

### 

### 

### GOUP MEMBERS: 1-Syeda Rameen Ezam(P19101073) 2-Alvina binte waqar(P19101005) COURSE INCHARGE:

### Dr Tehseen Ahmed Jilani

### COURSE:

### Data Warehousing and Data Mining CLASS MCS(Final)

### APPENDIX

### INTRODUCTION \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3

### Weather \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 4

### Weather Forecasting \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 4

### Types of Weather Forecasting \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 5

### Importance of Weather Forecasting \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 6

### Advantages of Weather Forecasting \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 6

### LITERATURE REVIEW \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 7

### Importing General Properties and Our Dataset \_\_\_\_\_\_\_\_\_\_ 7

### Statistics for all the Columns of The Dataset \_\_\_\_\_\_\_\_\_\_\_\_ 8

### Cleaning the Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 9

### Pre-Processing \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 11

### Data Scaling \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 14

### Models \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 14

REPORT:  
**INTRODUCTION**:  
We work on weather forecasting. We take data of Australia which is consist of almost one lac, forty-five thousand, six hundred. We have twenty-three columns, date, location, temperature, rainfall etc. Weather forecasting means the prediction of the weather through the application of the principles of physics, supplemented by a variety of statistical and empirical techniques. In addition to predictions of atmospheric phenomena themselves, weather forecasting includes predictions of changes on the Earth’s surface climate. These changes are caused by atmospheric conditions like snow and ice cover, storm tides, and floods.

The basis for weather prediction started with the theories of the ancient Greek philosophers and continued with Renaissance scientists. It was followed by the scientific revolution of the 17th and 18th centuries. The theoretical models of 20th- and 21st-century atmospheric scientists and meteorologists helped for the betterment in applications. The so-called synoptic weather map came to

be the principal tool of 19th-century meteorologists. This is used today in weather stations and on television weather reports all over the world. All can happen only through a comprehensive weather forecast. Any weather prediction needs a systematic collection of weather record of various places and proper analysis using the data for prediction.

**BACKGROUND:**

1

ABSTRACT

Weather forecasting is the application of science and technology to predict

the state of the atmosphere for a given location.Ancient weather forecasting

methods usually relied on observed patterns of events, also termed pattern

recognition. For example, it might be observed that if the sunset was particularly

red, the following day often brought fair weather.However, not all of these

predictions prove reliable.

Here this system will predict weather based on parameters such as

temperature, humidity and wind. User will enter current temperature; humidity and

wind, System will take this parameter and will predict weather(rainfall in inches)

from previous data in database(dataset). The role of the admin is to add previous

weather data in database, so that system will calculate weather(estimated rainfall in

inches) based on these data. Weather forecasting system takes parameters such as

temperature, humidity, and wind and will forecast weather based on previous

record therefore this prediction will prove reliable. This system can be used in Air

Traffic, Marine, Agriculture, Forestry, Military, and Navy etc.

1

ABSTRACT

Weather forecasting is the application of science and technology to predict

the state of the atmosphere for a given location.Ancient weather forecasting

methods usually relied on observed patterns of events, also termed pattern

recognition. For example, it might be observed that if the sunset was particularly

red, the following day often brought fair weather.However, not all of these

predictions prove reliable.

Here this system will predict weather based on parameters such as

temperature, humidity and wind. User will enter current temperature; humidity and

wind, System will take this parameter and will predict weather(rainfall in inches)

from previous data in database(dataset). The role of the admin is to add previous

weather data in database, so that system will calculate weather(estimated rainfall in

inches) based on these data. Weather forecasting system takes parameters such as

temperature, humidity, and wind and will forecast weather based on previous

record therefore this prediction will prove reliable. This system can be used in Air

Traffic, Marine, Agriculture, Forestry, Military, and Navy etc.

1

ABSTRACT

Weather forecasting is the application of science and technology to predict

the state of the atmosphere for a given location.Ancient weather forecasting

methods usually relied on observed patterns of events, also termed pattern

recognition. For example, it might be observed that if the sunset was particularly

red, the following day often brought fair weather.However, not all of these

predictions prove reliable.

Here this system will predict weather based on parameters such as

temperature, humidity and wind. User will enter current temperature; humidity and

wind, System will take this parameter and will predict weather(rainfall in inches)

from previous data in database(dataset). The role of the admin is to add previous

weather data in database, so that system will calculate weather(estimated rainfall in

inches) based on these data. Weather forecasting system takes parameters such as

temperature, humidity, and wind and will forecast weather based on previous

record therefore this prediction will prove reliable. This system can be used in Air

Traffic, Marine, Agriculture, Forestry, Military, and Navy etc.

1

ABSTRACT

Weather forecasting is the application of science and technology to predict

the state of the atmosphere for a given location.Ancient weather forecasting

methods usually relied on observed patterns of events, also termed pattern

recognition. For example, it might be observed that if the sunset was particularly

red, the following day often brought fair weather.However, not all of these

predictions prove reliable.

Here this system will predict weather based on parameters such as

temperature, humidity and wind. User will enter current temperature; humidity and

wind, System will take this parameter and will predict weather(rainfall in inches)

from previous data in database(dataset). The role of the admin is to add previous

weather data in database, so that system will calculate weather(estimated rainfall in

inches) based on these data. Weather forecasting system takes parameters such as

temperature, humidity, and wind and will forecast weather based on previous

record therefore this prediction will prove reliable. This system can be used in Air

Traffic, Marine, Agriculture, Forestry, Military, and Navy etc.

1

ABSTRACT

Weather forecasting is the application of science and technology to predict

the state of the atmosphere for a given location.Ancient weather forecasting

methods usually relied on observed patterns of events, also termed pattern

recognition. For example, it might be observed that if the sunset was particularly

red, the following day often brought fair weather.However, not all of these

predictions prove reliable.

Here this system will predict weather based on parameters such as

temperature, humidity and wind. User will enter current temperature; humidity and

wind, System will take this parameter and will predict weather(rainfall in inches)

from previous data in database(dataset). The role of the admin is to add previous

weather data in database, so that system will calculate weather(estimated rainfall in

inches) based on these data. Weather forecasting system takes parameters such as

temperature, humidity, and wind and will forecast weather based on previous

record therefore this prediction will prove reliable. This system can be used in Air

Traffic, Marine, Agriculture, Forestry, Military, and Navy etc.

Weather forecasting is the application of science and technology to predict

the state of the atmosphere for a given location.Ancient weather forecasting

methods usually relied on observed patterns of events, also termed pattern

recognition. For example, it might be observed that if the sunset was particularly

red, the following day often brought fair weather.However, not all of these

predictions prove reliable

Weather forecasting is the application of science and technology to predict

the state of the atmosphere for a given location.Ancient weather forecasting

methods usually relied on observed patterns of events, also termed pattern

recognition. For example, it might be observed that if the sunset was particularly

red, the following day often brought fair weather.However, not all of these

predictions prove reliable

Weather forecasting is the application of science and technology to predict

the state of the atmosphere for a given location.Ancient weather forecasting

methods usually relied on observed patterns of events, also termed pattern

recognition. For example, it might be observed that if the sunset was particularly

red, the following day often brought fair weather.However, not all of these

predictions prove reliable

Weather forecasting is the application of science and technology to predict

the state of the atmosphere for a given location.Ancient weather forecasting

methods usually relied on observed patterns of events, also termed pattern

recognition. For example, it might be observed that if the sunset was particularly

red, the following day often brought fair weather.However, not all of these

predictions prove reliable

Weather forecasting is the prediction of the state of the atmosphere for a given location using the application of science and technology. This includes temperature, rain, cloudiness, wind speed, and humidity. Weather warnings are a special kind of short-range forecast carried out for the protection of human life. Weather warnings are issued by the governments throughout the world for all kinds of threatening weather events including tropical storms and tropical cyclones depending upon the location. The forecast may be short-range or Long-range. It is a very interesting and challenging task. This report provides a basic understanding of the purpose and scope of weather forecasts, the basic principles and the general models developed for forecasting.  
Ancient weather forecasting methods usually relied on observed patterns of events, also termed pattern recognition. For example, it might be observed that if the sunset was particularly red, the following day often brought fair weather. However, not all of these predictions prove reliable.

**WEATHER:**Weather is the state of the atmosphere, describing for example the degree to which it is hot or cold, wet or dry, calm or stormy, clear or cloudy. On Earth, most weather phenomena occur in the lowest layer of the planet's atmosphere, the troposphere, just below the stratosphere.

Here this system will predict weather based on parameters such as

temperature, humidity and wind. User will enter current temperature; humidity and

wind, System will take this parameter and will predict weather(rainfall in inches)

from previous data in database(dataset). The role of the admin is to add previous

weather data in database, so that system will calculate weather(estimated rainfall in

inches) based on these data. Weather forecasting system takes parameters such as

temperature, humidity, and wind and will forecast weather based on previous

record therefore this prediction will prove reliable. This system can be used in Air

Traffic, Marine, Agriculture, Forestry, Military, and Navy etc.

Here this system will predict weather based on parameters such as

temperature, humidity and wind. User will enter current temperature; humidity and

wind, System will take this parameter and will predict weather(rainfall in inches)

from previous data in database(dataset). The role of the admin is to add previous

weather data in database, so that system will calculate weather(estimated rainfall in

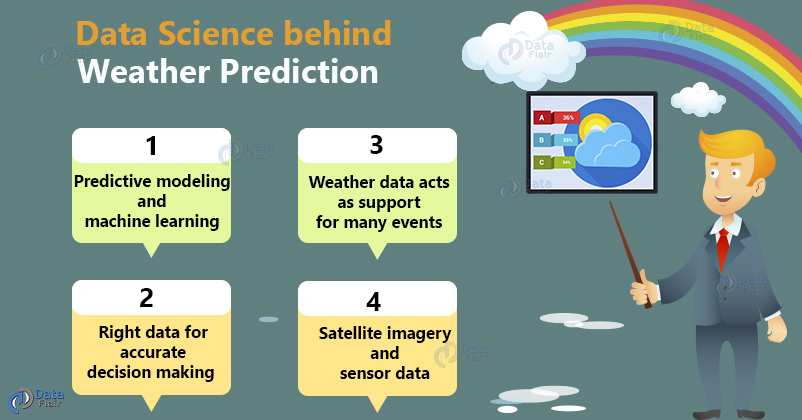
inches) based on these data. Weather forecasting system takes parameters such as

temperature, humidity, and wind and will forecast weather based on previous

record therefore this prediction will prove reliable. This system can be used in Air

Traffic, Marine, Agriculture, Forestry, Military, and Navy etc.

**WEATHER FORECASTING:**  
It cannot be denied that weather forecasting, i.e. predicting weather behavior, is a very challenging task, even with the rapid growth in science. Weather is known to be in the area of meteorology. The process is carried out by collecting data related to the current state of weather like rain, heat, wind, and fog. Data mining techniques in this field has increasingly developed over the last ten years.

Weather forecast is the application of science to predict the condition of atmosphere for a given city or area. It is made by collecting a maximum amount of data possible about the current condition of the atmosphere and using the understanding of atmospheric processes to determine how the atmosphere involve in the future.Weather forecast involves three steps:   
1-Observation  
2-Analysis  
3-Extrapolation   
  
  
  
Many businesses are depend on weather conditions for example agriculture activities, Construction work, Airport control authority etc.  
  
**TYPES OF WEATHER FORECASTING:**A daily weather forecast involves the work of thousands of observers and meteorologists all over the world. Modern computers make forecasts more accurate than ever, and weather satellites orbiting the earth take photographs of clouds from space.  
Forecasters use the observations from ground and space, along with formulas and rules based on experience of what has happened in the past, and then make their forecast.  
Meteorologists actually use a combination of several different methods to come up with their daily weather forecasts [HTTP1]. They are:

a) Persistence Forecasting

b) Synoptic Forecasting

c) Statistical Forecasting   
d) Computer forecasting

a- PERSISTENCE FORECASTING:  
The simplest method of forecasting the weather is persistence forecasting. It relies upon today's conditions to forecast the conditions tomorrow. This can be a valid way of forecasting the weather when it is in a steady state, such as during the summer season in the tropics. This method of forecasting strongly depends upon the presence of a stagnant weather pattern. It can be useful in both short range forecasts and long range forecasts. This assumes that what the weather is doing now is what it will continue to do. To find out what the weather is doing, meteorologists make weather observations.  
  
b- SYNOPTIC FORECASTING:  
This method uses the basic rules for forecasting. Meteorologists take their observations, and apply those rules to make a short-term forecast.

c- STATISTICAL FORECASTING:  
Meteorologists ask themselves, what does it usually do this time of the year? Records of average temperatures, average rainfall and average snowfall over the years give forecasters an idea of what the weather is "supposed to be like" at a certain time of the year.  
  
d- COMPUTER FORECASTING:

Forecasters take their observations and plug the numbers into complicated equations. Several ultra-high-speed computers run these various equations to make computer "models" which give a forecast for the next several days. Often, different equations produce different results, so meteorologists must always use the other forecasting methods along with this one. Using all the above methods, forecasters come up with their "best guess" as to what weather conditions will be over the next few days.  
Weather forecasting now has a wide range of operational products that traditionally are classified under the following groups:  
1-Very short-range forecast.

2-Short-range forecast.

3-Medium-range forecast.  
4-Long-range forecast.

Each weather forecast can be defined on the basis of the following criteria:  
a-Dominant technology.  
b- Temporal range of validity after emission.  
c- Characteristics of input and output time and space resolution.  
d- Broadcasting needs.  
e- Accuracy.

**IMPORTANCE OF WEATHER FORECASTING:**Weather forecasting is used in many situations like severe weather alerts and advisories, predicting the behavior of the cloud for air transport, prediction of waterways in a sea, agricultural development and avoiding forest fire.

### ADVANTAGES OF WEATHER FORECASTING:

Advantages of weather forecasting are:

* People are warned prior to what the weather will be like on a particular day.
* To help people take proper precautions to secure themselves and their families in case of unwanted occurrences.
* Organizations can work better with the help of accurate weather predictions.
* It helps to deliver visual forecasts by various methods that most companies prefer.
* Weather forecasting highly benefits the agriculture sector for buying/selling livestock.
* It also assists the farmers to decide when to plant crops, pastures, and when to irrigate.
* It is the best method for management of inventory, selling strategies and crop forecasts.
* It provides the business with valuable information that the business can use to make  
  decisions about future business strategies.

**LITERATURE REWIEW:**In our project we apply principal component analysis (PCA), logistic regression and random forest.

**ABOUT DATASET:**

This dataset contains about 10 years of daily weather observations from numerous Australian weather stations. Rain tomorrow is the target variable to predict. It means -- did it rain the next day, Yes or No?  
There are 23 attributes in this dataset.

**1-IMPORTING GENERAL PROPERTIES AND OUR DATASET:**

1-SEABORN:  
Used to make default matplotlib plots look nicer, and also introduces some additional plot types.

2-WARNINGS:  
Warnings are provided to warn the developer of situations that aren't necessarily exceptions.  
  
3-PANDAS:  
Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks.  
  
4-NUMPY:  
Numpy is a necessary package for scientific computation. It includes an incredibly versatile structure for working with arrays, which are the primary data format that scikit-learn uses for input data.  
  
5-MATPLOTLIB:  
Matplotlib is the fundamental package for data visualization in Python.

DF-HEAD:  
The head() function is used to get the first n rows.  
  
**2-STATISTICS FOR ALL THE COLUMNS OF THE DATASET:**

We write df-describe which is used for calculating some statistical data like percentile, mean and standard of the numerical values of the Series or Data Frame. It analyzes both numeric and object series and also the Data Frame column sets of mixed data types.

COLUMNS NAME IN DATASET:

We use col-names in python. Following are columns in our dataset:  
1-Date  
2-Location  
3-Min temp  
4-Max temp  
5-Rain fall   
6-Evaporation  
7-Sunshine  
8-WindGustDir  
9-WindGustSpeed  
10- WindDir9am  
11-WindDir3pm  
12-WindDir3pm  
13- WindSpeed9am  
14- WindSpeed3pm  
15- Humidity9am  
16- Humidity3pm  
17- Pressure9am  
18- Pressure3pm  
19- Cloud9am  
20- Cloud3pm  
21- Temp9am  
22- Temp3pm  
23- Rain Today  
24- Rain Tomorrow

**3-CLEANING THE DATA:**

Data cleaning is the process of preparing raw data for analysis by removing bad data, organizing the raw data, and filling in the null values. Ultimately, cleaning data prepares the data for the process of data mining when the most valuable information can be pulled from the data set.

NUNIQUE:  
nunique() function return number of unique elements in the object.  
  
INFO:  
The info() function is used to print a concise summary of a DataFrame. This method prints information about a DataFrame including the index dtype and column dtypes, non-null values and memory usage. Whether to print the full summary.

CONVERTING IN LOWERCASE:

To covert the columns names we use df.columns.

CHANGE DATE FORMAT:  
For changing date, day and year we use:  
df[['date', 'Year', 'Month', 'Day']]

df**.**drop:

The drop() function is used to drop specified labels from rows or columns. Remove rows or columns by specifying label names and corresponding axis, or by specifying directly index or column names. When using a multi-index, labels on different levels can be removed by specifying the level.

DEAL WITH NULL OR MISSING VALUES:

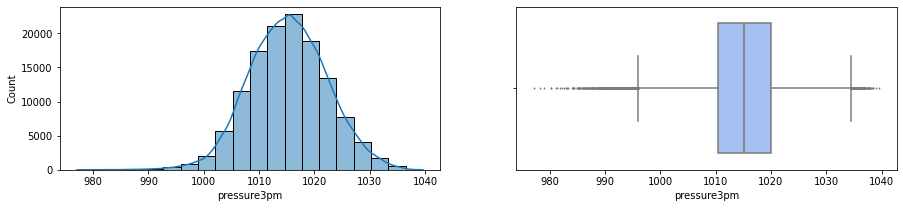
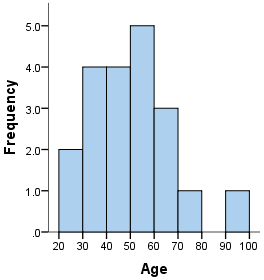
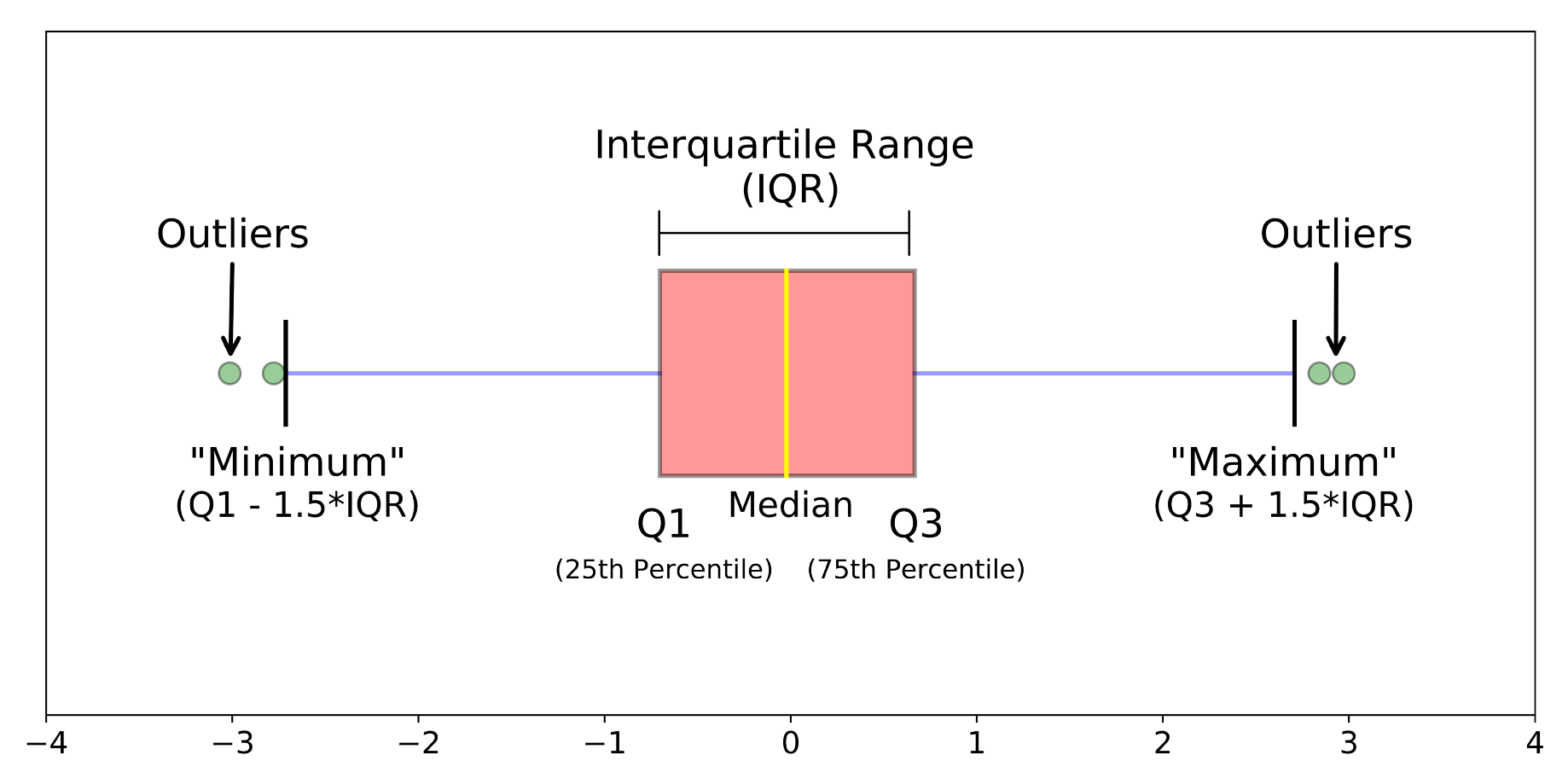
Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of the rows as null, then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped. We use:  
df**.**isnull()**.**sum( )  
df**=**df**.**drop(['sunshine','evaporation','cloud3pm','cloud9am'], axis**=**1)  
missing

Remove rows where target variables are missing:

df**.**dropna(how**=**'all', subset**=**['raintomorrow'], inplace**=True**)  
Now we have 15 columns remaining, and again we write df.head() for checking.

SORTING DATA INTO NUMERICAL AND CATEGORICAL FEATURES:

* We know that there are some different number of missing values to each features
* Depending on their distribution we are going to replace with median or mean.
* For categorical data we use cat\_col.

Now we plot histogram and box-plot for each column.  
  
HISTOGRAM:  
A histogram is a graphical representation that organizes a group of data points into user-specified ranges.  
  
  
BOX-PLOT:  
A box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.  


**4-PRE-PROCESSING:**

Before modeling we have to do two steps on our data, encoding and dealing with missing values.  
**1. DATA CLEANING:**   
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.   
We do encoding because we have categorical data.  
**(A) MISSING DATA:**   
This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:   
  
**i-IGNORE THE TUPLES:**   
This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.   
 **ii-FILL THE MISSING VALUES:**   
There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value. 

**(B)NOISY DATA:**   
Noisy data is a meaningless data that can’t be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways:  
i-BINNING METHOD:   
This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

ii-REGRESSION:   
Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

iii-CLUSTERING:   
s  **2. DATA TRANSFORMATION:**   
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:   
  
i-NORMALIZATION:   
It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)   
ii-ATTRIBUTE SELECTION:   
In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

iii-DISCRETIZATION:   
This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

iv-CONCEPT HIERARCHY GENERATION:   
Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

**3. DATA REDUCTION:**   
Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs. The various steps to data reduction are:   
  
**i-DATA CUBE AGGREGATION:**Aggregation operation is applied to data for the construction of the data cube.   
ii-ATTRIBUTE SUBSET SELECTION:   
The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute. The attribute having p-value greater than significance level can be discarded.   
iii-NUMEROSITY REDUCTION:   
This enable to store the model of data instead of whole data, for example: Regression Models.   
iv-DIMENSIONALITY REDUCTION:   
This reduce the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis).

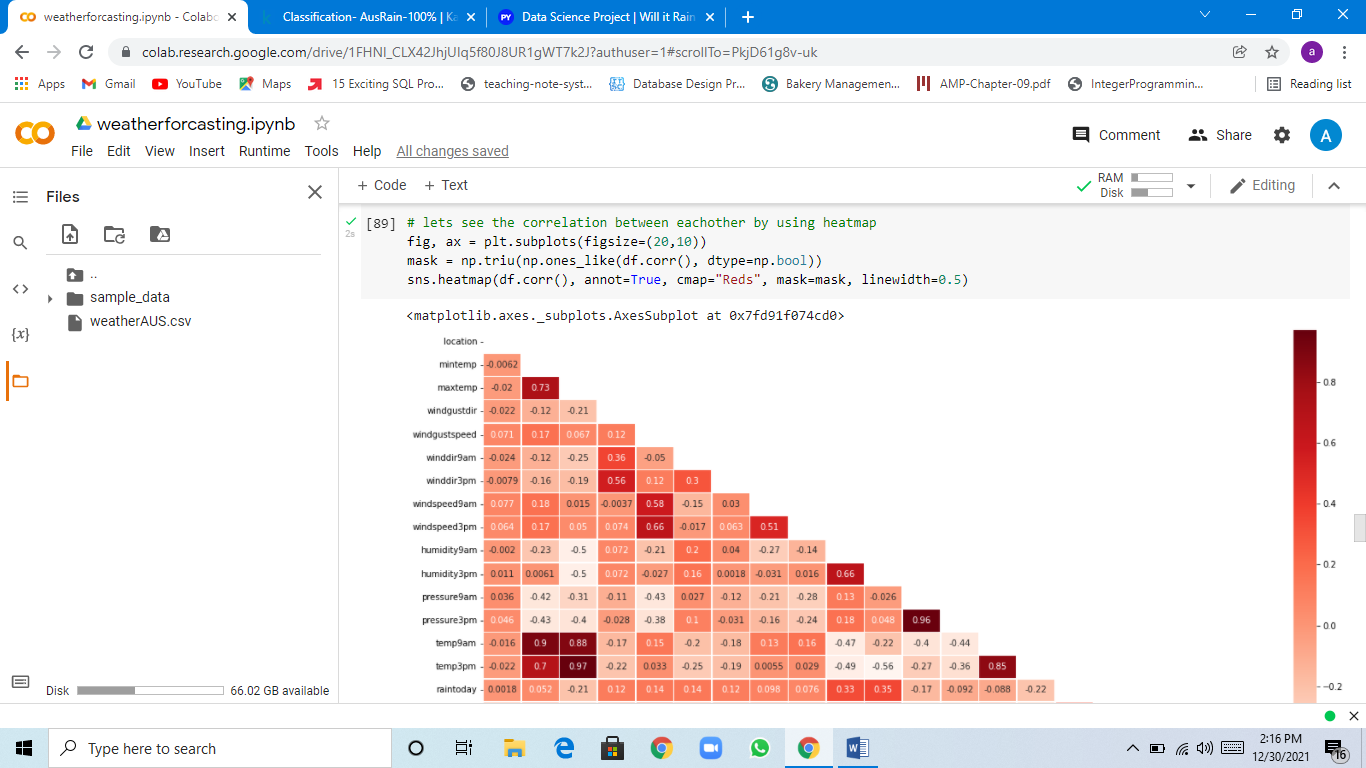
ENCODING:

It is a process by which categorical data converted into numerical for processing.

CORRELATION:

Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.

* POSITIVE CORRELATION: directly relation between two variables.
* NEGATIVE CORRELATION: inversely relation between two variables.
* ZERO CORRELATION: there is no relation between two variables.



**5-DATASCALING:**

Scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

THE DATA IS READY FOR MODELING:  
Train/Test is a method to measure the accuracy of your model. It is called Train/Test because you split the the data set into two sets: a training set and a testing set. 80% for training, and 20% for testing.  
The “training” data set is the general term for the samples used to create the model, while the “test” or “validation” data set is used to qualify performance.

**6-MODELS:**A model is an informative representation of an object, person or system. The term originally denoted the plans of a building in late 16th-century English, and derived via French and Italian ultimately from Latin modulus, a measure. Models can be divided into physical models and abstract models.

**A-LOGISTIC REGRESSION:**  
Logistic Regression is a classification algorithm. It is a predictive modeling algorithm that is used when the dependent variable(target) is categorical in nature.  
Logistic Regression is most commonly used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1. These types of problems are known as Binary Classification problems. When using logistic regression, a threshold is usually specified that indicates at what value the  
example will be put into one class vs. the other class. Although, sometime we come across more than 2 classes and still it is a classification problem. These types of problems are known as Multi class classification problems. We can accomplish this by applying a “one vs rest” strategy.

### https://ars.els-cdn.com/content/image/3-s2.0-B9780124166905000038-f03-26-9780124166905.jpg PURPOSE AND EXAMPLES OF LOGISTIC REGRESSION: Logistic regression is one of the most commonly used machine learning algorithms for binary classification problems, which are problems with two class values, including predictions such as “this or that,” “yes or no” and “A or B.” The purpose of logistic regression is to estimate the probabilities of events, including determining a relationship between features and the probabilities of particular outcomes. USES OF LOGISTIC REGRESSION:

Logistic regression has become particularly popular in online advertising, enabling marketers to predict the likelihood of specific website users who will click on particular advertisements as a yes or no percentage.  
Logistic regression can also be used in:  
1-Healthcare to identify risk factors for diseases and plan preventive measures.  
2-Weather forecasting [apps](https://searchsoftwarequality.techtarget.com/definition/application) to predict snowfall and weather conditions.  
3-Voting apps to determine if voters will vote for a particular candidate.  
4-Insurance to predict the chances that a policy holder will die before the term of the policy expires based on certain criteria, such as gender, age and physical examination.  
5-Banking to predict the chances that a loan applicant will default on a loan or not, based on annual income, past defaults and past debts.  
LOGISTIC REGRESSION VS. LINEAR REGRESSION:  
The main difference between logistic regression and linear regression is that  
logistic regression provides a constant output, while linear regression provides a continuous output. In logistic regression, the outcome, such as a dependent variable, only has a limited number of possible values. However, in linear regression, the outcome is continuous, which means that it can have any one of an infinite number of possible values. Logistic regression is used when the response variable is categorical, such as yes/no, true/false and pass/fail. Linear regression is used when the response variable is continuous, such as number of hours, height and weight.  
For example, given data on the time a student spent studying and that student’s exam scores, logistic regression and linear regression can predict different things. With logistic regression predictions, only specific values or categories are allowed. Therefore, logistic regression can predict whether the student passed or failed. Since linear regression predictions are continuous, such as numbers in a range, it can predict the student’s test score on a scale of 0 -100.

### 

**B-RANDOM FOREST:**  
Random forest is an ensemble technique which is better than decision tree. It always gives low variance. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.  
  
HOW RANDOM FOREST WORKS:  
Random forest is a [supervised learning algorithm](https://builtin.com/data-science/supervised-learning-python). The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the [bagging method](https://builtin.com/data-science/tour-top-10-algorithms-machine-learning-newbies) is that a combination of learning models increases the overall result.  
Random forest has nearly the same hyper parameters as a decision tree or a bagging classifier. Fortunately, there's no need to combine a decision tree with a bagging classifier because you can easily use the classifier-class of random forest. With random forest, you can also deal with regression tasks by using the algorithm's regressor.  
  
DIFFERENCE BETWEEN DECISION TREES AND RANDOM FORESTS:  
While random forest is a collection of decision trees, there are some differences. If you input a training dataset with features and labels into a decision tree, it will formulate some set of rules, which will be used to make the predictions.

## ADVANTAGES AND DISADVANTAGES OF THE RANDOM FOREST: One of the biggest advantages of random forest is its versatility. It can be used for both regression and classification tasks, and it’s also easy to view the relative importance it assigns to the input features. One of the biggest problems in machine learning is overfitting, but most of the time this won’t happen thanks to the random forest classifier. If there are enough trees in the forest, the classifier won’t overfit the model.



**C-PCA:**

PCA is a method used to reduce no. of variables in data by extracting important values from a large data. It reduces the dimension of data.   
The principal components of a collection of points is the direction of a line that best fits the data while being orthogonal to the first vectors. The fit process minimizes the average squared distance from the points to the best line.  
PCA can be thought of as fitting a p-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. If some axis of the ellipsoid is small, then the variance along that axis is also small.

STEPS:

1. Standardization.
2. Covariance matrix computation.
3. Compute the Eigen vectors and Eigen values.

GETTING THE PRINCIPAL COMPONENTS:

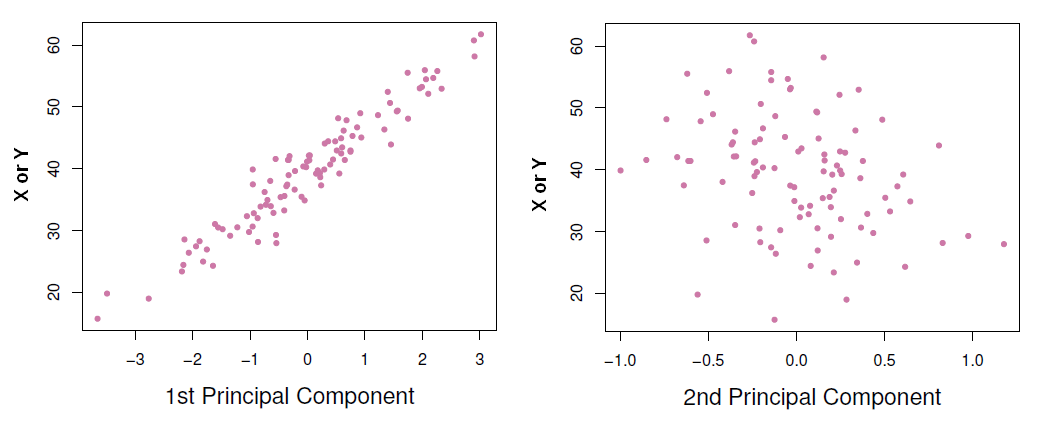
Getting the principal components of the data matrix x.

PROCEDURE:

* The first principle component is just the normalized linear combination of the variables that has the highest variance.
* The second principal component has largest variance, subject to being uncorrelated with the first.
* And so on.

#### The principal components produce a linear combinations or dimensions of the data that are really high in variance and that are uncorrelated. EXAMPLE: _

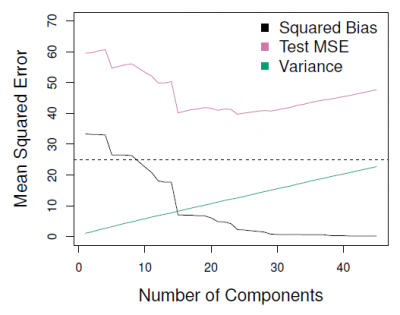
* The direction in which the data varies the most actually falls along the green line. This is the direction with the most variation in the data, this is why it's the first principal component (direction). The sum of square distances is the smallest possible.
* What's the direction along which the data varies the most out of all directions that are uncorrelated with the first direction? That's this blue dashed line. That's the second principle component in this data.

PRINCIPAL COMPONENT PLOT:  
A plot of the principle components against the variable helps to understand  
them better.  


As the first principle component is highly correlated with all variables, it means that it summarizes the data very well. Then instead of using the variables X or Y to make prediction, we can use just the first principle component.

When two variables are really [correlated](https://datacadamia.com/data_mining/correlation) with each other, one new variable (i.e. the first principle component) can really summarize both of those two variables very well.

FITTING THE MODELS WITH THE PRINCIPAL COMPONENTS:  
Perform least squares regression using those principal components as predictors.  
CHOOSING M, THE NUMBER OF PRINCIPAL COMPONENT (OR DIRECTIONS):When more and more components are used in the regression model, the [bias](https://datacadamia.com/data_mining/bias) will decrease (because we fit more and more [complex model](https://datacadamia.com/statistics/model/model#complexity)) but the variance will then increase.



where:

* [MSE](https://datacadamia.com/data_mining/mse) has a U shape and is smallest for a model with around 18 principal components.

Cross validation was chosen in order to choose the number of principal component directions.

**LINEAR MODEL:**It is simplest form of regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observe the data. LR attempts to find the mathematical relationship between variables. The relationship between dependent variable is given by straight line and it has only one independent variable.

**ACCURACY:**The accuracy of a classifier is given as the percentage of total correct predictions divided by the total number of instances.  
If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known.

**CONLUSION:**  
Weather forecasts still have their limitations despite the use of modern technology and improved techniques to predict the weather. Weather forecasting is complex and not always accurate, especially for days further in the future, because the weather can be chaotic and unpredictable. If weather patterns are relatively stable, the persistence method of forecasting provides a relatively useful technique to predict the weather for the next day. Weather observation techniques have improved and there have been technological advancements in predicting the weather in recent times. Despite this major scientific and technical progress, many challenges remain regarding long-term weather predictability. The accuracy of individual weather forecasts varies significantly.