IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# TEMPERATURE PREDICTION USING TIME SERIES MODELLING

[1]Khushi Bhuwania, [2]Hitesh Wadhwa, [3]Kavan Vasani, [4]Dr. Mukesh Israni
[1]Student, [2]Student, [3]Student, [4]Associate Professor
[1 2 3]Computer Engineering, [4]Information Technology
[1 2 3 4]Thadomal Shahani Engineering College, Mumbai, India

***Abstract:*** Temperature prediction is an attempt to forecast the temperature of a particular region at a particular time. Temperature prediction is very important as it helps to plan various activities. A prime example of its application is in the field of agriculture. Knowing the range of temperature on a particular day can help a farmer prepare for the next day in advance and plan his activities to obtain optimal output. Temperature forecasts are required by utility companies because they help in estimating demand [1]. An example would be in helping people decide the clothes they should wear on a particular day and hence predict the upcoming demand. If a person expects a sudden drop in temperature, he/she can dress accordingly. On top of that, it helps in various aerial activities such as planning flights and optimizing air traffic. It can be applied to several other areas of interest including marine and military applications. Hence it becomes an essential part of our daily lives. This paper uses four different Machine Learning algorithms - Linear regression, KNN, RandomForest, Neural network - to predict temperatures at a particular time. Proposed algorithms are evaluated using standard evaluation metrics. The results of all the algorithms are analysed and the paper also discusses the performance of the proposed algorithm. The goal of this paper is to accurately predict the temperature in a particular region at a particular time, given the temperature history of that region over the past few days.

***Index Terms*** **– KNN (K-Nearest Neighbors), MLP (Multi-layer Perceptron), Linear Regression, RandomForest, DateTime, $R^2$ Score.**

## I. INTRODUCTION

Temperature is a prominent parameter for various needs in the industrial, environmental and agricultural sectors among various other parameters, including humidity, precipitation and many others. The constant changes in climate have directly affected the yields of various crops, and global warming has greatly impacted our ecosystem making weather forecasting and temperature prediction an essential part of the informed decision-making process. The prediction of atmospheric parameters is vital for various applications. It covers monitoring of the weather, ascertaining droughts, severe weather detection, agriculture and production, preparation in aviation industry, energy industry, pollution, promulgation, communication, etc. [2]

Due to the dynamic nature of the atmosphere, accurate weather parameter prediction is challenging [3]. The sheer amount of air on this planet makes it nearly impossible to predict the temperature. On top of that, the non-uniformity and the dynamicity of the land and water surfaces makes it even more difficult. The change in even a tiny particle has the capability to change all the air in our atmosphere which further changes the pH level, moisture, dew points, etc. making the task of predicting temperature using humidity, pressure, etc. just another failed attempt.

So therefore, in this paper we are moving from the usual method and trying Time Series modelling for the prediction of the upcoming temperature. In this method, temperatures of the past few days in a given time are taken and passed to the machine learning model which will output the temperature of the next day. The number of past days that will be considered as the input will also be a feature to see the window of past data which can give us optimum results.

## II. METHODOLOGY

### A. Dataset Used

The dataset used in this paper is obtained from government sites. It has temperature records from January 2009 to January 2020. The dataset has over 20 features like the time of sunset, sunrise, visibility, temperatures for the past days and many more that could potentially help predict the temperature for the coming days.

## B. Proposed System

Below are the steps followed for cleaning and pre-processing raw data from different sources to gain the final data ready for applying machine learning models:

Step 1: Setting DateTime as the Index of our dataset.
Step 2: Cloning DateTime and Hourly Temperature variables into a new dataset and performing a downshift of 24 on the temperature column to gain yesterday's temperature and pasting it into a new column.
Step 3: Temperature difference to be obtained by performing a subtraction between adjacent rows of the 'Yesterday' column. All the rows with NA as value are dropped.
Step 4: Repeat Step 2 and 3 for the desired number of times depending on the number of days or the previous hour temperature that should affect the model. This step helps in selecting the window of past data that should be used to make predictions.
Step 5: All the different dataset are then sliced into a Training Set for the values till 2018 and Test set for values starting from 2019.
Step 6: All the training sets that are obtained, are passed through the train function which fits the model through the different machine learning algorithms returning graphs of time taken by each model and the accuracies by passing it through a cross-validation set.
Step 7: The best model is therefore selected based on the graphs and hyperparameter tuning using Grid Search is applied to that model.
Step 8: The model is then passed through the test set to provide predictions for years 2019-2020 and outputs the graph of the comparison between predictions and real values.

## III. MODELS USED:

1. **Linear regression** is a measurable test applied to a set of data to characterize and evaluate the correlation between the presumed variables. The correlation gives a quantitative approach to estimate the degree or strength of a relation between two factors. Regression study numerically portrays the relationship. Linear Regression permits anticipating the value of a dependent variable based on the value of atleast one independent variable. The utilization of Linear Regression model is significant for accompanying reasons: a. Descriptive - It helps in investigating the strength of the relationship between the result (subordinate variable) and indicator factors b. Adjustment - It adapts with the impact of covariates or the confounders. c. Predictors – It helps in analysing the essential risk factors that make an impact on the dependent variable. d. Extent – It helps in scrutinizing the extent of change in the prediction as to how modifying the independent variable by one "unit" would change the dependent variable. e. Prediction – It helps in estimating the new cases [4].

2. **Random Forest** is a "Tree"- based calculation that utilizes the characteristic elements of various Decision Trees for making decisions. Therefore, it very well may be alluded to as 'Forest' of trees and subsequently the name "Random Forest". The term 'Random' is because of the way that this calculation is a group of 'Arbitrarily made Decision Trees'. The Decision Tree calculation has a significant inconvenience in that it causes overfitting. This issue can be restricted by executing Random Forest Regression instead of the Decision Tree Regression [6]. Moreover, the Random Forest calculation is likewise extremely quick and more powerful than other regression models. Decision trees construct relapse or grouping models looking like a tree. They separate the information into subsets and simultaneously another decision tree, in light of the first one, is progressively evolved. The outcome is a tree which has decision nodes and leaf nodes. Leaf nodes address choices on a specific mathematical objective. The decision node that is at the top of a tree compares to the best indicator, called the root node. For each decision tree, the significance of a node is determined by utilizing Gini Importance.

3. **Multilayer Perceptrons (MLPs)** are the traditional kind of neural networks. They involve at least one layer of neurons. Information is taken to the input layer. There might be one or greater hidden layers giving degrees of abstraction. The output layer shows predictions, additionally called the visible layer. MLPs are needed for classification problems where input data is relegated to a class or label. They are additionally reasonable for the prediction of regression problems where a real world prediction is anticipated given a bunch of information. Data in the form of input is frequently given in a tabular configuration, for example, the one you would find in a CSV record or a spreadsheet. One can use MLPs for: 1.Datasets in the form of a table 2.Classification problems 3.Regression problems. They are truly adaptable and can be utilized by various domains. This adaptability permits them to be applied to different kinds of information. For instance, the pixels of a picture can be decreased down to one long line of information and put into a MLP [7]. The expressions of a record can likewise be decreased to one long column of information and fed to a MLP. Indeed, even the lag perceptions for a time series prediction problem can be decreased to a long line of information and provided to a MLP. All things considered, assuming your information is in a structure other than tabular form, like a picture, a document, or time series, it would be suggested testing a MLP on one's problem. One can therefore use MLPs on photos, text data, time series data, etc.

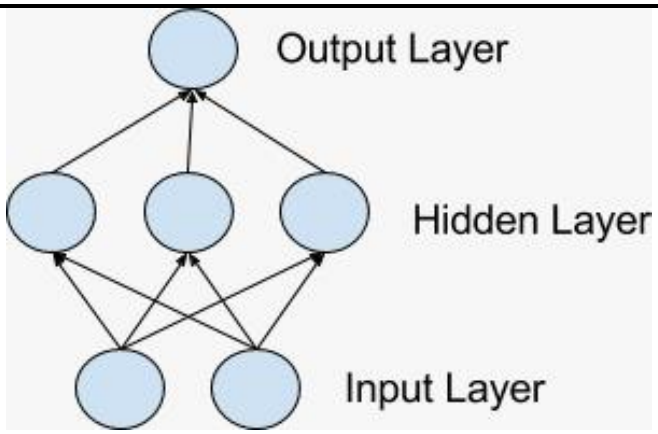

Fig 1 Basic MLP architecture

4. **K-Nearest Neighbors (KNN)** is a machine learning method and algorithm that can be utilized for both regression and classification tasks. K-Nearest Neighbors looks at the tags of a picked number of datapoints encompassing an objective element, to make an expectation about the class that the data point falls into. K-Nearest Neighbors (KNN) is a thoughtfully straightforward yet extremely strong calculation, and consequently, it's one of the most famous Artificial Intelligence algorithms [8]. For an information record T to be arranged, its k closest neighbors are recovered, and this frames a neighborhood of T. Majority-based voting amongst the neighborhood of T is done to classify in which group data point T will fall. This voting is actually the distance between the elements of a group. Nonetheless, to apply KNN we must decide a feasible value for k, and the accomplishment of grouping relies a lot on this value. Essentially, KNN calculation falls under the Supervised Learning class and is utilized for grouping and regression. It is a flexible algorithm likewise utilized for crediting missing values and resampling datasets. As the name (K Nearest Neighbor) proposes it presumes K Nearest Neighbors (Data points) to foresee the class or continual value for the new datapoint. The algorithm's learning is:
a. Instance-based learning: In this we don't learn weights from data which is trained to predict the output, however, it utilizes whole training data to foresee the output for unprecedented information.
b. Lazy Learning: Model isn't picked up utilizing training data earlier and the learning system is deferred to when prediction is required on the new example.
c. Non - Parametric: In KNN, there is no predetermined type of mapping function. Model portrayal for KNN is the whole training dataset.

## IV. RESULTS

Temperatures are predicted using the aforementioned machine learning models. The results of these models are then compared using a standard evaluation metric $R^2$ score.

### $R^2$ Score:

$R^2$ Scores are used to compute how well the values predicted fit the model which is calculated using the equation given below:

$$R^2 = 1 - \frac{SS_{regression}}{SS_{total}}$$

where, $SS_{regression}$ indicates the summation of squares of the regression result, and $SS_{total}$ indicates the complete sum of those squares [5].

### Feature and Model Selection:

An important feature to be decided on was the window of days considered for making predictions. In this paper, three different approaches are tried - using the past day, the past two days and the past 2 days along with the past hour.

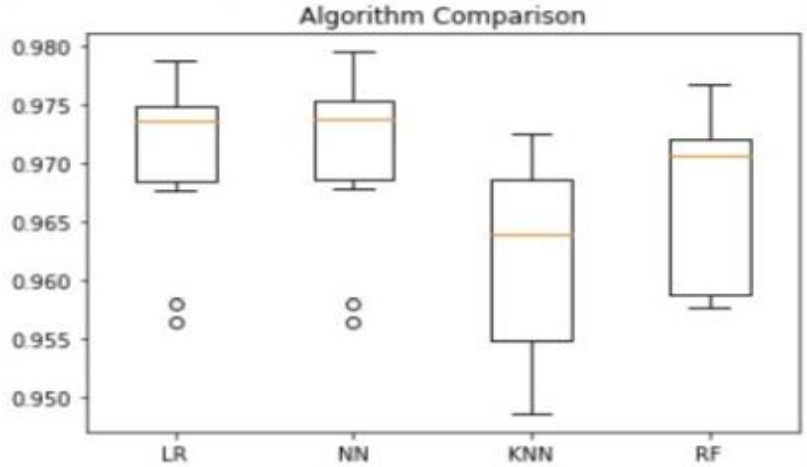


Fig 2 Comparison over the past 2 days along with the past hour

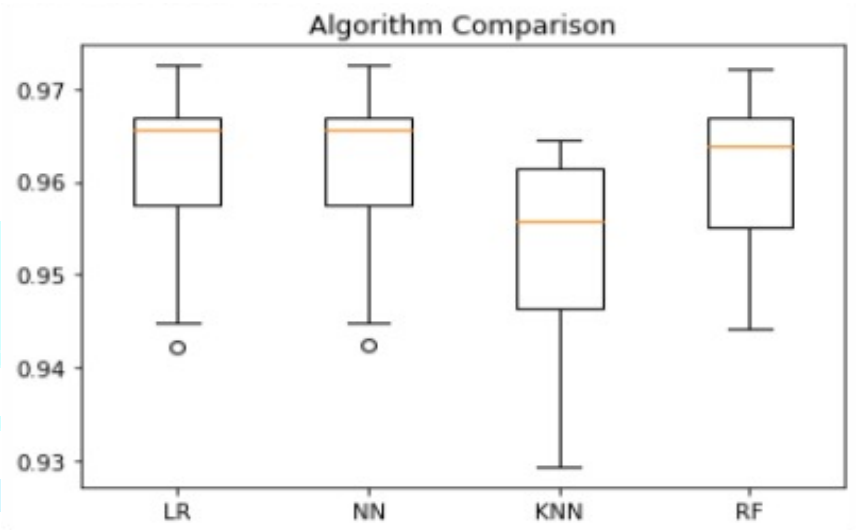


Fig 3 Comparison over the past day

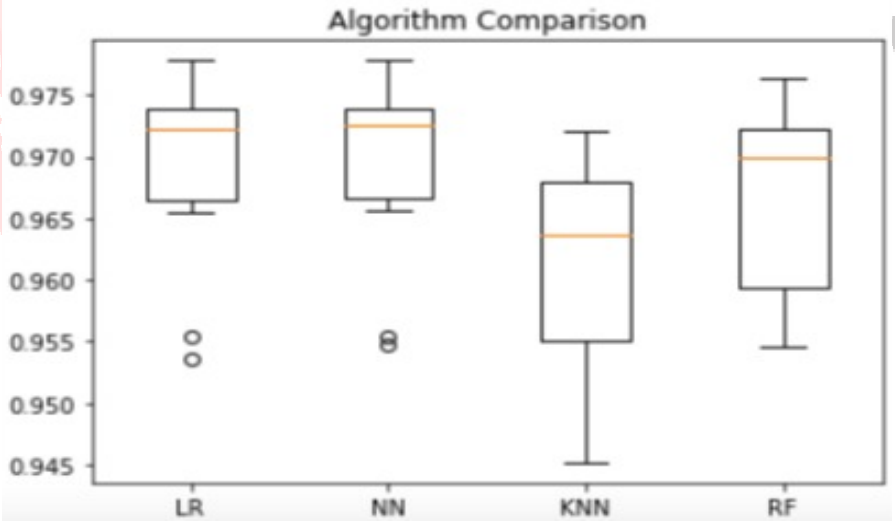


Fig 4 Comparison over the past 2 days

Based on these graphs, it can be concluded that the models work the best for a window containing the data of the past two days and the past hour. And using this window, it is evident that the MLP (or NN) is the model that gives the most accurate results. But the problem with using neural networks is that it is inefficient. On the other hand, if the results of Random Forest are considered, $R^2$ scores are very similar, but it uses a lot less time.

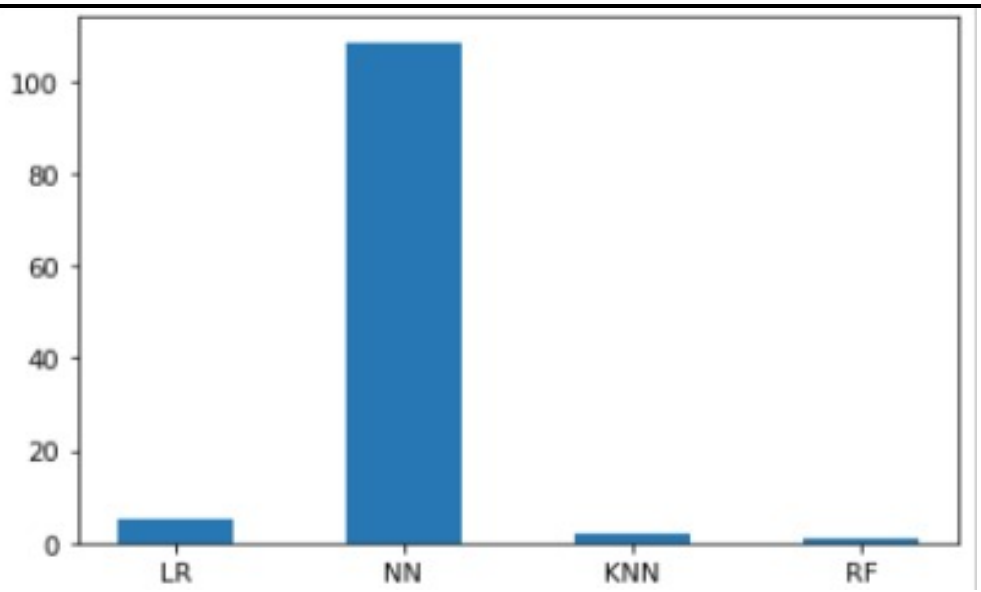


Fig 5 Comparison of time taken by algorithms

So considering the trade-off between the time needed to make the predictions and the accuracy of the results, Random Forest is the most useful model out of all the models considered in this paper. It predicts the temperatures with an $R^2$ Score of 0.967. Below graph compares the predicted values with the actual temperature values. This shows how accurately the Random Forest model predicts the temperatures of the upcoming dates.

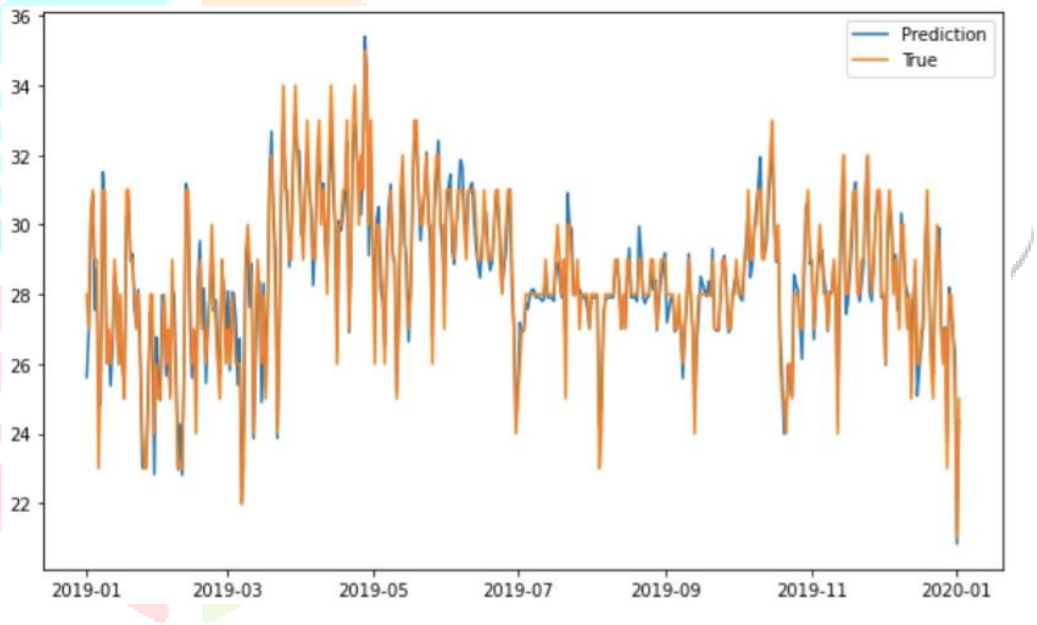


Fig 6 Graphical representation of the predicted and true values

## V. CONCLUSION

While during temperature prediction, common features like humidity, precipitation, rainfall, etc. are used, this paper aims to predict temperatures solely based on the previously recorded temperature points. This paper uses four different machine learning algorithms, namely, Linear Regression, Multi-Layer Perceptron, Random Forest Regressor and K-Nearest Neighbors. This paper performed regression analysis on these algorithms and found that Random Forest Regressor performed most efficiently on the considered dataset. Neural Network performed better than Random Forest Regressor, but was unable to give results in a feasible amount of time. Further research can be done by combining historical data with commonly used features for temperature prediction in tandem to try and predict temperatures for a larger time period.

## VI. REFERENCES

**[1]**https://en.wikipedia.org/wiki/Weather_forecasting#:~:text=Temperature%20forecasts%20are%20used%20by,wear%20on%20a%20given%20day.

**[2]** Neha Karna, Prem Chandra Roy, Subarna Shakya, "Temperature Prediction using Regression Model", Conference: 7th Online International Conference on Advanced Engineering and ICT-Convergence 2021 (ICAEIC-2021)

**[3]** JA Syeda. "Variability Analysis and Forecasting of Relative Humidity in Bangladesh", Journal of Environmental Science and Natural Resources, 2013

**[4]** Kumari, Khushbu & Yadav, Suniti. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences. 4. 33. 10.4103/jpcs.jpcs_8_18.

**[5]** Stephen Gbenga Fashoto, Elliot Mbunge, Gabriel Ogunleye and Johan Van den Burg, "Implementation of Machine Learning for Predicting Maize Crop Yields using Multiple Linear Regression and Backward Elimination", 2021 Malaysian Journal of Computing.

**[6]** Liaw, Andy & Wiener, Matthew. (2001). Classification and Regression by RandomForest. Forest. 23.

**[7]** Marius, Popescu & Balas, Valentina & Perescu-Popescu, Liliana & Mastorakis, Nikos. (2009). Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems. 8.

**[8]** Guo, Gongde & Wang, Hui & Bell, David & Bi, Yaxin. (2004). KNN Model-Based Approach in Classification.