CHAPTER – 2

STUDY AREA AND METHODOLOGY

2.0 GEOGRAPHICAL DESCRIPTION OF STUDY AREA

Chennai is the capital city of the Indian state of Tamil Nadu, located on the Coromandel Coast off the Bay of Bengal. Chennai was formerly known as Madras. It is the biggest industrial and commercial centre in South India and a major cultural, economic and educational centre. Chennai is known as the “Detroit of India” for its automobile industry (www.madrasi.info). Chennai is the fifth largest city and fourth most populous metropolitan area in the country and 31st largest urban area in the world. Chennai is situated in the Latitude 13.084°N and Longitude 80.27° E (en.wikipedia.org). The location map of the study area is shown in (Fig. 2.1).

Fig. 2.1 Shows the location Map of the Study Area
Chennai city had a population of 4.34 million in the 2001 census within the area administered by the Corporation of Chennai and an extended Metropolitan Population of 6.5 million (wikimapia.org). The urban agglomeration of metropolitan Chennai has an estimated population over 8.2 million people. Chennai is most famous for its beaches and culture.

2.1 HISTORY OF STUDY AREA

The origin of Madras as we know dates back to few centuries – about 350 years. Prior to that, small villages existed for well over 1000 years in a cluster of civilization. Villages around Temples like Parthasarathy in Triplicane and Kapaleeswarer temple in Mylapore near the southern coast and Marudheeswarer Temple in Thiruvanmiyur were in existence for several centuries long before the Europeans arrived here (en.wikipedia.org).

The first Europeans to reach the shores of Madras were, as in this part of the world, Portuguese. They built a church in Saint Thomas Mount enshrining the Bleeding Cross. And then they went further down to Little Mount where they built another small church in 1551 where St.Thomas, the disciple of Jesus Christ was hiding in a cave from his persecutors before being martyred in St.Thomas Mount.

One Francis Day, a member of the Masulipatnam council undertook a journey down the east coast of south India looking for a promising place. He found one near Poonamalee, ruled by one DamarlaVenkatapathyNayak, a local chieftain. He had chosen a village called as Madraspatnam. The fort built by Francis Day was called as Fort.St.George as it was finished at the festival of St.George - 23rd April, 1653 (en.wikipedia.org).

The settlement soon occurred around the Fort and the new town was called as Srirangarayapatnam, named after the nephew of VenkatapathyNayak, Srirangarayalu.
Some seek to distinguish the name of Chennaipatnam and Madraspatnam as two different settlements. Chennaipatnam was the predominantly Indian town adjacent to town around the Fort, which went under the name of Madraspatnam.

But the origin of the name Madras is mired in some folklore. Some attribute the name Madras to the small church built by the Portuguese dedicated to the Mother of God, whereas others ascribe the name to a local village chieftain Maddarasu.

Madras rose in stature and power with the rise of British Empire. Many colleges, hospitals and prominent landmarks were established in tune with the rising importance. Towards the end of 19th century (Wikipedia.com), Madras was clearly established as an important hub in South India.

Despite its hot and humid climate throughout the year, the British chose this city as it provided - a relatively docile populace still clinging to the old glory of its Tamil Kingdom, a natural harbour which made easy access to the maritime colonial power's reach and the position it occupied almost in the center of a region controlled by the British.

Madras population consisted of people from many parts of the South India and also those from North. The Tamil language they spoke borrowed many words from all these languages and became a dialect - Madras Tamil. Despite the amalgamation of so many cultures, Madras remains a conservative middle class south Indian city even to this day.

2.2 CLIMATE

Chennai has a tropical wet and dry climate. The city lies on the thermal equator and is also on the coast, which prevents extreme variation in seasonal temperature. The hottest part of the year is late May to early June, known regionally as Agni Nakshatram or as KathiriVeyyil, with maximum temperatures around 35-40°C (95-104°F). The coolest part of the year is January, with minimum
temperatures around 15-22 °C (59-72°F). The lowest recorded temperature was 13.8°C (56.8°F) on 11 December 1895 and 29 January 1905. The highest recorded temperature was 45°C (113°F) on 30 May 2003. The average annual rainfall is about 1400mm (55 in) (en.wikipedia.org). The city gets most of its seasonal rainfall from the north-east monsoon winds, from mid-October to mid-December. Cyclones in the Bay of Bengal sometimes hit the city. The highest annual rainfall recorded is 2570mm (101 in) in 2005. Prevailing winds in Chennai are usually southwesterly between April and October and north-easterly during the rest of the year. Historically, Chennai has relied on annual monsoon rains to replenish water reservoirs, as no major rivers flow through the area.

### 2.3 REVIEW OF LITERATURE

Many researchers have tried to use data mining technologies in areas related to meteorology and weather prediction. The emphasis in (Bilgin and Camurcu, 2004) is on using DBSCAN (Density Based Spatial Clustering of Applications with Noise) clustering algorithm to categorize Turkey into regions according to climatic characteristics. They use the daily maximum and minimum temperature records between 1930 and 1996 from 258 stations. They draw that this type of data mining application can help meteorological to create faster forecast and decisions and provide more performance and reliability than any other methods. Research (Kotsiantis et al., 2007) predict daily average, maximum and minimum temperature for Patras city in Greek by using six different data mining methods: Feed-Forward Back Propagation (BP), k-Nearest Neighbor (KNN), M5rules algorithm, linear least-squares regression (LR), Decision tree and instance based learning. They use four years period data [2002-2005] of temperature, relative humidity and rainfall. The results they obtained in this study were accurate in terms of Correlation Coefficient and Root Mean Square.

Using data mining in metrological application is not limited to prediction, but it also extends to participate in many important fields like water resource management (Jan et al., 2009) and air pollution management (Li and Shue, 2004).
Mining techniques also can be applied to various types of data like weather images and radar maps extend to characteristic features extracted from this weather images can be used to represent various weather patterns (Olaiya and Adeyemo, 2012). Data mining have been employed successfully to build an important applications in the field of meteorology like prediction abnormal events like hurricanes, storms and river flood prediction (Bartok et al., 2010; Mohammadi et al., 2006). These applications can maintain public safety and welfare. In this context, (Zhang et al., 2004) propose a new framework to discover dynamic inter dimension association rules for local scale weather prediction of Dallas City. The usefulness of applying association mining is to find a strong relation between severe conditions and the change tendencies of the measurements of the weather. The authors conclude with some predicates extracted from the obtained rules. Another contribution to detect severe events using data mining is by (Li et al., 2008) and (Peter et al., 2003) used the volumetric radar data to detect storm events and classify them into four types: hail, heavy rain, tornadoes and wind.

To predict the weather (Olaiya and Adeyemo, 2012) by numerical means, meteorologists have developed atmospheric models that approximate the atmosphere by using mathematical equations to describe how atmospheric temperature, pressure and moisture will change over time. The equations are programmed into a computer and data on the present atmospheric conditions are fed into the computer. The computer solves the equations to determine how the different atmospheric variables will change over the next few minutes. The computer repeats this procedure again and again using the output from one cycle as the input for the next cycle. For some desired time in the future (12, 24, 36, 48, 72 or 120 hours), the computer prints its calculated information. It then analyzes the data, drawing the lines for the projected position of the various pressure systems. The final computer-drawn forecast chart is called a prognostic chart, or prog. A forecaster uses the progs as a guide for predicting the weather. There are many atmospheric models that represent the atmosphere, with each one interpreting the atmosphere in a slightly different way. The forecaster learns the idiosyncrasies of each model and places more emphasis on the ones that do the best
job of predicting a particular aspect of the weather. Weather forecasts made for 12
and 24 hours are typically quite accurate. Forecasts made for two and three days are
usually good. Beyond about five days, forecast accuracy falls off rapidly. The most
commonly used techniques in data mining are: Artificial Neural Networks, Genetic
Algorithms, Rule Induction, Nearest Neighbor method, Memory-Based Reasoning,
Logistic Regression, Discriminant Analysis and Decision Trees.

In this work both Artificial Neural Networks (ANN) and Decision Trees
(DT) were used to analyze meteorological data gathered from the Ibadan synoptic
airport station over the period of ten years (2000 – 2009), in order to develop
classification rules for the weather parameters over the study period and for the
prediction of future weather conditions using available historical data. The targets
for the prediction are those weather changes that affect us daily like changes in
minimum and maximum temperature, rainfall, evaporation and wind speed.

2.4 DETAILED METHODOLOGY

Data Mining is not a random application of statistical, machine learning
and other methods and tools. It is not a random walk through the space of analytic
techniques but a carefully planned and considered process of deciding what will be
most useful, promising and revealing. It is important to realize that the problem of
discovering or estimating dependencies from data or discovering totally new data is
only part of the general experimental procedure used by scientists, engineers and
others who apply standard steps to draw conclusions from the data. The general
experimental procedure adapted to data mining problems involves the following steps:

2.4.1 Problem Statement And Hypothesis Formulation

Most data based modeling studies are performed in a particular
application domain. Hence, domain specific knowledge and experience are usually
necessary in order to come up with a meaningful problem statement. Unfortunately, many application studies tend to focus on the data mining technique
at the expense of a clear problem statement. In this step, a modeler usually
specifies a set of variables for the unknown dependency and if possible, a general form of this dependency as an initial hypothesis (Mohammadi et al., 2006). There may be several hypotheses formulated for a single problem at this stage. The first step requires the combined expertise of an application domain and a data mining model. In practice, it usually means a close interaction between the data mining expert and the application expert. In successful data mining application, this cooperation does not stop in the initial phase; it continues during the entire data mining process.

The aim of the research is to analyze the climatic data with the help of data mining technique and to perform visual representation of hidden information and prediction. The influence of Sea Level Pressure, Wind Speed, Visibility, Dew Point, Cloud Cover, Humidity, Wind Direction, etc., on Temperature has been studied with the help of Data Mining techniques. Analysis has been carried out if temperature can be predicted using other climatic parameters using Correlation analysis, Multiple Linear Regression, Classification Techniques, Clustering Techniques etc., on the climatic data collected. The steps for the research include Data collection, Data Preprocessing, Applying various data mining techniques and visual representation of hidden information from the processed data and prediction.

### 2.4.2 Data Collection

This step is concerned with how the data are generated and collected. In general, there are two distinct possibilities. The first is when the data generation process is under the control of an expert. This approach is known as a designed experiment. The second possibility is when the expert cannot influence the data generation process. This is known as the observational approach. An observational setting, namely, random data generation, is assumed in most data mining applications (Pang-Ning Tan et al., 2005). Typically, the sampling distribution is completely unknown after data are collected or it is partially and implicitly given in the data collection procedure. It is very important, however, to understand how data collection affects its theoretical distribution, since such a priori knowledge can
be very useful for modeling and later, for the final interpretation of results (Peter et al., 2003). Also it is important to make sure that the data used for estimation of a model and the data used later for testing and applying a model come from the same, unknown, sampling distribution. If this is not the case, the estimated model cannot be successfully used in a final application of the results.

In this research historical Climatic data of Chennai has been collected from Weather Underground(WU), the weather company, San Francisco, California, United States for a period of 2010, 2011, 2012 and 2013. Weather Underground was founded in 1995 as an offshoot of the University of Michigan’s, provides weather reports for newspapers at United States. Collected climatic data includes

- Temperature (High, Avg, Low) in °C

A temperature is a comparative objective measure of hot and cold. It is measured, typically by a thermometer, through the bulk behavior of a thermometric material, detection of heat radiation, or by particle velocity or kinetic energy. It may be calibrated in any various temperature scales, Celsius, Fahrenheit, Kelvin, etc.

- Dew Point (High, Avg, Low) in °C

The dew point is the temperature at which the water vapor in a sample of air at constant barometric pressure condenses into liquid water at the same rate at which it evaporates. At temperatures below the dew point, water will leave the air. The condensed water is called dew when it forms on a solid surface. The condensed water is called either fog or a cloud, depending on its altitude, when it forms in the air. It is calibrated in Celsius, Fahrenheit, etc.

- Humidity (High, Avg, Low) in %

Humidity is the amount of water vapor in the air. Water vapor is the gaseous state of water and is invisible. Humidity indicates the likelihood of precipitation, dew or fog. Higher Humidity reduces the effectiveness of sweating in
cooling the body by reducing the rate of evaporation of moisture from the skin. It is calibrated in (%) percentage.

- **Sea Level Pressure (High, Avg, Low) in (hPa)**

  It is the atmospheric pressure at sea level. This is the atmospheric pressure normally given in weather report on radio, television and newspaper or on the Internet. When barometers in the home are set to match the local weather reports, they measure pressure adjusted to sea level and not the actual local atmospheric pressure. It is calibrated in hPa (Hectopascals).

- **Visibility (High, Avg, Low) in KM**

  In meteorology, visibility is a measure of the distance at which an object or light can be clearly discerned. It is reported with surface weather observations and calibrated either in meters or miles, depending upon the country. Visibility affects all forms of traffic: roads, sailing and aviation.

- **Wind Speed (High, Avg, Low) in (KM/Hr)**

  Wind Speed is a fundamental atmospheric rate. Wind Speed is caused by air moving from high pressure to low pressure, usually due to changes in temperature. Wind Speed affects weather forecasting, aircraft and maritime operations, construction projects growth. It is calibrated in KM/Hr.

- **Precipitation in (mm)**

  Precipitation is any form of water in liquid or solid, falling from the sky. It includes rain, sleet, snow, hail and drizzle plus a few less common occurrences such as ice pellets and freezing rain. It is measured in mm (Millimeter).
2.4.3 Data Pre-Processing

Data pre-processing is a broad area and consists of a number of different strategies and techniques that are inter-related in complex ways. In general Aggregation, Sampling, Dimensionality reduction, Feature subset selection, Feature creation, Discretization and binarization and Variable transformation are the techniques of Data pre-processing. In this research we have adopted Discretization and binarization and Variable Transformation to apply over the collected continuous data. Some data mining algorithms, especially certain classification algorithms, require that the data be in the form of binary attributes. Thus it is often necessary to transform a continuous attribute into a categorical attribute (discretization), and both continuous and discrete attributes may need to be transformed into one more binary attributes (binarization) (Han and Kamber, 2006). And for some of the algorithms requires the data to be normalized or standardized to accomplish the similarity measures.

Data pre-processing includes several steps such as variable scaling and different types of encoding. For example, one feature with the range [0, 1] and the other with the range [-100, 1000] will not have the same weights in the applied technique. They will also influence the final data mining results differently (Rish et al., 2001). Therefore, it is recommended to scale them and bring both features to the same weight for further analysis. Also, application specific encoding methods usually achieve dimensionality reduction by providing a smaller number of informative features for subsequent data modelling.

Data pre-processing steps should not be considered completely independent from other data mining phases. In every iteration of the data mining process, all activities, together, could define new and improved data sets for subsequent iterations. Generally, a good pre-processing method provides an optimal representation for a data mining technique by incorporating a priori knowledge in the form of application specific scaling and encoding. We perform the following Variable transformation for the climatic data collected.
2.4.3.1 Z-Score Normalization

A Z-Score (also known as z-value, standard score or normal score) is a measure of the divergence of an individual experimental result from the most probable result, the mean (Kotuand Deshpande, 2014). Z is expressed in terms of the number of standard deviations from the mean value. In simple words, Z-Scores tell us whether a particular score is equal to the mean, below the mean or above the mean of a bunch of scores. Application of Z-Score Normalization on the data is elaborated with results in the Chapter 3.

2.4.3.2 Min-Max Normalization

In data transformation, the data are transformed or consolidated into forms appropriate for mining. We adopt Min-Max Normalization to transform all the collected climatic parameters in the range of 0 to 100, for performing the outlier analysis. Min-Max normalization performs a linear transformation on the original data. Suppose that \( \text{min}_A \) and \( \text{max}_A \) are the minimum and maximum values of an attribute, \( A \). Min-max normalization maps a value, \( v \), of \( A \) to \( v' \) in the range \([\text{new}_\text{min}_A, \text{new}_\text{max}_A]\) by computing

\[
v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} (\text{new}_\text{max}_A - \text{new}_\text{min}_A) + \text{new}_\text{min}_A.
\]

Min-max normalization (Peter et al., 2003) preserves the relationships among the original data values. It will encounter an “out-of-bounds” error if a future input case for normalization falls outside of the original data range for \( A \). Application of Min-Max Normalization on the data is elaborated with results in the Chapter 3.

2.4.3.3 Raw Data Analysis

In the present work, Raw Data analysis was applied to classify the climatic data of Chennai into groups based on the simple trend graphs generated. The climatic data of the years 2010, 2011, 2012 and 2013 were first analyzed to identify
the trends using Raw Data Analysis. The climatic data includes Temperature, Dew Point, Humidity, Sea Level Pressure, Visibility, Wind Speed, Precipitation, Wind Direction and Cloud Cover. Using Raw Data Analysis, the entire period was classified into 4 groups based on the increase and decrease of the temperature. Group 1 : (Jan – May), Group 2 : (Jun – Aug), Group 3 : (Sep – Oct) and Group 4 : (Nov – Dec). Group 1 across all the years 2010, 2011, 2012 and 2013 has been chosen to be analyzed using Data Mining techniques. Raw data analysis is elaborated with results in the Chapter 3.

2.4.3.4 Binning Method – Discretization of Data

Discretization reduces the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values (Li and Shue, 2004). Some classification algorithms only accept categorical attributes. Binning method is a type of discretization that has been adopted in this research. Binning method first sort data and partition into bins and then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc. There are two types of Binning method – Equal-Width Partitioning and Equal-depth partitioning. Equal-Width Partitioning method has been adopted for performing the data transformation.

If No. of Intervals = N

Width = (Max Value – Min Value) / N

Intervals are

< (Min Value + Width)

Betw. (Min Value + Width) and (Min Value + 2Width)

Betw. (Min Value + 2Width) and (Min Value + 3Width)

.

> (Min Value + (n-1)Width)
Application of Binning Method on the data is elaborated with results in the Chapter 3.

2.4.3.4 Outlier Analysis

Data examination is a time consuming, but necessary, initial step in any analysis that researchers often overlook. Here we evaluate the impact of missing data, identify outliers, and tests for the assumptions underlying most multivariate techniques. The objective of these data examination tasks is as much to reveal what is not apparent as it is to portray the actual data, for the “hidden” effects are easily overlooked (Pang-Ning Tan et al., 2005).

Outliers are observations with a unique combination of characteristics identifiable as distinctly different from the other observations. Outliers in weather data may be due to data entry problem, faulty data collection instruments, abnormal changes in weather such as tornadoes, hurricane, forest fires, etc. Such non-representative samples can seriously affect the model produced later.

There are two strategies for dealing with outliers:

a.) Detect and eventually remove outliers as a part of the preprocessing phase.

b.) Develop robust modeling methods that are insensitive to outliers.

Graphical techniques are used in identifying the outliers. Graphs are generated portraying the basic characteristics of individual variables and relationships between variables in a simple picture. For example, a simple scatter plot represents in a single picture not only the two basic elements of a correlation coefficient, namely the type of relationship (positive or negative) and the strength of the relationship, but also is a simple visual means for assessing linearity that would require a much more detailed analysis if attempted strictly by empirical means (Zhang et al., 2004). After applying the min-max normalization to all the climatic parameters, scatter plot graph has been plotted between Temperature and various parameters, through which the outliers are identified and removed for processing it separately.
2.4.4 Implementing and Estimating Model

The selection and implementation of the appropriate data mining technique is the main task in this phase. Implementation is based on several models and selecting the best one is an additional task. Various techniques are applied to perform a successful learning process from the data to develop an appropriate model.

2.4.4.1 Data Mining Techniques

As described already, Data Mining is the process of discovering insightful, interesting and novel patterns, as well as descriptive, understandable, and predictive models from large scale data. Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Following techniques has been adopted to analyze the climatic data.

Fig. 2.2 Shows the Graphs depicting Outliers

Detailed other Outlier analysis is discussed with the data and the results in the Chapter 3.
2.4.4.1 Correlation Analysis

Correlation between two data objects is a measure of the linear relationship between the attributes of the objects (Peter et al., 2003). More precisely, Pearson’s correlation coefficient between two data objects, x and y, is defined by the following equation

\[
\text{Corr}(x,y) = \frac{\text{Covariance}(x,y)}{\text{Standard Deviation}(x) \ast \text{Standard Deviation}(y)} = \frac{S_{xy}}{S_x \ast S_y}
\]

Where we are using the following standard statistical notation and definitions:

\[
\text{Covariance} (x,y) = S_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \text{Mean}(x)) (y_k - \text{Mean}(y))
\]

\[
\text{Standard Deviation} (x) = S_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \text{Mean}(x))^2}
\]

\[
\text{Standard Deviation} (y) = S_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \text{Mean}(y))^2}
\]

Correlation is always in the range -1 to 1. A correlation of 1 (-1) means that x and y have a perfect positive (negative) linear relationship. If the correlation is 0, then there is no linear relationship between the attributes of the two data objects. Correlation analysis has been done for the Correlation between transformed values of Air Pressure Vs Temperature, Humidity Vs Temperature, Wind Speed Vs Temperature, etc., are computed and studied to identify which climatic parameter strongly correlates with Temperature. Chapter 4 discusses more in detail on the Correlation analysis with results.

2.4.4.1.2 Multiple Linear Regression Analysis

Multiple regressions are the appropriate method of analysis when the research problem involves a single metric dependent variable presumed to be related to two or more metric independent variables. The objective of multiple
regression analysis is to predict the changes in the dependent variable in response to changes in the independent variables (Zhang et al., 2004). This objective is most often achieved through the statistical rule of least squares.

The multiple linear regression model with two regressors is represented as follows

\[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \]

Where \( Y \) represents the target function computed based on the regressor variables

\( \epsilon \) is a random error term. The parameter \( \beta_0 \) is the intercept of the plane, \( \beta_1 \) and \( \beta_2 \) are partial regression coefficients. By the method of Least Squares, we have the following equations to find the regression coefficients.

\[
\begin{align*}
    n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} &= \sum_{i=1}^{n} y_i \\
    \hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} &= \sum_{i=1}^{n} x_{i1}y_i \\
    \hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 &= \sum_{i=1}^{n} x_{i2}y_i
\end{align*}
\]

We perform variable transformation of Air Pressure, Humidity, Temperature, etc., to have its corresponding values between the range \([0,100]\) by applying the min-max normalization method. Followed by the variable transformation, raw data analysis as described in previous section has been done and group 1 data (Jan – May) was taken up for further analysis. Outlier Analysis as described in the previous section has been done to ignore the outliers from the collected data to avoid its impacts on the mining methods.

Multiple regression analysis has been carried out for the group 1 data (Jan-May) to derive multiple linear equations for prediction of temperature. Results have been validated to identify the accuracy of the predicted temperature. Chapter 4 discusses more in detail on the Multiple Linear Regression with results.
2.4.4.1.3 Classification Technique

Classification is the task of assigning objects to one of several predefined categories, is a pervasive problem that encompasses many of diverse applications. The input data for a classification task is a collection of records. Each record, also known as an instance or example, is characterized by a tuple \((x, y)\), where \(x\) is the attribute set and \(y\) is a special attribute, designated as the class label (also known as category or target attribute) (Han and Kamber, 2006). The class label, on the other hand, must be a discrete attribute. This is a key characteristic that distinguishes classification from regression, a predictive modelling task in which \(y\) is a continuous attribute. To define formally, Classification is the task of learning a target function \(f\) that maps each attribute set \(x\) to one of the predefined class labels \(y\). The target function is also known informally as a classification model.

A classification technique (or classifier) is a systematic approach for building classification models from an input data set. Examples include decision tree classifiers, rule-based classifiers, etc. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data (Jain and Dubes, 1988). The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records it has never seen before. Therefore, a key objective of the learning algorithm is to build models with good generalization capability i.e., models that accurately predict the class labels of previously unknown records (Han and Kamber, 2006).

We consider climatic parameters like Air Pressure, Humidity, Wind Speed, Visibility, etc., to predict Temperature with the help of Decision Tree Algorithm and Naïve Bayes Classification Algorithms. Chapter 5 discusses more on these techniques and results have been compared across.
2.4.4.1.4 Clustering Analysis Technique

Cluster analysis is an analytical technique for developing meaningful subgroups of individuals or objects. Specifically, the objective is to classify a sample of entities (individuals or objects) into a small number of mutually exclusive groups based on the similarities among the entities (Pang-Ning Tan et al., 2005).

1. Steps for Cluster Analysis:
   - Measurement of some form of similarity or association among the entities to determine how many groups really exist in the sample.
   - Actual clustering process, whereby entities are partitioned into groups (clusters).
   - Profiling the variables to determine their composition.

2. Standardizing the Variables:

   The most common form of standardization is the conversion of each variable to standard scores (also known as Z-Score transformation) by subtracting the mean and dividing by the standard deviation for each climatic variable (MacQueen, 1967). Outlier analysis has been conducted as described earlier. Similarity measures have been studied with the help of Euclidean Distance measure between the dates of the collected weather data (Pang-Ning Tan et al., 2005).

3. Similarity Measure:

   Euclidean distance is the most commonly recognized measure of distance, many times referred to as straight-line distance. An example of how Euclidean distance is obtained is shown geometrically below (Pang-Ning Tan et al., 2005). The Euclidean distance between the points is the length of the hypotenuse of a right triangle, as calculated by the formula under the Fig. 2.3. This concept is easily generalized to more than two variables.
Chapter 2

Fig. 2.3 Shows the Euclidean distance measure

The Euclidean distance between the two points is the hypotenuse of the triangle ABC:

\[ D(i,j) = \sqrt{A^2 + B^2} = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2}. \]

4. Applying Clustering Techniques:

Nearest-Neighbor method defines the similarity between clusters as the shortest distance from any object in one cluster to any object in the other (Beyer et al., 1999). This rule was applied and enables us to use the original distance matrix between observations without calculating new distance measures. Just find all the distances between observations in the two clusters and select the smallest as the measure of cluster similarity (Kotuand Deshpande, 2014).

K-Means and Orthogonal Partitioning Clustering algorithms have been adopted for clustering the collected climatic data. Both the algorithms and the results are discussed elaborately in Chapter 6.

2.4.5 Interpreting Model And Drawing Conclusions

In most cases, data mining models should help in decision making. Hence such models need to be interpretable in order to be useful because humans are not likely to base their decisions on complex black box models (Kaufman and Rousseeuw, 1990). Usually simple models are more interpretable, but they are also less accurate. Modern data mining methods are expected to yield highly accurate
results using high dimensional models. The problem of interpreting these models, also very important, is considered a separate task, with specific techniques to validate the results.

### 2.5 DATAWAREHOUSING OF THE CLIMATIC DATA

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources (Jan et al., 2009).

In this research we use Oracle 11g Enterprise Edition Database for data warehousing the collected climatic data. Oracle objects like tables, global temporary tables, procedures, functions, cursors, etc., were created for analyzing the climatic data.