From Data to Harvest: Leveraging Machine Learning for Wheat Yield Prediction in Kansas

1. Contribution

Fetching data

- Yield, Plant Area: Jinho Lee
- County: Dieu-Anh Le
- Soil: Fu-Hsin Liao
- Weather: Yin-Kai Huang
- Cleaning data: Fu-Hsin Liao, Yin-Kai Huang, Jinho Lee

Building ML model:

- Building ML model: Jinho Lee, Fu-Hsin Liao
- Accuracy Metrics Calculation: Yin-Kai Huang, Fu-Hsin Liao

NDVI:

- Building NDVI pipeline: Dieu-Anh Le, Fu-Hsin Liao, Jinho Lee

Background research:

- Relating to previous research: Jinho Lee
- Regional social, economics, and political information: Dieu Anh Le

Report writing: Dieu-Anh Le, Fu-Hsin Liao

Presentation: Dieu-Anh Le, Fu-Hsin Liao, Jinho Lee, Yin-Kai Huang

2. Introduction and Motivation

With the advancements in technology, AI and Machine Learning have become a crucial part of modeling and predictions across various fields. The scarcity of resources further emphasizes the importance of highly accurate predictive models for effective resource allocation and decision-making processes.

In this project, our focus was on developing a comprehensive wheat prediction model specifically tailored to the unique agricultural conditions of Kansas. Known as the "Wheat State" and the leading wheat producer in the US, Kansas serves as an ideal setting to explore the potential of machine learning in predicting wheat yields. We adopted the framework proposed in the paper "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt" [1] and collected data from diverse sources including USDA NASS, ISRIC, NASA POWER, among others. The subsequent section will delve into further details regarding our data acquisition process.

Wheat holds significant importance on a global scale, serving as a staple food, a valuable cash crop, and a key aspect of cultural diversity in various regions worldwide. In a narrower context, Kansas experiences extreme weather conditions as it falls within the infamous "Tornado Alley", which is a region prone to severe thunderstorms and tornadoes during the spring and early summer months. The wheat planting season for Kansas farmers typically begins around September, requiring the crop to endure harsh winter conditions and unpredictable spring weather before it can be harvested. Complicating matters, Kansas is situated in the convergence zone of

the Gulf of Mexico and the Rocky Mountains, further contributing to the volatile and unpredictable nature of its weather patterns. By constructing a reliable and robust predictive model, we aim to assist farmers in mitigating the impact of extreme weather events on wheat production, thereby enhancing their resilience and agricultural productivity.

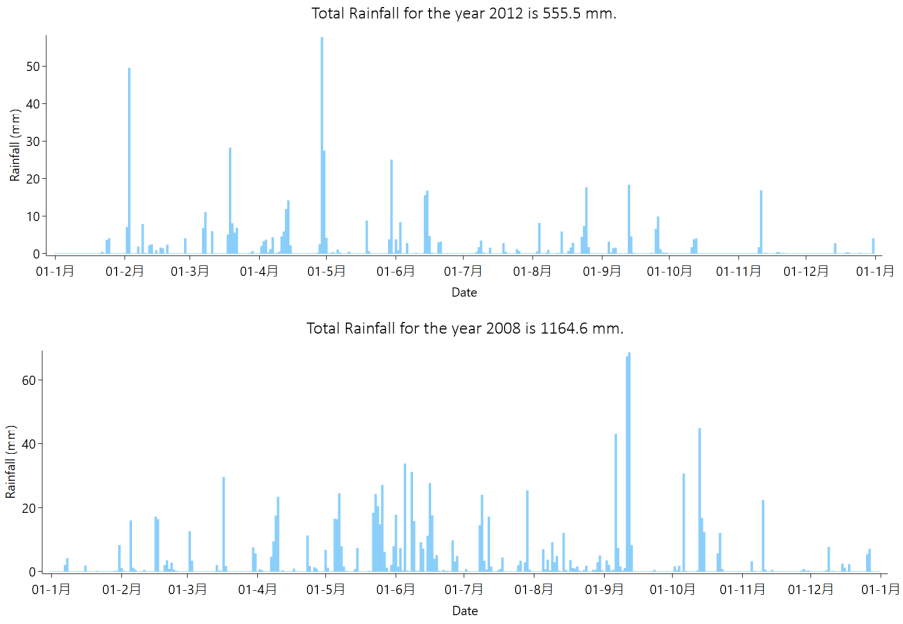
3. Mastery of Tools and Algorithms

Machine Learning, APSIMx

Initially, we obtained the wheat yield data from the USDA NASS, focusing specifically on Kansas, the leading wheat-producing state in the US. Our original dataset contained over 5000 values, but after thorough data cleaning and preprocessing, we were left with approximately 2500 values. Although the dataset size was not as extensive as anticipated, we believe it provides sufficient data to develop an accurate prediction model. To ensure a robust evaluation, we adopted an 80:20 ratio for splitting the dataset into training and testing sets.

To enrich our dataset, we proceeded to retrieve longitude and latitude information for the 105 counties in Kansas. Leveraging this geographical data, we obtained additional soil and weather information using resources such as ISRIC and NASA POWER, respectively. During the data cleaning process, we addressed certain inconsistencies in the soil data, specifically replacing the Soil Quality High and Low values with one-hot encodings to facilitate accurate analysis and model training. Additionally, we utilized APSIM, an agricultural simulation model, to generate precipitation patterns spanning 20 years, as shown in the two figures below. This aided us in visualizing and better understanding the temporal patterns of precipitation in Kansas.

By incorporating these diverse data sources, we aimed to capture the multi-faceted aspects influencing wheat production in Kansas, including soil conditions, weather patterns, and their interactions. The integration of this comprehensive dataset sets the foundation for developing a robust and reliable wheat prediction model tailored to the specific agricultural conditions of Kansas.



The process of gathering weather data for the growing season of wheat in Kansas involved fetching data for 12 months based on the longitude and latitude coordinates of the 105 counties. The data was obtained from NASA POWER and was initially available on a daily basis. To make it more manageable, the data was aggregated into weekly intervals. To organize the weather data, a dictionary was created where each row represented a dataframe containing weather information for a specific year, longitude, and latitude combination. This allowed for easier access and manipulation of the data during the modeling process.

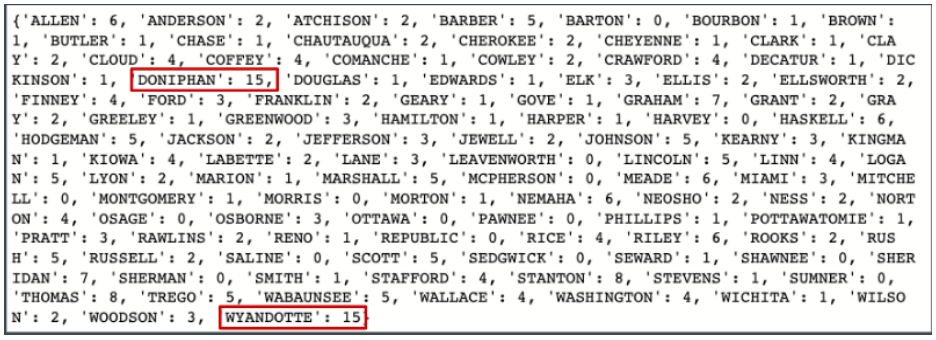
Normalization of the data was performed using the MinMaxScaler and StandardScaler from the Scikit-Learn library. This scaler transforms the data to a range between 0 and 1, taking into account the different scaling units of the attributes. Normalization helps to ensure that all the features are on a similar scale, which can improve the performance of the machine learning models.

Three different machine learning models were used in this project: Lasso, Normalized Linear Regression, and Random Forest. Each model was trained and evaluated using the normalized weather data and the corresponding wheat yield data. The results and findings of the models are discussed in the "Findings" section.

Overall, the process involved data collection, preprocessing, normalization, and model training using various machine learning algorithms. The results obtained from these models provide insights into the relationship between weather patterns and wheat yield in Kansas.

4. Data Cleaning

The data obtained from USDA-NASS are not complete because there were omitted yield values for each county. There are two possible outcomes (1) there were some harvest but not reported, or (2) no harvest at all. We took a further look at the dataset to deal with the NaN values. As shown in the figure below, many counties were missing yield value for at least one year, and some have up to 15 years of missing yield value such as Doniphan and Wyandotte counties.



```
{'ALLEN': 6, 'ANDERSON': 2, 'ATCHISON': 2, 'BARBER': 5, 'BARTON': 0, 'BOURBON': 1, 'BROWN':
1, 'BUTLER': 1, 'CHASE': 1, 'CHAUTAUQUA': 2, 'CHEROKEE': 2, 'CHEYENNE': 1, 'CLARK': 1, 'CLA
Y': 2, 'CLOUD': 4, 'COFFEY': 4, 'COMANCHE': 1, 'COWLEY': 2, 'CRAWFORD': 4, 'DECATUR': 1, 'DIC
KINSON': 1, 'DONIPHAN': 15, 'DOUGLAS': 1, 'EDWARDS': 1, 'ELK': 3, 'ELLIS': 2, 'ELLSWORTH': 2,
'FINNEY': 4, 'FORD': 3, 'FRANKLIN': 2, 'GEARY': 1, 'GOVE': 1, 'GRAHAM': 7, 'GRANT': 2, 'GRA
Y': 2, 'GREELEY': 1, 'GREENWOOD': 3, 'HAMILTON': 1, 'HARPER': 1, 'HARVEY': 0, 'HASKELL': 6,
'HODGEMAN': 5, 'JACKSON': 2, 'JEFFERSON': 3, 'JEWELL': 2, 'JOHNSON': 5, 'KEARNY': 3, 'KINGMA
N': 1, 'KIOWA': 4, 'LABETTE': 2, 'LANE': 3, 'LEAVENWORTH': 0, 'LINCOLN': 5, 'LINN': 4, 'LOGA
N': 5, 'LYON': 2, 'MARION': 1, 'MARSHALL': 5, 'MCPHERSON': 0, 'MEADE': 6, 'MIAMI': 3, 'MITCHE
LL': 0, 'MONTGOMERY': 1, 'MORRIS': 0, 'MORTON': 1, 'NEMAHA': 6, 'NEOSHO': 2, 'NESS': 2, 'NORT
ON': 4, 'OSAGE': 0, 'OSBORNE': 3, 'OTTAWA': 0, 'PAWNEE': 0, 'PHILLIPS': 1, 'POTTAWATOMIE': 1,
'PRATT': 3, 'RAWLINS': 2, 'RENO': 1, 'REPUBLIC': 0, 'RICE': 4, 'RILEY': 6, 'ROOKS': 2, 'RUS
H': 5, 'RUSSELL': 2, 'SALINE': 0, 'SCOTT': 5, 'SEDGWICK': 0, 'SEWARD': 1, 'SHAWNEE': 0, 'SHER
IDAN': 7, 'SHERMAN': 0, 'SMITH': 1, 'STAFFORD': 4, 'STANTON': 8, 'STEVENS': 1, 'SUMNER': 0,
'THOMAS': 8, 'TREGO': 5, 'WABAUNSEE': 5, 'WALLACE': 4, 'WASHINGTON': 4, 'WICHITA': 1, 'WILSO
N': 2, 'WOODSON': 3, 'WYANDOTTE': 15}
```

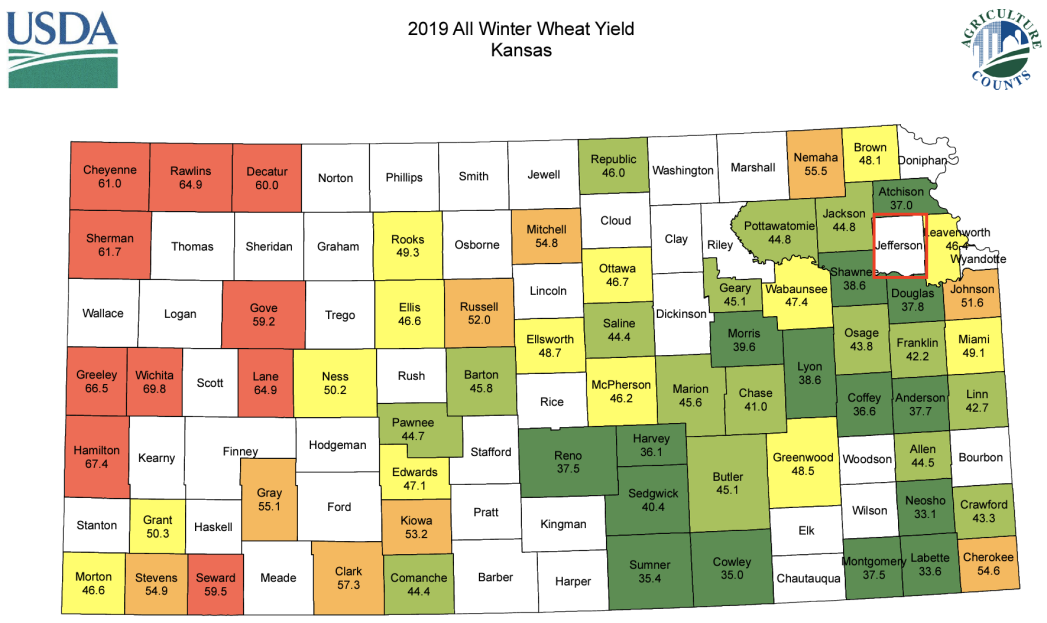
We hypothesized that planted area and yield value are positively correlated. That is, if there is a positive plant area, there must be a positive yield value, and vice versa. Hence, by looking through all the data points for each pair of county-year, we identified those that have positive planted area but zero for the yield values, which falls into the first outcome where there were harvest but not reported. Next, we cross-checked this dataset with the USDA CDL yield data where 40 random coordinates were generated for every county-year pair and only looked at the WinterWheat crop in that region. We kept track of the field count in the region and looked for

those with zero field count, which was an indication that there was no harvest, as shown in the figure below.

	year	state_name	county_name	yield	area_planted	field_count
0	2022	KANSAS	ALLEN	53.50	12100	0
1	2021	KANSAS	ALLEN	54.90	13800	2
2	2020	KANSAS	ALLEN	49.70	0	1
3	2019	KANSAS	ALLEN	44.50	4300	1
4	2018	KANSAS	ALLEN	43.65	0	0
...	...	...	...	...	...	...
2096	2007	KANSAS	WYANDOTTE	20.00	600	0
2097	2006	KANSAS	WYANDOTTE	43.00	400	0
2098	2005	KANSAS	WYANDOTTE	43.00	400	0
2099	2004	KANSAS	WYANDOTTE	49.00	800	0
2100	2003	KANSAS	WYANDOTTE	63.00	400	0

One short-coming with our data cleaning approach for the CDL dataset is that we randomly looked at 40 coordinates in each county, which may not be representative of the entire county, and hence the field_count may not be accurate. We could have increased the number of coordinates data but the tradeoff is time.

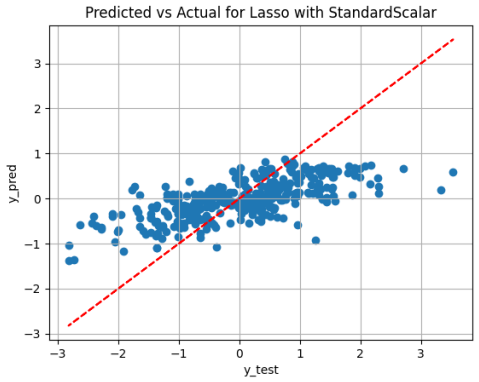
Next, we processed these data with NDVI computation by first generating valid county-year pairs, in which we omitted any data point that meets all of the following conditions: (1) plant-area is NaN, (2) yield is NaN, and (3) CDL field count is 0. As a result, we removed 49 rows without any harvest. Then, we set the period for getting sentinel image information to May because its NDVI value is significantly higher than other months in a year. Hence, our value of interest lies between 0.6 and 0.8, which resulted in 6 data points satisfying this range. We then compared them with the actual result and yet the prediction accuracy is not what we expected. The figure below showed different harvest data for different counties in Kansas, which matches with our data cleaning.

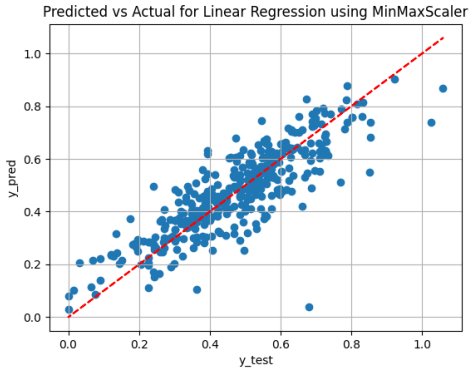


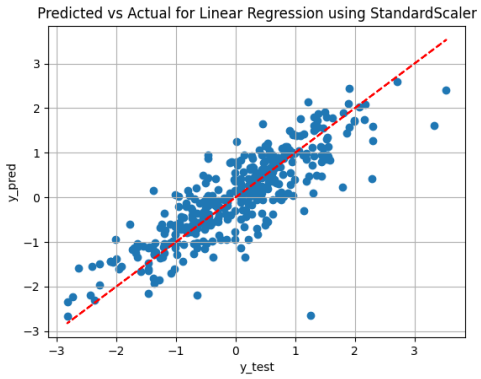
Finally, we do the last round of omission for any data points that do not have the NDVI values lie within the desired range of 0.6 and 0.8, where we set the yield values to 0. This is how we dealt

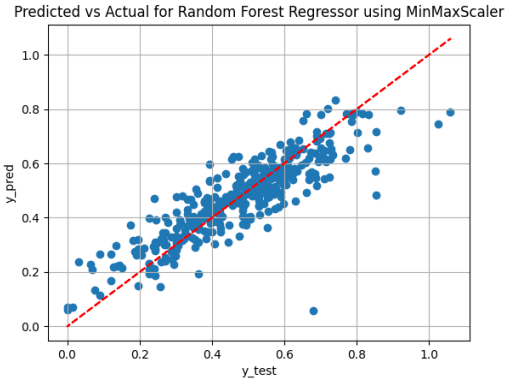
with the second outcome for a particular region without a report due to no harvest. Regarding the first outcome where there was a harvest but without report, we employed linear interpolation to replace the NaN values, which is taking the average of the two adjacent years.

5. Findings







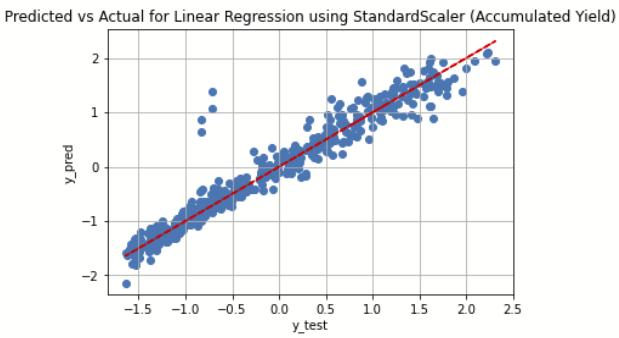


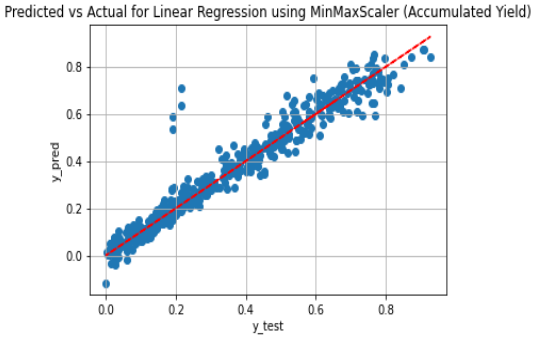
	Lasso	Linear Regression	Random Forest
R-squared	0.388	0.708	0.75
RMSE	9.137	6.313	5.839
Rank	3	2	1

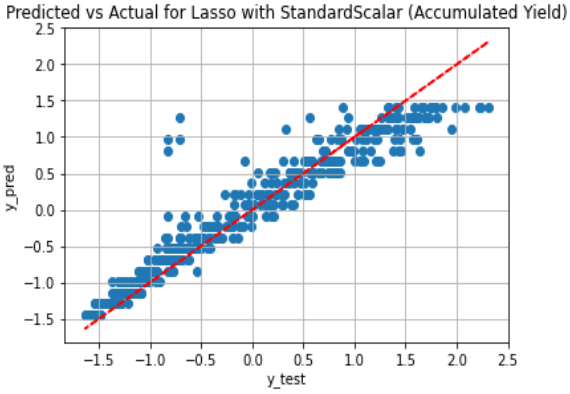
Table 1. Comparison among Three Machine Learning Models

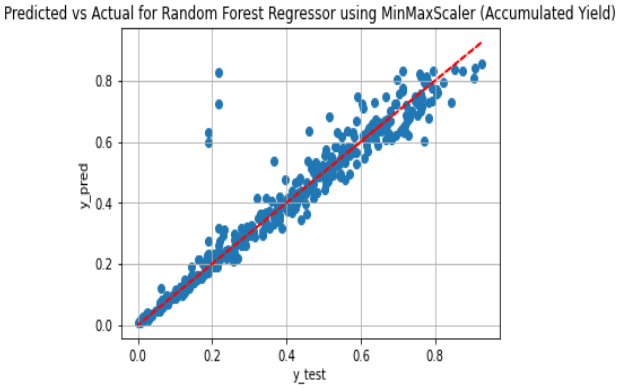
We employed three different Machine Learning models in our project, which are Lasso Regression, Linear Regression, and Random Forest. We evaluated the performance of these models using r-squared value and Root Mean Squared Error (RMSE). Typically, a model is considered poor when the r-squared value is below 0.3, while a value above 0.7 indicates a reliable performance. In Table 1., we observe that both Linear Regression and Random Forest achieved an r-squared value exceeding 0.7. However, Lasso Regression only obtained an r-squared value of 0.388, indicating a relatively weaker performance.To compare the RMSE

values across all models, we applied an inverse transform to the prediction results. As shown in Table 1, Random Forest exhibited the smallest RMSE, suggesting it outperformed the other models. Overall, our findings demonstrate that Random Forest is the most effective ML model for our dataset, while Lasso Regression performed the least favorably.
The graphs above are the outcome of the Machine Learning Prediction without data cleaning. The metrics are not as we expected so we did some data cleaning and re-train the models, which yielded us the result as below.









6. Limitations and Future Works

5.1. Small dataset

Issue: Our model was developed using a relatively small dataset, which may limit its representativeness across various wheat-producing regions with diverse attributes.

Future Work: To overcome this limitation, it is recommended to expand the dataset beyond the counties in Kansas and include neighboring states. This would provide a more comprehensive and diverse set of training and testing data. Some of the neighboring states to consider are Nebraska, Missouri, Oklahoma, and Colorado. Additionally, to capture a broader perspective, we could include the top 5 or 10 wheat-producing states in the analysis. These states may include North and South Dakota, Montana, Washington, Oklahoma, and others. By incorporating data from a wider range of regions, our model would be more robust and applicable to a larger geographic area.

5.2. Overfitting

Issue: Due to the limitation of our model outlined earlier, there is a possibility of overfitting the predicted outcomes to specific regions and relying on certain attributes available only in those regions.

Future Work: To address this issue, it is recommended to replicate the model for different regions. By using this as a base framework, it becomes possible to incorporate region-specific weather and soil conditions, enabling the development of predictive wheat models tailored to each region's unique characteristics. This approach would enhance the model's applicability and accuracy across diverse geographical areas, providing valuable insights for wheat production in various regions.

7. Conclusion

In conclusion, by using data established by various well-known sources, we have created a predictive model for wheat in Kansas. Despite the challenges posed by extreme weather, Kansas remains the top wheat producer for a significant number of years. Overall, the development of a comprehensive wheat prediction model for Kansas holds significant promise for addressing the unique challenges faced by farmers in the state. By leveraging the power of AI and Machine Learning, we can contribute to improving wheat production, increasing resilience to extreme weather events, and ultimately supporting the agricultural sector and food security in Kansas.

8. Reference

a. Shahhosseini, Mohsen, et al. "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt." Scientific reports 11.1 (2021): 1-15.
b. Masialeti, Iwake, Stephen Egbert, and Brian D. Wardlow. "A comparative analysis of phenological curves for major crops in Kansas." GIScience & Remote Sensing 47.2 (2010): 241-259.
c. Shroyer, James P., et al. "Kansas Crop Planting Guide." L-818. Manhattan, Kansas: Kansas State University Cooperative Extension Service, 1996.
d. United States Department of Agriculture, National Agricultural Statistics Service. "Kansas County Estimates." 19KSww.pdf. Accessed 7 June 2023. URL: https://www.nass.usda.gov/Statistics_by_State/Kansas/Publications/County_Estimates/19KSww.pdf.