Disclaimer

A report is submitted to Dublin City University, School of Computing for module Practicum in course Data Analytics.

We understand that the University regards breaches of academic integrity and plagiarism as grave and serious. We have read and understood the DCU Academic Integrity and Plagiarism Policy. We accept the penalties that may be imposed should we engage in practice or practices that breach this policy.

We have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the sources cited are identified in the assignment references.

We declare that this material, which we now submit for assessment, is entirely our own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of our work.

By signing this form or by submitting this material online we confirm that this assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study. By signing this form or by submitting material for assessment online we confirm that we have read and understood DCU Academic Integrity and Plagiarism Policy.

Date: 17th April 2021

Github link: https://github.com/prasad1825/Irish-rainfall-Prediction.git

Irish Rainfall Analysis and Prediction Using Data Mining Techniques.

Prasad Govardhankar
prasad.govardhankar2@mail.dcu.ie
Dublin City University (Msc. Computing)
20210305

Pracheta Munje
pracheta.munje2@mail.dcu.ie
Dublin City University (Msc. Computing)
20214193

Rasika Mehta
rasika.mehta3@mail.dcu.ie
Dublin City University (Msc. Computing)
20210036

Shubham Bhosale
shubham.bhosale2@mail.dcu.ie
Dublin City University (Msc. Computing)
20210879

***Abstract*— This research discusses many aspects that contribute to rainfall prediction. Despite, advancements in technology, it is quite difficult to predict accurate weather. As we all know how it is important in our day-to-day life to check the weather before going outside these days. In Ireland, the unpredictability of the weather is very likely, but particularly accurate rainfall prediction is very much difficult. The research is based on data which is gathered from the Met Eireann website which is readily available on the Kaggle website from 1990 to 2020. An exploratory analysis was conducted to assess the most predictive variables for rainfall prediction. Weather prediction is done previously using many techniques like observing wind, changes in air pressure, and computerized methods. As a result, this assignment aims to use various data mining and analytics techniques to obtain useful information trying to predict the accurate rainfall prediction for Irish weather. In this paper, we have used the CRISP-DM model and its six methodologies are implemented to get the rainfall predictions. Based on the comparison of various machine learning models, we concluded that Random forest and XGB boost are the best performing models for our prediction.**

Keywords- Random Forest, CRISP-DM, Extreme Gradient Boosting, Logistic Regression, Decision Tree, K-nearest neighbors,Root mean square error

I. Introduction

Rainfall is the most significant climate factor affecting the majority of Ireland's livelihood and well-being. The development of a predictive model for accurate rainfall is one of the most difficult challenges for researchers from a variety of fields, including weather data mining, environmental machine learning, functional hydrology, and numerical forecasting. How to infer past predictions and make use of future predictions is a common question in these problems. Early warning of severe weather will help prevent natural disaster accidents and damage if timely and reliable predictions are made, so accurate precipitation forecasting has been a major problem in hydrological science. There have been much research and implementations done in the field of weather predictions and their forecasting. Rainfall is also a significant factor in evaluating water resources, irrigation, habitats, and hydrology. Weather prediction is considered as the most delicate research field which confronting a ton of ongoing issues like false forecasts, absence of taking care of in immense information volume, and insufficient innovation headway. The most commonly used computational approaches for weather prediction are regression, Artificial Neural Network (ANN), Decision Tree algorithm, Fuzzy logic, and team data handling processes. In the proposed system, rainfall forecasting is done for Ireland's 15 different counties. In this, data from 1st Jan 1990 to 1st June 2020 is used. Hourly prediction of rainfall is done using the above-mentioned data.

II. RELATED WORK

A study was done by Ramesh Medar, Akshata B. Angadi, Prashant Y. Niranjan, and Pushpa Tamase to compare application suitable Data Mining techniques, Regression approaches, and Artificial Neural Network models to predict weather parameters.[1] The main aim of this study was to compare and recognize precise models for weather forecasting. forecasting models used in this paper are namely: Data Mining, Multivariate Linear Regression, Autoregressive Integrated Moving Average Model(ARIMA), Artificial Neural Network Model(ANN), Hybrid Model of Multiple Linear Regression and Artificial Neural Network (MLR ANN). An Artificial Neural Network is an amazing data mining tool that gives a strategy to tackling numerous kinds of issues that are hard to settle by customary procedures. Neural Network makes low amount presumptions, rather than ordinariness usually found in measurable strategies a few works have been done, and distinctive Artificial Nural Network(ANN) models, have been tried. Kaur[2] and Maqsood[3] depicts a model that predicts the hourly temperature, wind speed, and relative mugginess 24 hours ahead. Above mentioned approaches helped us for clarification of methods used, evaluation of methods, and parameters on which those should be compared.

Bin Wang and co-ordinates have provided a new method called a deep uncertainty quantification (DUQ) method to create a weather forecast.[4] For us to use Keras, this paper has helped in ways to implement different methodologies to predict the forecast. Essentially scalable solutions square measure needed in forecasting to cope up with the ever-increasing size of data at hand. Namrata, Jayapraya, and Santosh have implemented Artificial Neural Network on Map-reduce framework for weather prediction.[5] This research helped us to answer our question of handling big-data problems for processing. Implementation over Hadoop was done to get more scalable results. Mingming Huang, Runsheng Lin, Shuai Huang, and Tengfei Xing have used an improved version KNN algorithm which offers performance against completely different decisions of the neighborhood size k.[6] We have researched upon this paper, but decided not to move forward with this due to lack of method explanation was given. The research proposed by S. C. Sreenivasa contrasts the proposed model and ANN, ANFIS models for the momentary forecast of wind speed.[7] Limitation for this comparative study was the small amount of data provided to the algorithm. To use XG Boost for weather forecasting we have referenced *machinelearningmastery* website which gives a detailed explanation of how to use XGB and which things to be considered to get the appropriate results.[8] Also, for regression analysis we have referenced one of the articles proposed by Jim Frost where he has explained different approaches, explained how analysis models can be performed under different conditions such as limited data.[9] Accessing the precision score using various methods has been useful from this article. A study was done by Palla Ratna Sai Kumar, Mylapalle Yeshwanth, Dr.G. Mathivanan shows us that yearly information was gathered to predict Indian rainfall. Also, they have used multiple linear regression, random forest regressor, and AdaBoost regressor to predict precipitation for the next day. However, the multiple linear regression model outperforms the other two models.[10]

This research paper helps us in predicting Indian rainfall prediction using deep learning approaches, for example, Artificial neural network, Multilayer perceptron, and Auto-encoder neural network. But, from this paper, it is evident that the ANN algorithm outperforms the other deep learning algorithms.[11]The research proposed by Moulana Mohammad and Roshitha kolapalli predicts rainfall basically in two categories like short-term rainfall and long-term rainfall prediction using machine learning techniques. Also, they have used three prediction models: support vector machine, Lasso regression, and multiple linear regression but the SVM algorithm outperforms the other two models.[12]In this paper, neural networks such as multi-layer perceptron, radial basis function network this model are used for the prediction of local rainfall of Japan. The meteorological dataset is used and backpropagation and random optimization methods are used to train 3LP model.[13]

A. Research Questions:

1. What factors play a role in analyzing and predicting rainfall?

2. Which machine learning algorithm is best for analyzing rainfall data and predicting precipitation?

The primary goal of this study is to find answers to the questions mentioned above. The first research issue concerns the collection of precise or improvised attributes from a column of options. The second question is based on the research conducted for this report, which found that the best-fitting model is among regression algorithms. The CRISP-DM approach can be used to find the answers to the above questions.

III. DATASET AND EXPLORATORY ANALYSIS

A. Dataset:

Dataset used in this study is taken from kaggle.com. The dataset is in CSV format consisting of 1.8 million rows and 18 columns or attributes. To keep the project simple we have added one target column i.e. Rain or not column and using this target column we have predicted the rainfall at a binary level i.e. rain(1) and for no rain(0).

B. Data Pre-processing

1) Data Collection: Met. ie screens, investigates, and predicts Ireland's climate and environment, and gives a scope of great meteorological and related data. This site provides data which is hourly, daily, and monthly collected from across 16 weather stations of Dublin county. As per our approach, we have taken the dataset from Kaggle which was originally taken from Met. ie. In this dataset, hourly data is considered from over 15 counties from Ireland dating from the year 1990 to 2020.

2) Data Understanding: Historical data is just estimations or occasions that are followed, observed, down-sampled, and aggregated after some time. For weather prediction, data needs to validate around 3 properties which are *Volume, Variety, and Velocity*. To correctly predict future points, the data needs to be complete, for that all the null values, NaN values, missing values needs to be removed and processed.

3) Data Preparation: Data gathered was discrete and uncleaned. Data were collected and processed in Google Co-laboratory using *Jupyter Notebooks*. The following figure shows the data gathered:

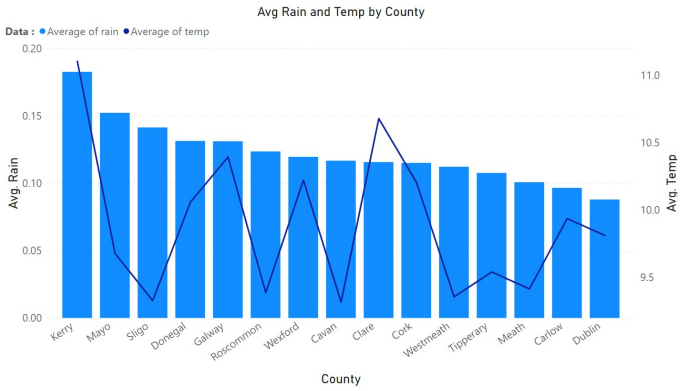

Fig. 1. Data Gathering

4) Data Sampling: To address the imbalance issue in the dataset, this analysis uses a resampling approach. The count of 0 in the goal column was charged off as 80 percent, while the 20 percent count was for 1 as shown in the figure below. There are several approaches to dealing with unbalanced data. We have examined that and used undersampling to sample the data.

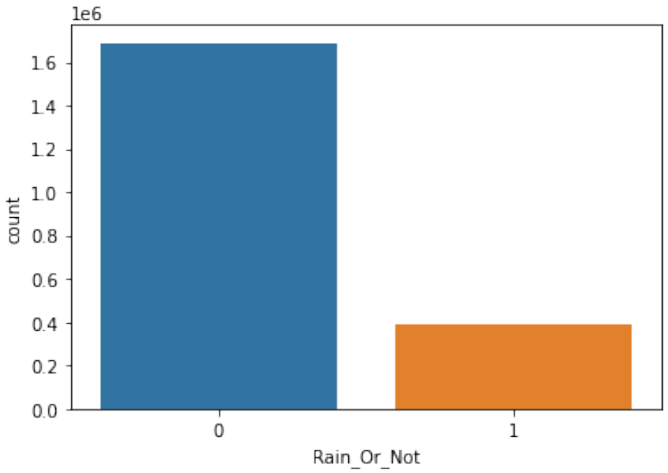

Fig. 2. Data Imbalance

The data is split and sampled into two groups, 0 and 1, for no rain and rain, respectively, as shown in the figure below and then modeled with Random forest and XGB boost, and its performance factors are verified using ROC.

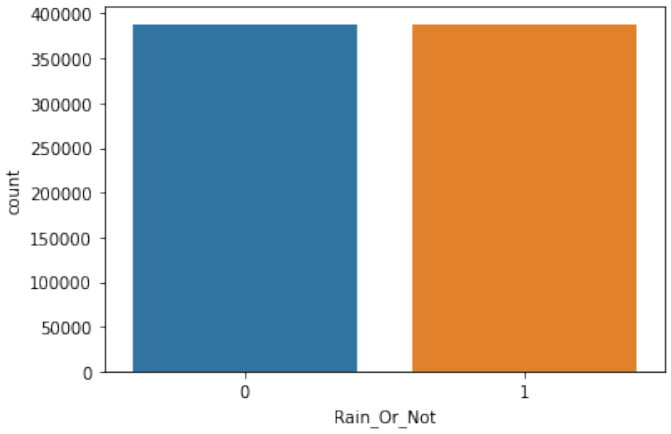

Fig. 3. Balanced Data

C. Machine Learning Approach

Several steps are followed in implementing the Machine Learning algorithm. These steps are as follows:

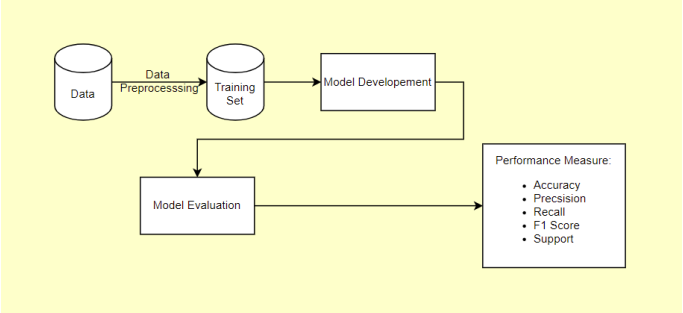

Fig. 4. Architecture Followed

- Data Preparation: This step will be used to gather, analyze, pre-process and clean the data as per the requirements.
- Model Definition: Defines and trains the model of implementation for weather prediction and sets the scope of comparison.
- Performance Measure: Evaluation of models used in this research based on Accuracy, Precision, Recall, F1 Score, and ROC.

D. Feature Selection

Feature selection can be done in multiple ways but there are broadly 3 categories of it:

- Filter Method
- Wrapper Method
- Embedded Method

In our research, we used **Wrapper method** which can be derived as *RFE(Recursive Feature Elimination)* which works by first removing the attributes and then generating a model on those filters/attributes. Afterwards the relevant features are kept and others are discarded.

E. Final Parameters Used:

Following parameters are derived after feature selection and used for implementation:

- Temperature(temp)
- Wet Bulb Temperature(WetB)
- Dew Point(Dewpt)
- Vapour Pressure(Vappr)
- Relative Humidity(RHum)
- Mean Sea Level Pressure(MSL)
- Wind Speed(Wdsp)
- Wind Direction(Wddir)
- Date(DD, YY, HH)

IV. Evaluation and Results

In this study, we have built two models: Random Forest Regressor and Extreme Gradient Boosting Regressor. In comparison, other algorithms such as KNN, Nave Bayes, Logistic Regression, and Decision Tree are not considered because their success was insignificant and their observations were trivial. We used Python 3.7 on the Anaconda distribution with Google Collab Notebooks to implement the models.

Furthermore, we have used sklearn's train, test, and split method to divide the dataset for training and testing. We have divided 65% for training and 25% for testing as the training model on more than 65% was overfitting the models. The final input columns taken are temp, wetb, dewpt, vappr, rhum, msl, wdsp, wddir, day, year, hour and the target column given is 'Rain Or Not'. For evaluation, we have used sklearn's accuracy score metrics and plotted ROC for each model.

1) Random Forest: Random decision forest is a Machine Learning algorithm for data processing, analysis, and prediction. Since the production and execution of tree works occur concurrently, random forest is called a Bagging technique. During training, decisions are made by creating a large number of decision trees, and the performance is in the form of a class. The key feature of random forest architecture

is the hyperparameter. Any small percentage of the total number can be divided into more nodes in each node. This algorithm extracts k – data points from the training dataset and constructs the decision tree for these k data points. The preceding step is repeated with several trees equal to the n estimators parameter specified during model fitting. At each iteration, it produces a new data point and predicts the value of the goal variable for that data point. Finally, the average of new data points is used to assign all of the expected goal values. For training the model, some of the features of this dataset are used such as temp,wetb,dewpt,vappr,rhum,msl,wdsp,wddir,day, year, and hour. The model is trained using the RandomForestRegressor function of python which is available in sklearn library. Below are the scores achieved through it.

0.7905491980736093

	precision	recall	f1-score	support
0	0.81	0.76	0.78	135520
1	0.78	0.82	0.80	136076
accuracy			0.79	271596
macro avg	0.79	0.79	0.79	271596
weighted avg	0.79	0.79	0.79	271596

Fig. 5. Random Forest Scores

The below diagram represents the roc curve for it.

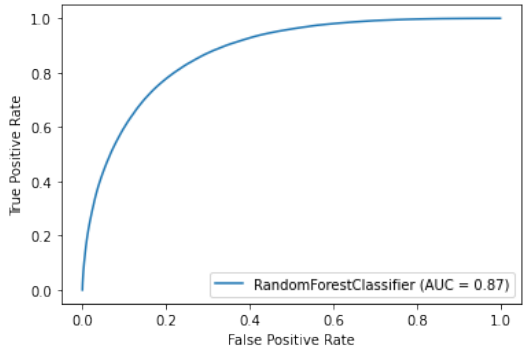

Fig. 6. Random Forest ROC

2) Extreme Gradient Boosting: Extreme Gradient Boosting is a Python machine learning library that can perform regression, classification, and ranking. The loss function is reduced at each test value as the model's pseudo residuals are reduced. It is a more sophisticated version of the Gradient Boosting model. XGBRegressor is for continuous outcome variables. These are sometimes referred to as "regression issues." Since this algorithm was originally written in C++, it is faster than other ensemble models. It is also a parallelizable model, making it possible to train on massive datasets. It has internal parameters for cross-validation, regularization, missing values, and tree parameters, among other things. The parameters used in the Random Forest classification are the same as the input vector used for training this model.

0.7748530906198913

	precision	recall	f1-score	support
0	0.79	0.74	0.77	135520
1	0.76	0.81	0.78	136076
accuracy			0.77	271596
macro avg	0.78	0.77	0.77	271596
weighted avg	0.78	0.77	0.77	271596

Fig. 7. XGB Accuracy scores

The below diagram represents the roc curve for it.

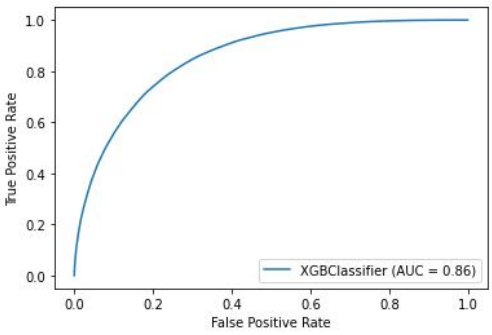

Fig. 8. XGB ROC

V. CONCLUSIONS

To summarize the study, we can confidently state that there is no significant difference in the accuracy of models logistic regression, KNN, Decision tree, and Navies Bayes, except for XGB and random forest, which predict with higher accuracy on the same features as you can see in figure 8.

We also enhanced the Random forest's accuracy by using the RandomizedSearchCV process. The classification accuracy is a troublesome measure for imbalanced classification, which is a common scenario in rainfall prediction, which may lead to incorrect decisions. The ROC curve, which

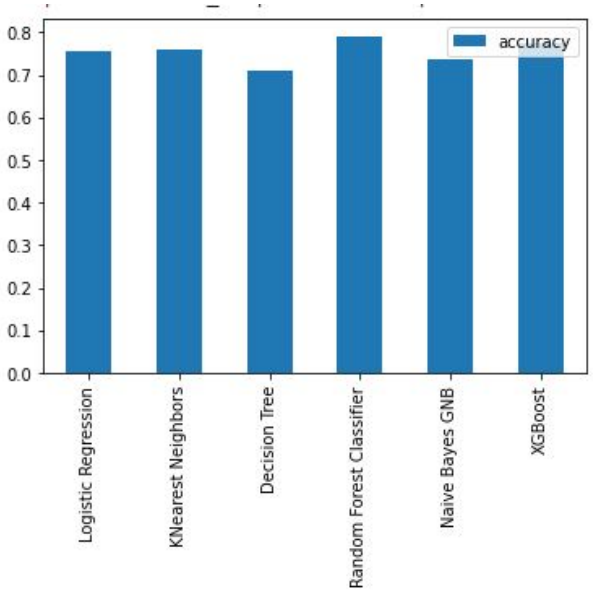

Fig. 9. Comparison of various models

compares true positive rates to false positive rates, is a better measure for these types of models. The research' source code can be found on the shared collab. In the future, CNN can be used to investigate techniques such as Feedforward and backpropagation on the dataset "Irish weather." CNN can detect hidden layers and patterns, which can increase predictive ability and assist us in making more accurate predictions.

REFERENCES

[1] R. Medar, A. B. Angadi, P. Y. Niranjan and P. Tamase, *"Comparative study of different weather forecasting models,"* 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 2017, pp. 1604-1609, doi: 10.1109/ICECDS.2017.8389719.

[2] Amanpreet Kaur, J K Sharma, and Sunil Agrawal.,*"Artificial neural networks in forecasting maximum and minimum relative humidity"*, International Journal of Computer Science and Network Security, 11(5):197-199, May 2011.

[3] Imran Maqsood, Muhammad Riaz Khan, and Ajith Abraham, *"An ensemble of neural networks for weather forecasting: , Neural Computing and Applications"*, pp. 112–122, 2004.

[4] Bin Wang, Jie Lu, Zheng Yan, Huaishao Luo, Tianrui Li, Yu Zheng, and Guangquan Zhang. 2019. Deep Uncertainty Quantification: A Machine Learning Approach for Weather Forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19).* Association for Computing Machinery, New York, NY, USA, 2087–2095. DOI:https://doi.org/10.1145/3292500.3330704

[5] K. Namitha, A. Jayapriya, and G. Santhosh Kumar. 2015. *Rainfall Prediction using Artificial Neural Network on Map-Reduce Framework.* In Proceedings of the Third International Symposium on Women in Computing and Informatics (WCI '15). Association for Computing Machinery, New York, NY, USA, 492–495. DOI:https://doi.org/10.1145/2791405.2791468

[6] Mingming Huang, Runsheng Lin, Shuai Huang, and Tengfei Xing. 2017. *A novel approach for precipitation forecast via improved K-nearest neighbor algorithm.* Adv. Eng. Inform. 33, C (August 2017), 89–95. DOI:https://doi.org/10.1016/j.aei.2017.05.003

[7] S. C. Sreenivasa, S. K. Agarwal and R. Kumar, *"Short term wind forecasting using logistic regression driven hypothesis in artificial neural network,"* 2014 6th IEEE Power India International Conference (PIICON), Delhi, India, 2014, pp. 1-6, doi: 10.1109/POWERI.2014.7117710.

[8] *"How to Use XGBoost for Time Series Forecasting"* by Jason Brownlee on August 5, 2020.

[9] *"Making Predictions with Regression Analysis"* by Jim Frost.

[10] M. Yeshwanth,P. R. S. Kumar, Dr. G. M. M. E. Ph.D and Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India*"Comparative study of machine learning algorithms for rainfall prediction, "*IJTSRD, vol. Volume-3, no.Issue-3,pp.677-681, Apr.2019,doi:10.31142/ijtsrd22961

[11] C. Z. Basha, N. Bhavana, and P. Bhavya*"Rainfall Prediction Using Machine Learning and Deep Learning Techniques,"*p.6,2020

[12] M. Mohammed, R. Kolapalli, N. Golla, and S. S. Maturi*"Prediction of Rainfall using machine learning techniques,"*vol. 9,no. 01, p. 5,2020

[13] T. Kashiwao, K. Nakayama, S. Ando, K. Ikeda, M. Lee, and A. Bahadori*"A neural network-based local rainfall prediction system using meterological data on the internet:"*A case study using data from the Japan Meteorological Agency,"p.43