National College of Ireland

Project Submission Sheet

Student Name:	Ciara Fallon, Cormac Hill & Ailis Curran		
Student ID:	x23170964, x23239252 & x23165669		
Programme:	Postgraduate Diploma in Data Analytics	**Year:**	Year 1 Semester 3
Module:	Domain Applications of Predictive Analytics		
Lecturer:	Vikas Sahni		
Submission Due Date:	29/07/24		
Project Title:	Predictive Analytics of floods in Bangladesh		
Word Count:	2,527		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: CF, CH, AC

Date: 28/07/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	

Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

I. Domain Application of Predictve Analytics

Predictive Analytics of floods in Bangladesh

Your Name/Student Number	Course	Date
Ciara Fallon (x23170964), Cormac Hill (x23239352), Ailis Curran (x23165669)	Postgraduate Diploma in Data Analytics	28/07/24

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

II. AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool

III. Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

IV. Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

V. Additional Evidence:

[Place evidence here]

Predictive Analytics of floods in Bangladesh

Ailis Curran - x23165669 Ciara Fallon - x23170964 Cormac Hill - x23239352

***Abstract*—The battle with climate change and the effects of global warming grow increasingly prevalent with each passing year. Climate disasters like flooding continue to grow in frequency and intensity. Take for instance the record flooding that ravaged regions in the south of Ireland in the wake of Storm Babet in October of 2023. Another region that is heavily effected by flooding is Bangladesh, with the country being bombarded by extreme flood waters every year with the impact of each flooding event growing increasingly menacing. This paper outlines research undertaken to develop a K-Nearest Neighbours machine learning model for predicting the occurrence of flooding events in Bangladesh. In doing so, key weather data factors that can lead to flooding events were also identified to aid in development and deployment of similar flood prediction models worldwide. The model was trained on historic Bangladeshi weather station data and the final model achieved a mean accuracy of 88.9% in predicting flooding events.**

***Index Terms*—predictive analytics, flood prediction, machine learning, binary classification, K-Nearest Neighbours**

I. Techniques

As outlined in the project proposal, the goal of this project is to develop a machine learning model that can provide robust and accurate prediction of flooding events occurring to mitigate damage to infrastructure, minimise threats to the population, and potentially prevent flooding entirely in some cases. Given that the model being trained for this project aimed to predict flooding events based on weather station data, the two most immediately obvious options for techniques were clustering and time series methods. Clustering is a strong machine learning method for this topic as it is almost certain that the same weather patterns will result in the same outcome with regard to flooding, for example if a metre of rainfall lead to a flood, than a future weather forecast of a metre of rainfall will nearly guarantee another flood occurring. The main clustering method considered for this was K-Nearest Neighbours (KNN), as seen in other papers for similar purposes such as the work of Zhou et al. [1] where they utilised a KNN algorithm to rapidly forecast flash flooding, and the paper by Gauhar et al. [2] where they also trained a KNN model for predicting flooding events in Bangladesh. As mentioned, the other technique considered for implementation was a time series model. This has also previously been employed for flood prediction, such as in the model developed by Damle et al. [3] which made use of time series data mining to predict future discharge of rivers to infer the likelihood of flooding.

While both methods we researched held weight, we determined that time series analysis methods would not be sufficient for the goals of our project. Many implementations of time series analysis allow for use of just one feature, which would more than likely be the rainfall measurement for this project as this is typically the most commonly correlated attribute with regard to flooding occurring. By utilising only one feature to forecast flooding events, the remaining features of the dataset we had elected to use for training and testing would become redundant. As our research also aims to determine the various key weather factors that contribute to flooding, use of a more robust multi-feature prediction method, in this case KNN, better aligned with achieving the desired outcome of the project. The existing research in the field of flood prediction using KNN also showed great promise which we hope to build upon in our own research, achieving high accuracy of 99% in the work of Zhou et al. [1] and a mean accuracy of 92% in the paper by Gauhar et al. [2]. Reviewed literature detailing time series analysis showed very few concrete results for the models, and some did not even directly predict flooding events but simply inferred it based on prediction of other factors such as the work by Damle et al. [3]. All of these factors combined lead to the decision to use a KNN model as the technique to be implemented for this research paper.

II. Implementation

With KNN chosen as our technique, we began investigating methods of implementing such an algorithm. Python is among the most commonly utilised programming tools for machine learning model development as it provides a litany of easy-to-use and incredibly effective libraries for doing so. One such example is the SciKit-Learn library, which not only provides a vast array of machine learning algorithms and tools, including KNN, but it also has seamless integration with other key Python libraries such as Pandas and NumPy for handling data and complex mathematics respectively. SciKit-Learn has also been cited as the foundational library of multiple flood prediction research papers, including a general investigation of flood prediction using various machine learning models such as Binary Logistic Regression and KNN [4], and a case study which built models for flood prediction in Kebbi State Nigeria [5]. With these studies supporting the use of SciKit-Learn, we opted to utilise the library for development of our KNN model. Having chosen the technique to be implemented and the core tools for creating said implementation, work began on development of the model.

A. Data Exploration

The first step undertaken was to clean the dataset to prepare it for use in training the model. This included dropping features that contained redundant information, replacing the dependent variable values with 1 and 0 to indicate a flood occurring or not occurring, and re-coding categorical features

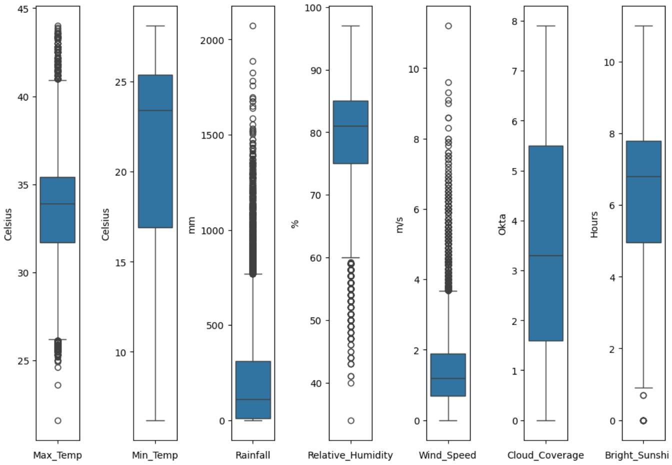

Fig. 1. Box plot of all numeric variables

such as the station names with numeric identifiers for later use. Following initial data cleaning, more advanced data exploration began. We first verified that there were no null values in any variable to ensure that all of the dataset was usable for training the model. We then determined the number of positive and negative cases there were in the dataset and found there to be 16 412 cases in which no flood occurred and 4 132 cases in which there was a flood, combining for a total of 20 544 cases.

To identify any outlying values or cases, a box plot was created for all numeric variables in the dataset (see Fig. 1). While the results of the box plot may look somewhat alarming initially, especially with regard to the maximum temperature, rainfall, and humidity, some thorough investigation revealed that there was no true concerning values in the data. The findings of this investigation include:

- Max_Temp: Highest ever recorded temperature in Bangladesh was 45.1°C [6], thus the range seen was deemed acceptable.
- Rainfall: Bangladesh's annual average rainfall ranges between 2 000mm and 5 000mm [7], which is far greater than the upper bound of any values in this dataset, so this was deemed acceptable.
- Humidity: Relative humidity in Bangladesh typically ranges between roughly 50% and 85% [8] and there are very few values exceeding these limits in the dataset, this was also deemed acceptable.

After determining there were no outlying variables, a correlation matrix of all features in the dataset was made to find any colinearity between them. From this there was a clear strong correlation between the features denoting the X and Y coordinates of each station and the longitude and latitude of each station. This is not unexpected as the feature pairs do contain near identical information despite the difference in their values so we elected to drop the X and Y coordinate features. The station number feature was also found to be strongly correlated with some of the aforementioned features and was found to be redundant anyway given that we had previously coded each station name as a number, so this feature was also dropped. A strong correlation was also observed between the features regarding hours of sunshine and cloud coverage, which is to be expected given that they are intrinsically linked in reality. We determined a principal component analysis (PCA) would be carried out on the features later in the project to attempt to combine these features. Lastly, while not of concern, it was noted from the correlation matrix that rainfall was highly correlated with flood occurrence, as anticipated in the outset of this paper.

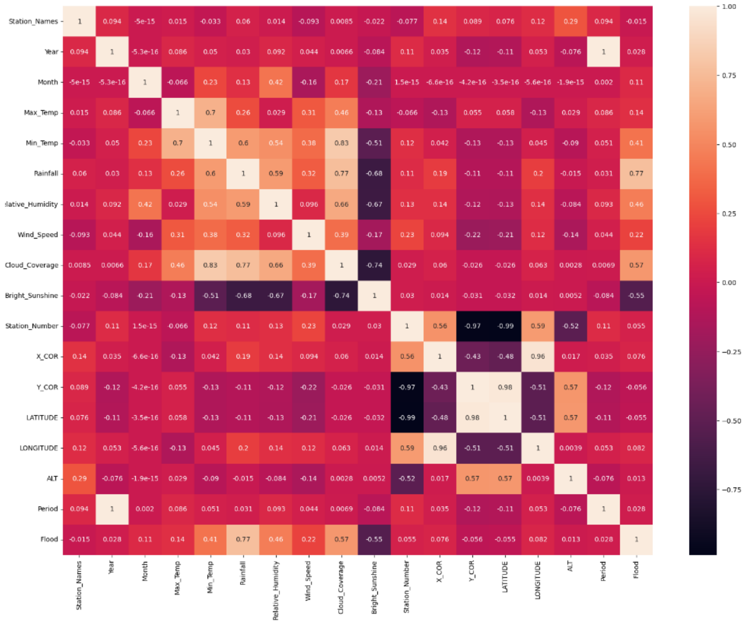

Fig. 2. Initial correlation matrix of all features

B. Feature Engineering

It is vital for clustering based machine learning methods such as KNN that the features be standardised to a similar scale before training. This was done for all numeric variables in the dataset. A random forest classifier was then fit on the dataset to test for feature importance. This allowed us to determine if there were any redundant features in the dataset that could be removed to reduce the dimensionality. From this analysis, we saw that the period and the year of each case were the two least important features of all fourteen features, while the month was the sixth most important indicating that the seasonality of the data, that is to say the time of year, was the only truly important aspect with regard to the effect that time has on flooding events occurring. This information is vital, especially in a region such as Bangladesh due to the country having a monsoon season from June to October every year which blankets the country and surrounding areas with torrential rainfall [9]. After identifying this, we elected to drop the period and year features.

As mentioned previously, a PCA was to be performed for the features detailing cloud coverage and hours of sunshine

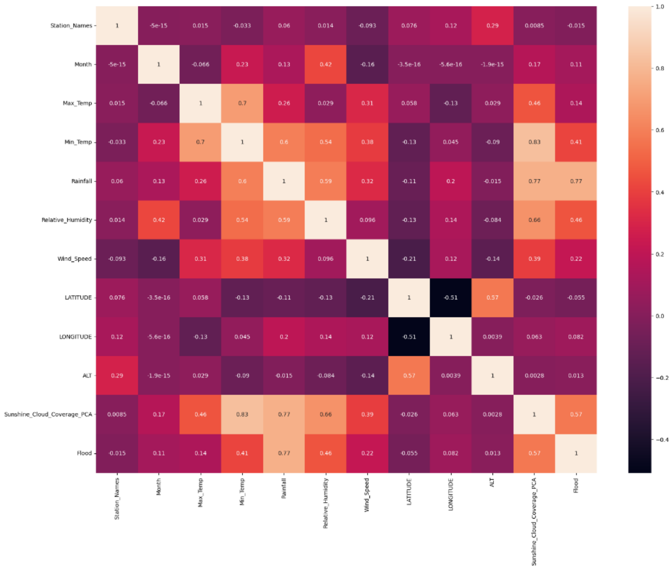

Fig. 3. Reduced correlation matrix of engineered dataset

due to the strong correlation between them. In the initial feature importance analysis, these features were ranked second and third most important respectively, and following a PCA to reduce them into a single feature, the resulting feature remained the second most important of the features. As the newly combined feature maintained a similar importance as that of the features used in the PCA it was deemed acceptable to maintain the new feature in place of the existing features.

Following this, a second correlation matrix was plotted with the engineered dataset. Fig. 3 shows a significant reduction in strongly correlated features throughout the dataset. With the results of this correlation matrix, we determined it was acceptable to begin training the model with the engineered dataset.

C. Training and Evaluation

Once the data had been sufficiently cleaned and engineered, work began to train the KNN model. To ensure the model was working as intended and to minimise the risk of overfitting, the model was fit once using an 80/20 training and test data split. This initial model achieved an incredible accuracy of 95.3% when predicting the occurrence of a flood. This initial iteration of the model was fit with an arbitrary K value of 10. To determine the optimal value for K the model was iteratively fit with K values ranging from 1 to 20 and tested for it's accuracy with each value for K. The accuracy achieved by each iteration was then plotted (see Fig. 4) and the most promising K value was chosen as the value to be utilised in further training of the model. The K value chosen for this model was four as it achieved the highest accuracy and also provides an acceptably sized neighbour cluster to ensure the model makes correctly informed predictions.

Purushotham et al. [10] suggests that stratified 10-fold cross-validation is among the optimal methods for testing the accuracy of a model. This validates that the model is truly generalised and robust in it's decision making when classifying unseen cases, and that it does not overfit the data provided to it in training. The KNN model was thus cross-validated using this method and the suggested parameters. The F1 score of the model was also evaluated to determine its ability of precision and recall.

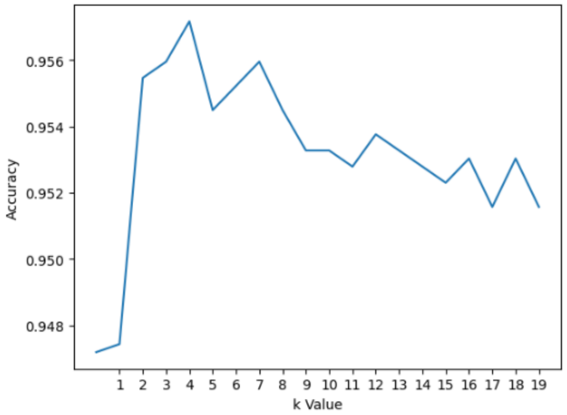

Fig. 4. Plot of accuracy achieved by K values from 1 to 20

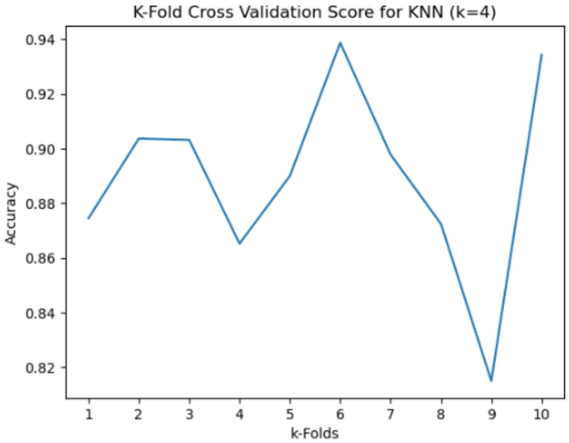

Fig. 5. Accuracy after each fold of stratified cross-validation

III. Results

After evaluating the KNN model with 10-fold cross-validation, as seen in Fig. 5 the model achieved a high accuracy of 94% and a mean accuracy of 88.9%. As the mean accuracy remained high throughout the cross-validation, we can see it is likely that the model is not overfitting the training data and is simply well generalised and capable of accurately predicting flooding events. While this high mean accuracy is a great achievement, we would ideally expect the model to achieve an accuracy of greater than 90% given the high risk nature of flood prediction and the implications it could potentially have for peoples safety.

The model was also tested for an F1 score after each fold of cross-validation. The mean F1 score across all folds of the stratified cross-validation was 71.1%. This indicates that the model achieved a relatively good balance between precision and recall, meaning it is capable of confidently identifying true positives and avoiding any false classifications. Similar to the raw accuracy of the model, this is a good score but it may not necessarily be deemed acceptable given the critical context of this particular model. As mentioned previously, flood prediction is an incredibly high risk area of prediction given the implications and impact it may potentially have for the safety of the people in the affected areas. This is certainly something that could be improved upon in future work in this field.

IV. Business Value

Broadly speaking, the business value of predictive analytics is that it allows for deeper understanding and more enhanced insights into a given topic. In the case of this project, the aim was to investigate the ability to predict flooding occurring and through doing so identify high contributing factors to flooding. By developing the model for predicting flooding in Bangladesh, key insights gained during the development such as the core contributing weather factors could be utilised for flood prediction on a global scale. The ability to predict flooding can provide meaningful insights for emergency response organisations to use in efficient planning and action, both preventative and reactive, which in turn may save lives and protect critical infrastructure from suffering from excessive damages. The ability to predict future flooding also lends itself to urban and financial planning, as determining the risk of flooding can allow planners to install preventative measures to minimise flood risks in areas that may be predicted as flood prone.

As outlined in the results section, both the accuracy and the F1 score achieved by the model are strong. The raw accuracy of the model alone is incredibly promising and the high balance achieved between precision and recall simply adds to this achievement. With those results it is clear that the model can deliver upon the desired business value, that being the ability to predict flooding events in Bangladesh. As anticipated, training the model also helped in identifying key weather features that are strongly correlated with flooding occurring such as rainfall, the hours of sunshine, and the level of cloud coverage. Identifying these factors may be of great use in future work in this field across the globe, not just for predicting floods in Bangladesh. By identifying these key factors, weather stations at any location in the world may now gather similar data points and features from past flooding and non-flooding events in order train a similar model to predict flooding events for their locality. By doing so, they may be more suitably equipped to preempt and react to predicted potential flooding events.

With that said and despite the achievements of the model, as we have detailed previously, the topic of this prediction model is incredibly high risk and thus it may not be advisable to deploy this model for true use as it has not yet achieved an acceptable accuracy and thus may require further training and testing.

References

[1] N. Zhou, J. Hou, H. Chen, G. Chen, and B. Liu, “A rapid forecast method for the process of flash flood based on hydrodynamic model and KNN algorithm.” Research Square Platform LLC, Oct. 07, 2022. doi: 10.21203/rs.3.rs-2118609/v1.

[2] N. Gauhar, S. Das, and K. S. Moury, “Prediction of Flood in Bangladesh using k-Nearest Neighbors Algorithm,” 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE, Jan. 05, 2021. doi: 10.1109/icrest51555.2021.9331199.

[3] C. Damle and A. Yalcin, “Flood prediction using Time Series Data Mining,” Journal of Hydrology, vol. 333, no. 2–4. Elsevier BV, pp. 305–316, Feb. 2007. doi: 10.1016/j.jhydrol.2006.09.001.

[4] M. M. A. Syeed, M. Farzana, I. Namir, I. Ishrar, M. H. Nushra, and T. Rahman, “Flood Prediction Using Machine Learning Models,” 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). IEEE, Jun. 09, 2022. doi: 10.1109/hora55278.2022.9800023.

[5] Z. K. Lawal, H. Yassin, and R. Y. Zakari, “Flood Prediction Using Machine Learning Models: A Case Study of Kebbi State Nigeria,” 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). IEEE, Dec. 08, 2021. doi: 10.1109/csde53843.2021.9718497.

[6] Bangladesh sizzles in highest temperature recorded in 52 ..., https://www.dhakatribune.com/bangladesh/bangladesh-environment/345286/highest-temperature-recorded-in-country-in-52.

[7] W. K. Ho, “The looming threat of sea level rise in Bangladesh,” Earth.Org, https://earth.org/sea-level-rise-in-bangladesh.

[8] [1] N. Sharif and S. K. Dey, Average relative humidity of the cities per week in ..., https://www.researchgate.net/figure/Average-relative-humidity-of-the-cities-per-week-in-Bangladesh_fig4_348309553.

[9] MOAS, “Monsoon season in Bangladesh” MOAS, https://www.moas.eu/monsoon-season-in-bangladesh/.

[10] S. Purushotham and B. K. Tripathy, “Evaluation of classifier models using stratified tenfold cross validation techniques,” Communications in Computer and Information Science, pp. 680–690, 2012. doi:10.1007/978-3-642-29216-3_74.