جامعة الحسين التقنية
Al Hussein Technical University

Artificial Intelligence

Name	Hamza Muhsen
Student Number	20110044
Course Tutor	Dr.Rami AlOuran

Jordanian Strategic Executive Plan For Artificial Intelligence

In this project, as an AI engineer, my job is to benefit from my experiences to help and improve my country, so I selected one card from 61 cards of the Jordanian strategic executive plan for artificial intelligence. The card I selected to implement it is related to the agriculture sector, and I chose it because of the importance of agriculture to the country. The card is shown below:

بطاقة مشروع

الهدف الخامس : تطبيق أدوات الذكاء الاصطناعي لرفع كفاءة القطاع العام والقطاعات ذات الاولوية

الرقم	القطاع	اسم المشروع	المدة الزمنية المقدرة
43	الزراعة	إنشاء نظام تنبه مبكر للمزارعين	2024-2023

وصف المشروع:

إنشاء نظام تنبه مبكر يحذر المزارعين من اخطار قد تصيب محاصيلهم مثل الصقيع والافات.

أهداف المشروع:

الحفاظ على الإنتاج الوطني وتنمية المحاصيل الزراعية

الجهات المسؤولة عن تنفيذ المشروع:

وزارة الزراعة/ وزارة الاقتصاد الرقمي/ الارصاد الجوية /طقس العرب

المخرجات والنتائج ومؤشرات الاداء الرئيسية:

نسبة الاستفادة من التنبؤات

Related work and User Target in this project

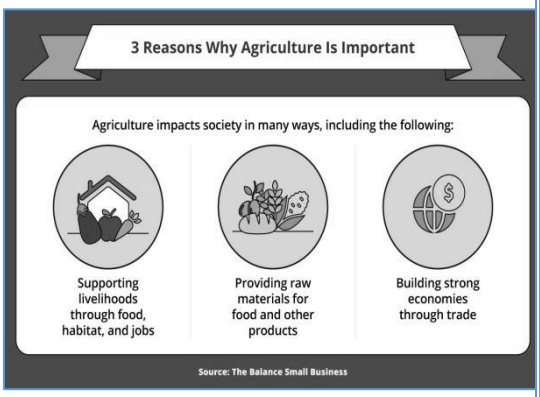
I'm not the only one who made an AI system for farmers and weather; there are many papers and authors who made it before me. An example of one of these papers is "Machine Learning Applied to Weather Forecasting" made by Mark Holmstrom, Dylan Liu, and Christopher Vo; they made an AI system to predict the weather, but with other data than in this project.**(Machine learning applied to weather forecasting - Stanford University)[2]**

The Ministry of Agriculture in the countries, the companies specialized in agriculture and farmers are the base and the targets of this project. I hope that the system gives a good effect and get more agricultural crops.

Weather prediction

For a long time now, many farmers in Jordan have faced risk factors that negatively affected and damaged their agriculture crops, and one of these factors is the weather. Cold weather conditions such as frost, wind, and heavy rains affect the crops and destroy them, causing losses to farmers in Jordan. **(Why is agriculture important? benefits and its role 2022)[1]**

This problem does not only affect farmers but also the economy of the country, which is Jordan in my current project. To elaborate, if there are insufficient agricultural crops, it will have an impact on citizens' mental health as well as livelihoods and the Jordanian economy, so we must pay attention to this issue and find a solution. As shown in Figure 1, it shows the importance of agriculture.**(Why is agriculture important? benefits and its role 2022)[1]**



In this project, based on a dataset of weather information, I will make an AI system that predicts the weather by using machine learning models in various locations to help maintain national production and develop agriculture crops. By using this system, the farmers can see the accurate weather status in the coming days, and if the weather is rainy, the system will remind the farmers, and they will do some procedures and plans to avoid any damage that will happen to the crops, and this will help the economy and keep supporting livelihoods.

When I first looked at the dataset, it became clear to me the most important features in this dataset , such as: minimum temperature,maximum temperature,rainfall,wind gust direction ,wind gust speed, rain today and rain tomorrow.

In this project, the problem statement is "How to implement accurate and fast weather forecasts for Jordanian farmers in order to prevent crop losses or damage"

Data used in this project:

To implement the system, I must find a dataset related to weather information to build models and predict the weather, I found the dataset with help Kaggle to predict the weather and make this project, the link of the data is here:
(Chakraborty, Australia Weather Data 2022)[8]

Also I read the data card that is in Kaggle, and the publisher showed the data source of this dataset which is the Australian government website (Bureau of meteorology), here is the link of the data in their website:
(Climate Data Online)[9] ----> (1)
(Daily weather observations)[10]----->(2)

B. From the notes of the dataset on the Australian government website (Bureau of Meteorology), they just mentioned that these observations were taken from the Bureau of Meteorology's real-time system. I tried to find papers that used the dataset to see if they mentioned some information about how the Australian government got these values, but I didn't find any papers. **(Notes to accompany Daily weather observations)[3]**

C. This dataset speaks in particular about the daily weather for a group of places in the country of Australia during the past ten years. In Section 1, I mentioned some elements in the data, but in fact, the dataset contains more than what I mentioned in Section A. The dataset contains 22 elements and 99 thousand rows, which helps the system increase the accuracy of the machine learning models. (The elements I mentioned in Section A are the most important elements from my point of view). Here is all elements and the description of these elements:

Heading		Meaning	Units
Date		Day of the month	
Day		Day of the week	first two letters
Temps	**Min**	Minimum temperature in the 24 hours to 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
	Max	Maximum temperature in the 24 hours from 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
Rain		Precipitation (rainfall) in the 24 hours to 9am. Sometimes only known to the nearest whole millimetre.	millimetres
Evap		"Class A" pan evaporation in the 24 hours to 9am	millimetres
Sun		Bright sunshine in the 24 hours to midnight	hours
Max wind gust	**Dirn**	Direction of strongest gust in the 24 hours to midnight	16 compass points
	Spd	Speed of strongest wind gust in the 24 hours to midnight	kilometres per hour
	Time	Time of strongest wind gust	local time hh:mm
9 am	**Temp**	Temperature at 9 am	degrees Celsius
	RH	Relative humidity at 9 am	percent
	Cld	Fraction of sky obscured by cloud at 9 am	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 9 am	compass points
	Spd	Wind speed averaged over 10 minutes prior to 9 am	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 9 am	hectopascals
3 pm	**Temp**	Temperature at 3 pm	degrees Celsius
	RH	Relative humidity at 3 pm	percent
	Cld	Fraction of sky obscured by cloud at 3 pm	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 3 pm	compass points
	Spd	Wind speed averaged over 10 minutes prior to 3 pm	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 3 pm	hectopascals

The source of this figure :**(Notes to accompany Daily weather observations)[3]**

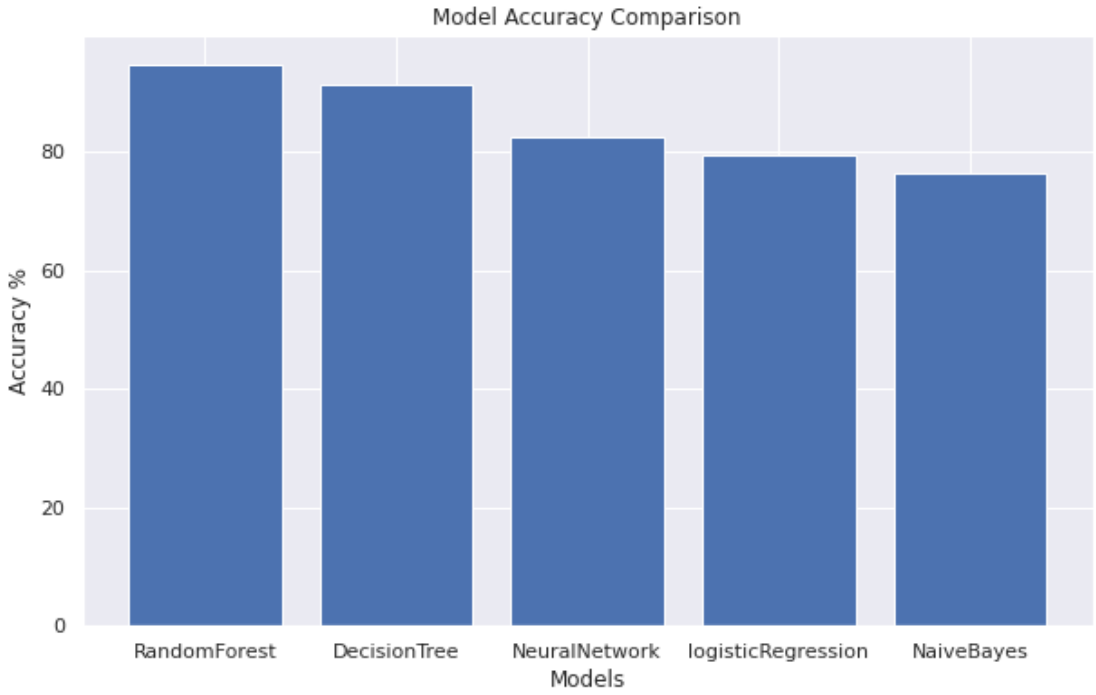
Method and models Section :

A. In the system, I implemented five models to see which model gave the best reasonable prediction and the best performance and time. The models that were implemented in the system are:

- Decision tree.
- Random forest.
- Naive bayes.
- Logistic regression.
- Neural network.

When I implemented the system, I divided the features (columns) into X and Y and decided that the Y was "Rain Tomorrow" and its values were 0 (no rain) and 1 (rain), so I used the above-mentioned classification models.

After building every model, I used the accuracy to measure the performance of each model to select the best model to implement in the system. In our case, the best performance model was the random forest, with around 95% accuracy. Here is the accuracy comparison between each model implemented in the system:



So, in the system, I will use the random forest model to make a reliable system that the farmers can depend on to save the crops. Furthermore, the random forest model is simple to understand, deal with, and interpret, and it performs admirably.

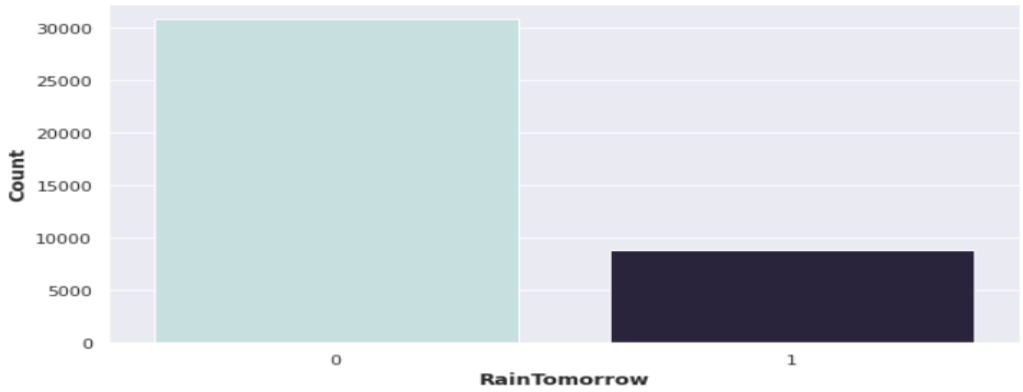
B. The AI tools and frameworks that I used in the system:

- **Scikit Learn**

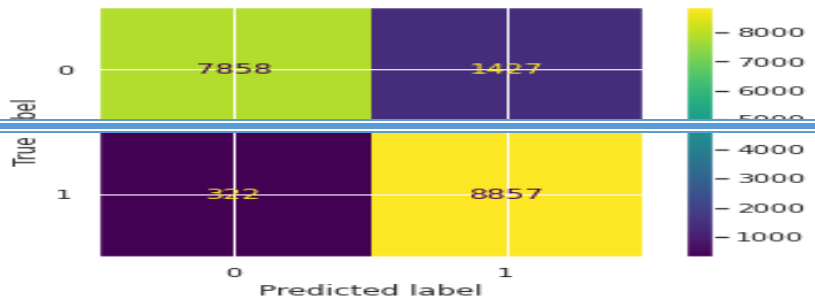
I used Scikit Learn for many things, and I considered it the secret behind the success of the system. It is a machine learning library that most of the AI engineers use to make systems, projects, or whatever. It's extremely useful because it contains numerous libraries that aid the system, such as those for scaling, splitting, and processing data.

I use Scikit-Learn in the system to import libraries such as:

I. Processing data such as I used **Label Encoder** to convert the data type of features from object to integer to prepare the data before prediction.
II. I **oversampled** the data by using the resample library because the values of "rain tomorrow" were unbalanced and I needed to make the number of (1) values equal to the number of (0) values. The figure depicts the imbalance:



III. Splitting the data into a training set and a testing set is an important step in building machine learning models. A training set is used to fit and train the model, and a testing set is used to evaluate the model.
IV. Processing data: I also used the **MinMax scaler** library to scale the data into a range that is between 0 and 1. It may help increase the performance of the system.
V. Build models such as the random forest model, decision tree, naive bayes model, logistic regression model, and others, but I used the previously mentioned models.
VI. Show the accuracy of the model implemented by using **metrics** ; it is useful, and it helped me compare the performance between the models I selected.
VII. Show the confusion matrix as shown in the figure below:

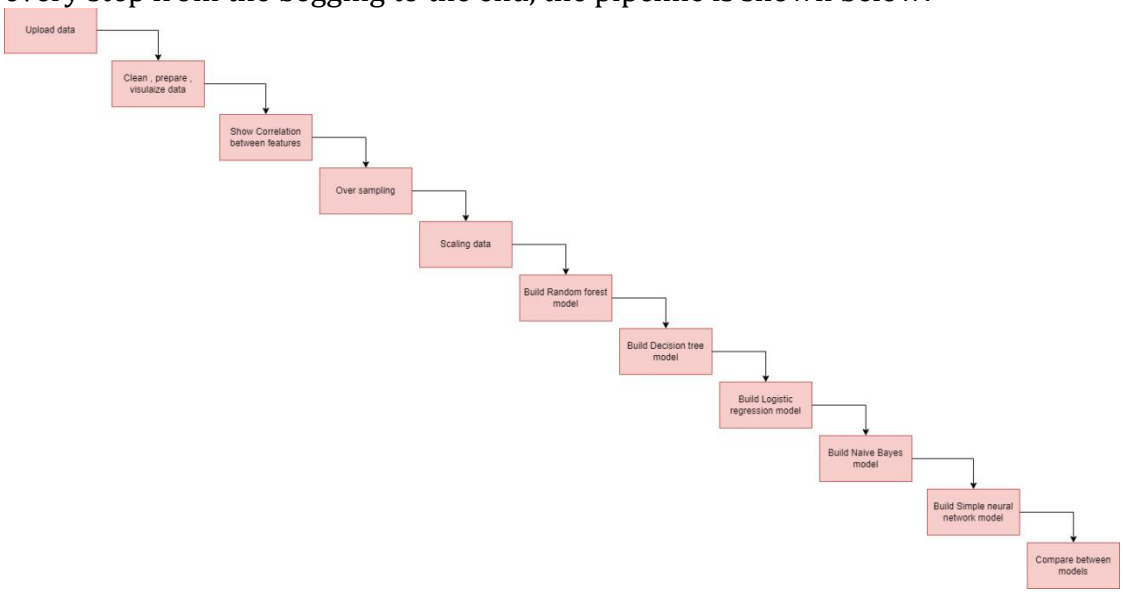


- **Keras**

I used it when I implemented a neural network; it is considered a deep learning API and was developed by Google for use in the implementation of a neural network model. I imported two things from Keras:

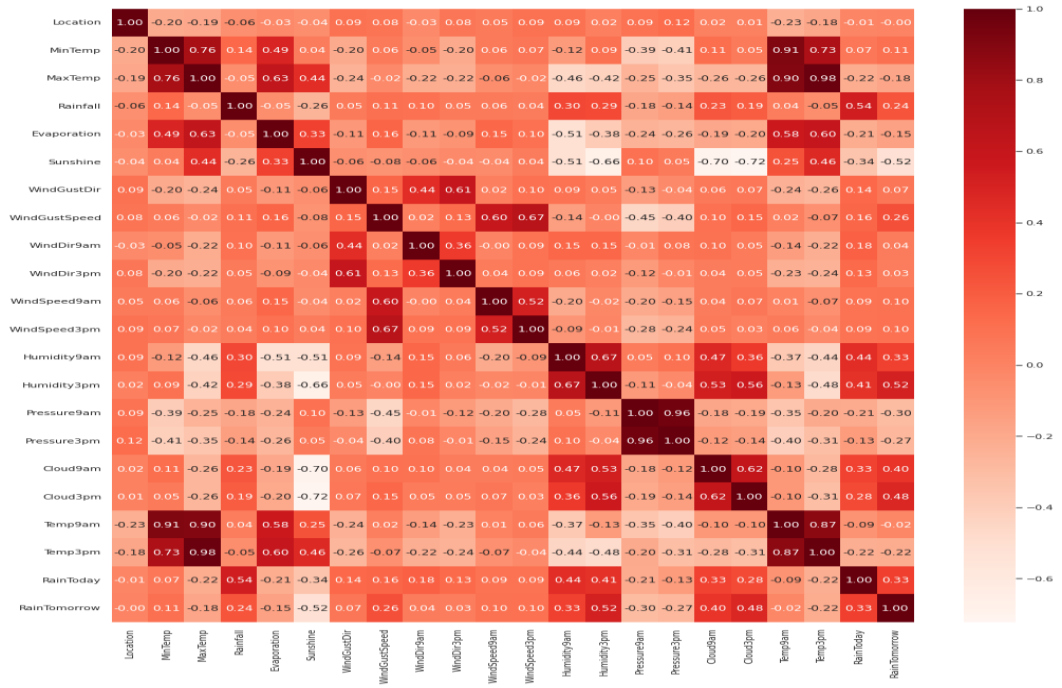
I. Sequential
II. Dense

(Simplilearn, What is Keras and why it so popular in 2021: Simplilearn 2022)[4]
And both used to layers. In this project, I used a specific pipeline to implement the system, through pipeline , it helps me to implement the system by doing every step from the begging to the end, the pipeline is shown below:



C.
When I implemented the models, I used these steps:

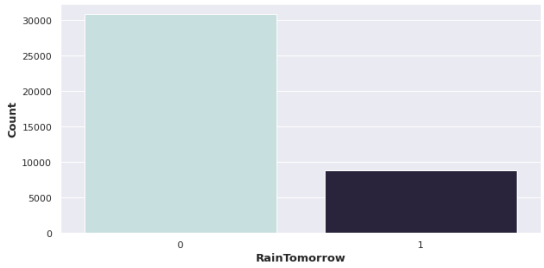
- I uploaded the dataset in **colab** (the description of the data is mentioned above). This step was also greatly aided by using Pandas library.
- Clean, and visualize the data before choosing the models.
- Before I predicted the values, I made a feature selection because there were many high correlations between features, such as "Temp9am" with "Temp3pm," "Mintemp", and "Maxtemp," so I removed Temp9am and Temp3pm because Min temperature and Max temperature have higher feature importance than Temp9am and Temp3pm. The figure below describes the correlation between features:

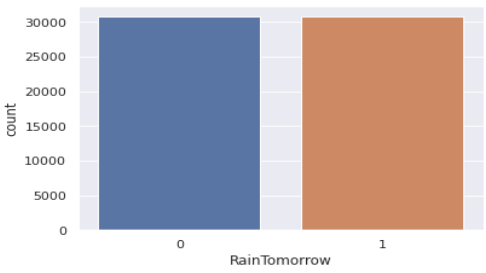


- X and Y were defined , all the columns except the "rain tomorrow" was X and Y is "rain tomorrow" only.
- splitting the data into training and testing sets (I put 80% on the training set and 20% on the testing set).
- Build a random forest model with Scikit-Learn, and the n_estimators value was 200.
- Predict the model and see if the predicted values are logical.
- Evaluate the model and see the performance of the model (random forest).
- I did these steps for all the suggested models in this project.

After I built the system, I think that I used all the necessary steps to make the system, and I used all the necessary libraries to make and improve the system. Here are some examples of how I increased the performance of the system:

- Oversampling data, I used it because my data was imbalanced as shown in the figure below, one figure shows the imbalance data and the another after using oversampling:

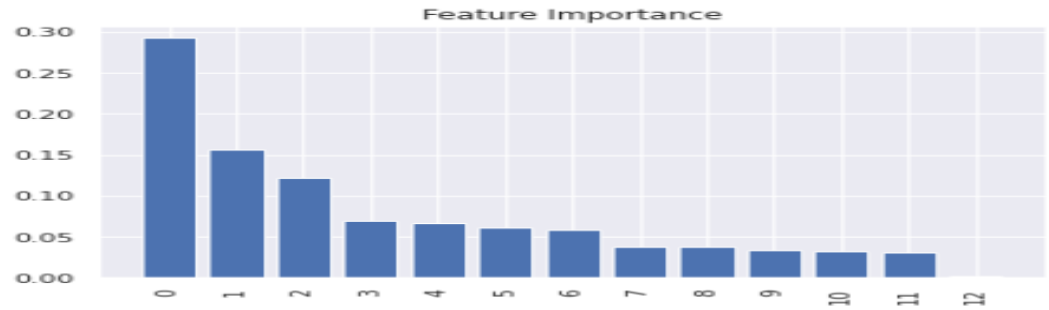




- Remove null values. In the dataset, there were over 5000 null values, so I removed them by using the PANDS library.
- Remove highly correlated features because they nearly have the same values, the correlation between features is shown in the figure above.
- Increase the training set data because big data means increasing the efficiency of the model and making it capable of solving complex patterns.

- I knew what the most important features that affect the model were by using future importance, for example, in the system, I used future importance for the decision tree (first figure) and the random forest (second figure):

```
 1) Sunshine                       0.292505
 2) Humidity3pm                    0.156547
 3) Pressure3pm                    0.121660
 4) WindGustSpeed                  0.068752
 5) MinTemp                        0.066630
 6) MaxTemp                        0.061204
 7) Evaporation                    0.058841
 8) WindDir9am                     0.037918
 9) WindDir3pm                     0.037765
10) Rainfall                       0.034149
11) WindGustDir                    0.031496
12) Cloud3pm                       0.030650
13) RainToday                      0.001883
```



Random Forest:


```
 1) Sunshine                       0.174956
 2) Humidity3pm                    0.171547
 3) Pressure3pm                    0.105745
 4) Cloud3pm                       0.082236
 5) WindGustSpeed                  0.075179
 6) MinTemp                        0.067781
 7) MaxTemp                        0.066874
 8) Evaporation                    0.057215
 9) Rainfall                       0.053875
10) WindDir9am                     0.041066
11) WindDir3pm                     0.040697
12) WindGustDir                    0.039480
13) RainToday                      0.023348
```

I tried many steps and libraries to increase the performance of the model, but I always got an accuracy of 94 or lower. After all these steps, I think that I can't do anything to increase the performance of the models, but still, the models need to increase their performance to make the system more reliable when the farmers or any other stakeholder use it.

D. As I mentioned above, I used classification models in the system, and I used accuracy to measure the performance of the models. Accuracy is used in classification problems to evaluate the models. Accuracy is calculated by using this equation:

$$\text{Accuracy} = \frac{\text{Number of correct prediction}}{\text{Total number of prediction}}$$

Also, we can calculate the accuracy in binary classification by using classification evaluation metrics. For example, I will calculate the accuracy in the random forest model in the system by using the following classification evaluation metrics:

$$\text{Accuracy} = \frac{5947+5730}{12309} = 94.86\%$$

Which is the same as in the colab.**(Classification: Accuracy | machine learning | google developers)[5]**

From the research I did, I found that accuracy is recommended when the data is balanced and that it has weaknesses when the data is imbalanced, and in the system in this project, I did oversampling to make the data balanced because the data was imbalanced before oversampling. **(Bressler, How to check the accuracy of your machine learning model 2022)[6]**

Evaluate:

I built the random forest without any hyper parameters and calculated the performance using accuracy; the accuracy was 93.7%. After that, I changed the n_estimators from 100, which is the default value, to 200 and calculated the accuracy again; the accuracy increased to 94.86%.

Also, not only the n-estimators were the factor that enhanced the model; I also did feature selection by removing the high feature correlation and low feature importance, which also helped enhance the model, and I modified the size of the training and testing sets to define the best size, which I defined using accuracy, and that also increased the performance of the model.

So I used accuracy to compare the changes I made and then selected the best change that improved both performance and accuracy. So by using accuracy I enhanced the system.

Result :

A. After implementing all of the models and determining that the random forest is the most accurate model, I found it logical because, prior to implementing the models, when I first saw the dataset, I decided to implement the random forest because it fit the dataset.

Also, I think random forest got better accuracy than decision tree model because it creates numerous decision trees and combines their predictions to increase resilience and decrease overfitting, Random Forest often produces forecasts that are more accurate than those of a single decision tree.

Not only that, but when I implemented the models and determined their future importance, I looked at the values for all of them and saw that the random forest values are more logical than decision tree values, and the difference between the values in random forest is not as great as the difference between decision tree and other values.

In my opinion, we can implement the system in Jordan and develop it because, by getting strong, relevant, and up-to-date data from the Jordanian Ministry of Agriculture, we can implement this system in Jordan and develop it after many years. Also, in terms of cost, I believe that it will not be prohibitively expensive to implement in Jordan, but in return, as discussed in the introduction about the issues and problems that farmers face, the system will improve agricultural efficiency and increase revenue for Jordanian farmers and the economy.

Maybe the only problem we will face is how to convince the farmers and other stakeholders about the AI system and convince them to depend on it. For this problem, I will recommend making videos that discuss the agriculture problems, the solution, which is the system, how it is used, and the benefits of using the system. And we can make a deal with the Jordanian ministry of agriculture to convince the farmers to use the system.

Finally, every system needs support and maintenance, so before implementing the system, we must make a maintenance schedule to prevent any technical problems.

B. I think by using more parameters in random forest model or the others models, it can enhance the model, such as : in the system , if I used max_depth in random forest , it can enhance the model and increase the accuracy. Also, I think that feature engineering can enhance the model.

About tools and frameworks, i think we can use **LightGBM** framework, which is a gradient boost that can be used to increase the efficiency of models and decrease the memory usage. It can be used in our system to increase the efficiency of the random forest model. **(Lightgbm (Light Gradient Boosting Machine) 2021)[7]**

C. In this project, as an AI engineer, I learned new concepts and methods to build and improve an AI system. Before I did this system, I didn't know what outliners and oversampling meant, but when I worked on this system, I understood what these concepts meant. I also knew how to improve the system , what the performance measures were , and how to deal with data.

But on the other hand, I faced some issues, such as the fact that the dataset used was large and contained many null values. So firstly, I didn't know if I should remove these values or replace these values with 0, but I decided to remove the null values. Also, in the neural network model, it took more than 15 minutes to run the model one time because of the large dataset used in the system, so I didn't run the model more than five times because it took a long time.

References

Why is agriculture important? benefits and its role (2022) Maryville Online. Available at: https://online.maryville.edu/blog/why-is-agriculture-important/ (Accessed: January 23, 2023). [1]

Machine learning applied to weather forecasting - Stanford University (no date). Available at: http://cs229.stanford.edu/proj2016/report/HolmstromLiuVo-MachineLearningAppliedToWeatherForecasting-report.pdf (Accessed: January 29, 2023). [2]

(no date) Notes to accompany Daily weather observations. Available at: http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml (Accessed: February 1, 2023). [3]

Simplilearn (2022) What is Keras and why it so popular in 2021: Simplilearn, Simplilearn.com. Simplilearn. Available at: https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-keras (Accessed: February 1, 2023). [4]

Classification: Accuracy | machine learning | google developers (no date) Google. Google. Available at: https://developers.google.com/machine-learning/crash-course/classification/accuracy (Accessed: February 1, 2023). [5]

Bressler, N. (2022) How to check the accuracy of your machine learning model, Deepchecks. Available at: https://deepchecks.com/how-to-check-the-accuracy-of-your-machine-learning-model/ (Accessed: February 1, 2023). [6]

Lightgbm (Light Gradient Boosting Machine) (2021) GeeksforGeeks. GeeksforGeeks. Available at: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/ (Accessed: February 1, 2023). [7]

Chakraborty, A.K. (2022) Australia Weather Data, Kaggle. Available at: https://www.kaggle.com/datasets/arunavakrchakraborty/australia-weather-data (Accessed: February 1, 2023). [8]

Climate Data Online (no date) Climate Data Online - Map search. Available at: http://www.bom.gov.au/climate/data/ (Accessed: February 1, 2023). [9]

(no date) Daily weather observations. Available at: http://www.bom.gov.au/climate/dwo/ (Accessed: February 1, 2023). [10]