	10	2.2	5.1	rain
26	25-01-2012	8.1	8.9	4.4	5.4	rain
27	26-01-2012	4.8	8.9	1.1	4.8	rain
28	27-01-2012	0	6.7	-2.2	1.4	drizzle
29	28-01-2012	0	6.7	0.6	2.2	rain
30	29-01-2012	27.7	9.4	3.9	4.5	rain

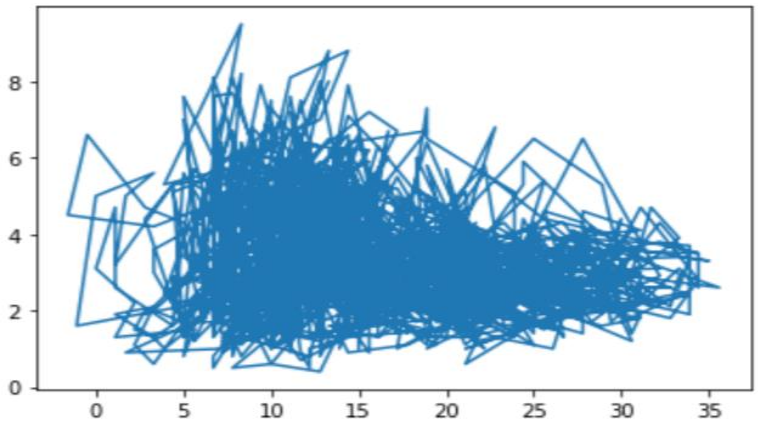
Machine Predictive Maintenance Classification Dataset

Since real predictive maintenance datasets are generally difficult to obtain and in particular difficult to publish, the data provided by the UCI repository is a synthetic dataset that reflects real predictive maintenance encountered in industry to the best of their knowledge.

Precipitation

```
plt.plot(x2,y)
```

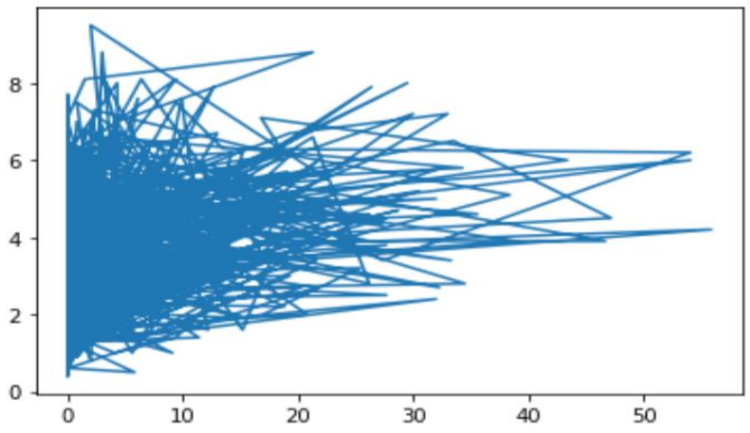
```
[<matplotlib.lines.Line2D at 0x7f90ed990220>]
```



Temp_max

```
plt.plot(x1,y)
```

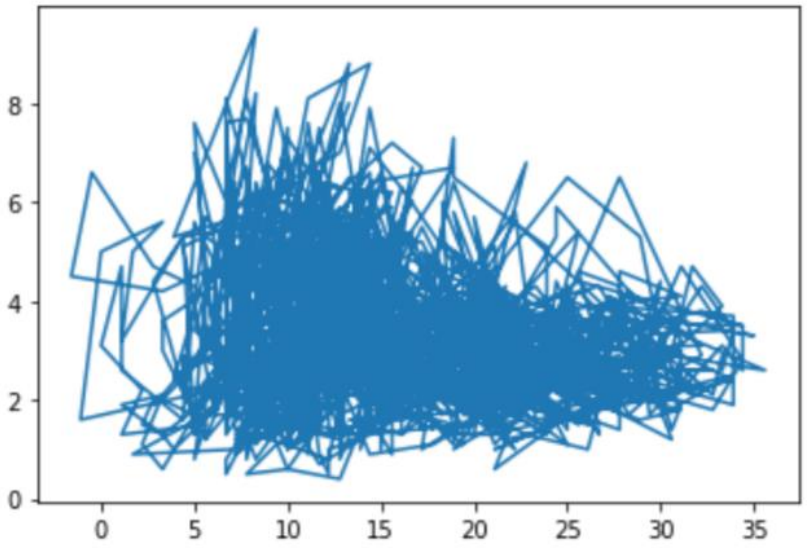
```
[<matplotlib.lines.Line2D at 0x7f90edeb9f70>]
```



Temp_min

```
plt.plot(x2,y)
```

```
[<matplotlib.lines.Line2D at 0x7f90ed990220>]
```



DATA CLEANING

DATAPREPROCESSING:

Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine model. In this particular section we re-label & convert some categorial features into numeric values. This is crucial for training machine learning models since machine learning models accepts the numeric values. We have a total of 1463 rows and 6 columns(attributes) in the dataset.

Enough methods are performed on the data to evaluate the dataset and gather knowledge about the data. Let's perform some Machine Learning model and Experimentation to create a model that helps us to achieve our goal we state in the problem definition.

CHAPTER 4
METHODOLOGY

This section talks about the algorithms used for the project. We are used algorithm logistic regression and KNN for classification.

Logistic regression:

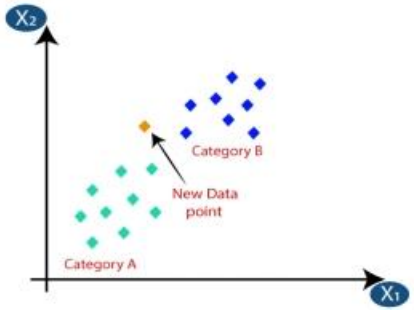
Logistic Regression estimates the probability of a dependent variable as a function of independent variables. The dependent variable is the output that we are trying to predict while the independent variables or explanatory variables are the factors that we feel could influence the output. For its simplicity and interpretability, we decide to use Logistic Regression as a Benchmark model, a basic model that represents the starting point for comparing the results obtained from other models.

MODEL ARCHITECTURE:

PROCEDURE TO SLOVE THE GIVEN PROBLEM

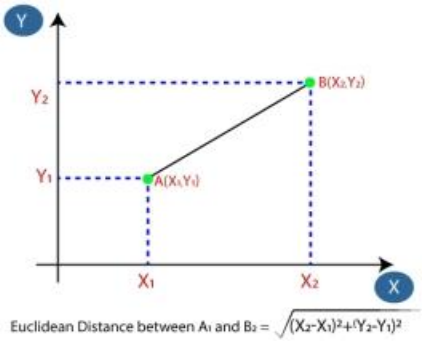
Using KNN

The K-NN working can be explained on the basis of the below algorithm:
Step-1: Select the number K of the neighbors
Step-2: Calculate the Euclidean distance of K number of neighbors
Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
Step-4: Among these k neighbors, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
Step-6: Our model is ready.
Suppose we have a new data point and we need to put it in the required category. Consider the below image:



Firstly, we will choose the number of neighbors, so we will choose the k=5.
Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry.

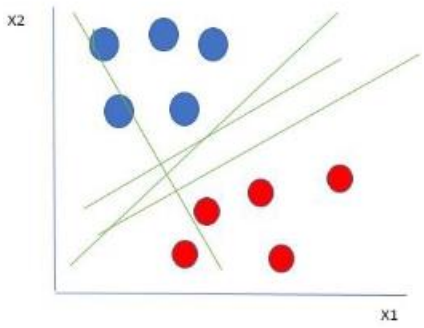
By calculating the Euclidean distance, we get the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B.



As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

Using SVM

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.



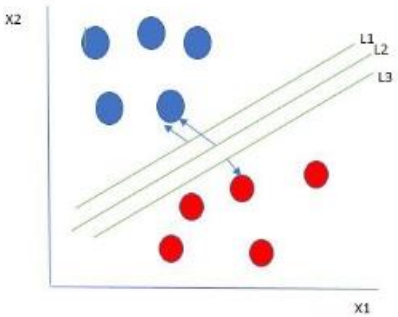
Let's consider two independent variables x1, x2 and one dependent variable which is either a blue circle or a red circle.

Linearly Separable Data points

From the figure above its very clear that there are multiple lines (our hyperplane here is a line because we are considering only two input features x1, x2) that segregates our data points or does a classification between red and blue circles. So how do we choose the best line or in general the best hyperplane that segregates our data points.

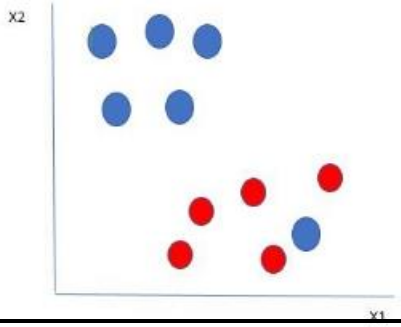
Selecting the best hyper-plane:

One reasonable choice as the best hyperplane is the one that represents the largest separation or margin between the two classes.



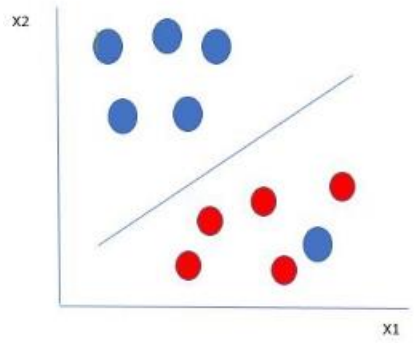
So, we choose the hyperplane whose distance from it to the nearest data point on each side is maximized. If such a hyperplane exists it is known as the maximum-margin hyperplane/hard margin. So, from the above figure, we choose L2.

Let's consider a scenario like shown below



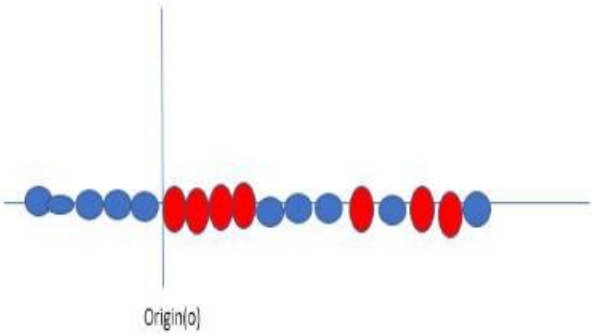
Here we have one blue ball in the boundary of red ones is an outlier of blue balls. The SVM algorithm has the characteristics to ignore the outlier and finds the best hyperplane that maximizes the margin. SVM is robust to outliers.

how does SVM classify the data? It's simple! The blue ball boundary of the red ball. So

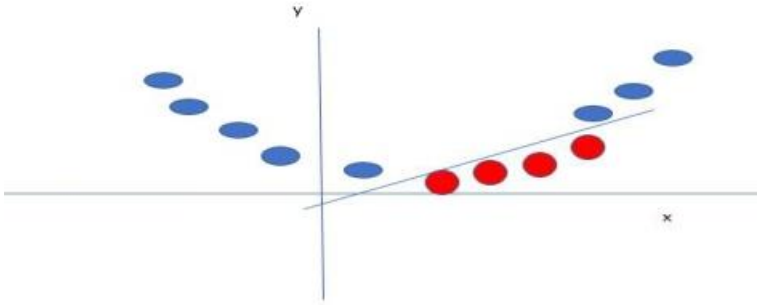


So, in this type of data points what SVM does is, it finds maximum margin as done with previous data sets along with that it adds a penalty each time a point crosses the margin. So, the margins in these types of cases are called soft margin. When there is a soft margin to the data set, the SVM tries to minimize *(1/margin+∧(∑penalty))*. Hinge loss is a commonly used penalty. If no violations no hinge loss. If violations hinge loss proportional to the distance of violation.

Till now, we were talking about linearly separable data (the group of blue balls and red balls are separable by a straight line/linear line). What to do if data are not linearly separable?



Say, our data is like shown in the figure above. SVM solves this by creating a new variable using a kernel. We call a point x_i on the line and we create a new variable y_i as a function of distance from origin o.so if we plot this, we get something like as shown below

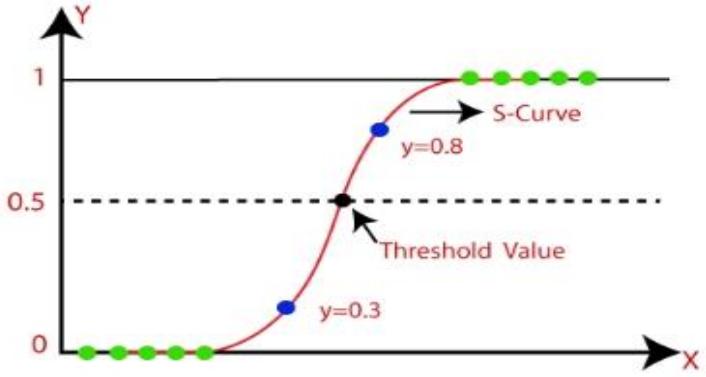


In this case, the new variable y is created as a function of distance from the origin. A non-linear function that creates a new variable is referred to as kernel.

Using Logistic Regression

Logistic regression uses the concept of predictive modeling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm.

Logistic regression is used when the dependent variable is binary such as click on a given advertisement link or not, spam detection, Diabetes prediction, the customer will purchase or not, an employee will leave the company or not.



Logistic regression uses Maximum Likelihood Estimation (MLE) approach i.e., it determines the parameters (mean and variance) that are maximizing the likelihood to produce the desired output.

Logistic Regression uses a sigmoid or logit function which will squash the best fit straight line that will map any values including the exceeding values from 0 to 1 range. So, it forms an "S" shaped curve.
Sigmoid function removes the effect of outlier and makes the output between 0 to 1.

The logistic function is of the form:

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

where μ is a location parameter

s is a scale parameter.

Software description:

HARDWARE REQUIREMENT:

System : Intel core i3,i5,i7 and 2ghz minimum
RAM : 4gb or above
Hard disk : 10GB OR above
Input : keyboard and mouse
Output : monitor or pc

SOFTWARE REQUIREMENTS:

OS : windows 8 or higher version
Platform : juptyer notebook,goolgle colab
Program language : python

RESULTS:

Accuracy and Model score for KNN, SVM and Logistic Regression

ML model	Accuracy	Model Score
KNN	0.932471473354126	0.9239441204131186
SVM	0.899564272413793	0.9054613542158956

Accuracy and Model score for Logistic Regression on Partial Data

ML model	Accuracy	Model Score
Logistic Regression	0.8222222222222223	0.912346542136454

CONCLUSION:

So, from the above table (accuracy), we can predict that the linear regression model has shown accurate results for the model. We have used linear regression model to get 82.22% accuracy rate in linear regression.

FUTURE SCOPE:

As there is a lot of possibility of improvement in this based on the data as modern real time data can be collected which can be used to test all the different models that are present and to create a new accuracy based on this. Another thing that can be done is to test the model made by the authors and also create a comparison on the new data that is there. The data collection would take a long time hence till then multiple times the data should be collected from different sources.

REFERENCES:

[1].https://www.investopedia.com/terms/m/mlr.asp
[2].https://www.sciencedirect.com/topics/social-sciences/multiple-linear-regression
[3] .https://www.kaggle.com
[4] https://www.karger.com/Article/Fulltext/505021
[5] https://www.geeksforgeeks.org/support-vector-machine-algorithm/
[6] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
https://onlinelibrary.wiley.com/doi/full/10.1111/exsy.12899