CIND820: Big Data Analytics Project

**Literature Review**: Predictive modeling based on weather and soil quality for crop yield in Ontario

Student name: Long, Nguyen

ID: 501214226

Supervisor: Dr. Ceni, Babaoglu

Submission date: Feb 21, 2023

Contents

[I. Abstract: 3](#_Toc131440250)

[II. Literature review: 3](#_Toc131440251)

[a) Critical attributes in predicting crop yield: 3](#_Toc131440252)

[b) Suitable algorithm for this purpose: 4](#_Toc131440253)

[c) How to improve model accuracy: 4](#_Toc131440254)

[d) Implementation of this research after finishing the model: 5](#_Toc131440255)

[III. Finding and reasoning during development: 5](#_Toc131440256)

[a) Data pre-processing 5](#_Toc131440257)

[b) Feature selection: 8](#_Toc131440258)

[c) Model training and selection: 9](#_Toc131440259)

[d) Model optimization: 9](#_Toc131440260)

[e) Process flow chart: 10](#_Toc131440261)

[IV. Conclusion 10](#_Toc131440262)

[V. Link to Github 11](#_Toc131440263)

[VI. Reference: 11](#_Toc131440264)

# Abstract:

Recently, using a predictive model to support decision-making in the production environment is not a new topic but I have not had a chance to see its implementation in the mass production environment. My target for this project is to get to know the tools for Classification and Predictive Analytics and how to use them so I can implement them later. However, due to the lack of actual data in production, I choose a common theme in which information is more accessible: agriculture in Ontario province. The topic is building a predictive model to forecast crop yield using weather in Ontario. Based on all of the findings, two models using Decision trees and Support Vector Machines will be built so that the accuracy can be compared, as well as the practicality in implementing it in real life in terms of computing requirements.

# Literature review:

## Critical attributes in predicting crop yield:

Farming is an important part of many countries’ economies, especially developing countries. Thus, being able to predict crop yield would tremendously help the government to make the appropriate action. Shrivastava, N., & Shukla, K. (2020) indicated that the ability to predict crop yield would help farmers plan and optimize their planting and harvesting, increasing the yield and reducing the cost.

According to Rao, M., S., (2022), crop yield is affected by many factors, including humidity, temperature, fertilizers, and soil quality. Thus, not only a good prediction model is needed, but the most impact attributes should also be determined. Xu X. (2019) indicated that crop yield is very sensitive to weather and Ray et al. (2015) also identified that changes in weather contributed roughly 33% of the crop variability globally. Lobell et al. (2007) also agreed that the variability in temperature can have a significant impact on the crop. This thinking was supported and enhanced by A. Rosenzweig and M. Parry (1994), who stated that the two most important attributes were temperature and precipitation.

Based on this, it can be understood that it would be great to have data for all variables when building the model, but it is not the case all the time due to the limitation of available data. Thus, it is critical to identify significant attributes and then focus on collecting and cleaning them, instead of trying to cover everything, which is unrealistic. Even within a higher level of control environment like manufacturing, many variables can impact the quality of the products or the output of the production lines.

## Suitable algorithm for this purpose:

N. Shrivastava and K. Shukla (2020) conducted research on this subject and they used only weather attributes like temperature, humidity, rainfall, and wind speed. Using K nearest neighbor (KNN), random forest, decision tree, and support vector machine (SVM), they managed to build a model that accurately predict crop yields in three locations in India. The support vector machine model had the highest accuracy at 90.5% for one of the locations, while random forest and KNN had 86.9% and 83.7% respectively.

Wang et al. (2020) did research potential machine learning models that can be used for this purpose and also provided their pros and cons. While an Artificial neural network (ANN) is considered great at capturing complex relationships between attributes, it requires extensive training data and ample computation power, which was the same for the random forest model and SVM. On the other hand, Decision trees are simple and easy to use but they may ignore the relationship between attributes and are vulnerable to overfitting. This conclusion was quite similar to Karimzadeh, A., & Abbaspour-Gilandeh, A. (2020), and Olaniyan et al (2019). Thus, it could be understood that decision trees, random forests, ANN, SVM, and KNN are reliable methods used in building a predictive model for crop yield using weather data.

## How to improve model accuracy:

In their paper, N. Shrivastava and K. Shukla (2020) used several methods to improve their model accuracy:

* + Feature selection: Recursive Feature Elimination (RFE) was used, combined with the correlation-based feature selection method to identify the most important attributes.
  + Ensemble learning: by combining the prediction of different models, the authors managed to increase the overall prediction.
  + Hyperparameter tuning: one of the examples given in the paper was adjusting the learning rate and the number of hidden layers in the neural network. This is an area for me to focus on since this is an area, I have not had any experience with.
  + Data augmentation: random oversampling and data normalization were used to increase the size of training set

In another paper published by Kern et al. (2019), the authors managed to increase their model accuracy by applying several techniques:

* + Feature selection: stepwise regression was used to select the most relevant attributes
  + Data preprocessing: including data cleaning, normalization, standardization, and transformation (by taking the logarithm of some attributes)
  + Cross-validation: the author used k-fold cross-validation to test the performance of different models

Additionally, there are other researches indicating that using real-time data can help to capture the complex and dynamic relationships between attributes, as well as indicate the most significant one. For example:

* + De Castro et al. (2018) set up a wireless sensor network to capture different variables including soil moisture and temperature. This enabled the author to analyze and identify the intriguing relationships between them, which in turn enhanced the performance of the model.
  + Yao et al. (2019) analyzed different data sources such as satellite imagery, soil sensors, and climate data.
  + In another research, Liu et al. (2018) also use satellite imagery and climate data to identify the spatial and temporal variability of the crop.

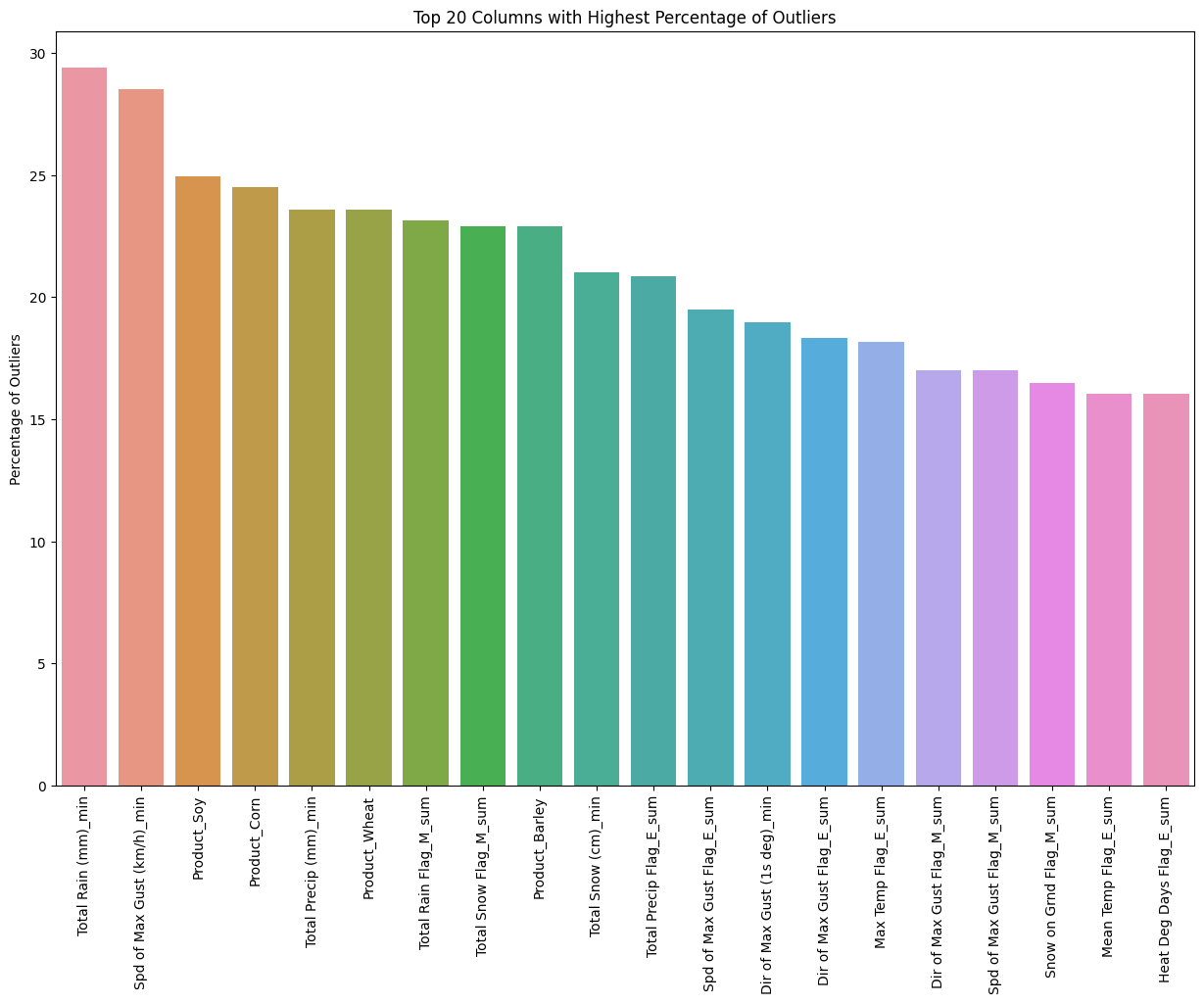
## Implementation of this research after finishing the model:

A quick search on Google scholar shows more than 100 thousand results and the number of models had been developed for this purpose is numerous. That’s the reason why my intention for this project is to learn the whole process of building a predictive model which I can implement later in my company. In the overall method, I will go through the suggest approach that using all potential techniques.

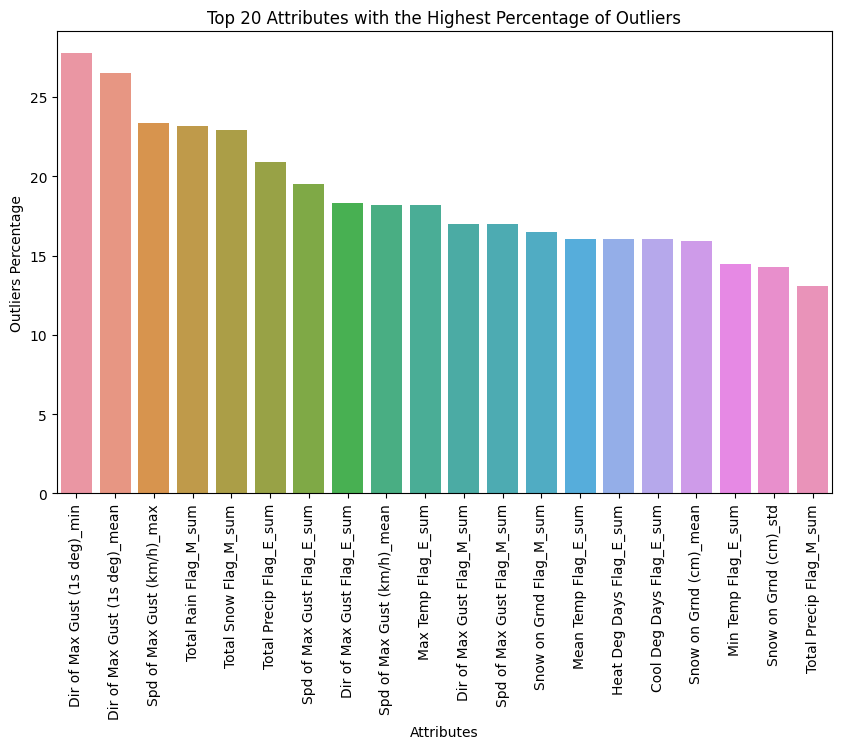
# Finding and reasoning during development:

## Data pre-processing

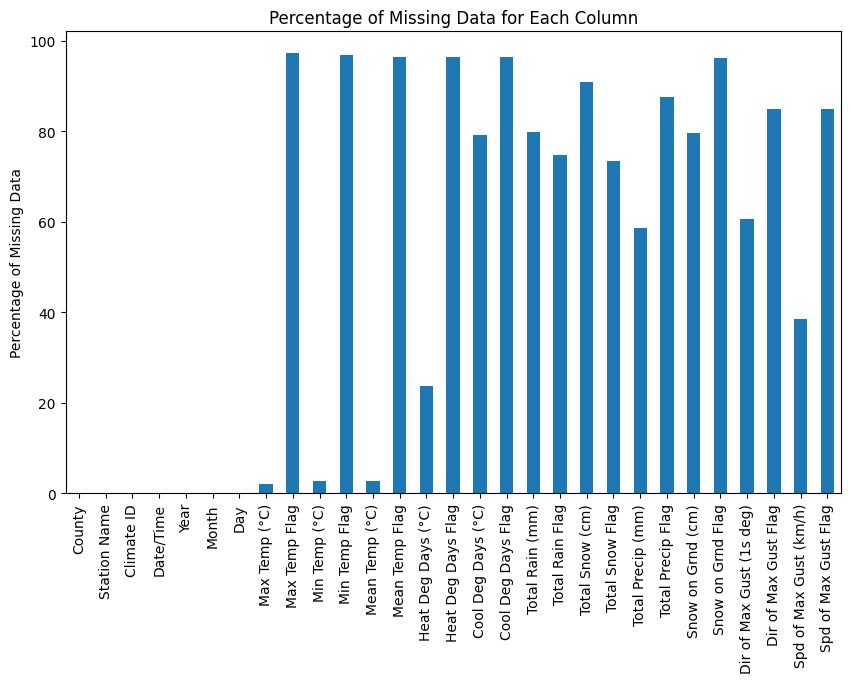
* + Initial approach:
    - First try: keep all original weather data and apply Regression Imputation with max iteration 200 🡪 data return with high level of outliers.

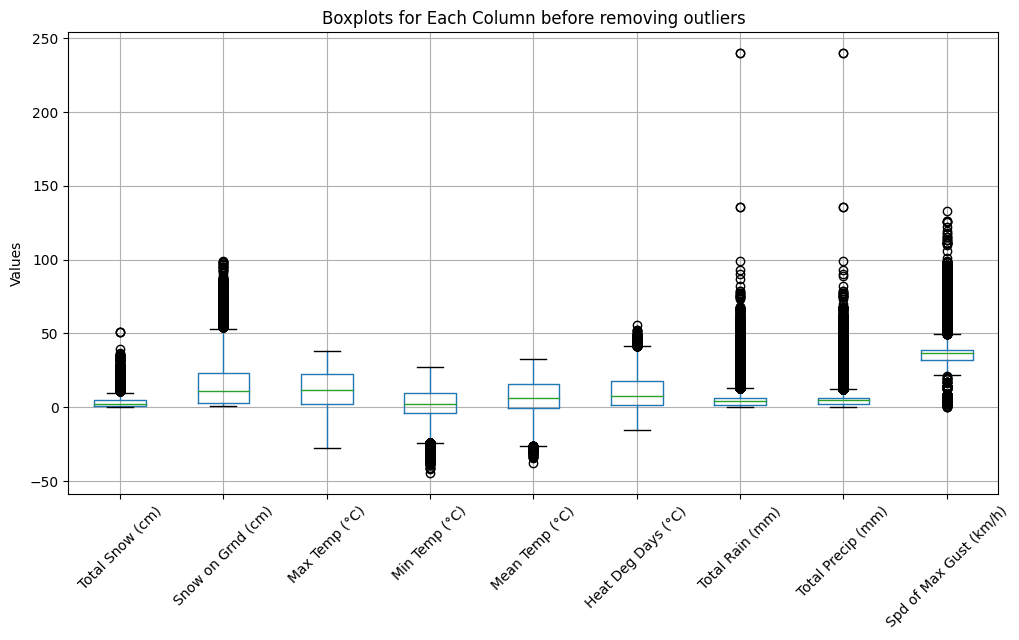
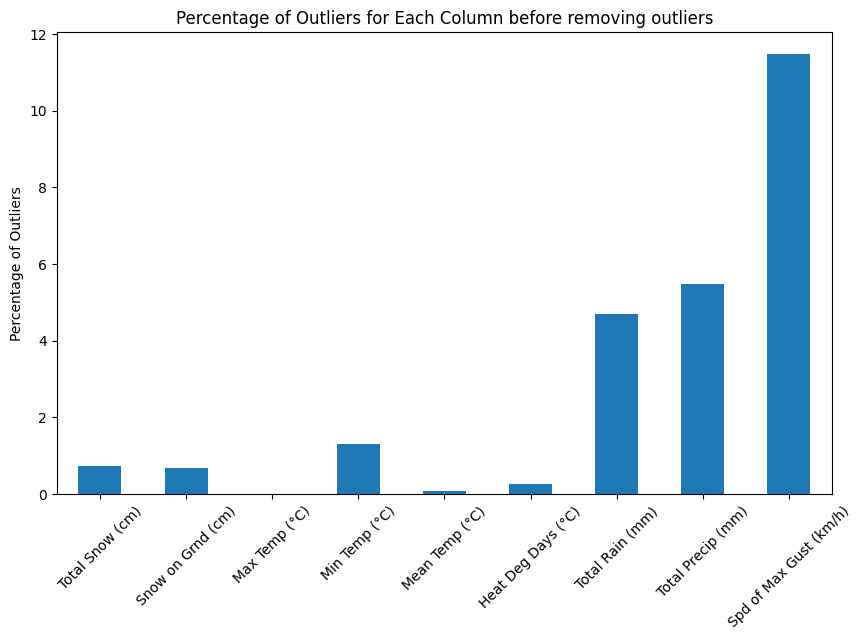


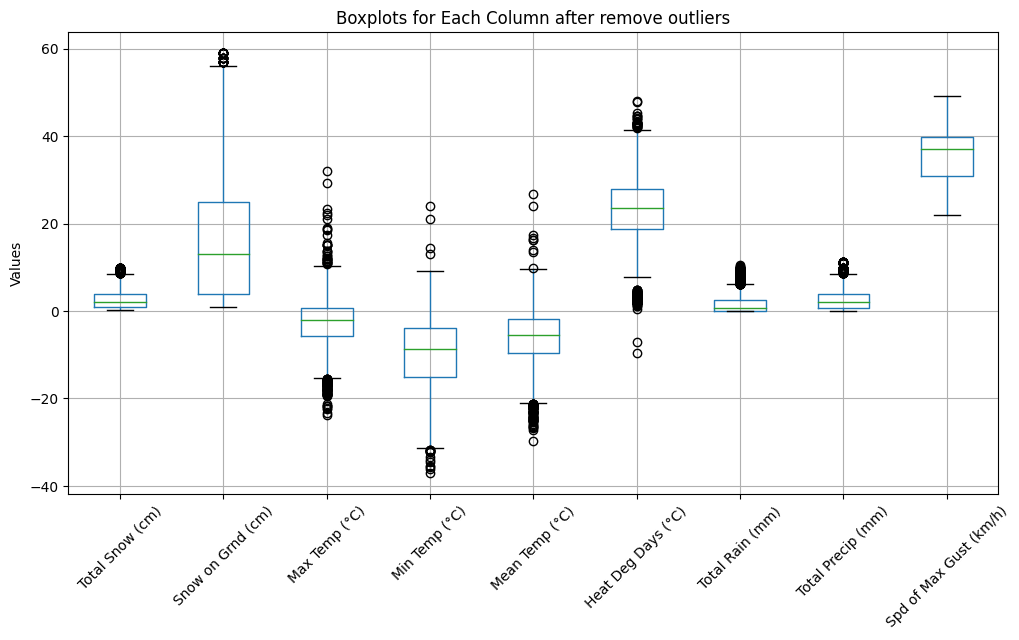
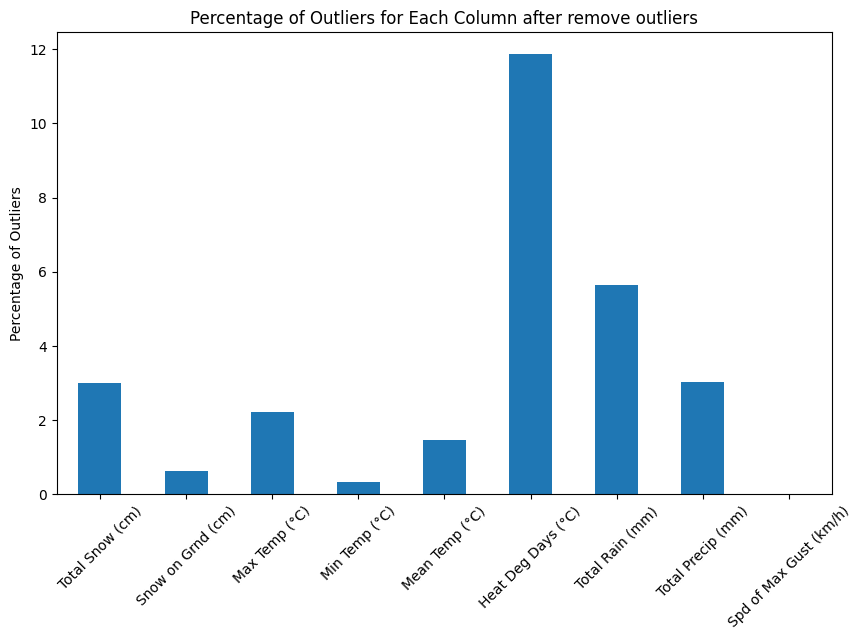
* + - Second try: group data by County and Year, then apply Regression Imputation with max iteration increase from 200 to 10000 🡪 still get the warning Early stopping criterion not reached and introduced less outliers but still very high

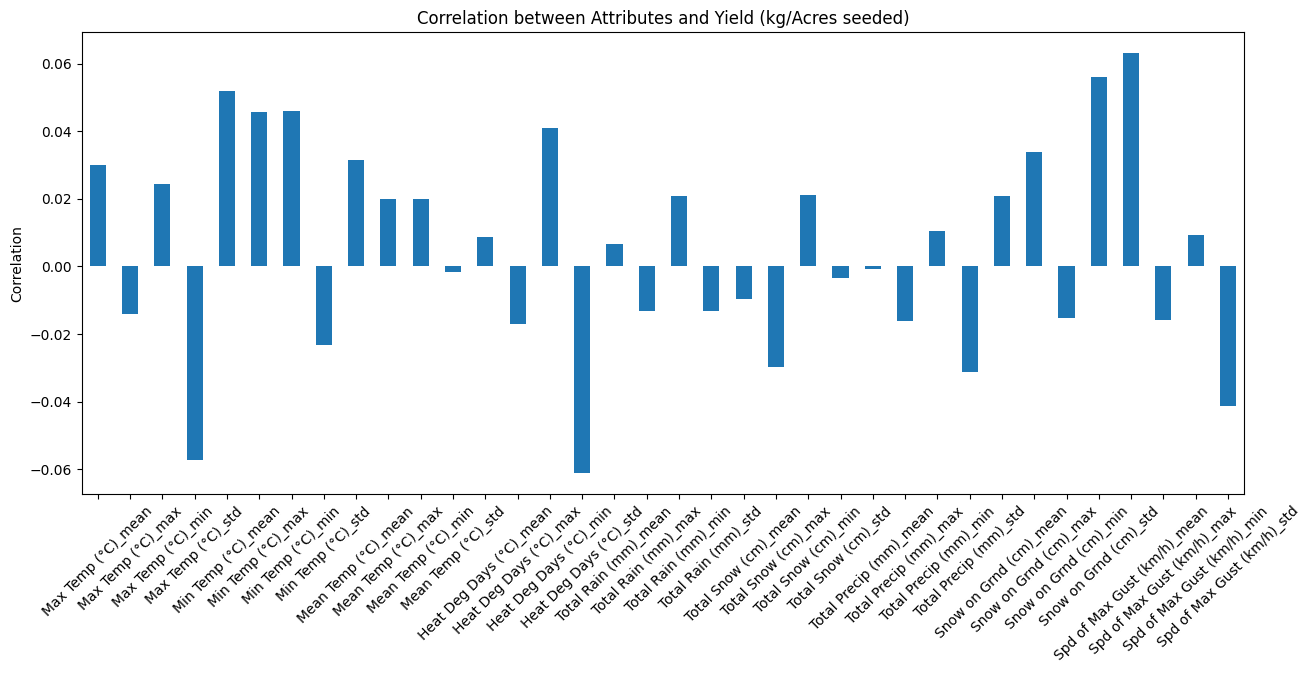


* + Final approach:
    - Read through the data and understand how each attribute mean, combine with the % of missing data helped to delete non-value added and duplicated attributes.
    - Apply Regression Imputation on the remained attributes and remove outliers for attributes that possess high % of missing data
    - Group by Year and provide statistical summary, then combine with Crop data and remove observations with 0 value in target variable







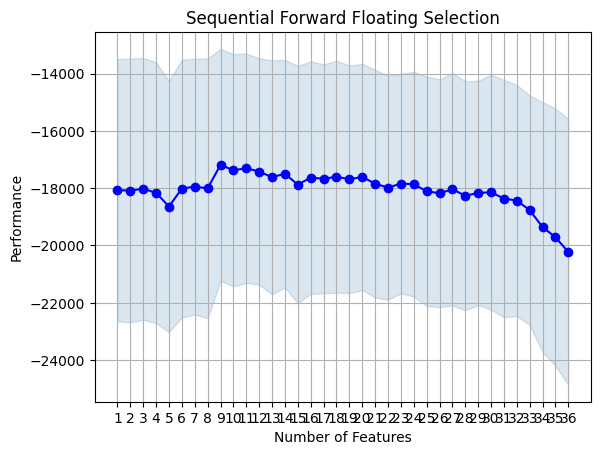


* + Using Regression Imputation created invalid outliers which can be observed in the Boxplots, so outliers in Attributes that have the highest missing values were removed. Also, due to low correlation between those attributes with the target variable, RandomForestRegressor eliminator and 10 fold validation will be used during the Feature selection phase

## Feature selection:

* + Technique will be use is RFE and bi-directional stepwise selection:
    - RFE result: Optimal number of features: 5

Selected features: Index(['Max Temp (°C)\_mean', 'Max Temp (°C)\_min', 'Min Temp (°C)\_max', 'Spd of Max Gust (km/h)\_max', 'Spd of Max Gust (km/h)\_std'], dtype='object')

* Stepwise selection: first run took 4 hours without result, needed to adjust n\_jobs parameter to enable parallel processing. It took 41 minutes with 12 CPU cores.   
  Selected features: Index(['Min Temp (°C)\_mean', 'Mean Temp (°C)\_std', 'Total Rain (mm)\_min', 'Total Snow (cm)\_max', 'Total Snow (cm)\_min', 'Total Precip (mm)\_min', 'Snow on Grnd (cm)\_max', 'Snow on Grnd (cm)\_min', 'Spd of Max Gust (km/h)\_min'], dtype='object')
  + Since 2 techniques suggest 2 different sets of attributes, a third set which is the combination of those 2 will be used in the next step to build the models.
  + 

## Model training and selection:

* + Using the 3 datasets along with 2 models (Decision Tree and Support Vector Machines) give below result:
    - Set 1 - Decision Tree Mean Squared Error: 21505.456
    - Set 1 - Support Vector Machine Mean Squared Error: 18179.727
    - Set 2 - Decision Tree Mean Squared Error: 19224.939
    - Set 2 - Support Vector Machine Mean Squared Error: 18179.319
    - Set 3 - Decision Tree Mean Squared Error: 20251.899
    - Set 3 - Support Vector Machine Mean Squared Error: 18179.269
  + Based on this result, Set 3 work well with SVM model, and all result provide much higher accuracy comparing with the initial result.

## Model optimization:

* + Using GridSearchCV to identify best hyperparameters to be used. The options were considered are

param\_grid = {

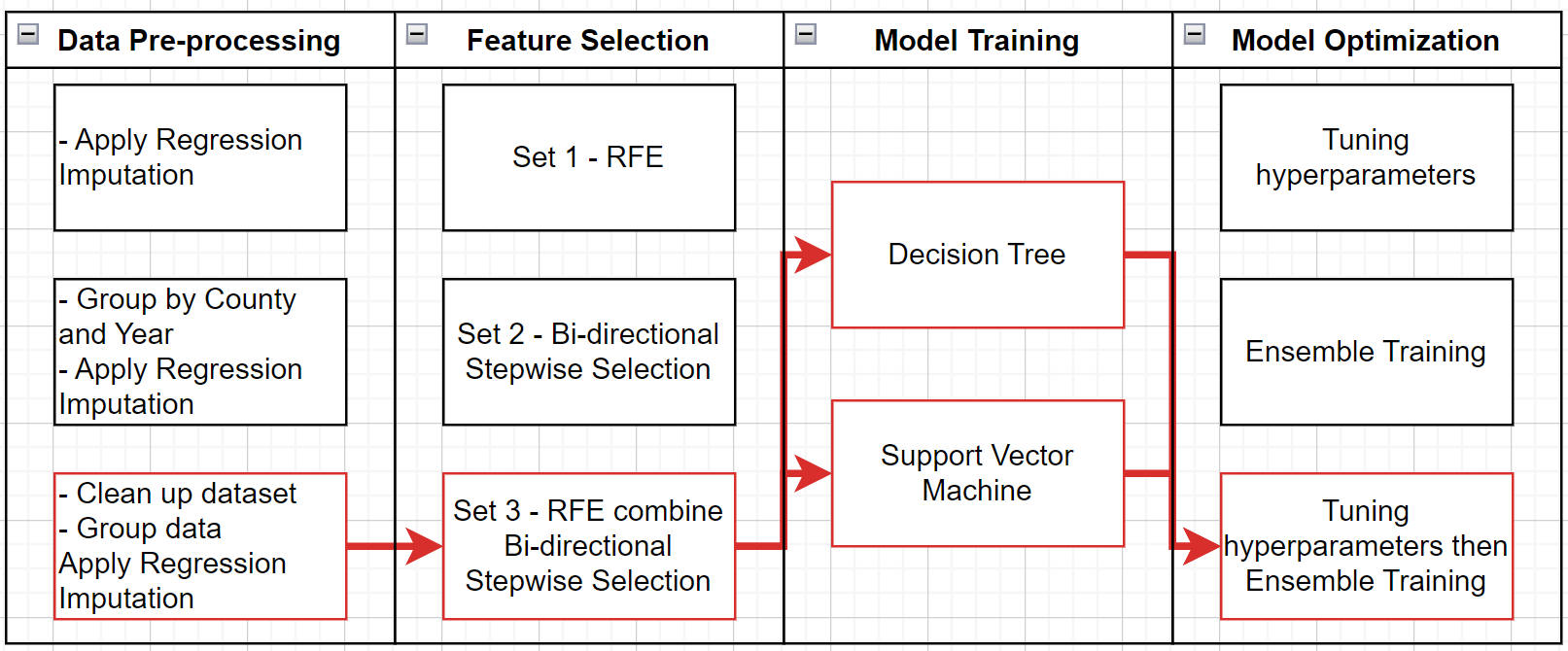
'C': [0.1, 1, 10],

'kernel': ['linear', 'rbf' , 'poly'],

'gamma': ['scale', 'auto']}

* + Approach 1: adjusting hyperparameters ('C': 10, 'gamma': 'auto', 'kernel': 'rbf') of the SVM model using set 3 attributes, this return the MSE of 18177.824
  + Approach 2: using ensemble learning for Decision Trees set 2 and SVM set 3, which return the MSE of 17594.462
  + Approach 3: final option is adjusting the hyperparameters of the SVM model then apply ensemble learning between it and the Decision Trees set 2. The MSE is 17333.503 which is our best result

## Process flow chart:



# Conclusion

In conclusion, this study focus on make predictive model for forecast crop yield in Ontario, Canada, by using weather data. We look at many machine learning algorithm, and choose Decision Trees and Support Vector Machines (SVM) for our models. This research show importance of data pre-processing, feature selection, and model optimization for get better accuracy in predict crop yield.

The result show that SVM with specific hyperparameters and combination of attribute sets work well for accuracy. Also, using ensemble learning between Decision Trees and SVM models make Mean Squared Error (MSE) better. This study gives good understanding about process of building predictive model, and author wants to use this knowledge in other areas.

This research has some limitations, like not having real-time data and needing more validation with more data sources. But this project can be starting point for using machine learning in the author's work, especially in supply chain and inventory management.

# Link to Github

<https://github.com/LukeHNguyen/CIND820.git>

# Reference:

De Castro, R., Kebbe, A., Manzoni, L., & Farina, A. (2018). Wireless sensor network data for crop yield prediction: A machine learning approach. Sensors, 18(7), 2339. doi: 10.3390/s18072339

I. Supit (1997). “Predicting national wheat yields using a crop simulation and trend models”. *Agricultural and Forest Meteorology*. <https://doi.org/10.1016/S0168-1923(97)00037-3> [Volume 88, Issues 1–4](https://www.sciencedirect.com/journal/agricultural-and-forest-meteorology/vol/88/issue/1), Pages 199-214

Karimzadeh, A., & Abbaspour-Gilandeh, A. (2020). Crop Yield Prediction using Machine Learning: A Review. International Journal of Agricultural and Biological Engineering, 13(4), 1-17. doi: 10.25165/j.ijabe.20201304.5702

Kern, A., Barcza, Z., Marjanović, H., Árendás, T., Fodor, N., Bónis, P., Bognár, P., & Lichtenberger, J. (2019). Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices. Agricultural and Forest Meteorology, 276-277, 107617. <https://doi.org/10.1016/j.agrformet.2019.107617>

K.O. Adekalu and H.O. Fapohunda (2006) “A numerical model to Predict Crop Yield from Soil-Water Deficit”. <https://doi.org/10.1016/j.biosystemseng.2006.03.012>

Khaki, S. and Wang, L. (2019) “Crop Yield Prediction Using Deep Neural Networks”. *Sec. Computational Genomics.* <https://doi.org/10.3389/fpls.2019.00621> Volume 10 - 2019

Klompenburg, T., Kassahuna, A., and Catal, C. (2020). “Crop yield prediction using machine learning: A systematic literature review”. <https://doi.org/10.1016/j.compag.2020.105709>

Liu, B., Xu, Y., Xu, X., Wang, H., & Cheng, Q. (2018). Sugarcane yield prediction using a machine learning approach with a combination of satellite imagery and climate parameters. Computers and Electronics in Agriculture, 155, 222-233. doi: 10.1016/j.compag.2018.10.002

Lobell, D. B., Field, C. B., & Cahill, K. N. (2007). Climate change and crop yields: The importance of temperature variability. Environmental Research Letters, 2(1), 014002. doi: 10.1088/1748-9326/2/1/014002

Olaniyan, T. O., Agbele, K. K., & Adeyemo, S. A. (2019). A review of crop yield prediction models with emphasis on machine learning techniques. Journal of Agricultural Science, 11(11), 129. doi: 10.5539/jas.v11n11p129

Rao M., S., et al, 2022, J. Phys.: Conf. Ser. 2161 012033, **DOI** 10.1088/1742-6596/2161/1/012033

Ray, D.K., Gerber, J.S., MacDonald, G.K., West, P.C., 2015. Climate variation explains a third of global crop yield variability. Nat. Commun. 6, 5989.

Rosenzweig, C., & Parry, M. L. (1994). Climate change impacts on agriculture: Implications for crop production. Climate Research, 3(2), 149-160. doi: 10.3354/cr003149

Shrivastava, N., & Shukla, K. (2020). Crop Yield Prediction: A Machine Learning Approach. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

V.R.Reddy and Ya.A.Pachepsky (2000). “Predicting crop yields under climate change conditions from monthly GCM weather projections”. <https://doi.org/10.1016/S1364-8152(99)00011-0>

Wang, J., Huang, G., Su, Y., Chen, X., Wang, G., & Zhang, J. (2020). A review of crop yield prediction methods for sustainable agricultural production management. Sustainability, 12(21), 8992. doi: 10.3390/su12218992

Xu. X., et al, 2019, Design of an integrated climatic assessment indicator (ICAI) for wheat production: A case study in Jiangsu Province, China, DOI <https://doi.org/10.1016/j.ecolind.2019.01.059>

Yao, X., Zhang, L., Chen, C., Ye, Y., & Zhang, H. (2019). Combining multi-source data for crop yield prediction using deep learning: A case study in China. Precision Agriculture, 20(5), 986-1003. doi: 10.1007/s11119-019-09670-4