|  |  |
| --- | --- |
| How IoT infrastructure allows for more accurate weather forecasting | Perle  News  **[Weather forecasting using machine learning models]**  [Research proposal] | Dr. Mohammad yahya  [Research mentor]  AHMAD OBEID  [Researcher] |

Table of Contents

**Introduction2**

**Literature review3-5**

**Research objectives & questions6**

**Research methodology6-**

Research approaches6-7

Saunders’ research onion7-8

Flowchart9

Data collection10-18

Data analysis13-20

Result20-21

**Conclusion & recommendations22**

**References23-24**

**Introduction**

The world has become connected after the internet, along with it the computers were invented with memories in it to store the data. After this, almost everything has been digitalized and the amount of data in the world has increased nine times within a short period of time, so we started to deal with huge amounts of data daily. Big data refers to enormous, complicated, and varied data sets that are challenging to store, analyze, and visualize for use in subsequent procedures or outcomes. Big data analytics is the process of analyzing data to uncover hidden patterns and obtain better understanding. Big data implementations are important for this reason, and they require the most precise analysis. (IEEE, 2013)

Five key phrases that encapsulate the characteristics of big data are useful for describing big data. The first word, volume, describes the amount of data. The second phrase, variety, describes the sources and types of data, while the third term, velocity, refers to the speed at which data is sent and received. Veracity, which stands for both the messy and reliable nature of data, was included by IBM and Microsoft as the fourth V in their definition of big data. The term "value," which describes the value of undiscovered thoughts within big data, was also added by McKinsey & Co. (Yaqoob, 2016)

Big data entered almost every field in life, since all devices and sensors record data each minute and second. Weather forecasting is one of the fields that the big data is involved in deeply and it’s so important to predict the weather status for some region. Accurately predicting air temperature promotes efficient and waste-free use of resources including electricity, water, and agricultural inputs.

Large volumes of data from many sources, such as weather stations and satellite imaging, must be analyzed to produce weather forecasts. By allowing for the real-time analysis of this data and the integration of high-resolution data from numerous sources, big data analytics improves forecast accuracy. Big data facilitates the use of machine learning algorithms in weather forecasting, which may find patterns, correlations, and trends in massive datasets to better comprehend intricate atmospheric interactions and produce more accurate forecasts.

As I mentioned earlier, conducting precise analysis on a dataset will enable us to investigate the hidden patterns and insights. By doing this, we can extract useful information from the weather dataset, recognize the erratic patterns of climatic conditions, and then predict the air temperature using a machine learning model (IEEE, 2018).

**Purpose statement**

The purpose of this research is to create a machine learning model for weather forecasting that can predict air temperature. The objective of this research is to ascertain the optimal regression-based machine learning method for this task, as well as the dataset required to produce precise predictions. This project aims to analyze air temperature forecasts in weather forecasting applications by examining several machine learning models.

**Literature review**

Meteorologists have historically used statistical techniques and numerical models to forecast weather trends. Nonetheless, the increase of advanced methods is required because of the growing of aerial processes. Combining deep learning (DL) and machine learning (ML) techniques has become a applicable way to improve the reliability and effectiveness of weather forecasts in recent years. This literature review aims to give a thorough overview of the latest developments in weather forecasting using ML and DL.

The first research paper talked about machine learning-based alignment correction of air temperature forecasts and is titled "Comparative Assessment of Various Machine Learning-Based Bias Correction Methods for Numerical Weather Prediction Model Forecasts of Extreme Air Temperatures in Urban Areas". They talk about how biases in the Numerical Weather Prediction (NWP) model can be corrected through machine learning to increase the tuning of air temperature forecasts. There were four different machine learning models that have been employed by the authors: a multi-model ensemble (MME), random forest (RF), support vector regression (SVR), and artificial neural network (ANN). They discovered that the accuracy of the NWP model forecasts may be increased by all four models. The RF, SVR, ANN, and MME models were the next most accurate models. (Cho et al., 2020).

The authors produced NWP forecasts of the maximum and lowest air temperatures for the following day using the Local Data Assimilation and Prediction System (LDAPS) model. The biases in the LDAPS model forecasts were then corrected using the machine learning models. An archive of past NWP forecasts and recorded air temperatures served as the training dataset for the machine learning models. The outcomes show that the LDAPS model forecasts' accuracy might be increased by the machine learning models. The best technique for adjusting the LDAPS model forecasts was to use the MME model. Also, for both the maximum and minimum air temperatures, the mean absolute error (MAE) of the LDAPS model projections was lowered by 0.5°C and 0.4°C. The authors concluded that the machine learning is a useful technique for raising air temperature predictions' accuracy. (Cho et al., 2020).

The subject of the second research article, "Smart Weather Forecasting Using Machine Learning: A Case Study in Tennessee," involves machine learning-based weather forecasting. It talks about the shortcomings of normal weather forecasting techniques. The authors suggest a version way for weather forecasting that makes use of machine learning. The approach is a strategy for predicting the weather that trains basic machine learning models using past data from multiple weather stations. Then, in a short period, these models can be utilized to produce useful forecasts regarding specific weather conditions for the near future. The accuracy of this method exceeds that of conventional procedures. The authors also demonstrate the advantages of utilizing information from several weather stations. A review of the suggested technique's drawbacks is absent from the paper alongside detailed discussion about the result. (Jakaria et al., 2020).

The topic of deep learning-based weather prediction is covered in the third research paper, "Deep Learning-Based Weather Prediction: A Survey." The author discussed the disadvantages of conventional weather forecasting techniques and how deep learning could be used to overcome them. Deep learning has the potential to increase weather forecast accuracy, according to the authors of this article, because of deep learning's ability to learn from massive volumes of data, which can support in seeing patterns that are challenging to find using more conventional techniques. (Ren et al., 2020).

The difficulties with deep learning-based weather prediction were also covered by the author. Deep learning models can be very expensive to train. Furthermore, it can be challenging to interpret deep learning models, which can make it challenging to comprehend the reasoning behind a model's specific predictions. The authors of this paper think that deep learning can completely change weather prediction, despite its difficulties. (Ren et al., 2020).

This research paper summarizes several experiments that have utilized deep learning to predict weather. Numerous deep learning models, such as autoencoders, recurrent neural networks, and convolutional neural networks (CNNs), were employed in the experiments. A range of datasets, such as radar data, satellite photography, and meteorological measurements, were also used in the investigations. The conclusion of the experiment is that deep learning can forecast weather as accurately more than conventional techniques. (Ren et al., 2020).

The purpose of machine learning to enhance numerical weather forecasting was discussed by the authors of a research paper titled "Machine Learning to Improve Numerical Weather Forecasting." Trends in numerical weather prediction are covered, along with challenges and potential solutions. In order to improve the accuracy of 2m temperature forecasts, the authors suggest using a neural network design. This architecture enables the prediction of atmospheric model forecast errors because of their additional corrections. The purpose of the two experiments in this research is to evaluate the performance of a suggested neural network. Using a dataset of previous regional model errors, the neural network was trained in the first experiment. In the second experiment, a fresh dataset of regional model errors was used to assess the trained neural network. The studies' findings demonstrated that, in about 50% of the situations, the suggested neural network architecture was able to enhance a 2 m temperature forecast. (Doroshenko et al., 2021).

The fifth research paper is titled 'Multi-class Weather Forecasting from Twitter Using Machine Learning Approaches'. It involves classifying weather using machine learning on Twitter data. Three machine learning algorithms were tested by the researchers to compare their performances. They employed Logistic Regression (LR), Multinomial Naive Bayes (MNB), and Support Vector Machine (SVM). They gathered a dataset of tweets classified with various weather conditions, such as cloudy, rainy, or sunny. The three algorithms were then evaluated on a fresh set of tweets after being trained on the dataset. With a 93% accuracy rate, they discovered that SVM performed better than the other two methods. (Pardamean et al., 2021)

To find out how various features affected the algorithms' accuracy, the researchers also ran experiments. They discovered that the phrases used in the tweets, their location, and the time of day they were tweeted were the most crucial characteristics. (Pardamean et al., 2021)

The sixth study investigates whether deep learning can improve or replace conventional numerical weather prediction (NWP) techniques. The title of the research paper is "Can deep learning beat numerical weather prediction?" Although NWP is an essential tool for comprehending and forecasting weather patterns, it has drawbacks, including expensive computing expenses and a hard time capturing nonlinear correlations in meteorological data. While deep learning, an artificial intelligence area, has demonstrated impressive results in a few applications, including natural language processing and picture identification. (Stadtler et al., 2021).

The authors compete that by capturing complicated nonlinear correlations in meteorological data, lowering computational costs through more effective algorithms, and enhancing forecasting accuracy, particularly for short-term forecasts, deep learning could overcome the constraints of NWP. One of the paper's limitations is the absence of case examples illustrating the use of DL in weather prediction. (Stadtler et al., 2021).

The paper's conclusion declares that it is too soon to determine if deep learning will outperform NWP. But the authors think that deep learning will eventually be able to significantly improve weather forecasting. (Stadtler et al., 2021).

The seventh study examined the use of machine learning in climate analysis and weather forecasting. The authors talk about how machine learning is used in climate research and weather forecasting. Additionally, they go over a range of machine learning applications for climate analysis and weather prediction, such as: 1) Wind power forecasting: wind farms can operate more efficiently if machine learning is applied to forecast wind power. 2) Ensemble forecasting: To produce forecasts that are more accurate, the results of several weather models are combined using machine learning. (Bochenek & Ustrnul, 2022).

The difficulties of using machine learning to climate analysis and weather prediction are also covered in the article, including the complexity of weather systems and the requirement for large volumes of data. Information regarding the accuracy of machine learning techniques in comparison to conventional techniques is absent from the paper. (Bochenek & Ustrnul, 2022).

**Research questions**

* What is the process of predicting air temperature using machine learning model?
* What is the best regression-based machine learning model to use? What is the needed dataset?

**Research objectives**

* To conduct qualitative research (interview) with an expert with the concept of weather forecasting.
* Discuss with the expert what feature will affect the result of the prediction the most. Also, finding out what is the best regression-based machine learning model.
* Understanding the process of weather forecasting using machine learning models and meet our research predictions by doing a comparison between different machine learning models.

**Research methodology**

**Quantitative approach**

The quantitative approach is research where numerical data and closed-ended questions' answers are collected to be analyzed for addressing correlations and some theories. Surveys and experiments are usually implemented in this approach to collect data. The approach’s purpose is to find results and present them typically in the form of graphs and statistics. (Streefkerk, 2023)

**Qualitative approach**

Qualitative approach is a research where non-numerical data like words or images and open-ended questions' answers is collected to be analyzed for addressing correlations, theories, patterns, and insights but interpreting process will be more complex unlike qualitative approach. interviews, case studies, and focus groups are usually implemented in this approach to collect data. The approach purpose is to find result and present it typically in form of narrative. (Streefkerk, 2023)

**Mix approach**

Mix approach contain both quantitative and qualitative approaches together. The purpose of using this approach is to benefit from the strengths of both approaches for better problem understanding. By using this strategy, researchers can pair up their findings—that is, utilize the results of one method to confirm or enhance the conclusions of another. When researching complex problems that neither methodology can fully capture, mixed methods are especially helpful. (Streefkerk, 2023)

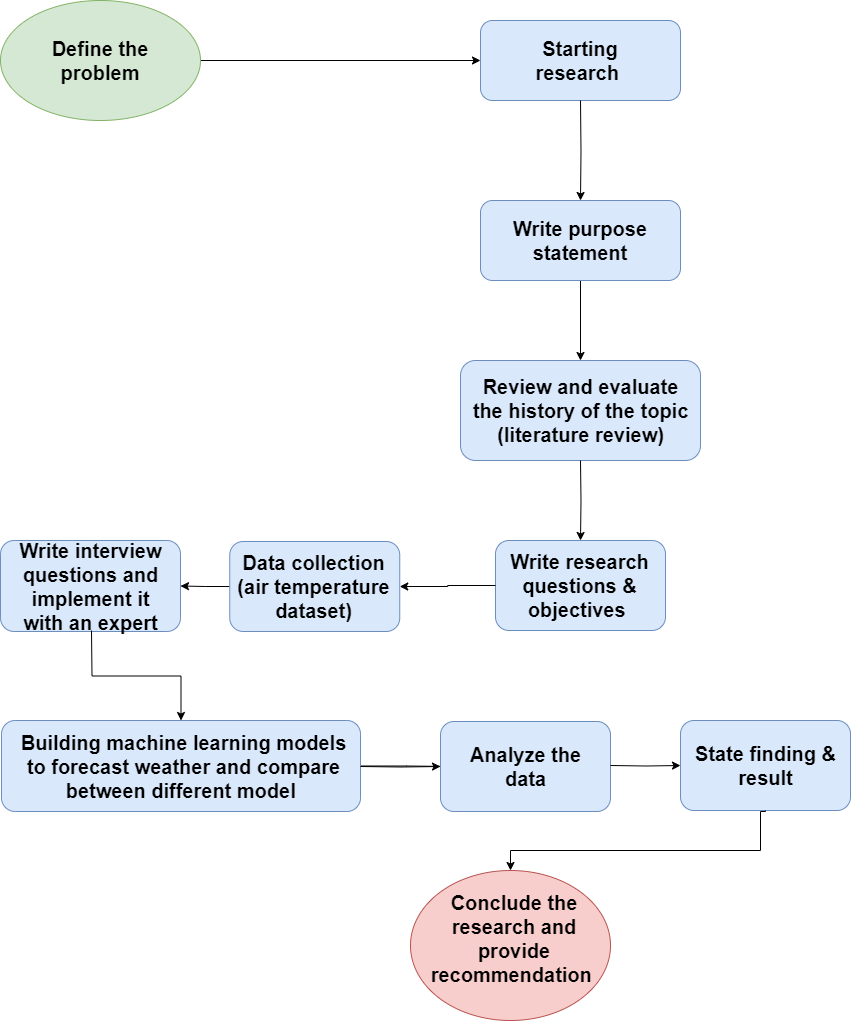
**The followed approach in the research**

In my research, I followed the qualitative method. I conducted an interview with open-end question with an expert in Arab weather company. I got non-numerical data like words and analyzed to get insights. Following the qualitative approach help me a lot in my research and was better option than quantitative approach, because my research is constructionism, and the research nature requires experts’ opinion to get some advice and have answers for questions which are significant in the research. But I faced a challenge in doing more than one interview with different experts, so I decided to get an air temperature dataset and build machine learning models to predict data and for another two reasons. The first one is to make up for the lack of interviews. Second, to meet the objectives of my research and answer the research question which these two goals could not been done without machine learning models. Next time if I will do research paper, I suggest that I use mix approaches (qualitative & quantitative), because its beneficial to send surveys to the people who work in companies in this field and get their opinion, and closed-end questions might help alongside with open-end questions.

**Saunders’ research Onion**

1. **Research philosophy:** there are two types of philosophy in this layer, positivism and constructionism, I chose the second type which constructionism because positivism philosophy is about constant facts, unlike constructionism philosophy which is about that the people have different point of view and different opinions about reality and also this philosophy usually followed in qualitative research. My research is qualitative, and I conducted an interview and get an opinion from an expert about weather forecasting which his opinion may differs from other experts and his answers for the interview question may differs also.
2. **Research approach:** deductive and conductive are the two approaches in this layer. I followed the deductive research because the first thing I did was come up with theory that regression-based machine learning models can effectively predict air temperature. This theory assumes that historical air temperature data, combined with relevant meteorological variables, can be leveraged to train accurate predictive models. The research prediction was that using machine learning models in weather prediction can lead to accurate air temperature predictions. Also, I predicted that random forest model will be the model with the highest prediction accuracy among the rest of regression-based models. The first part of the experiment in the research was conducting the interview, and through the interview I found that the random forest will probably be the best model. The second part of the experiment ensured that, which was preprocessing the air temperature dataset, training different models on it, predict, and then compare the result.
3. **Research strategy:** I followed experimental research strategy because it involves manipulating the independent variables to observe a change in the dependent variable which is in my research (air temperature). In other words, to assess the relationship between variables. The purpose of experimental research is to support, refute or validate a research hypothesis. I followed the experimental research strategy to test my theory and this strategy is deductive in nature. In my research, I wanted to predict air temperature using machine learning models, so by preprocessing the dataset I found the relations and correlations between the features and knew what the most important features in the prediction process. Also, by training the models I predict the dependent variable from the independent variables and improving the model’s accuracy to validate my theory and meet my predictions.
4. **Research methodology choice:** there are three research methodology in this layer which are mono method (using either qualitative or quantitative method), mixed method, which is using both methods, and multi method which is using wider range of quantitative and qualitative methods. I chose the first methodology which is mono method and specifically I used the qualitative research method only because my research is constructionism which is about people opinion. I was investigating weather forecasting, and definitely there are many opinions and many ways of weather forecasting.
5. **Time horizon:** I will conduct different samples over fixed point of time and this type called cross-sectional. Because I will not stick to only air temperature dataset that I have, conducting different samples and training the models on different dataset will increase the accuracy and open my eyes on new insights for improvements.
6. **Data collection & analysis:** I decided to undertake the research by doing both primary research and secondary research for several reasons. First, the primary research is required in my research, and it helped me in getting deeper in the weather forecasting field by conducting an interview with an expert in the field from Arab weather company. Second, the primary research was important in the experimental research strategy that I followed in my research, I was able through the interview to answer my research question, which lead me to ensure the answer by building machine learning model and meet my prediction. Third, the secondary research was significant because I needed an already collected air temperature dataset to train the models on which this dataset is considered as secondary research. Also, after collecting the data I analyzed it using thematic method and building machine learning models to gain insights with interpretations to answer my research questions and meet my research objectives and predictions.

**Flowchart**



First, I define the problem and what my research will be about. Second, I started the research and wrote the purpose statement, which is the purpose of the research. Third, I reviewed the history of the topics and saw the past experiment and what the opinions of the authors. Then I wrote my research questions and objective and made sure that I would answer the questions and meet the objectives. After that the data collection process has started, I conducted an interview, got an air temperature dataset, and build machine learning models, then I analyzed the data and stated finding and result with making sure that they answer and meet the research questions and objectives. Finally, I wrote the conclusion and gave recommendations.

**Primary research**

Primary research involves the direct collection of distinct data that is appropriate for a given study goal. A range of methods are employed in this type of study to obtain direct data from people or sources, such as surveys, interviews, experiments, and observations. While interviews allow participants to have more in-depth conversations, surveys can be used to get input on a certain issue by having respondents respond to some closed-ended questions. Gaining current, focused insights and finding answers to novel research questions are two benefits of conducting primary research. However, it can cost more money and time. (George, 2024)

**Secondary research**

Secondary research utilizes data that has already been collected by others for different purposes. In this study, among other sources, I look over and assess previously gathered data from published literature, market research studies, and online databases. This method saves time and money by providing a comprehensive synopsis of a topic and its historical context. It does, however, have some shortcomings, such as the potential for being outdated or unrelated to the objectives of the study. Despite these drawbacks, secondary research can be useful for getting early insights and laying the groundwork before engaging in more concentrated primary research. (George, 2024)

**The implementation of primary & secondary research in my research**

First, I reach out an expert in Arab weather company to conduct an interview with him. I prepared the questions (open-ended questions) and ask him about them. I got no-numerical data like words, and he answered all my questions. I recorded the interview and then analyzed the data using thematic method to gain insights and start answering the research questions. I wrote a literature review to see the history of weather forecasting, alongside with getting an air temperature dataset to build and train machine learning models to see their performance and meet the research predictions and objectives, and this was the implementation of secondary research in my project.

**Ethical & access issues of data collection process and its limitations**

1. Data privacy: I contacted several sources before Arab weather company to obtain data, but they clarified that this kind of data is confidential and private.
2. Data ownership: Arab weather clarified that they have the ownership right of the data and I can use it only to continue my research in this course.
3. Data bias: I found a dataset on the internet about air temperature, but the issue was that the dataset was relied on a region that is outside Jordan.
4. Data accessibility: accessing the data and reaching out to the expert to do the interview was an issue in the beginning. Because the data is private, I needed to contact a few people to get to the dataset and access it in equitable way.
5. Big dataset: the dataset that I collected has a large volume and it took me a long time to analyze and process it.

**Merits of data process**

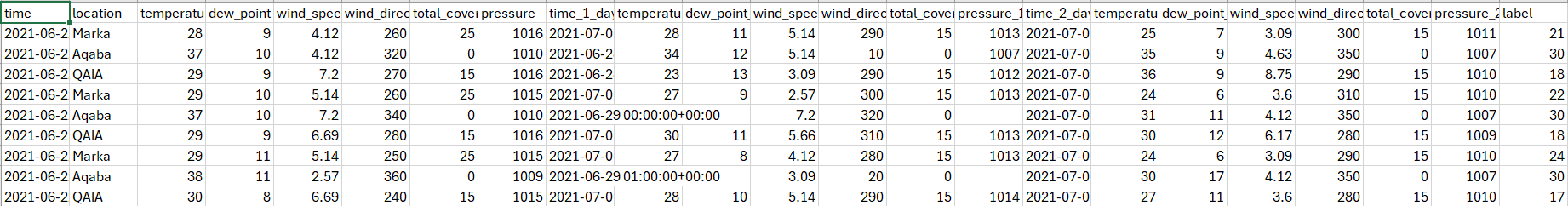
1. High accuracy: access a dataset and collecting data from an interview with an expert in Arab weather company helped in improving the accuracy of my result and the models that I built to forecast air temperature.
2. Weather change: my data collection process supported research on long-term weather patterns and better understanding of air temperature changes in different places.
3. Impact assessment: the data collection process was detailed, so it allowed me to answer my research questions in an efficient way.

**Data collection and analysis tools:**

1. Interview with an expert: in this approach I used google meet to do the interview and record it.
2. Coding (thematic analysis method): involves methodically classifying different parts of the interview material that have been gathered according to recurrent themes or concepts. This helps me to spot trends and learn more about the viewpoints and experiences of the participants.
3. Google Colab: I used this tool to analyze the dataset and build a model to predict air temperature.

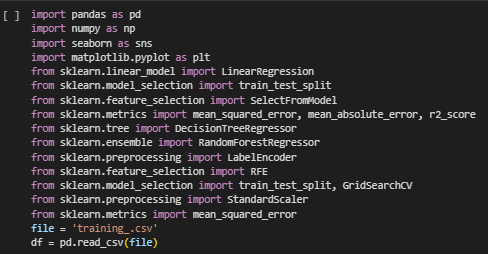
**The dataset**

The dataset I used in the research is an air temperature dataset which contain large amount of air temperature records in different regions in Jordan, with time, location, and other features. I got this dataset from Arab weather company. This dataset considered secondary research and it helped me a lot.



**The code and the machine learning models**

Pre-Processing:

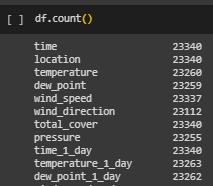


This first cell of code is to combine all the imported libraries that I have used for the preprocessing steps as well as the building of my models.

A black and white text

Description automatically generated

This shape attribute is used to discover the total number of rows and columns in the dataset.



This count method is used to find the total number of non-values in each feature in the dataset.

A screenshot of a computer

Description automatically generated

This method is used to explore the statistics for each numeric feature.

A screen shot of a computer program

Description automatically generated

This method is used to drop the null values in each of the selected features.

A screen shot of a computer

Description automatically generated

This type of preprocessing is to encode the columns of type objects(categorical).

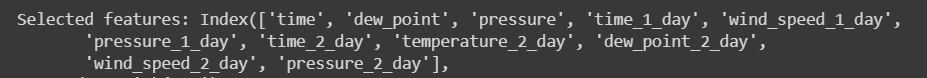
A screen shot of a computer

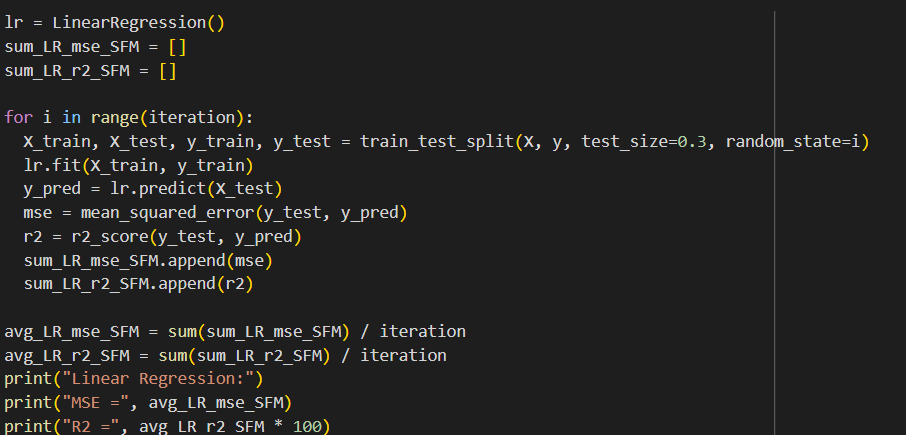
Description automatically generated

This type of preprocessing is to detect outliers in the specified columns using the IQR method which will help me to see any abnormal behavior and following that I will remove the columns that have outliers.

A computer screen shot of white text

Description automatically generated

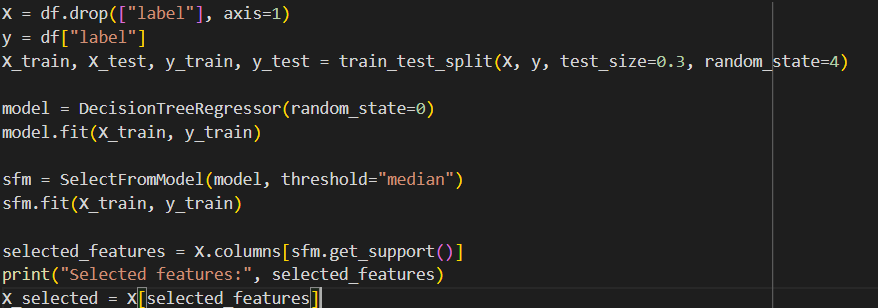
Applying select from model feature selection technique with linear regression. This chooses the best features to train the linear regression model to get better results and simpler model, the selected features are then stored in ‘X’ variable, which contains the best set of features displayed in the figure below that the model will be trained and tested on. 



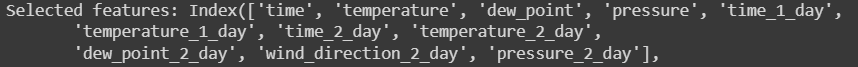
The first model I tried was linear regression, which is a simple regression model that assumes the data is linearly separable and tries to find the best fit linear line for the data. The process of building the model involved defining the regressor, creating lists to store the MSE and R squared from each iterations, after that, in each iteration the data is split into training and testing, training the model, making predictions, and calculating the MSE and R squared, and then appending their values to the lists, this process is repeated 30 times, and then the average for both metrics is printed as follows:

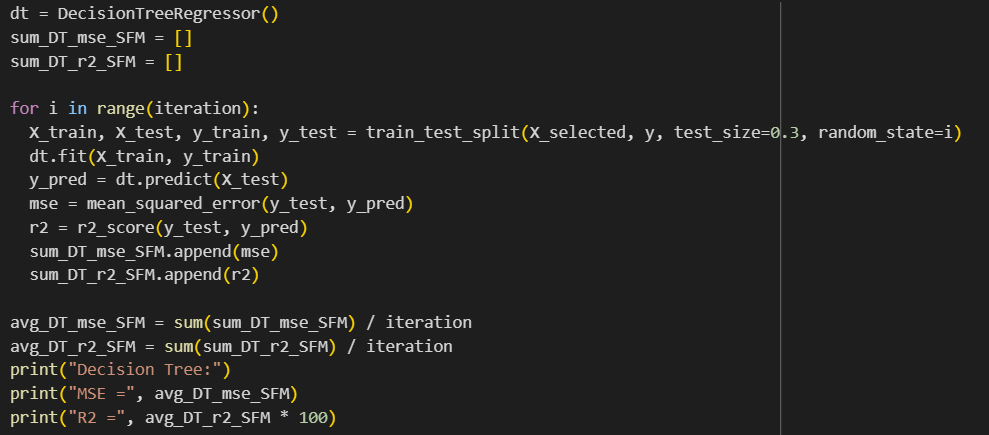
A black background with white text

Description automatically generated



The same feature selection technique which is select from model was used with decision tree for equal comparison between the models, the above snip of code as the same as the one in the linear regression but with changing the model to decision tree, so now the technique select features assuming that the model is decision tree, and there exist a difference between the selected features for both models.





The second model I tried was decision tree regressor, which is a model that works on both regression and classification problems, it involves building a tree that separates the hyperplane according to specific decision boundaries, each layer in the tree represents a decision boundary and each leaf node is a result. The process of building the model followed the same steps in building the linear regression but with changing the regressor model to decision tree. Comparing the two model’s performance, linear regression values for both R squared and MSE were better.

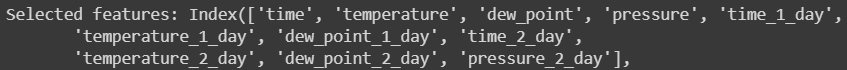
A black background with white text

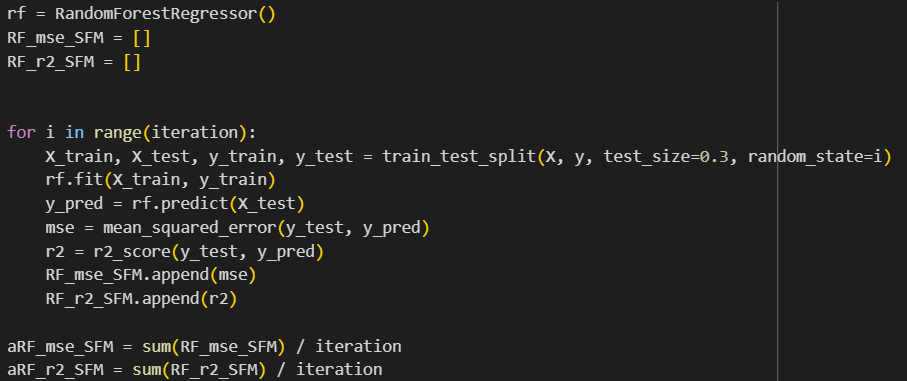
Description automatically generated

A computer screen shot of text

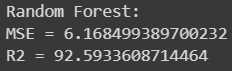
Description automatically generated

As mentioned earlier, to achieve equal comparison between models, I also applied the select from model feature selection technique with the random forest. The selected features were very similar to the selected features from the decision tree model, this is because both models are similar to each other’s.





Th last model I tried to implement is Random Forest regressor, it’s an ensemble learning algorithm that combines multiple decision trees and gets the final decision according to the average between all of them. The model was built in the same way that the previous 2 models were built, and the results were as follows:



**The interview**

* Could you explain the main technique that is employed in your company to predict air temperature?

Weather prediction go through different stages, first stages, numerical weather prediction models (NWPS) which is physics equations, local model are implemented and give different parameter for air temperature and forecasting for multiple days, then we do forecasting correction by getting the model forecasting result in past 5 days (for example) and the model forecasting result for the next 5 days, and observation (actual result) and then calculate the model error percentage and changing the forecasting result for next 5 days depending on the model errors, my dataset was recorded based on actual result of air temperature in different places in different days.

* How do you evaluate temperature forecasting models for accuracy?

Air temperature forecasting will be more accurate using NPWS technique.

* About machine learning different models, could you give examples of strong machine learning models that have the best results in air temperature forecasts so far?

Linear regression, random forest, and decision tree are examples of machine learning models to build a simple model to predict air temperature using the dataset. These models used to make time series forecasting to predict air temperature for the next day. Using deep learning for time series forecasting is not efficient as machine learning. Moreover, random forest will probably be the best model among the rest of regression-based model.

* What is the essential feature of a dataset used for temperature forecasting, and how important are they to the precision of the predictions?

In weather forecasting, the important feature is the previous records for the temperatures. The second thing is the future records which are not available in my dataset. The expert said that in this situation we do something called auto correlation which is the relation between the past records of temperature and the current readings. Let’s say the current reading for temperature is 17, what is it relation with the past two hours reading which is for example 15, the reading for the past day is 16, the reading for the past three days is 14, and the readings for the past one week is 8. The reading of the past week is far away from the current reading which tell us that the relation between them is weak, but the reading for the past three days is closer to current reading than the past week reading, also the reading for the past day is much closer to current reading, and the reading for the past two hours is even closer than the rest. That’s tell us that auto correlation illustrates how the close past readings influence the current and the future reading more than the far reading, also the relation between the closed reading with the future (predicted) reading is strong.

* What are the primary obstacles in the forecasting of air temperature?

The first obstacle is data preprocessing. Second one, handling the models’ errors. Third one, keep the accuracy high as much as possible.

**Primary research analysis**

in the analyzing process, I will use the thematic analysis method. This method is about themes and the analysis for these themes. The interview will be analyzed by putting several themes or in other word important points and topics that were in the interview, and then analyze each point.

**NWPS Technique.**

The expert first explains how important the NWPS method is for assessing models that predict temperature. This method compares observed data collected over a given period of time with model forecasts. The topic emphasizes a useful and data-driven review process by implying that forecasting models' accuracy should be confirmed by experimental data.

**Combining models with Correction procedures.**

The Expert outlines several steps in the followed air temperature forecasting process in Arab weather company, beginning with the use of NWPS, then, observational data and both past and current model forecasts are used to apply correction process. To make more accurate forecasts, this point highlights the significance of integrating observations from the real world with models.

**Time series forecasting using machine learning models.**

The expert suggests decision trees, random forest, and linear regression models as examples of methods for forecasting air temperature. Time series forecasting is the focus, and past temperature readings are significant to this process. In this case, machine learning is preferred over deep learning, indicating that efficiency and model fit are important considerations.

**Importance of Time Series, past records Data and Autocorrelation.**

The expert highlights the importance of time series data while outlining a dataset's key components. One important method for comprehending the connections between past temperature readings and the current forecast is the introduction of autocorrelation. The theme emphasizes how crucial it is to take temporal dependencies into account to make precise forecasts. Moreover, we understood through the interview how strong the relation between the past records and the future records (predicted records).

**Data Preprocessing.**

Managing model mistakes, preserving high forecasting accuracy, and the complexity of data preprocessing are the challenges that have been discovered during the meeting with the expert. In order to improve the overall accuracy of temperature predictions, this theme highlights the practical difficulties associated with the forecasting process and emphasizes the necessity of strong data preparation methods and efficient error-handling procedures. The analysis of this theme also tells us and explains the process of weather forecasting using machine learning models in clear way.

**My research finding results and how they align with my research objective and answering my research questions.**

My research questions and objectives revolve around knowing and understanding how to forecast air temperature using machine learning model, which is the best machine learning model in performance metrics and in predicting air temperature, and discussing which features affect the predicting process the most. First, I preprocessed the dataset to filter, clean, encode, and to allow the machine learning model to select the features with most effect on the prediction, which is one of these features according to the interview is the past record of air temperature. Then, I start building the model to predict, I tried many models like Linear Regression, decision tree, and Random Forest. The founded results from building these models were that the best model in forecasting air temperature was Random Forest, because it gave the highest accuracy and lowest mean squared error in performance metrics, and this model provide deep prediction for the next day, that’s what confirmed the expert opinion about machine learning models in weather forecasting. Also, I found how important the NWPS method is for assessing models that predict temperature. This method compares observed data collected over a given period with model predictions. This insight helps me in understanding the modern and following weather forecasting method in big companies like Arab weather. One of the insights that I gained from the interview is that the feature that the predicted air temperature records based on, is past records, and its logical because I found in the interview, after the expert explain to me, how the past records affect the future records and how the past records usually be similar to the closest future records to it. I also see that while training the models. I did not do complex preprocessing for the data, but I found it harder than building and training the model itself, and the result of preprocessing was significant in the model’s result.

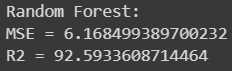
After comparing the machine learning models, the results I have gained from each model will help me to decide which model performed well and underperformed. The results were gained from the Linear Regression model, Decision Tree regressor model, and Random Forest regressor model. The evaluation metrics used to represent the model's performance are Mean Square Error and R-squared error.

Comparing the models based on the average, the random forest regressor has shown a higher average of the R-squared error by 92% followed was the linear regression at 87%, and finally the decision tree with 85%, the Mean Squared Error has also shown that the random forest outperforms both the linear regression and decision tree as the results were as follows 6, 10, and 12. These result also shows that the select from model feature selection technique performs better on Random forest than other algorithms.

A screenshot of a table

Description automatically generated

As we can see in the image above, on the left there are the actual air temperature records on the dataset, while on the right the predicted ones. This predicted result is the result of the random forest model which had the heist accuracy, and the results show that the distance between the predicted results and the actual results is proportional to the accuracy rate (R2) of the random forest model.



**Conclusion & Recommendations**

In the beginning of the research proposal, I put a purpose statement which contains research questions which are what the process of weather prediction using machine learning models, what are the best machine learning model, what is the needed dataset and the features to select. After implementing primary and secondary research, following qualitative approach, conducting an interview, gathering dataset and build few machine learning models, and then analyzing all the data the I have come up with, I state the finding and result by the end of my proposal and they answer my research questions, align with research objectives, and meet my research predictions. The process of weather forecasting is collecting dataset about what I want to predict, which in this research air temperature records in different regions, understanding the dataset and implement preprocessing techniques on it, build a machine learning models and train them to come up with the result (MSE and R2). The best model was Random Forest with the highest R2, and the essential features were selected by random forest. The predicted air temperature records were based on the past air temperature records in the dataset. Here are the recommendations that I suggest as I future steps which I believe that by implementing them, there will be an enhancement in the research proposal. The recommendations:

1. **Gathering more data by following mixed approaches (qualitative & quantitative)**: this recommendation will help me to dive deeper into air temperature data set and provide better understanding of the features correlations with each other and how exactly some features have more impact than others on the result.
2. **Doing more preprocessing techniques on the dataset before I start training the models**: this recommendation is significant because it allows me to understand, explore, clean, and filter in better way which lead to more accurate result and efficient model to predict.
3. **Doing more interviews with different experts**: this future step is important because it will open my mind to different and efficient opinions that come from experts, which will make me improve my research in the future.
4. **Expand the forecasting:** expand the future that I will predict, like predicting if it will rain in the next day or not based on the status of the past few days.

**References**

Yaqoob, I. (2016) Big data: From beginning to future, International Journal of Information Management. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0268401216304753 (Accessed: 14 November 2023).

(2013) Big Data: A review | IEEE conference publication | IEEE Xplore. Available at: https://ieeexplore.ieee.org/abstract/document/6567202/ (Accessed: 14 November 2023).

(2017) Big Data in weather forecasting: Applications and challenges | IEEE ... Available at: https://ieeexplore.ieee.org/document/8070824 (Accessed: 14 November 2023).

(2018) Analysis of weather prediction using Machine Learning &amp; Big Data | IEEE ... Available at: https://ieeexplore.ieee.org/document/8441679 (Accessed: 14 November 2023).

Fathi, M. et al. (2021) Big Data Analytics in weather forecasting: A systematic review - archives of computational methods in engineering, SpringerLink. Available at: https://link.springer.com/article/10.1007/s11831-021-09616-4 (Accessed: 14 November 2023).

Salcedo-Sanz, S. et al. (2015) Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms - theoretical and applied climatology, SpringerLink. Available at: https://link.springer.com/article/10.1007/s00704-015-1480-4 (Accessed: 14 November 2023).

Cifuentes, J. et al. (2020) Air temperature forecasting using Machine Learning Techniques: A Review, MDPI. Available at: https://www.mdpi.com/1996-1073/13/16/4215 (Accessed: 14 November 2023).

Cho, D., Yoo, C. and Cha, D.-H. (2020) Comparative assessment of various machine ... - wiley online library. Available at: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019EA000740 (Accessed: 25 November 2023).

Jakaria, A.H.M., Hossain, M.M. and Rahman, M.A. (2020) Smart Weather Forecasting using machine learning:A case study in Tennessee, arXiv.org. Available at: https://arxiv.org/abs/2008.10789 (Accessed: 25 November 2023).

Ren et al. (2020) Deep learning-based Weather prediction: A survey, Big Data Research. Available at: https://www.sciencedirect.com/science/article/abs/pii/S2214579620300460 (Accessed: 25 November 2023).

Doroshenko, A., Shpyg, V. and Kushnirenko, R. (2021) Machine learning to improve numerical weather forecasting | IEEE ... Available at: https://ieeexplore.ieee.org/document/9349325 (Accessed: 25 November 2023).

Stadtler, S. (2021) Can deep learning beat numerical weather prediction? Available at: https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0097 (Accessed: 25 November 2023).

Bochenek, B. and Ustrnul, Z. (2022) Machine learning in weather prediction and climate analyses-applications and perspectives, MDPI. Available at: https://www.mdpi.com/2073-4433/13/2/180 (Accessed: 25 November 2023).

Pardamean et al. (2021) Multi-class weather forecasting from Twitter using Machine Learning Approaches, Procedia Computer Science. Available at: https://www.sciencedirect.com/science/article/pii/S187705092032442X (Accessed: 25 November 2023).

Streefkerk, R. (2023) Qualitative vs. quantitative research: Differences, examples & methods, Scribbr. Available at: https://www.scribbr.com/methodology/qualitative-quantitative-research/ (Accessed: 09 December 2023).

Research methods: Definitions, types, examples (no date) Scribbr. Available at: https://www.scribbr.com/category/methodology/ (Accessed: 10 December 2023).

George, T. (2024) Primary research: Definition, types, & examples, Scribbr. Available at: https://www.scribbr.com/methodology/primary-research/ (Accessed: 09 February 2024).