Introduction

The project aims to utilize data analysis tools and methodologies to present insights in Saskatchewan crops. It is the scope of this project to know which crops are harvested the most in Saskatchewan and which area are these to be found.

Saskatchewan is among the world's largest exporter of various field crops. With crop yield analysis, growth pattern can be seen across the different geographical location over time. With these insights, farmers can make better decisions on resource management, including optimal crop type and use of land.

But first, what is crop yield? Crop yield is a standard measurement of the amount of agricultural production harvested—yield of a crop—per unit of land area, which in our dataset is bushels per acre.

Data Collection and Preprocessing

The dataset given covers the period from 1938 to 2021 from 299 unique rural municipalities (RM). It has crop yield data for Winter Wheat, Canola, Spring Wheat, Mustard, Durum, Sunflowers, Oats, Lentils, Peas, Barley, Fall Rye, Canary Seed, Spring Rye, Tame Hay, Flax and Chickpeas. There were 25017 rows of data.

The geographic data has geometry information for 298 rural municipalities. There are no geometry information for Kutawa 278, Prairie No. 408 and Greenfield No. 529. These are dissolved rural municipalities. Greenfield No. 529 was absorbed by RM of Mervin 499.

Data types are converted to match between main data frame and geographic data frame. Data columns are dropped on a necessity basis.

Exploratory Data Analysis (EDA)

Top crops over the years are Oats, Winter Wheat, Barley, Peas, Durum and Spring Wheat. I chose to look into Spring Wheat because it has the most number of records of data. As can be seen, there are certain years, like 1961, 1988, 2002, 2021 where the yield dropped. There are records of drought in those years. I looked into the municipalities that have the highest yield of Spring Wheat. Flett's Spring (RM 429) is one of the top producer. Looking at the group of crops that this municipality produced, it confirms the correlation between some crops like Spring Wheat, Barley, Oat and Flax, as determined by the correlation matrix. Canola, being a newly introduced crop in the 1960s became a part of their crop line too.

Methodology

Within the field of machine learning, there are two types: supervised and unsupervised. Supervised learning uses labeled datasets whereas unsupervised learning uses unlabeled datasets.

K-means Clustering

One technique of unsupervised learning is clustering. Here, I used the K-means clustering algorithm. This algorithm used an iterative process to find similarities between data points and group them into clusters.

Clustering Spring Wheat, as we can see from the map, it is in the northern part of Saskatchewan where municipalities have higher yields for Spring Wheat, which is the similar case for Barley and Oats, while for Canola, it's a bit distributed in location.

Time Series Analysis

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. Time series forecasting is a technique that utilizes historical and current data to predict future values over a period of time or a specific point in the future.

Autoregression is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step.

Rather than using past values of the forecast variable in a regression, a moving average model uses past forecast errors in a regression-like model.

Exponential Smoothing Methods are a family of forecasting models. They use weighted averages of past observations to forecast new values. The idea is to give more importance to recent values in the series. Thus, as observations get older in time, the importance of these values get exponentially smaller.

Simple/single exponential smoothing: This smoothing can be used for making forecasts based in a time series that has no trend and seasonality. This method is suitable for forecasting data with no clear trend or seasonal pattern.

Double exponential smoothing: This type of exponential smoothing comes with the support for trend components of time series.

Triple exponential smoothing: This type of exponential smoothing comes with the support for trend and seasonality components of time series.

A forecast "error" is the difference between an observed value and its forecast. Mean Absolute Error, MAE, is simply, as the name suggests, the mean of the absolute errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value. The Mean absolute error is calculated by adding up all the absolute errors and dividing them by the number of errors.

Mean Absolute Percentage Error, MAPE, is the sum of the individual absolute percentage errors divided by the number of errors.

MAPE is more understandable than MAE for end users as it is given as a percentage. MAE varies in scale depending on the target you are predicting for, making it difficult to compare across models. This is a problem that MAPE does not have as it is given as a percentage

As you can see from the graph, the double exponential smoothing model has the lowest mean absolute percentage error.

Conclusion

The availability of the recent data analytics tools and methodology coupled with availability of data will greatly help farmers in the near future as more targeted improvement initiatives be it in technology, infrastructure, equipment can be developed and more fit-for-purpose pilot runs can be deployed to maximize value / returns.

Future Work

Additional information like weather, pests incidents, soil type, topography as weighted parameters can provide a better analysis of crop yield. With these additional data, better time analysis forecasting models can be tested as well.