Precipitation Prediction System

Yash Gupta
0869229
Software Engineering
ygupta1@lakeheadu.ca

Shaun Cyr
0688734
Software Engineering
sbcyr@lakeheadu.ca

Emmanuelle Trudel
0687776
Software Engineering
etrudel@lakeheadu.ca

***Abstract*—Weather patterns are dynamic and difficult to predict; extensive literature has attempted to model and predict weather patterns with limited success on a large scale. The application of machine learning enables unique automated solutions to predict unfolding weather events. The successful prediction of weather events offers large financial savings and improvements in community safety.**

I. Introduction

According to the World Meteorological Organisation, weather forecasting is a vital element in order to meet the needs for the growing population. Weather forecasting helps a lot of industries. Being able to forecast and plan for the future when it comes to the local climate is a major advantage in planning tourism facilities. Transportation infrastructure can be set up to measure road surface conditions to improve traffic safety. Weather forecasting can also help in travel preparations by knowing about a storm that is going to happen at the destination that can potentially pose a threat to lives of people. Weather uncertainty can also pose a threat to the construction business, so having prior knowledge of the weather conditions can help construction companies to establish a more efficient schedule. The agricultural industry is one of the biggest industries that relies on suitable weather conditions for production. Having an idea about what the weather is going to be like in the future enables them to protect their plants from bad weather and make sure that their crops receive enough water in hot, dry weather. They can also treat their plants according to what the weather is going to be like so that they are safe from any detrimental conditions.

A. Definition of Problem

Many conditions contribute towards the weather patterns that are exhibited at surface level. Predicting the surface level conditions can be difficult and involve an incredible number of mathematical characteristics that are not possible for humans to calculate themselves. Existing models have been developed using deterministic calculations or machine learning models, each with strengths and limitations that make an accurate prediction difficult long term forecast. The introduction of additional data can help to make predictions more accurate, but they introduce computational complexity that must be managed to allow frequent updating of predictions using the latest data available.

B. Motivation

Weather events can have lasting damage on the lives of people all around the world. Developing a model that can accurately predict weather events with sufficient accuracy and time to enable preparations can help to reduce the impact of these weather events. There are many industries that rely upon specific weather conditions ranging from construction all the way to tourism. Existing weather prediction models typically face a constraint between computational complexity, prediction accuracy, and how far in time it is able to accurately predict.

C. Organization Of Paper

The paper will be divided into 7 sections.

- Abstract
- Introduction
- Literature Review
- Materials and Methods
- Results & Discussion
- Conclusion
- References

II. Literature Review

Weather prediction is very important for people around the world, but it is known to not be incredibly accurate. Luckily, there have been many attempts to solve this problem in a variety of ways; there have been many software researchers that have applied machine learning techniques to solving this as well.

Some of the most simple methods of solving this problem is by using numerical analysis methods. For example, Krasnopolsky and Rabinovitz[1] attempt to improve the accuracy of existing weather prediction models by incorporating known relationships between climate conditions and the resulting weather patterns. The team incorporates geophysical fluid dynamics equations with respect to time in order to provide a rigid framework for models to build upon when predicting weather events; other deterministic equations are also used. This method has many benefits such as being less complex compared to other methods, and this method also allows models to be trained and predicted faster than standard ML models. Unfortunately, this model makes it difficult to gather training data on rare events in order to produce an effective prediction model. Zhang, Chen, Xu, and Ou [2] also

make use of numerical models by using a distributed lag non linear model(DLNM) to predict precipitation conditions up to half a year in advance. These researchers aimed to predict ordinal categories corresponding to the Standardized Precipitation Evapotranspiration Index (SPEI) by first using stepwise regression to select predictors for study. It was found that it was able to have the highest prediction accuracy for droughts. Unfortunately, this study did not have much information surrounding how they implemented this method and determined this.

The models above, even though they are found to be fairly accurate, have many problems surrounding them. Wei, Yan, and Jones[3] found that these numerical/linear models had a common problem; the models rapidly deteriorate in skill for predicting the forecast as they move outside a given time window. To combat this, the authors put forward a decision tree approach to predict the extreme weather in eastern Asia and compare it with binary regression models. They found that the decision tree method outperformed the other methods; the proposed method had "great potential of skillful seasonal prediction of the regional extreme precipitation, with quite consistent performance even with limited samples"[3]. There have been others that have had success with decision tree models such as Chauhan and Thakur[4] who believe that the decision tree is the most promising method for classification and prediction in the field of data mining. This paper compares three different decision tree algorithms, namely C4.5, CART, and LMT with and without using AdaBoost. It was found that AdaBoost with an iteration of 15(the max they tested), the accuracy of the decision tree was the best, and when applied to the 3 decision tree methods, LMT gave the most accurate results. Thus, LMT with AbaBoost with an iteration of 15 was the best algorithm they tested. Unfortunately, there are many others who believe that even decision tree methods are not as accurate as desired.

So there have been others that have attempted to use methods such as the K-Nearest Neighbour that is beneficial because it is a simple algorithm that gives fairly accurate results. Authors such as Huang, Lin, Huang, and Xing[5] proposed an improved KNN algorithm for weather prediction. They discuss the related WKNN, where there are weights assigned to neighbours, and DWKKN, which extends and enhances the linear mapping used in WKNN. The accuracy of the author's proposed algorithm is compared to the accuracy of the modified KNN algorithms, WKNN and DWKNN. It is found that in most instances, the proposed algorithm gave a more accurate prediction of the weather. Unfortunately, authors such as Prasetya and Ridwan[6] compare multiple methods of weather prediction using the probabilistic brier score, confusion matrix and ROC. They compared classification tree, naive bayes, and k-nearest algorithms; they found that after using the confusion matrix that the naive bayes method provided the highest precision, recall, and accuracy compared to the other methods.

Since these methods proposed have been fairly simple methods, there have been many who attempt to use more complex algorithms such as neural network models. These algorithms are still fairly new and attempt to find relationships in a way that is similar to the human brain. Since this is fairly new technology there are many researchers such as Hashim, Duad, Ahmad, Adnan, and Rizman [7] who attempt to determine whether artificial neural networks(ANN) would be a suitable and accurate model for prediction weather, specifically precipitation. They determine that ANN models are suitable methods that could be used for weather prediction which should be further researched. With this, researchers such as Fente and Singh[8] who implemented long short-term memory(LSTM) neural networks and were able to achieve very high accuracy. There were also Salman, Kanigoro, and Heryadi[9] that compared recurrent neural networks(RNN), conditional restricted Boltzmann machines(CRBM) and convolutional neural networks(CNN). The study experimented with all three models and realised that RNN showed the most promising results because of its advantage of working with time series data and as weather is time dependent. They even go as far as to say that it worked perfectly. Unfortunately, authors like Cho, Yoo, and Im[10] knew that models such as random forest, support vector regression, and neural networks had a problem with bias. Luckily, they were able to create a strategy that was able to increase the R2 and reduce the bias.

As described, there are a wide variety of methods to predict weather, but each of these methods all had some positives and negatives. It is important to develop a single model that is able to incorporate the insights from all of these studies to produce a model that provides sufficient accuracy. This project aims to apply these insights to weather data collected in Szeged, Turkey using the methods learned in class in addition to research conducted in the duration of this project. This project will build upon the successes of other researchers to incorporate an effective implementation to their lessons learned.

III. Material and Methods

A. The Proposed Method

Many factors contribute to the weather conditions experienced at the surface. These conditions follow patterns that can be predicted and understood using machine learning models that can process highly detailed and accurate weather data. It is understood that the outcome of weather is dependent on a huge number of factors, however, predictions can still be made using a limited number of attributes. These include conditions such as temperature, wind, and humidity as the primary attributes that we have access to. The different models that will be implemented are decision tree, random forest, KNN, naive bayes, and ANN; they are explained in greater detail below. The success of this project will be based upon its ability to generate an understanding of the relationships that exist between the characteristics of weather. To this extent a

variety of predictor and predicted variables will be used to gain a deep understanding of the data. Since the maximum accuracy is what we are attempting to determine, many different models will be implemented then hyper-parameter tuning will be applied to each; this is shown in figure 1. Once this is complete, the different models will be compared to one another to determine which model had the highest accuracy. Since weather is highly volatile, accuracy of the predictions has been a problem; therefore, it is the most important factor in order to make this model useful.

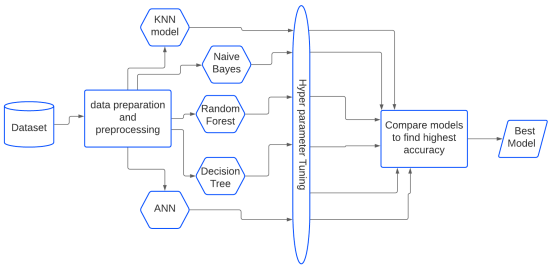

Fig. 1. Workflow of the Project

1) Data Set: The dataset is collected from Kaggle. The original dataset consists of 12 variables/columns.

- Formatted date - Date and time when the observation was made.
- Summary - Summary of the weather observation made.
- Precip Type - Informs whether there was rain or snow at the time of observation.
- Temperature - Temperature measured in Celsius.
- Apparent Temperature - The perceived temperature in degrees Fahrenheit derived from either a combination of temperature and wind (Wind Chill) or temperature and humidity (Heat Index) for the indicated hour.
- Humidity - Humidity in the air.
- Wind Speed - Wind Speed measured in km/h.
- Wind Bearing - Direction of the wind.
- Visibility - Visibility is a measure of the distance at which an object or light can be clearly discerned, measured in km.
- Cloud Cover - Cloud cover refers to the fraction of the sky obscured by clouds when observed from a particular location.
- Pressure - Atmospheric pressure, measured in millibars.
- Daily Summary - Weather summary of the entire day.

The development of this model first required a number of preprocessing steps. The first of these preprocessing steps requires that the three columns of Formatted Date, Daily Summary, and Loud Cover are removed since these do not provide any meaningful data. The Loud Cover attribute is a constant column and therefore provides no useful insights into the relationships between the variables. The daily summary is a text based attribute that cannot be processed or understood by any of our models, and must be removed. Similarly none of the models developed at this stage require information about

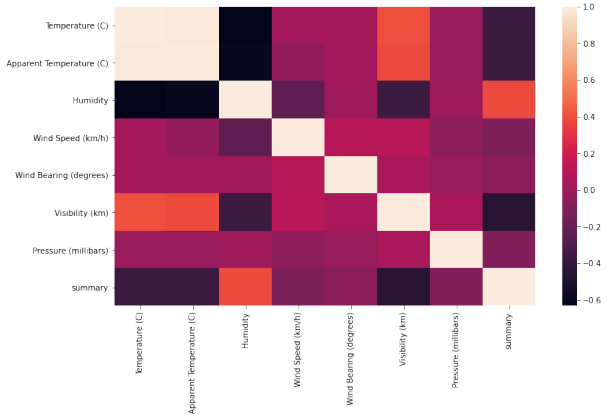

Fig. 2. Correlation Heatmap

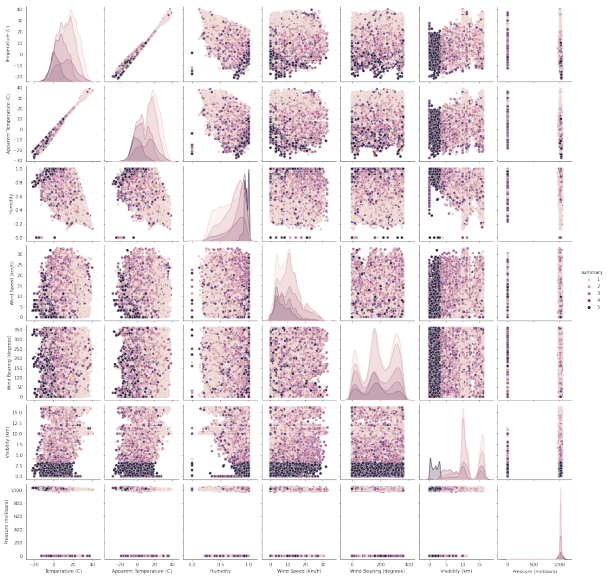

Fig. 3. Scatter Plot with Labels as hue.

the date, so this attribute should be dropped.

After undesirable attributes are removed in the preprocessing steps outlined below, the next step is to establish the target attribute. We develop a new column in our data frame called SummaryCat that, and set this attribute equal to the numerical categorical code established for our frequent and selected conditions of ”Partly Cloudy”, ”Mostly Cloudy”, ”Overcast”, ”Clear”, and ”Foggy”. From the box plots, it was observed that the dataset had many outliers which were removed as a part of cleaning the dataset.

In the future, recurrent neural networks like the LSTM Neural Network can be explored and implemented. The model will be highly suitable for our project due to its special memory cells which allows it to look in the recent and the distant past values in order to make decisions. For our project, the dataset has a time series component, which can be further explored with LSTM. The input will be a window of recent weather observations which can be used by the model to

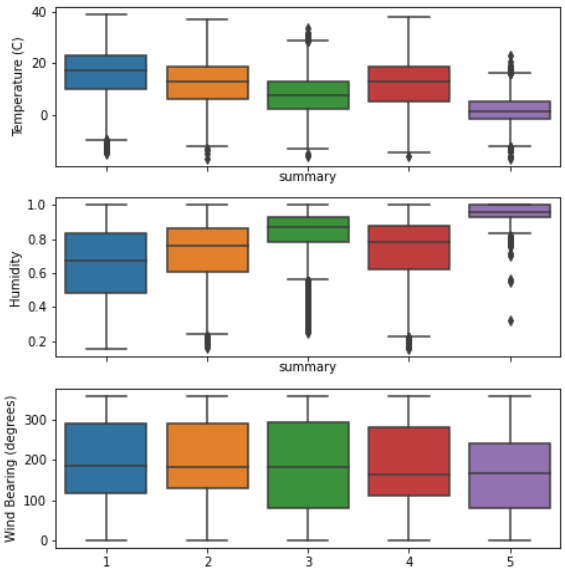

Fig. 4. Box plots

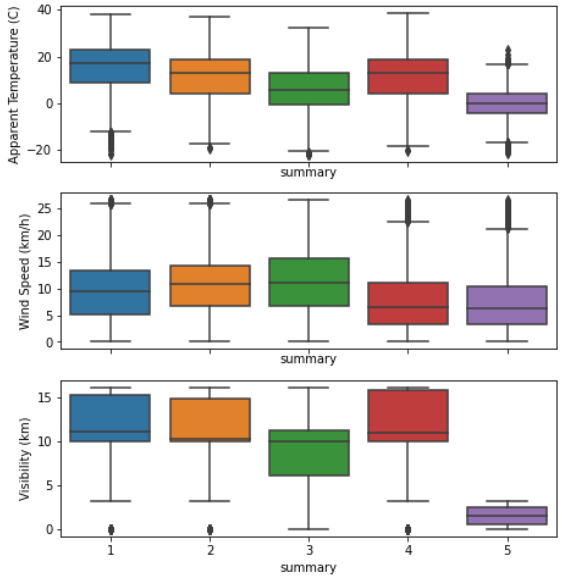

Fig. 5. Box plots

predict the future observations. [9]

B. Experimental Setup

Since the members for this project are unable to meet and run the program together on one device, each member will be working with their own device at home. Emmanuelle's computer is running Windows 10 Home edition with an Intel core i7-8700 with a 3.20 GHz processor and 8GB of RAM. Shaun's computer has Windows 10 Education and has an i7-2600 Intel core, 3.40GHz processor, and 16GB of RAM. Yash's computer is also running Windows 10 Education and has an Intel core i7-7700HQ with a 2.80 processor and 16GB of RAM.

1) Model Descriptions:

- Decision Tree - With the preprocessing steps completed a layered training approach is used to fit the DecisionTreeClassifier to our dataset. First, the target columns are listed, and every combination of this is formulated. There are 7 predictor attributes, providing for 127 combinations of attributes to be considered. For each of these combinations, a 5-fold cross validation procedure is used to split the dataset into a test and train set, therefore developing 5 separate Decision Tree models on this set of attributes. The model with the highest accuracy of these 5 is selected. From there, each of the 127 attribute combinations will produce their highest model and its accuracy, and the most accurate model is selected from these options. This is the most accurate Decision Tree that can be generated from the dataset. This model is shown in figure 6.

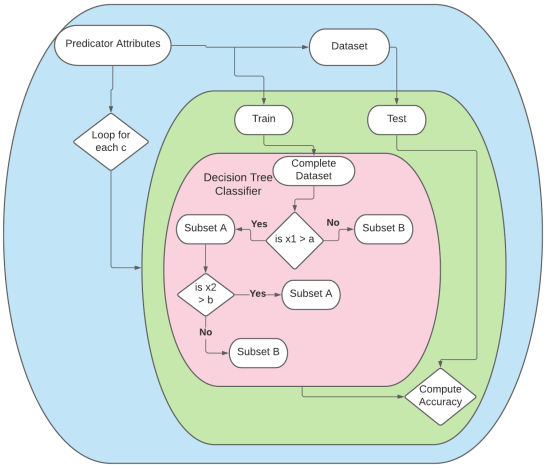

Fig. 6. Decision Tree Model

- Random Forest - Due to the similarities between this model and the Decision Tree described above, much of the implementation is copied from that model. In this situation a 5 fold cross validation approach is used, and is repeated on each combination of possible input predictors.
- K Nearest Neighbour - To begin, this algorithm will read the csv files containing the preprocessed test and train data. For the purposes of this project there were 3 different sets for the train/test datasets: preprocessed, preprocessed with normalization, and preprocessed with standardization. The model will be tested with all three of these train/test sets to determine which one receives the highest accuracy. Once that model/dataset is determined, hyper-parameter tuning will begin to determine the optimal values for the leaf size, k value, and find the best distance calculation method. The leaf size will be changed to optimize the construction time, query time, and memory required.

- Naive Bayes - The naive bayes model will have the same methodology as the k nearest neighbour. It will determine the model that provides the highest accuracy when using the test and train models that are preprocessed, preprocessed and normalized, and preprocessed and standardized. For naive bayes, there is only one parameter that can be tuned: variation smoothing. The best value for this parameter will be determined. Once the best var_smoothing parameter was decided, the correlations between each of the different columns were determined. Since the naive bayes model will give lower accuracies when there are multiple columns with high correlation, some columns may be removed.
- Multi-Layer Perception (ANN) - The MLP is a deep learning model. At its core, it is a vanilla artificial neural network which is being used for multivariate classification. The MLP model when compared to classic machine learning models performs better when provided with a bigger dataset. The MLP model learns through back propagation and can also perform feature extraction during training. Hyper-parameters for the MLP include hidden layers, neurons per hidden layer, activation function and a loss function for back propagation.

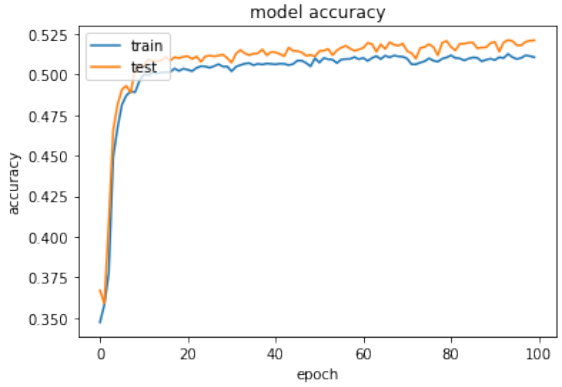

Fig. 7. Model Accuracy for ANN

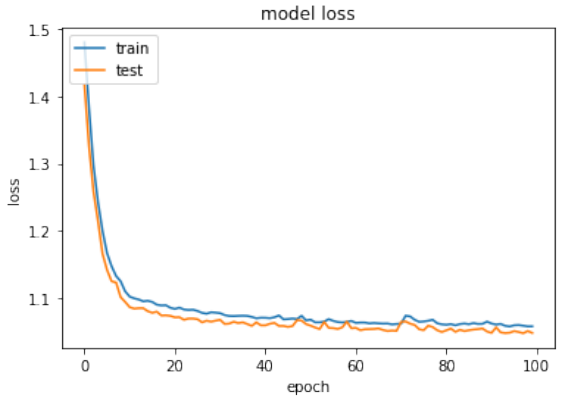

Fig. 8. Model Loss for ANN

IV. Results & Discussion

Each of the different models were implemented and compared to one another. The comparison between the models looked at both the elapsed time for the training of the model and the maximum accuracy the model was able to predict.

TABLE I
ACCURACIES AND TRAINING TIMES FOR EACH MODEL

Model	*Max Accuracy*	*E.T. On Individual Computers*
Naive Bayes	47.2%	140.918s - Emma
Decision Tree	47.7%	241.810s - Shaun
ANN	50.1%	10.865s - Yash
KNN	53.1%	1746.833s - Emma
Random Forest	53.9%	1490.393s - Shaun

As you can see from the table, the model with the highest accuracy of 58% was ANN; this just so happened to be the model that also required the least amount of training time as well. So, clearly, this model is the best for our purposes.

The literature review performed above was ordered in a specific way, mainly from least complex (numerical methods) to most complex (neural networks and LSTM); this just happens to usually have a positive correlation with the accuracy of the model. Unfortunately, this is not the case for this project. After the numerical models were determined to not be accurate enough, researchers attempted to use naive bayes, then decision trees. These models being the least complex implemented for this project yielded the lowest accuracies of 47 and 48 respectively. The decision tree model implemented here had one of the lowest accuracies. Chauhan and Thakur[4] said that they believed decision trees were the most promising method of classification and prediction. This is not true here, but we were able to make models that yielded better results such as the ANN model and KNN model. The ANN model should usually yielded the best results in literature since it is the most complex, but we were unable to perform hyper-parameter tuning on this model due to a limitation on computational power. With hyper-parameter tuning, the model would probably receive the highest accuracy. As explained by authors Cho, Yoo, and Im[10], neural networks have a problem with bias. Since the model was built in an isolated environment, there should be no bias. Our KNN model is unfortunately not as good as the models implemented by Huang, Lin, Huang, and Xing[5], since they were able to compare their KNN model to multiple different modified models, WKNN and DWKNN. Modifying the KNN model was outside the scope of this project, so the basic KNN model was used even though it received lower accuracy. In terms of complexity, we believe random forest to be under neural networks; this model received the highest accuracy at 57. Unfortunately, this model also had one of the longest training times, but this should not influence the fact that it is the best model since this would change depending on

the data and the technology the model is ran on. Therefore, the random forest model is the best model implemented.

In our experiment, different classification models were used which helped gain better insight into the dataset and its hidden features. Our method was highly detailed, thoroughly researched, and formed based on the previous work done in the field. Our method incorporated the best qualities observed from different implementations in the field. This was done to mitigate maximum number of errors and risk as early as possible. For most of the models, rigorous hyper-parameter tuning was done to receive the best results.

Weather has always been very difficult to predict since many surface features such as temperature and precipitation can change drastically and quickly; it is a very inaccurate science that a wide variety of people from different domains have attempted to solve. Even today with plenty of satellites, weather meters, and much more, the accuracy of the weather forecasting is still quite low. Machine learning brings in a whole new method of predicting weather which can yield fairly accurate results(60%+), so we implemented many different algorithms to predict the summary of the day. Originally, the precipitation was what we anticipated predicting, but even the basic prediction models were giving 100% accuracy. This is clearly incorrect; this was because it was a binary column with just “rain” and “snow”, and the dataset itself was dramatically skewed towards “rain”. In order to overcome this difficulty, the summary of the day was predicted instead. In the summary of the day column, there were 8 values, but since the majority of the data was in the first 5 output labels, only they were used: “partly cloudy”, “mostly cloudy”, “overcast”, “clear”, and “foggy”. Using these as the values for prediction worked well and gave us fairly accurate predictions.

The models developed in this project show that there are characteristics that are related in weather patterns. Using data from some elements of weather characteristics such as pressure, temperature, wind speed, direction, and others, it is possible to predict the climate conditions that affect the visibility. Cloudy or foggy weather conditions can be predicted from the data with the accuracies shown above.

The results of the models developed can be used in a wide range of applications beyond basic weather forecasting by supporting areas that are impacted by climate conditions. Accurate ability to forecast weather events before they happen will enable construction companies to better plan their operations to take advantage of warm and dry weather conditions when it is needed, and thus can prioritize other tasks for days that are anticipated to have suboptimal conditions. Other industries have a more direct impact from weather and can make use of more reliable predictions in order to better use their resources; with better forecasts snow removal can be optimized by using the correct amount of salt or sand on roads closer to the time that it will be required to prevent any from being scattered by being driven over repeatedly. The models developed in this project predict the conditions that affect visibility and are directly related to the amount of sun exposure experienced by the surface. This is a vital attribute to understand in order to be able to predict the effect that the sun will have on warming the surface and supporting agricultural products.

While performing the experiment, some of the models implemented had over fitting issues. These issues arose as a result of high number of iterations and using a lot of features. These issues were later fixed by dropping some features and by performing hyper-parameter tuning using GridSearch methodology to obtain the best results from the implemented models.

Since our project compared the accuracies of multiple different models to determine the best one for weather prediction, a larger system with similar column variables would be able to know which model to use without having to implement them. In the future, a more in depth analysis of the models could be performed with more datasets with a variety of column variables that could be created in order to determine which model is best overall for predicting the weather. The use of multiple different datasets would also help improve the accuracy of the models above since there are more training values; model training and the amount of data used to train the model are directly proportional to each other. So, the more training done with more datasets, the higher the accuracy of the model.

V. Conclusion

This project analysed the performance and accuracies of several different machine learning models and attempted to combine their results using Ensemble Learning to determine if a more effective model could be developed using these procedures. Using the dataset available to this project, Ensemble Learning was not able to make a positive contribution to the accuracy of the predictions, instead yielding an accuracy that was between the best and worst models developed during the project.

It is an important point to note that the Random Forest was able to develop the most accurate predictions for this data. The Random Forest model used in this project incorporates elements of Ensemble Learning in its algorithm to combine many random decision trees and develop the necessary weights and voting mechanisms to determine a single output from this collection. This suggests that it may be possible to develop an Ensemble Learning mechanism that could produce effective results, but further research would be required to determine how this can be done.

Further consideration could also be given to the use of hybrid models that are able to incorporate known relationships between the data to allow for additional input features to be provided to the machine learning models to develop more

accurate predictions with. This would be a difficult undertaking that would require gathering highly detailed data with sufficient data points to be of use in the training and testing of the models developed.

Accurate weather forecasting would allow for better planning within a range of commercial applications to make better use of clear weather and ensure preparations are made for less favourable weather conditions. Many construction activities are limited to taking place in conditions that are free of precipitation while others can happen during any conditions. Being able to plan activities that can only happen on clear days for when that will be available will allow for significant cost savings and accelerated project time lines. Further consideration in hybrid models using more detailed information in conjunction with more complex Ensemble Learning models have the potential to make this a reality.

REFERENCES

[1] V. Krasnopolsky & M. Fox-Rabinovitz, "Complex Hybrid Models Combining Deterministic and Machine Learning Components for Numerical Climate Modeling and Weather Prediction," Neural Networks, 2006, vol. 19, no. 2, pp. 122-134, doi: 10.1016/j.neunet.2006.01.002

[2] R. Zhang, Z. Chen, L. Xu & C. Ou, "Meteorological Drought Forecasting based on a Statistical Model with Machine Learning Techniques in Shaanxi Province, China," The Science of the Total Environment, 2019, vol. 665, pp.338-346, doi: 10.1016/j.scitotenv.2019.01.431

[3] W. Wei, Z. Yan & P. Jones, "A Decision Tree Approach to Seasonal Prediction of Extreme Precipitation in eastern China," International Journal of Climatology, 2019, vol. 40, n.1, pp. 255-272

[4] D. Chauhan & J. Thakur, "Boosting Decision Tree Algorithm for Weather Prediction," Journal of Advanced Database Management & Systems, 2014, vol. 1, no.3,

[5] M. Huang, R. Lin, S. Huang & T. Xing, "A Novel Approach for Precipitation Forecast via improved K-Nearest Neighbor Algorithm," Advanced Engineering Informatics, 2017, vol. 33, pp.89-95, doi: 10.1016/j.aei.2017.05.003

[6] R. Prasetya & A. Ridwan, "Data Mining Application on Weather Prediction Using Classification Tree, Naive Bayes and K-Nearest neighbor Algorithm With Model Testing pf Supervised Learning Probabilistic Brier Score, Confusion Matrix and ROC," Journal of Applied Communication and Information Technologies, 2019, vol. 4, no. 2, doi: 10.32497/jaict.v4i2.1690

[7] F. Hashim, N. Duad, K. Ahmad, J. Adnan & Z. Rizman, "Prediction of Rainfall Based on Weather Parameter Using Artificial Neural Network," Journal of Fundamental and Applied Sciences, 2017, vol. , no. 3S, doi: 10.4314/jfas.v9i3s.38

[8] D. N. Fente and D. Kumar Singh, "Weather Forecasting Using Artificial Neural Network," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2018, pp. 1757-1761, doi: 10.1109/ICICCT.2018.8473167.

[9] A. Salman, B. Kanigoro & Y. Heryadi, "Weather Forecasting using Deep Learning Techniques," 2015 International Conference on Advanced Computer Science and Information Systems, 2015, pp. 281-285, doi: 10.1109/ICACSIS.2015.7415154

[10] D. Cho, C. Yoo, J. Im & D. Cha, "Comparative Assessment of Various Machine Learning-Based Bias Correction Methods for Numerical Weather Prediction Model Forecasts of Extreme Air Temperatures in Urban Areas," Earth and Space Science, 2020, vol. 7, no. 4, doi: 10.1029/2019EA000740