**Final Manuscript**

**Sravan Kumar Borra**

**Abstract**

Now days all the research has been concentrated on Big data and Internet of things (IoT). Big data is a vast concept used for analyzing and studying the data collected from different sources. IoT is concept about connectivity where everything is connected to internet. All these concepts are digitally revolutionized and the research is in nascent stage. One of the most important concepts which were based on IoT is Smart city.

Smart city are typically based on connectivity where everything is connected to internet. Insights for these cities are based on big data. The data is collected from different sources such as traffic, library events, pollution etc. All this data is collected and analyzed to provide better services to the people living in the city. Mainly insurance companies are looking forward to collect the data and analyze it.

Our aim in this project is to provide an efficient model which helps to build a Smart city. Smart city mostly runs on connectivity. This connectivity is established by using big data. We are using big data techniques, machine learning techniques, data mining techniques to build a model that will be used for the prediction of weather which helps in building a smart city. There are various machine learning techniques that are used for predicting the weather. We can use Support vector machines method; Decision trees techniques, neural networks.

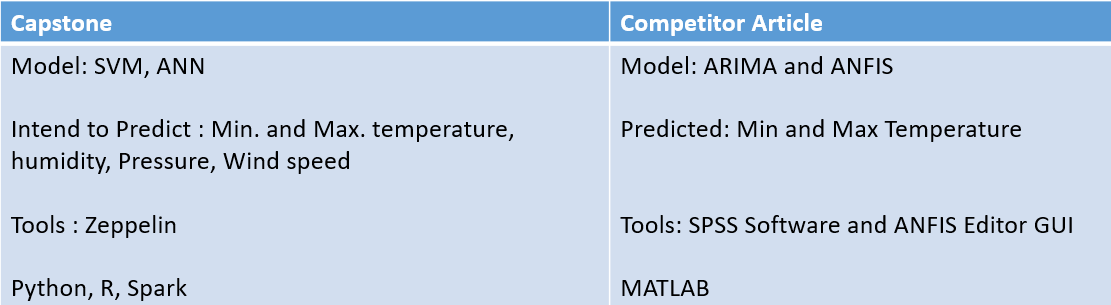
We can predict the climate changes seasonally. We can predict whether the summer season is going to be with high temperatures. We predict winter season whether it is going to have heavy snow fall or low temperatures so that people can be alerted if there is any problem. Natural calamities can be predicted if we can build a model using our machine learning technique methods. Now days it has become very important for detecting the blizzards and storms so that people can take shelters.

**Work by Competitors**

There are so many research papers which were already published based on this topic. All those papers are being used for predicting climatic changes and weather conditions. But this paper particularly concentrates on the results and models which will help to build a smart city. All the predictions and predictive models will help to build a smart city which will run based on internet and connectivity. There are so many papers that used techniques for predicting weather conditions especially machine learning techniques such as SVM, linear regression, neural networks etc.

SVM works on statistical learning and it is also known as a kernel machine. It is a machine learning algorithm used for classification purpose and regression purpose. If we have n observations in a data set, Support vector machines can be explained as the simple co-ordinates which is related to each and every observation. Support vector machines segregates very two important parts of a plot. They are hyper plane and line.

In an article named “An efficient approach for Weather forecasting using Support Vector Machines” used support vector machines technique to predict maximum and minimum temperatures for a particular day. This article has been published in International conference in 2012. In this paper we are taking inspiration from this particular technique to perform the predictive model. This forecasting helps in building the smart city efficiently and helps the people who live in those cities. By taking inspiration form this paper we are also going to use decision tree model and neural network model to predict the temperatures. As mentioned above we are dividing the data into two sets one is for training and other is for testing the data. This approach will help us in predicting accurate maximum and minimum temperatures.



In an IEEE article named “Unsupervised and supervised learning concept for 24-hour loading” used supervised or unsupervised learning techniques to predict the load and temperature for a given season. They used the algorithms which are supervised and help in predicting the temperature based on the 24 hour loadings of the day and these predictions are done seasonally. In this paper we are taking inspiration from this particular technique to perform the predictive model. This forecasting helps in building the smart city efficiently and helps the people who live in those cities. This can help to predict the seasonal changes in hourly timings. If we take USA into consideration Day time saving will affect the behavior of temperatures etc. This type of prediction is very useful for building smart cities.

In an IEEE paper named “Incremental learning for place recognition in dynamic environments” explains about the learning techniques which are used for a city in different learning techniques. For building a smart city, these types of predictions are very useful. To build a smart city we need very good weather conditions where there is seasonal rainfall which helps for farmers and agriculture sector of that city. There should not be a frequent change in weather conditions which can destroy the seasonal crops.

In an article named “Modeling and Forecasting the Daily Maximum Temperature Using Abductive Machine Learning” used AML to predict the maximum and minimum temperatures of a city in a particular day. Here AML is a different machine learning technique that has been used for the prediction purpose. We are using Zeppelin as a platform to perform all our predictions and using Spark language. Our ways of technique is very unique and predict more accurately. We are using our test data and training data to predict the temperatures of a particular day. This will help us in finding more accurate results. Accurate results will help us to build most efficient smart city.

In an IEEE article named “Unsupervised and supervised learning concept for 24-hour loading” used supervised or unsupervised learning techniques to predict the load and temperature for a given season. They used the algorithms which are supervised and help in predicting the temperature based on the 24 hour loadings of the day and these predictions are done seasonally. In this paper we are taking inspiration from this particular technique to perform the predictive model. This forecasting helps in building the smart city efficiently and helps the people who live in those cities. This can help to predict the seasonal changes in hourly timings. If we take USA into consideration Day time saving will affect the behavior of temperatures etc. This type of prediction is very useful for building smart cities.

An article named “Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction” used most of the complex hybrid models to predict the climatic changes and change in weather conditions. They used deterministic machine learning techniques to build the predictive models and to perform the techniques. They produced phenomenal results of predictions and that model help us to take the inspiration and build a similar model. We can predict most of the data temperatures and climatic changes by replicating and producing the same predictive model. The results of this model are more accurate than the original model. The accuracy of the model helps in building efficient smart city. Building efficient smart city requires perfect statistical models which predict accurate climatic changes and helps in predicting temperature changes.

**Contribution**

Our aim in this project is to provide an efficient model which helps to build a Smart city. Smart city mostly runs on connectivity. This connectivity is established by using big data. We are using big data techniques, machine learning techniques, data mining techniques to build a model that will be used for the prediction of weather which helps in building a smart city. There are various machine learning techniques that are used for predicting the weather. We can use Support vector machines method; Decision trees techniques, neural networks.

In competitor article they performed ARIMA and ANFIS statistical methods. We are using SVM and ANN. We out performed them as these statistical models gives better results. In the competitor article, they predicted only Maximum and minimum temperature. We are predicting Max and Min temperature, wind speed on a particular day, dew point at a particular time. The tool used in competitor article is SPSS software. We used Zeppelin in our project. We again outperformed in the selection of tools as Zeppelin if far more efficient than SPSS. For excluding null values their code is lengthy and not accurate.

All these methods are used in machine learning for predicting using the given data set. SVM works on statistical learning and it is also known as a kernel machine. It is a machine learning algorithm used for classification purpose and regression purpose. If we have n observations in a data set, Support vector machines can be explained as the simple co-ordinates which is related to each and every observation. Support vector machines segregates very two important parts of a plot. They are hyper plane and line. For this particular paper, we selected weather data of Aarhus city located in Denmark to predict the climatic changes. This prediction model helps in building a Smart city which runs on connectivity.

The techniques mentioned above are used in machine learning. For using those techniques we need big data set. To perform those machine learning techniques we divide the data set into two parts. One part is used for testing data where the machine learning technique is applied to the data set. Using the prediction results obtained from training set we will run a prediction on the remaining data set which is known as test data set. We already know the results for the test data set, so if the predicted results match with the original set of results then we can use our model to predict the weather for upcoming years.

We can predict weather or climatic changes for quarterly, yearly, seasonally, daily time periods for a year. We can compare our results after testing the data set with the original data set. This comparison is done by using a set of topics or metrics such as MAE and MSE. MAE is known as Mean absolute error and MSE is known as Mean square error. Among the machine learning techniques, the predictions done by using SVM (Support Vector Machines) technique turns out to be more accurate and reliable. We are comparing our predicted results with the results obtained from multiple linear regression models. It is a machine learning technique which is used for every type of prediction for comparing the results. Forecasting accurate weather conditions can also help in transportation. For flights and airplanes which are flying mostly in air perfect weather conditions are must and important. If we forecast the weather in advance it helps the aviation department to plan according and it also helps in damage control.

The main idea behind this project is to predict the temperatures to their maxima and minima for a particular day. We are going to predict wind speed and humidity for a particular day. Predicting climatic changes seasonally based on the data that we already have. We can also predict some of the areas in a city or a place where the solar energy production is suitable. We can also find the places which are helpful for producing wind energy. These predictions are done by using wind speed data. These are natural energies which are helpful for building a smart city as the city is going to be pollution free city. We can also predict the amount of rainfall that is going to come for a particular season which helps the farmers.

All the data sets provided are in JSON format. So I cleaned the data and saved to CSV file. There are sevendifferent data sets in weather data. So cleaning the data for each data set will consume time and makes the process slow. So write a loop for loading the data set where the code is optimized and took less than half of the time for loading the data separately. This makes my code optimized and efficient.

While plotting time series, the values presented in the data are not sorted. So the results are not accurate and it will not give sorted values. So I separately sorted the data set in CSV file separately and converted the strings into Date Time format. All this data manipulation will be useful for time series plotting and performing regression, SVM in the future.

We had the correlation results based on the experiments done in class using zeppelin. I performed precision correlation method to find the correlation between the variables. It shows that Temperature and density highest percentage of correlation.

**Data**

Smart city are typically based on connectivity where everything is connected to internet. Insights for these cities are based on big data. The data is collected from different sources such as traffic, library events, pollution etc. All this data is collected and analyzed to provide better services to the people living in the city. Mainly insurance companies are looking forward to collect the data and analyze it.

A data set with weather observations has been selected for this project. Data set observations are collected from the city of Aarhus in Denmark. This data set has been fetched form City pulse website, which is trying to provide models to build smart city. Data set has different types of fields such as Dew point, humidity, pressure, temperature, wind direction, wind speed. All these parameters are used for predicting weather conditions in different times of the year. We have two sets of data. One is from February-June 2014 and other set is from August-September 2014. These data sets were provided in JSON file format. We use Zeppelin platform to study the data. We initially load the data and study about the different fields in the data.

This connectivity is established by using big data. We are using big data techniques, machine learning techniques, data mining techniques to build a model that will be used for the prediction of weather which helps in building a smart city. There are various machine learning techniques that are used for predicting the weather. We can use Support vector machines method; Decision trees techniques, neural networks. All these methods are used in machine learning for predicting using the given data set. For this particular paper, we selected weather data of Aarhus city located in Denmark to predict the climatic changes. This prediction model helps in building a Smart city which runs on connectivity. There are five scenarios in understanding Support Vector Machines. In scenario one, or in step one we need to first classify the hyper plane. A plot can have two or more hyper planes. We need to classify those hyper planes and we have to select the best two planes which can classify our observation given in data.

Weather data set of Aarhus has seven sub data sets. They are humidity, pressure, wind speed, wind direction, temperature etc. Each data set contains the readings from February 2014 to June 2014 and august 2014 to September 2014. All these data sets were provided in JSON format. We are using Zeppelin to perform our project and predict the weather conditions. We are going to build a model which will help us to predict the weather for Aarhus city in Denmark. We are going to use Artificial Neural Networks and Support vector machines to predict the weather for the Aarhus city.

We can compare our results after testing the data set with the original data set. This comparison is done by using a set of topics or metrics. MAE is known as Mean absolute error and MSE is known as Mean square error. Among the machine learning techniques, the predictions done by using SVM (Support Vector Machines) technique turns out to be more accurate and reliable. We are comparing our predicted results with the results obtained from multiple linear regression models. It is a machine learning technique which is used for every type of prediction for comparing the results.

**Method**

Our aim in this project is to provide an efficient model which helps to build a Smart city. Smart city mostly runs on connectivity. This connectivity is established by using big data. We are using big data techniques, machine learning techniques, data mining techniques to build a model that will be used for the prediction of weather which helps in building a smart city. There are various machine learning techniques that are used for predicting the weather. We can use Support vector machines method; Decision trees techniques, neural networks. All these methods are used in machine learning for predicting using the given data set. For this particular paper, we selected weather data of Aarhus city located in Denmark to predict the climatic changes. This prediction model helps in building a Smart city which runs on connectivity.

Artificial Neural Networks and Support vector Machines are the methods used in Data mining and Machine learning. We are going to use those techniques to predict the maximum and minimum temperatures for a particular day, for a particular hour a day. We are also going to predict the humidity scale for a particular day. We are going to predict the wind speed and direction of a particular day.

We have two sets of data. We use one of the data set to train the model and use another set to test the model and we can also test the model by predicting for upcoming days. Data Cleaning, implementation of methods on data set are going to be performed in Zeppelin platform. We are going to use Python language in Zeppelin to clean the data and run the methods on data set.

The techniques mentioned above are used in machine learning. For using those techniques we need big data set. To perform those machine learning techniques we divide the data set into two parts. One part is used for testing data where the machine learning technique is applied to the data set. Using the prediction results obtained from training set we will run a prediction on the remaining data set which is known as test data set. We already know the results for the test data set, so if the predicted results match with the original set of results then we can use our model to predict the weather for upcoming years.

* **Support Vector Machines**

SVM works on statistical learning and it is also known as a kernel machine. It is a machine learning algorithm used for classification purpose and regression purpose. If we have n observations in a data set, Support vector machines can be explained as the simple co-ordinates which is related to each and every observation. Support vector machines segregates very two important parts of a plot. They are hyper plane and line.

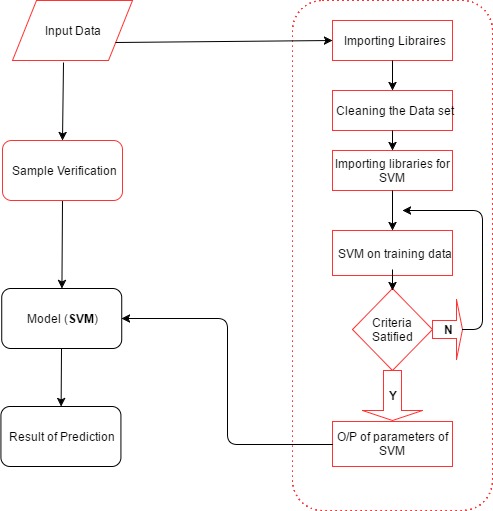
There are five scenarios in understanding Support Vector Machines. In scenario one, or in step one we need to first classify the hyper plane. A plot can have two or more hyper planes. We need to classify those hyper planes and we have to select the best two planes which can classify our observation given in data.

In scenario 2, we might have lines and hyper planes. All the lines are segregating the planes equally and well. Now we have to select which hyper plane is better to segregate our observations. Now we have to select a plane by using the distance which is known as Margin. The distance has to be maximized between the hyper planes and nearest data point. The closest hyper planes can be excluded from the equation and we have to select the planes which is not so nearest to the observations when compared to the other two hyper planes. There is another reason to select the hyper plane which has margined higher than other two planes is robustness. Miss classification might be a problem if we select the hyper plane with low margin and this will affect our prediction.

In scenario 3, we need to understand that selecting a hyper plane based on its margin is not only the reason to select. A hyper plane should classify the classes equally. If a hyper plane having higher margin but not classifying the classes as efficiently as other hyper plane, then it is not right hyper plane to select. A hyper plane which classifies the classes correctly and have has better margin when compared to the other hyper planes is better for selection.

In scenario 4, we will learn about outliers. If the observations have outliers then it will be difficult to classify by using hyper plane. It is because we cannot find a plane which can classify the classes accurately. Support vector machines have a feature to eliminate and ignore the outliers. It will exclude the outliers out of the equation and divide the classes into hyper plane ignoring the outliers.

In scenario 5, we will see about the classes where we cannot divide those using linear hyper planes. Support vector machines will form an equation where the distances are squared. The new equation will be the sum of square of both distances. The trick of forming a new equation is known as Kernel trick. That is the reason why support vector machines are also known as Kernel method. It is an inbuilt function in support vector machines. These types of methods will make the lower dimensional space to a higher dimensional which will help us to classify the classes using linear classification or linear hyper planes.



We can predict weather or climatic changes for quarterly, yearly, seasonally, daily time periods for a year. We can compare our results after testing the data set with the original data set. This comparison is done by using a set of topics or metrics such as MAE and MSE. MAE is known as Mean absolute error and MSE is known as Mean square error. Among the machine learning techniques, the predictions done by using SVM (Support Vector Machines) technique turns out to be more accurate and reliable. We are comparing our predicted results with the results obtained from multiple linear regression models. It is a machine learning technique which is used for every type of prediction for comparing the results.

There are several ideas which are useful to predict the weather conditions, climatic changes, forecasting the weather which can be used for daily predictions, half yearly predictions. We can predict the temperatures changes such as maximum temperature of a particular day and minimum temperature of a particular day. We can also predict the weather condition for a particular day. We can predict whether the day is going to be sunny or it is going to rain. We can find out or predict the wind speed for a specific day so that people can take safety measures if there are high speed winds which can even destroy the buildings.

We can predict the climate changes seasonally. We can predict whether the summer season is going to be with high temperatures. We predict winter season whether it is going to have heavy snow fall or low temperatures so that people can be alerted if there is any problem. Natural calamities can be predicted if we can build a model using our machine learning technique methods. Now days it has become very important for detecting the blizzards and storms so that people can take shelters.

We had the correlation results based on the experiments done by using zeppelin. I performed precision correlation method to find the correlation between the variables. It shows that Temperature and density highest percentage of correlation. Our aim in this project is to provide an efficient model which helps to build a Smart city. Smart city mostly runs on connectivity. This connectivity is established by using big data. We are using big data techniques, machine learning techniques, data mining techniques to build a model that will be used for the prediction of weather which helps in building a smart city. There are various machine learning techniques that are used for predicting the weather. We can use Support vector machines method; Decision trees techniques, neural networks.

For building a smart city, these types of predictions are very useful. To build a smart city we need very good weather conditions where there is seasonal rainfall which helps for farmers and agriculture sector of that city. There should not be a frequent change in weather conditions which can destroy the seasonal crops. Our model is useful to find a place where the weather conditions are suitable for building a smart city. We can use the weather data of particular places to study and predict the weather changes and prepare a report about climatic changes in future for that particular area.

We can also build a model which is useful after building a smart city. We can predict the climatic changes, prediction of blizzards and storms. Due to that prediction one can take safety measures which helps in decreasing the losses which can be caused by a storm or blizzard. Now days due to pollution, the climate is changing without any warning. In those conditions these types of predictive models will help for damage control.

**Results**

We had the correlation results based on the experiments done by using zeppelin. I performed precision correlation method to find the correlation between the variables. It shows that Temperature and density highest percentage of correlation. Our aim in this project is to provide an efficient model which helps to build a Smart city. Smart city mostly runs on connectivity. This connectivity is established by using big data. We are using big data techniques, machine learning techniques, data mining techniques to build a model that will be used for the prediction of weather which helps in building a smart city. There are various machine learning techniques that are used for predicting the weather. We can use Support vector machines method; Decision trees techniques, neural networks.

|  |  |  |  |
| --- | --- | --- | --- |
| **Non-Seasonal Lags** |  | **Estimated** | **Standard Error** |
| AR1 | 1.553 | 0.026 |
| AR2 | -0.559 | 0.025 |
| MA | 0.867 | 0.017 |

The results provided above are the results obtained by applying ARIMA on the data set. It explains the estimated and standard error obtained by applying ARIMA.

|  |  |  |
| --- | --- | --- |
| **Criteria** | **SVM** | **ARIMA** |
| SSE | 2.0151 | 0.2133 |
| R^2 | 1.0151 | 0.7867 |
| RMSE | 5.0808 | 1.6594 |
| MAE | 4.1040 | 1.2008 |

Above results explains the comparison of results between the two methods we have applied on our data set. We have applied Support vector machines and ARIMA on our data set. We have compared the results for SSE, R^2, RMSE and MAE. By comparing the table we can observe that the value of SSE is better when compared to the SSE value of ARIMA method. This explains that the method we chose is better than the method applied by the competitor. When we compare the R^2 values of SVM and ARIMA the results are better with the SVM and here also the method applied by us is better than the competitors method. If we compare RMSE and MAE values we can observe that the values obtained by SVM method are better than the values obtained by ARIMA. By this we can conclude that SVM method results are much better than the results obtained by SVM.

**Discussion**

In competitor article they performed ARIMA and ANFIS statistical methods. We are using SVM and ANN. SVM works on statistical learning and it is also known as a kernel machine. It is a machine learning algorithm used for classification purpose and regression purpose. If we have n observations in a data set, Support vector machines can be explained as the simple co-ordinates which is related to each and every observation. Support vector machines segregates very two important parts of a plot. They are hyper plane and line.

We out performed them as these statistical models gives better results. In the competitor article, they predicted only Maximum and minimum temperature. We are predicting Max and Min temperature, wind speed on a particular day, dew point at a particular time. The tool used in competitor article is SPSS software. We used Zeppelin in our project. We again outperformed in the selection of tools as Zeppelin if far more efficient than SPSS. For excluding null values their code is lengthy and not accurate. It consumes more time. In our article, it took almost less than 10 lines of code to exclude null values as we are using spark. Pearson Correlation for temperature and dew point- more than 70%. Our data has less outlier as we excluded the null values before performing the statistical analysis. The regression value of temperature and humidity grouped by pressure is -7.500000 and the Intercept is 123.500. We still do not have the R^2 value for our data as we did not performed SVM on our data. The work is still going on. In the competitor article, people have not performed correlation on their data. So they don’t know the relation between their data variables. We performed correlation on our data. We have a 70% relation between temperature and dew point. With this result we can predict the temperature based on dew point variable.

**Conclusion**

We have predicted the wind speed and humidity for a particular day. Predicting climatic changes seasonally based on the data that we already have. We can also predict some of the areas in a city or a place where the solar energy production is suitable. We can also find the places which are helpful for producing wind energy. These predictions are done by using wind speed data. These are natural energies which are helpful for building a smart city as the city is going to be pollution free city. We can also predict the amount of rainfall that is going to come for a particular season which helps the farmers.

Forecasting accurate weather conditions can also help in transportation. For flights and airplanes which are flying mostly in air perfect weather conditions are must and important. If we forecast the weather in advance it helps the aviation department to plan according and it also helps in damage control.

We can also build a model which is useful after building a smart city. We can predict the climatic changes, prediction of blizzards and storms. Due to that prediction one can take safety measures which helps in decreasing the losses which can be caused by a storm or blizzard. Now days due to pollution, the climate is changing without any warning. In those conditions these types of predictive models will help for damage control.