# **Part II**

## Introduction:

My project is meant to help with weather predicting as it uses the data that is collected from the elements such as a the min and max temperatures, and the humidity levels recorded in a day to predict the possibility of getting rain fall in the next day, this can be used in weather forecasting and it can be helpful for people to prepare for their day accordingly (like if a building company wanted to lay concrete they can use it to see if there is rain or not to be able to lay it). The idea of the project is not that new but this project uses more that wind movement that is typically used to predict rain and weather so it going to be more accurate.

You need to explain the problem you are trying to solve, and what your project is going to do for solving this problem (i.e., impact on users and organizations). Also, describe the novelty/contribution of your project.

## Deployment

### *Materials:*

* What is the Source/description of the dataset?

First I tried to use my own data but I soon knew that it was too small to be used so I used Kaggle as a source for my data.

* How was the dataset collected?

There are a number of ready to use data set but it was hard to find one that can be helpful, after I found a suitable dataset I deleted a couple of columns because they weren’t useful for rain predicting or they had 45% of the rows not filled and I also changed any cell that said NA to be empty.

* What are the dataset attributes?

They were the min and max temperature of that day and how much it rained that day in (mm), the humidity levels at different times at that day, wind speed at different times of the day, the temperatures at different times of the day, and if there is rain tomorrow.

* What is the size of the dataset?

The shape was (135643, 8)

And the size is 1085144

* In how many projects was the dataset used?

According to Kaggle it was used 1563 times.

* How was the dataset pre-processed?

Firstly I used pandas profiling report to give me an over view to the data that I have and it gave me aa couple of alerts most of them where about coloration but it was not a big problem because they were mostly under 90% of coloration, and there were a row that had a lot of zeroes.

Secondly, I checked for Num values and there were some and I chose to delete them instead of predicting their values because the data set was very big so they won’t matter a lot, then I checked for duplicated rows and there were not any.

Then I used label encoding to make everything numerical numbers because machine learning can only accept numerical inputs.

After that I checked for correlation between the columns and two columns had 98% correlation so I deleted one of them.

Then I normalized the data set using min max scaler to make all the data in the data set to be between 0 and 1.

* What is the exploratory analysis you have made to understand your data?

The pandas profiling report gave me some alerts but they were all normal but one, most of them told me that there were a high correlation between some columns but they were under 90% so I left them, the other two warnings were about having a lot of zero values in two of the columns but they were normal because the columns were about rain amount in (mm)and it can be zero if it did not rain and the other one was about the wind speed and it can be zero if there was no winds in that day.

### *Methods:*

* What models you selected for your project?

I choose the Logistic Regression model

And the Neural Networks model

* Why did you select these models?

I used logistic regression because my goal was to predict if there is rain the next day or not so it’s a 0 or 1 and this model is used for binary classification, so it was the most suitable model for the job, and I am most familiar with it.

Neural networks is a complex model inspired by the human brain and it is good at learning and representing the relation ships between the features and the targets, so it can be used for predicting according to the features.

* Draw pipeline (architecture) for each model:

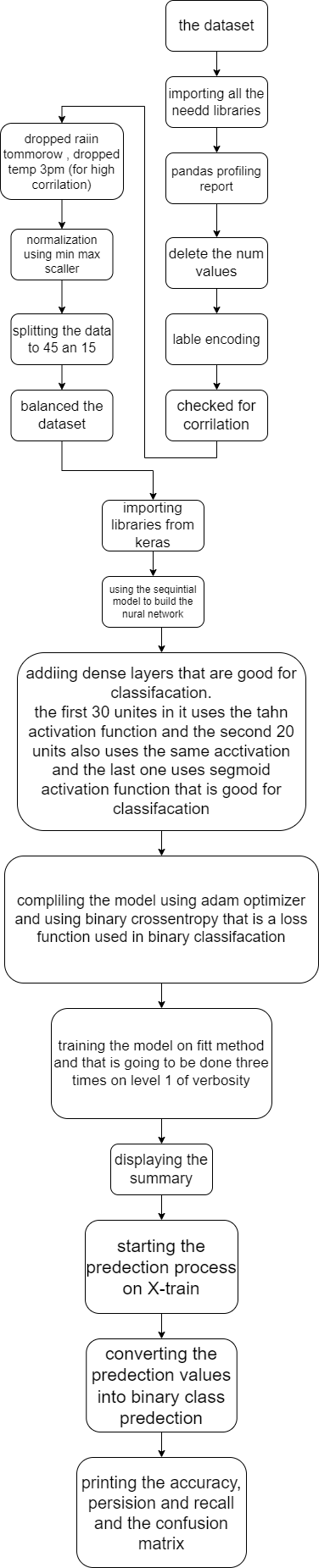
***Model#1 Pipeline chart***

You can build it with any drawing tool (e.g., drawio) and paste the generated image here.

A picture containing text, font, diagram, document

Description automatically generated

***Model#2 Pipeline chart***



* Explain the technical implementation of each model:

***Model#1 Technical implementation***

You must write the steps you have done in your Python notebook to implement your first model. (the steps were written on the code its self)

***Model#2 Technical implementation***

You must write the steps you have done in your Python notebook to implement your second model. (the steps were written on the code its self)

* Explain how your models can work together with other models/tools/approaches in the organization?

You need to describe the big picture and how your project can be integrated with the organization to improve its business and performance.

This project is used to predict the possibility of having rain the next day or not

And it uses a lot of variables, and it takes all those variables to get more accurate prediction over the traditional and old ways they are using, and it can be combined with several prediction models for other environmental behaviour to make the weather prediction automated and more accurate, and they have to be connected to automated weather recording devices to get their data from.

* What measures of performance you used to evaluate your model and what is the rationale for using such metrics?

Discuss performance metrics used to evaluate your models. And why did you use these metrics?

Because the goal for the models were classification, I used classification metrics which are combustion matrix, accuracy, precision, recall and f1score.

-Confusion matrix: it provides a visual, tabular representation of the model performance, it shows the numbers of true positive, true negative, false positive and false negative.

-Accuracy: it gives an overall correctness to the model predictions, it is the ratio between the positive prediction out of all the predictions.

-Precision: it gives the ratio between correctly predicted positive predictions out of all the positives it predicted.(is the model good at avoiding FP)so it is not useful alone.

-Recall: it gives the ratio of correctly predicted positive predictions out of the total number of actually positive.(is the model good at finding all positive instances) so it is not useful alone.

Precision and recall opposite to each other; if one goes up the other goes down

-F1 Score: it uses both the Precision and recall and it balances the trade off between them.

* Evaluate how based on the performance measures you were able to enhance the model?

Describe the techniques/approaches you followed to improve the performance of your models.

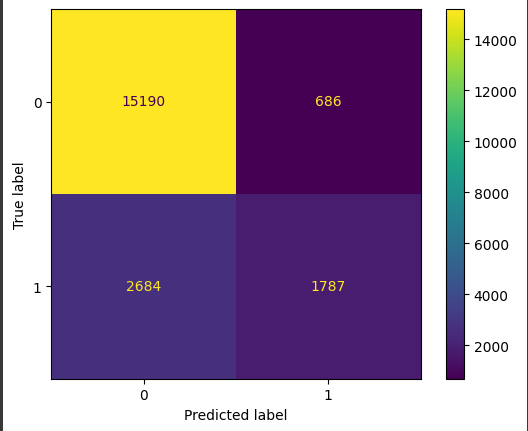
The performance measures that I concentrated on are the confusion matrix and the accuracy and the F1 score, and so to improve those I first pre processed the data in a good way to make it easier for the model to interact with, and I also made feature selection. When using logistic regression I used the normal logistic regression model then I the underfitting logistic regression and the over fitting logistic regression and that is spouse to give me more performance it the data was imbalance but because I did deal with it before it did not matter.

Also when using ANN and to improve the performance of the model by changing the number of units in every dense layer and it improved then I tried different activation functions but they did not add any thing so I returned them.

### *Results and Discussion:*

* Analysis of the results.

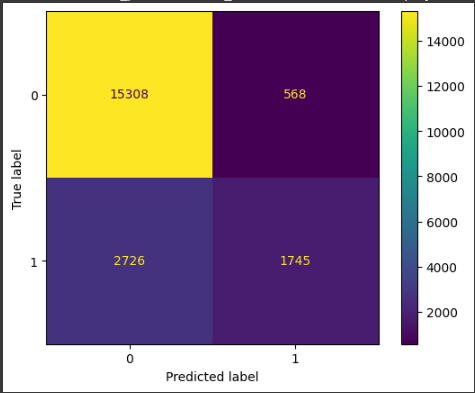
Write a detailed analysis of your results.

Using logistic regression the model gave me good performance 

The accuracy of the normal logistic regression was 83.44% and the precision was 72.2% and the recall was 39% now recall and precision made the f1 score for the (0 & 1 ) classes, 90% for the Zero and 51% for the (1)

The macro average was 71% and the weighted average was 82%

Using the ANN model



The accuracy of the ANN model was 83.39% and the precision was 67.5% and the recall was 47% now recall and precision made the f1 score for the (0 & 1) classes, 90% for the Zero and 55% for the (1)

The macro average was 73% and the weighted average was 82%

Although the model's overall accuracy is respectable, there is still space for improvement, particularly with attention to recall for the positive class.

I think one of the reasons that ANN was able to give more performance is because the data set is very big and that enable it to give more performance but if the data were smaller the logistic regression would have given better performance because it is designed for classification.

* Possibility of implementing this project in Jordan.

Based on the results of your model, are there any benefits of implementing your project in Jordan?

The possibility is very high because most of the weather recording stations in Jordan are already automated to record the weather at different times of the day and to send those data to the labs that has expert that is in charge of predicting, but if they use this model along with other models for other weather elements, the prediction becomes automated and more accurate and much faster.

* What further enhancements can be done in the future?

Discuss any limitations and future improvements of your project.

The limitation is that it is one model to predict only the rain and there must be more models to predict all the weather elements.

The biggest enhancement it to give the model more features to add into the prediction so it relies on more various types of data to get better understanding of the data thus becoming more accurate, also the data that is going to be collected is going to be huge and it needs big storage devices and very capable computers to analyse those data.

* What is your role in building/improving this project?

Describe what you plan to do to improve your project and mention specific future directions.

May be to make the a CNN model that analyses satellite images to predict storms and other natural phenomenon’s and when are they going to hit a cerine place.

## References

You need to write the references here using Harvard style of referencing.