CMPT 318 Project Proposal

Name: Jillian Watson

Student Number: 11253320

NSID: jiw248

Dataset Description & Licensing --The dataset for this project is available from the Genomes to Fields 2022 Maize Genotype by Environment (G x E) Prediction Competition. The Competition site provides multiple datasets covering phenotype, genotype, soil, weather, and environmental attributes related to maize crop cultivation. This project is released under an open data license with the Public Domain Dedication and License v1.0.

Motivation & Importance – The primary motivation behind the analysis of the Genomes to Fields Competition datasets is to develop a predictive model for the maize grain yield based on various environmental attributes, namely all-sky surface PAR total, which measures the total amount of solar radiation for the specified area, profile soil moisture, which measures the water content of the soil profile, and plant height, which is calculated by measuring the distance between the base of a plant and the ligule of the flag leaf in centimeters. This analysis is important to solve for the agricultural domain as it attempts to provide insight into the relationships between environmental conditions and crop yield. Locally, the information and results collected from this analysis will aid farmers in optimizing their practices for their fields. On a global scale, this analysis is important for meeting the demands of growing populations by mitigating food insecurity.

Challenges – The task presented by the competition dataset for crop yield prediction is challenging due to the complexity of the datasets and features, year-to-year variability, and lack of control over environmental factors. There is difficulty predicting crop yield given the many variables that can influence it, making it complex to pinpoint the most significant factors. While some factors can significantly influence crop yield in one year, they may be less significant the following year(s) due to changing and migrating weather patterns, precipitation, and temperature. The data collected in the maize prediction competition datasets were collected in an outdoor agricultural field environment. It is unlike a controlled environment in a lab where scientists can control and mitigate pests, temperature, and exposure to the outdoor elements. It can be challenging to collect sufficient data when scientists cannot control natural disasters, pest and disease outbreaks, and damaging weather patterns.

Task Classification – The task being presented is a regression problem. The aim is to predict maize grain yield based on a set of input features that include temporal information, all-sky surface PAR total, profile soil moisture, and plant height. The output of the task will be a continuous numerical value that represents the predicted grain yield.

Task Transferability – The primary domain for the task is agricultural crop yield prediction. The techniques of the project may be transferable to other domains of regression problems that involve environmental factors as a significant contributor.

Evaluation Metrics – I am unable to locate metrics defined by the competition itself. For this model, various metrics can be considered to determine the goodness or badness of outputs, such as the R-squared value and Mean Square Error.

Fig. 1

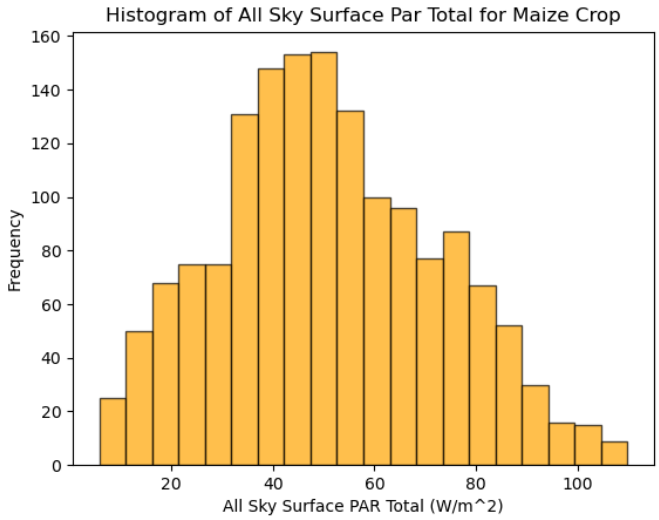

Fig. 1-- A histogram depicting the All-Sky Surface PAR (Photosynthetic Active Radiation) Total over various locations of maize crops in the 2022 season. The histogram depicts a normal distribution curve, peaking at around 50 Sky Surface PAR total, measured in Watts per Meter squared.

Fig. 2

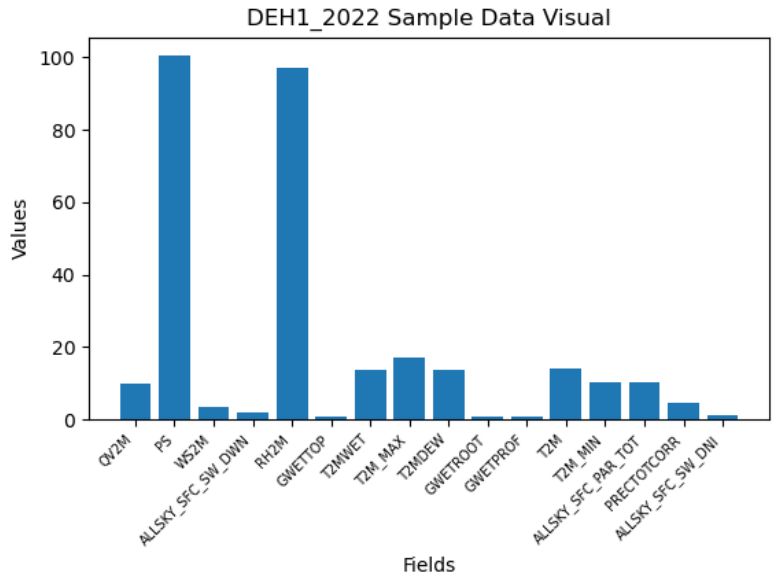

Fig. 2-- A bar chart from the Weather Data Set depicting the test sample DEH1_2022 and its values.

Fig. 3

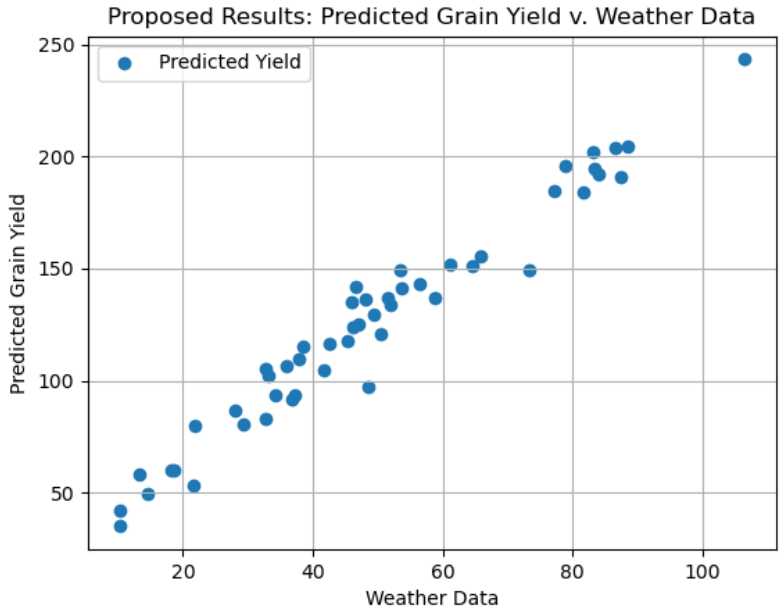

Fig. 3-- An engineered proposed results scatter plot for the Sky Surface PAR Total data and predicted grain yield using random normal distribution. Data was downsized to a sample of 300. A possible significant linear relationship is depicted.