Stephanie Benavidez

DSC680-T301 (2241-1)

Applied Data Science

Week 8: Milestone 3 – Project 2: Final White Paper

Introduction

For this second project of this semester, I thought it would be neat to do a decision tree classifier model for Texas crops, primarily Texas pecans. According to Texas Agriculture, “Texas Pecans hold a special place in the hearts and kitchens across the state. Pecans are the only major tree nut truly native to the United States. Archaeologists have found evidence of pecan seed and leaf fossils formed along the Rio Grande dating as far back as 6100 B.C. Today, Texas is a leading pecan grower and the top producer among states where the tree is native. It is no wonder that the pecan is recognized as the State Tree (pecan tree), the State Health nut (pecan nut), and the Texas State Pie (pecan pie) designations” (Miller, n.d.). Growing up in the big state of Texas, pecan trees are quite common to have in your backyard. As a child, they were one of the best trees to climb, especially since the best pecans were always at the top. So, you have to be dangerous and get a treat at the same time.

According to recent studies and research, pecans can help prevent obesity, diabetes, and inflammation. “Pecans, along with other nuts of similar composition, like walnuts and hazelnuts, are rich in polyunsaturated fats, dietary fiber, and polyphenols, substances known for their antioxidant and anti-inflammatory properties” (Curley, 2023).

Pecans also have nutrients like “monounsaturated fats, vitamin E, and fiber” (Curley, 2023).

Data Explanation

With this second project, I will be focusing on predicting pecan crop yield and potential profits. There are multiple questions I would like to try to answer with more coming through the progress of the project. Some of the research questions that I would like to try to explore are as follows:

- Are pecan crop yields increasing or decreasing over time?
- If costs have increased and cut into profit margins over time, are there examples of what caused these outliers?
- Do the temperatures show any effect on the trend lines within the project data timeline?
- What year had the highest crop yield for Pecans?
- What models will be used for this project?
- Will the model be able to predict over 75%? If so, what models predicted over 75%?

Data Source

For the datasets, the data will come from the Food and Agriculture Organization (FAO) of the United Nations (Food and Agriculture Organization, 2023). This organization’s mission is to tackle world hunger by achieving food security and access to high-quality food for everyone. The data the organization collects measures

important markers of food production, needs, and costs as well as the population characteristics of different countries.

The data of focus for this project is contained within several different datasets offered on the FAO website. I will be pulling variable data for the United States other nuts, in shell, from the following datasets:

- Production – Crops and livestock products (Crops and Livestock Products Data, n.d.)
    - *Crop Yield* – the amount produced per hectare.
    - *Area Harvested* – number of hectares harvested.
- Climate Change Indicators – Temperature change on land (Temperature Change on Land Data, n.d.)
    - *Temperature Change* – degrees Celsius of difference from the previous meteorological year.
- Prices – Producer Prices (Producer Prices Data, n.d.)
    - *Producer Price* – the amount paid to growers when selling crops.
- Producer Price Trade – Crops and Livestock Products (Producer Price Trade Data - Crops and livestock products, n.d.)
    - *Export Value* – total value of exported crop that year.
- SDG Indicators - SDG Indicators (SDG Indicators, n.d.)
    - *Level of Water Stress* – the proportion of water withdrawn from available sources.

The format that the FAO puts the data in is more descriptive than immediately useful for analysis. Many columns contain repetitive filter data that is not relevant to the analysis and different variables often end up as rows within the same generalized table. The easiest way to download the data is to put one variable at a time into a CSV via the FAO website. Unfortunately, the FAO website did not have pecans by themselves, so I used the data from other nuts in a shell since that is the closest data set, I could find. Kaggle and other datasets did not have anything pecan related.

Data Cleanup

There was a total of six different datasets that will need to be processed before they are ready to be used. Not only will I need to remove unnecessary columns, but will also need to remove all duplicate values, rename value columns, and replace missing values with a zero or to be blank.

Next, I will merge the different target variable columns into a single master data frame based on the year value. This will enable me to look for trends over time. After looking over the final merged dataframe, I have decided to focus on the years 2000-2017, since these years had data across all data variables being used in this project. Therefore, creating a final dataset focused on this time range.

```
## Extract rows from mergeddata to show only years 2000-2017
newdf = mergeddata[mergeddata['Year']<=2017]
newdf = newdf[mergeddata['Year']>=2000]
## make sure variables are numeric
newdf['Yield100GperHa'] = pd.to_numeric(newdf['Yield100GperHa'])
newdf['HectaresHarvested'] = pd.to_numeric(newdf['HectaresHarvested'])
newdf['ExportValue1000USD'] = pd.to_numeric(newdf['ExportValue1000USD'])
newdf['ImportTons'] = pd.to_numeric(newdf['ImportTons'])
newdf['ProdUSDperTonne'] = pd.to_numeric(newdf['ProdUSDperTonne'])
newdf['WaterStressPerct'] = pd.to_numeric(newdf['WaterStressPerct'])
newdf.head()
```

	Year	TempChangeC	Yield100GperHa	HectaresHarvested	ExportValue1000USD	ImportTons	ProdUSDperTonne	WaterStressPerct
39	2000	1.543	26189.0	45000.0	107315.0	39130.61	2513.0	10.54
40	2001	2.055	25941.0	69000.0	82715.0	27963.00	1310.0	10.51
41	2002	2.783	25628.0	40000.0	86719.0	36969.00	2105.0	10.47
42	2003	2.167	26215.0	58000.0	89797.0	44374.00	2169.0	10.43
43	2004	0.116	25871.0	42500.0	124934.0	58973.00	3880.0	10.39

Graphical Analysis

Since the final dataset is focused on the year, I thought it would be good to do some column analysis to see if it progressed or declined over the years utilizing line plots, box plots, and a correlation heatmap.

From some of the graphical analysis done, which is shown in the appendix section, you will be able to see that variable names with potential outliers in box plots were: Yield Over Time, and Hectares Harvested Over Time. For the line plot analysis, Temperature Change Over Time had a fluctuation throughout plot; Yield Over Time had a decline after 2013; Hectares Harvested Over Time had an increase after 2013; Export Value gradually increased; Imports Over time and Producer Price Over Time had a fluctuation throughout plot; and Agricultural Water Stress Over Time had a decline in 2010 and increased to flatten out after 2015.

On the correlation heatmap on the final dataset, you can see that ‘Year’ on the x-axis and ‘ExportValue1000USD’ on the y-axis had the highest correlation with each

other with a value of 0.96. Some of the other higher correlations with each other starting on the x-axis, and then the y-axis are as follows:

- ‘Yield100GperHa’ and ‘HectaresHarvested’ with a value of 0.92.
- ‘Yield100GperHa’ and ‘ExportValue1000USD’ with a value of 0.87
- ‘Year’ and ‘ImportTons’ with a value of 0.86.
- ‘ImportTons’ and ‘ProdUSDperTonne’ with a value of 0.85

Method/Analysis

The analysis methods used in this project are a Neural Network, a Random Forest, and a Decision Tree Analysis. To train and evaluate the dataset, I needed to remove the “Year” column. Thresholds also needed to be set up to categorize them in the training and testing data. After training all three models, both the Random Forest, and Decision Tree had a training accuracy of 1.0, and the Neural Network had a training accuracy of 0.86. After evaluating all three models, the Neural Network had a testing accuracy of 1.0, and both the Random Forest and Decision Tree had a testing accuracy of 0.75.

From the model’s accuracy scores, I created a plot to demonstrate both the training and test accuracies. For each model, the classification report was printed out to show the differences between each model. For the Random Forest model, I wanted to plot a bar graph of all the importance within the model, in which you could tell that the “TempChangeC” showed the most importance followed by the “Yield100GperHa.”

Lastly, for the Decision Tree Model, I wanted to see the visualization in a dot format using graphviz. In the visual graph, you can see the high importance of

“TempChangeC” with a greater than 0.83, and the medium importance of “Yield100GperHa” with a greater than 0.68. From this information, I did want to visualize the dot format in the visualization from dtreeviz, unfortunately, I was not able to replicate the code to work and kept giving errors.

Overall, the results for all of the models seemed to be higher than I originally expected them to be and am pleased with the outcomes of these three models.

Ethical Assessment

The ethical considerations to consider are the data sets not having all the variables in the data, therefore, not allowing the project to reach the potential that is needed. There may also not be enough resources available to have the proper data sets to have a successful project. Any missing variables can make an impact on the final results are predicted. Hopefully, I will be able to find all of these variables on the website, and if not search for another dataset that has the missing variables.

Lastly, some of the ethical questions asked during this project were:

- Is there anything wrong with the data in the final CSV file?
- Did it contain all the variables needed to reach the desires of the model’s analysis?
- Any issues finding a dataset suitable for this project?
- Is there anything that was tried, but wasn’t able to produce results?

Challenges/Limitations/Assumptions

For the challenges or issues that were faced within this project, the main concern is the data. Data always plays a huge role in the outcome of the project. From everything I have learned through this program you spend at least 80% of the time with the data to prepare it to use for the predictions and models. Secondly, make sure your coding to clean and prepare the data, as well as each model is correct. Being new to some of these models, the coding can be a little tricky, not allowing the desired results to be formulated. Lastly, a lack of resources can be difficult in any project, which also makes it difficult to get the desired results.

Recommendations

The recommendations that I have for myself are to get a better understanding of what I am trying to evaluate and train within models. Possibly try other ways to evaluate and train the dataset to see if the results change. Make any adjustments as needed throughout the project, so that all visualizations with plots, matrixes, models, classification reports, and graphs are presentable. The one visualization that I would have liked to have included was a visualization graph of the dot graph for a decision tree.

Conclusion

In conclusion, based on the models' results, both the Random Forest and Decision Tree models had better results. For both the importance and dot graph of these two models, both the "TempChangeC," and "Yield100GperHa" had the highest values.

With each project this semester, my focus was to use models or predictive analyses that are out of my comfort zone or that I have not used that much. However, in the end, the more we use these types of analysis the better we get and the more comfortable we are with them. In the end, I just want this project to provide decent results.

Lastly, for the research and ethical questions asked throughout this project, the answers are as follows:

- Are pecan crop yields increasing or decreasing over time? After 2013 the crop yield seemed to decrease.
- If costs have increased and cut into profit margins over time, are there examples of what caused these outliers? In this dataset, the Export Value, Imports, or Producer price did not show any outliers.
- Do the temperatures show any effect on the trend lines within the project data timeline? Yes, you can see the fluctuations of the change.
- What year had the highest crop yield for Pecans? From the year 2000-2008, the crop yield stayed pretty steady before decreasing around 2008.
- What models will be used for this project? A Neural Network, Random Forest, and a Decision Tree Classifier.
- Will the model be able to predict over 75%? If so, what models predicted over 75%? According to the accuracy plot, all three models had over 0.85 for the training dataset. For the test accuracy, only the Neural Network had over 75%.
- Is there anything wrong with the data in the final CSV file? No.

- Did it contain all the variables needed to reach the desires of the model's analysis? I believe so.
- Any issues finding a dataset suitable for this project? Yes, unfortunately, the dataset was based on other tree nuts since there were no datasets available focused on pecans.
- Is there anything that was tried, but wasn't able to produce results? Yes, I was trying to get a better visualization of the dot graph from the Decision Tree Classifier, however, the coding kept giving me errors.

Though I have been able to answer the questions that I have thought about through this project, there are different possibilities to try to get better results. From the expectations from each project, I have come to learn that there is always room for improvement, and the results could get better each time with more time.

References

Climate Change Indicators: Weather and Climate. (2022, August 1). Retrieved from www.epa.gov: https://www.epa.gov/climate-indicators/weather-climate

Crops and Livestock Products Data. (n.d.). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data/QCL

Curley, C. (2023, 8 8). *Pecans may have protective effects against obesity, and diabetes.* Retrieved from https://www.medicalnewstoday.com: https://www.medicalnewstoday.com/articles/eating-pecans-may-prevent-obesity-diabetes

Food and Agriculture Organization. (2023). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data

Miller, C. S. (n.d.). *Texas Pecans. Texas Heart.* Retrieved from https://www.texasagriculture.gov: https://www.texasagriculture.gov/Grants-Services/Marketing-and-International-Trade/International/Buyers-and-Consumers/Pecans

Producer Price Trade Data - Crops and livestock products. (n.d.). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data/TCL

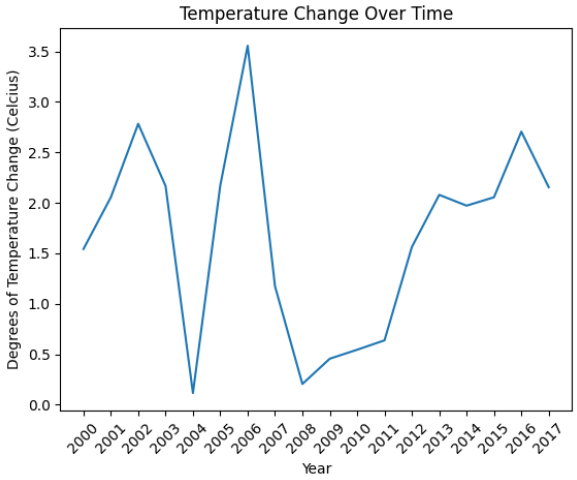
Producer Prices Data. (n.d.). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data/PP

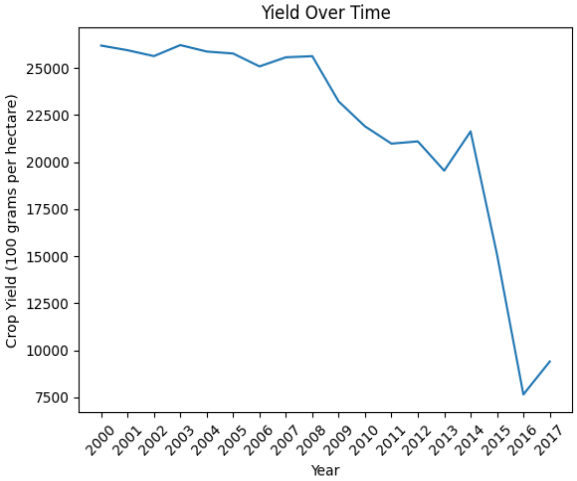
SDG Indicators. (n.d.). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data/SDGB

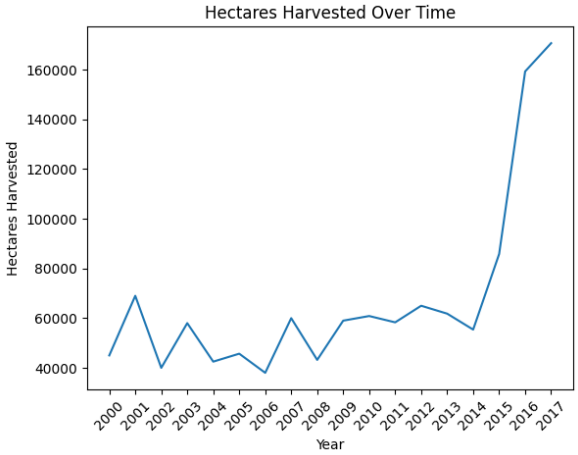
Temperature Change on Land Data. (n.d.). Retrieved from www.fao.org: https://www.fao.org/faostat/en/#data/ET

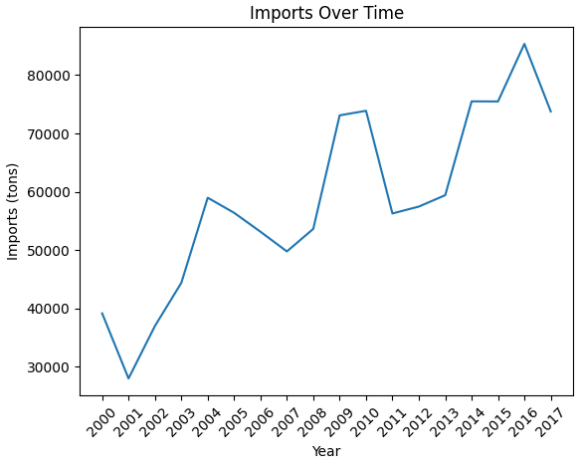
Appendix

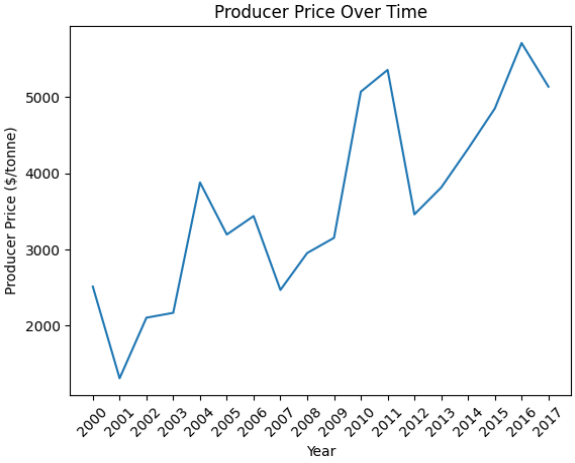
1. Variable Trends Over Time

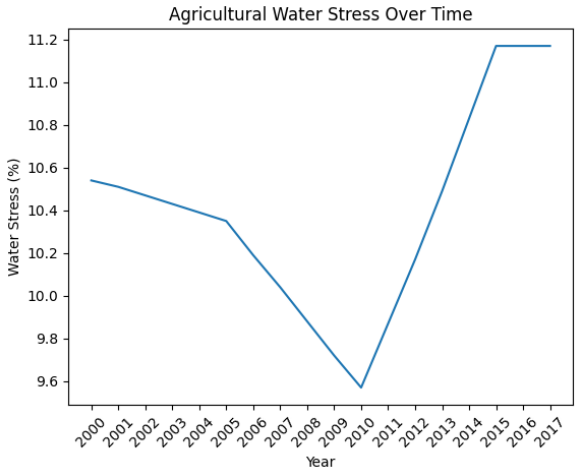




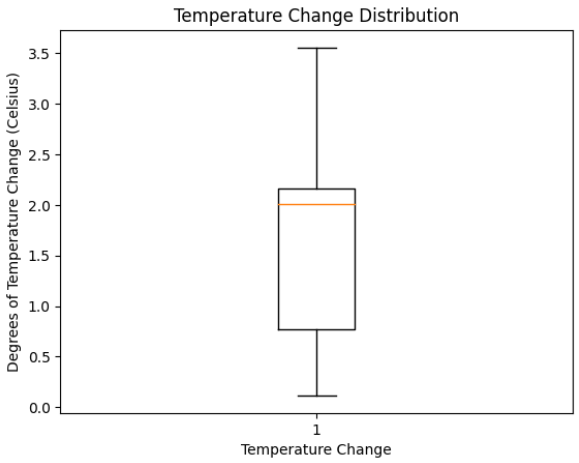


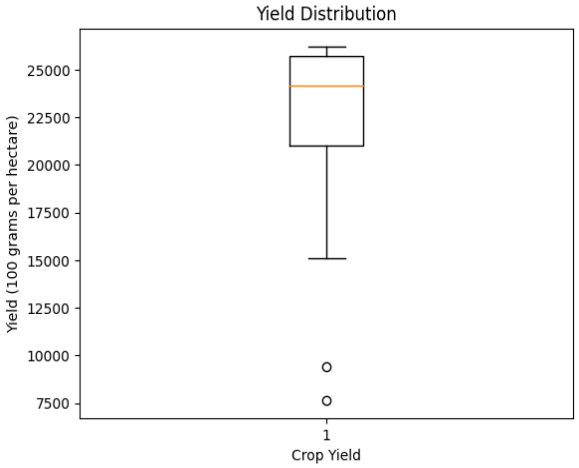


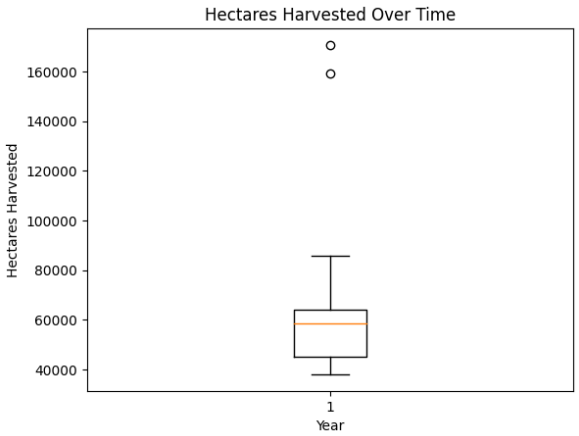


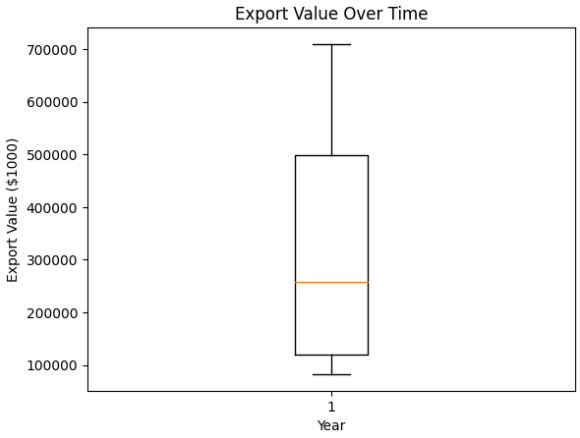


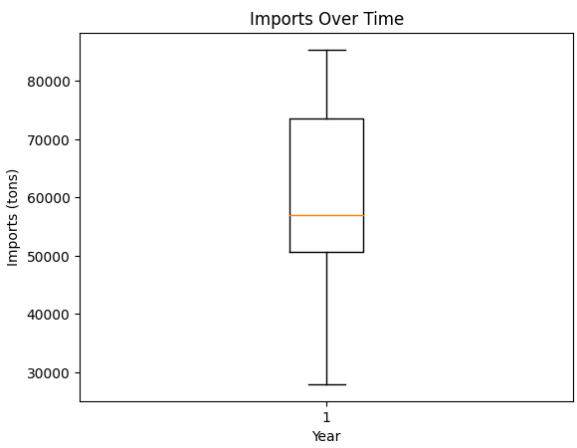
2. Examination of Outliers via Boxplots

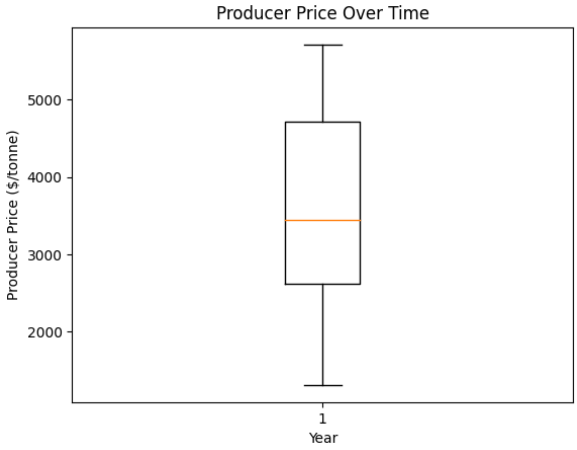


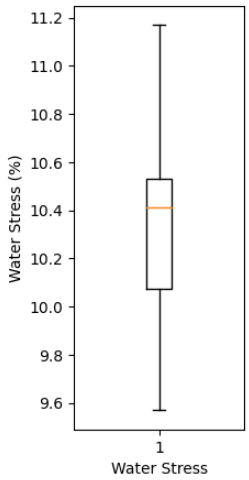




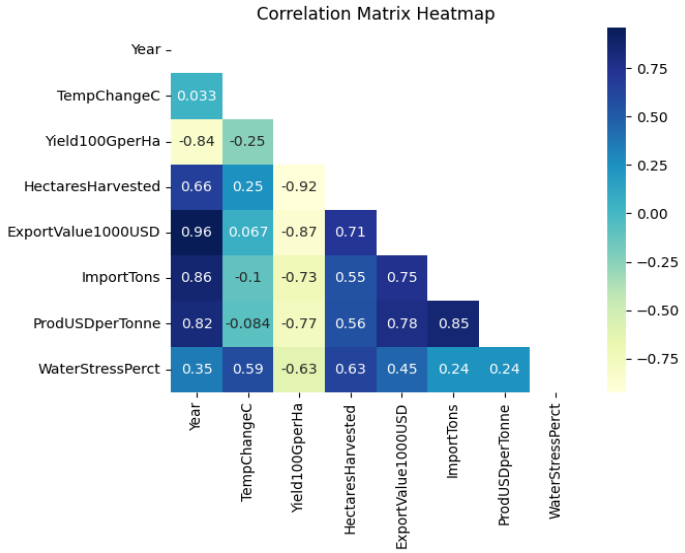








3. Correlation Matrix on data



4. Neural Network

Neural Network Training Accuracy: 0.8571428571428571

Neural Network Testing Accuracy: 1.0

Classification Report for Neural Network:

	precision	recall	f1-score	support
High	1.00	1.00	1.00	3
Medium	1.00	1.00	1.00	1
accuracy			1.00	4
macro avg	1.00	1.00	1.00	4
weighted avg	1.00	1.00	1.00	4

5. Random Forest

Random Forest Training Accuracy: 1.0
Random Forest Testing Accuracy: 0.75

Classification Report for Random Forest:

	precision	recall	f1-score	support
High	0.75	1.00	0.86	3
Medium	0.00	0.00	0.00	1
accuracy			0.75	4
macro avg	0.38	0.50	0.43	4
weighted avg	0.56	0.75	0.64	4

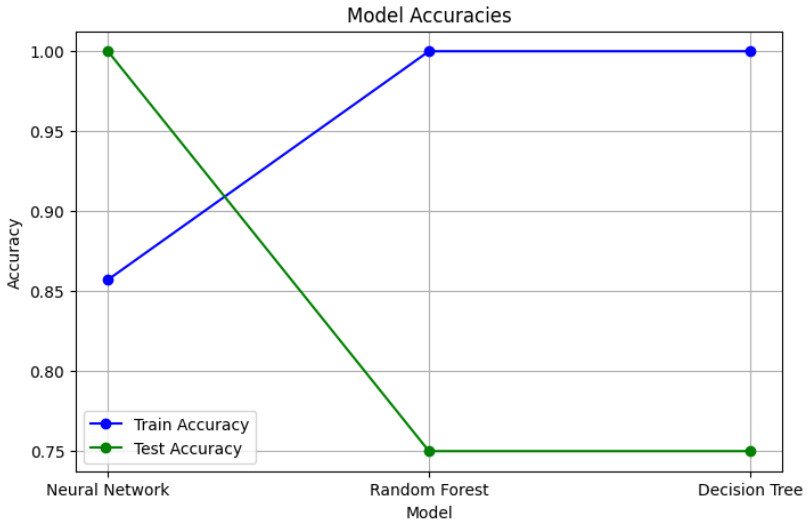
6. Decision Tree Classifier

Decision Tree Training Accuracy: 1.0
Decision Tree Testing Accuracy: 0.75

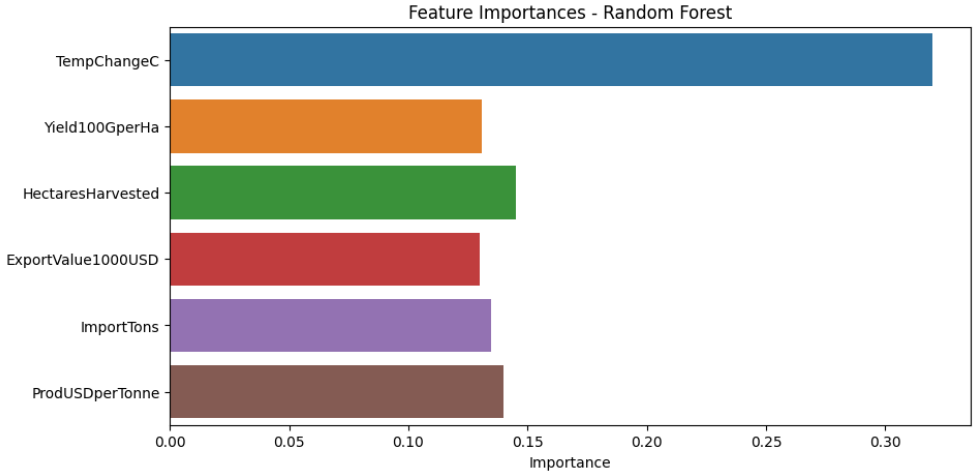
Classification Report for Decision Tree:

	precision	recall	f1-score	support
High	0.75	1.00	0.86	3
Medium	0.00	0.00	0.00	1
accuracy			0.75	4
macro avg	0.38	0.50	0.43	4
weighted avg	0.56	0.75	0.64	4

7. Model Accuracies Plot



8. Feature Importance's for Random Forest



9. Visualization of the Decision Tree Model