MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY BHOPAL (M.P.), INDIA
मौलाना आज़ाद राष्ट्रीय प्रौद्योगिकी संस्थान भोपाल (म. प्र.), भारत

Department of Computer Science and Engineering

Presentation for M. Tech. Thesis Final Evaluation (Semester - IV)

Feature-Driven Crop Yield Prediction using Tuned Tree-based Regressors

Under the supervision of :

"Dr. Akhtar Rasool"
"Dr. Rajesh Wadhvani"

Presented By:

Priyanshu Kumar (2321203119)

M. Tech. (Advanced Computing)

Outline

- Introduction
- Literature Review
- Research Gaps
- Problem Landscape
- Research Objectives
- Methodology
- Results
- Deployment
- Conclusion
- Limitations and Future Work
- References
- Publications

Introduction to Crop Yield Prediction

1. Overview:

- Crop yield prediction forecasts agricultural output using data-driven methods to improve decision-making and optimize resource management. [1]
- **Agricultural Impact:** It aids farmers, governments, and agribusinesses in making informed decisions on planting strategies, resource allocation, market planning, and food security.[2][4]
- **Growing Importance:** With increasing global food demand due to population growth and changing diets, optimizing agricultural productivity is crucial.
- Forecasting helps reduce economic losses by better predicting market supply and demand. [3][5]
- **Emergence of Data-Driven Solutions:** Data-driven methods are key to achieving food security and efficient agricultural management by offering more accurate and adaptive crop yield predictions. [3]

2. Applications:

Decision Making: Farmers can plan irrigation, fertilization, and crop rotation strategies based on predictive insights.

Supply Chain Optimization: Yield predictions help in managing storage, distribution, and pricing strategies for both local and global markets.

Insurance: Agricultural insurance companies use predictions for risk assessments and to design suitable insurance policies for farmers.

Policy Making: Governments and NGOs can use crop yield predictions to anticipate food shortages and plan relief efforts.

Climate Change Adaptation: Machine learning models can be used to predict the impacts of climate change on agricultural productivity, helping in developing mitigation strategies.

3. Challenges:

- The **complexity of ML algorithms** requires expertise in data science, making implementation challenging for farmers.[1][4]
- **Data Availability & Quality**: Limited access to reliable and complete datasets, especially in developing regions.[2][6]
- **High Computational Costs**: Training large models requires significant computational resources.[2][6]
- **Data Heterogeneity**: Integration of diverse data types (weather, soil, satellite) is complex.[3][5]
- **Environmental Uncertainty**: Unpredictable weather or pests reduce model reliability.[3][6]
- **Lack of proper infrastructure** to collect and process data in rural areas limits model effectiveness.[4]

Literature Review

S. No.	Title	Publication	Authors	Overview
1.	**Incorporating Meteorological Data and Pesticide Information to Forecast Crop Yields using Machine Learning** [1]	IEEE Access *(SCIE)* *2024*	MD Jiabul Hoque, MD Saiful Islam, Jia Uddin, MD Adbus Samad, Beatriz Sainz de Abazo	• **Methods:** Used Gradient Boosting, KNN, and Multivariate Logistic Regression with GridSearchCV & k-fold cross-validation. • **Results:** Gradient Boosting achieved $R^2 \approx 99.94\%$, outperforming other models. • **Dataset:** Combined FAO & World Bank data (**1990–2013**) on rainfall, temperature, pesticides, & yields for 6 crops (rice, wheat, potato, soybean, sweet potato, sorghum) across India (~3142 cleaned rows). • **Features:** Included rainfall, avg temp, pesticide use; performed IQR-based outlier removal, standard scaling & one-hot encoding of crops.
2.	**Ensemble learning prediction of soybean yields in China based on meteorological data** [2]	Journal of Integrative Agriculture (SCIE) *Elsevier*, 2023	Ll Qian-chuan, XU Shi-wei, ZHUANG Jia-yu, LIU Jia-jia, ZHOU Yi, ZHANG Ze-xi	• **Methods:** Stacking ensemble with RF, SVR, KNN as base models & Ridge Regression as meta-learner; PCA used for dimensionality reduction. • **Results:** Achieved MAPE < 5% on 5-year sliding predictions for soybean yield across 173 counties, showing strong spatiotemporal prediction. • **Dataset:** 2.29 million daily records (**1980–2013**) of temperature, precipitation, sunshine over 173 counties in China's two major soybean regions. • **Features:** Averaged meteorological factors over 6 soybean growth stages, plus lagged meteorological yields (past 5 years) to capture trends.
3.	**Combining Multi-Source Data and Machine Learning Approaches to Predict Winter Wheat Yield in the Conterminous United States** [3]	Remote Sensing, (SCIE) *MDPI, 2020*	Yumiao Whang, Zhou Zhang, Luwei Feng, Qingyun Du, Troy Runge	• **Methods:** Used AdaBoost, Random Forest, SVM, DNN, LASSO, OLS to predict winter wheat yield at county level; AdaBoost performed best. Used correlation filtering & GridSearchCV for hyperparameter tuning. • **Results:** AdaBoost achieved R^2=0.86, RMSE=0.51 t/ha, accurate even 2.5 months before harvest. • **Dataset:** Combined multi-source data for CONUS: MODIS VIs, PRISM climate data, soil grids, USDA yields over **2008–2018**, ~300+ features (VIs, climate, soil at 7 depths, 2-year yield history). • **Features:** NDWI & EVI (VIs), LST_D, Tmean, VPDmx, PPT (climate), SOC & CC (soil), historical yields. Used PCA & correlation to select features; showed soil data was most important single source.

Research Gaps

- **Data gaps:** Hard to get clean, combined weather, soil, and yield data; remote sensing is noisy, yield monitors inconsistent.
- **Coverage issues:** Many datasets lack fine spatial detail or timely updates, hurting early-season or real-time predictions.
- **Model limits:** ML often overfits, struggles on new data, or is hard to interpret—especially deep learning.
- **Technique gaps:** Few studies fully explore ensemble setups, advanced deep models, or explainability like SHAP.
- **Real-world use:** Most models aren't built for on-farm use, lack IoT links, or decision tools that farmers can actually use.

This work aims to address these gaps by developing a structured pipeline with robust tuning and detailed feature impact analysis.

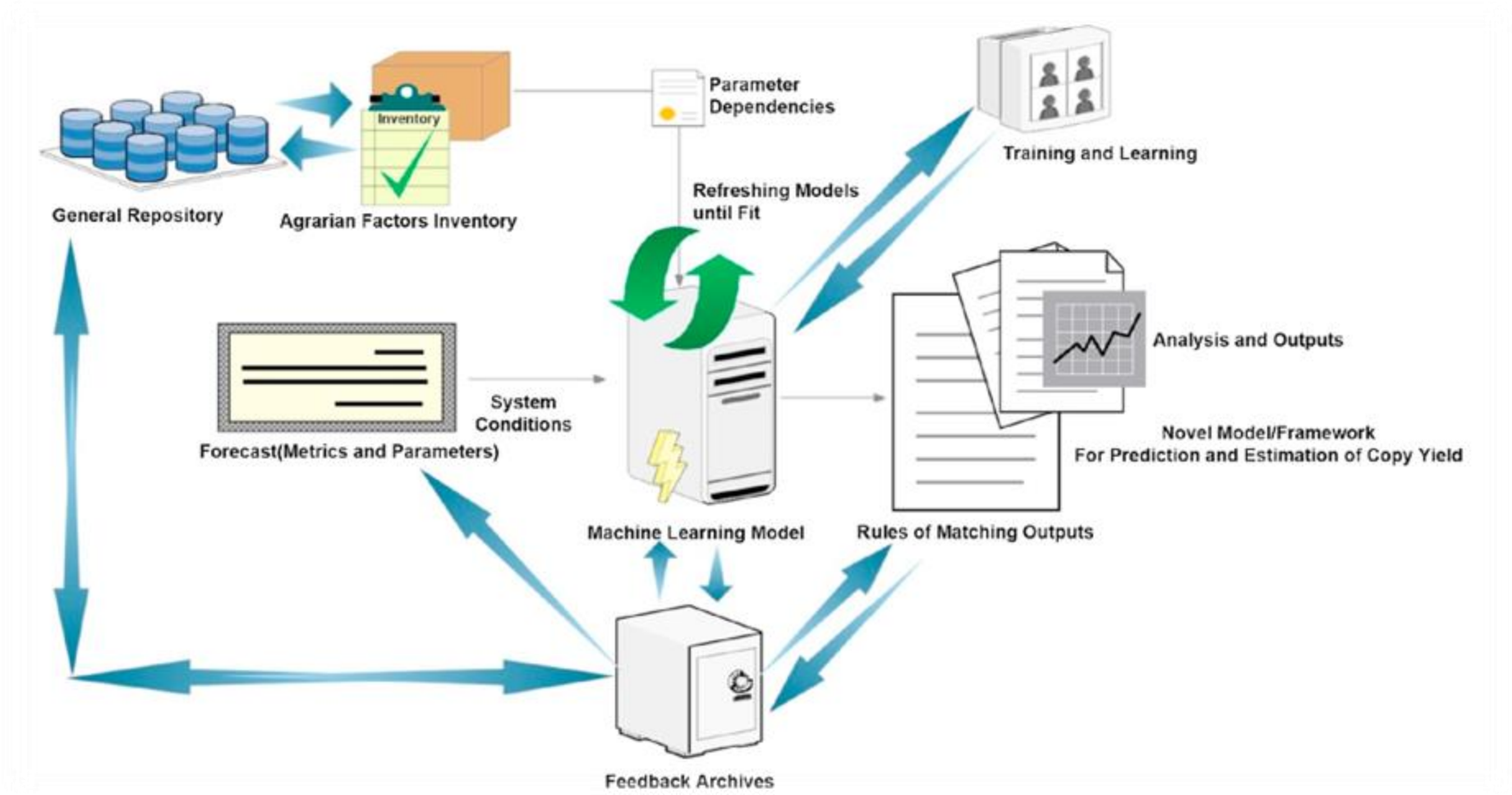
Problem Landscape

Accurate crop yield prediction is vital for food security, economic stability, and sustainable farming. Yet it remains challenging due to erratic weather, climate change, and diverse on-field conditions. Traditional statistical methods often fail, especially in data-scarce or highly variable regions. While machine learning offers promise, it demands large, clean datasets and struggles with noise, outliers, and missing values. Integrating weather, soil, and satellite data can improve predictions but adds complexity, making models harder to interpret. Climate variability further undermines generalizability, and many advanced models operate as black boxes, limiting trust and adoption. There is a clear need for prediction systems that are not only accurate, but also interpretable, scalable, and practical for real-world agricultural use.

Research Objectives

- ✓ Develop an effective machine learning framework to predict crop yield using a combination of spectral, environmental, and soil features.
- ✓ Benchmark multiple tree-based models including Decision Tree, Random Forest, and XGBoost to identify the most suitable approach.
- ✓ Systematically perform hyperparameter tuning using techniques like Grid Search to optimize model performance.
- ✓ Analyze the impact of different features on prediction outcomes through feature importance scores.
- ✓ Validate the robustness of models using cross-validation and comparative performance metrics.
- ✓ Generate insights that can support precision agriculture practices and decision-making for stakeholders.

Methodology (Outlook)



Courtesy: Google

An exemplary illustration of the workflow and operational model that can be applied through different machine learning solutions while focusing on the forecasting of crop yield.

Pipeline Overview

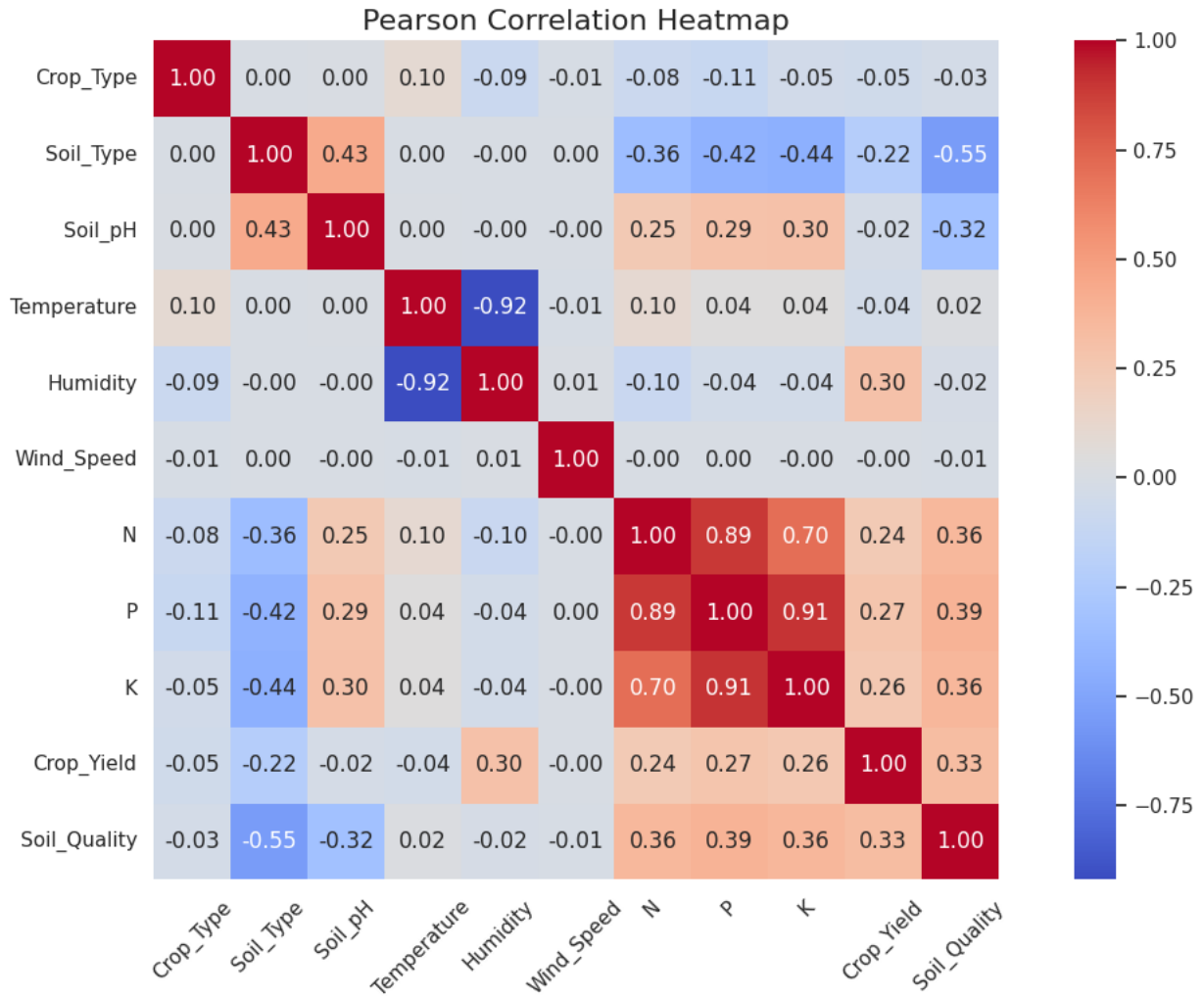
Dataset Description

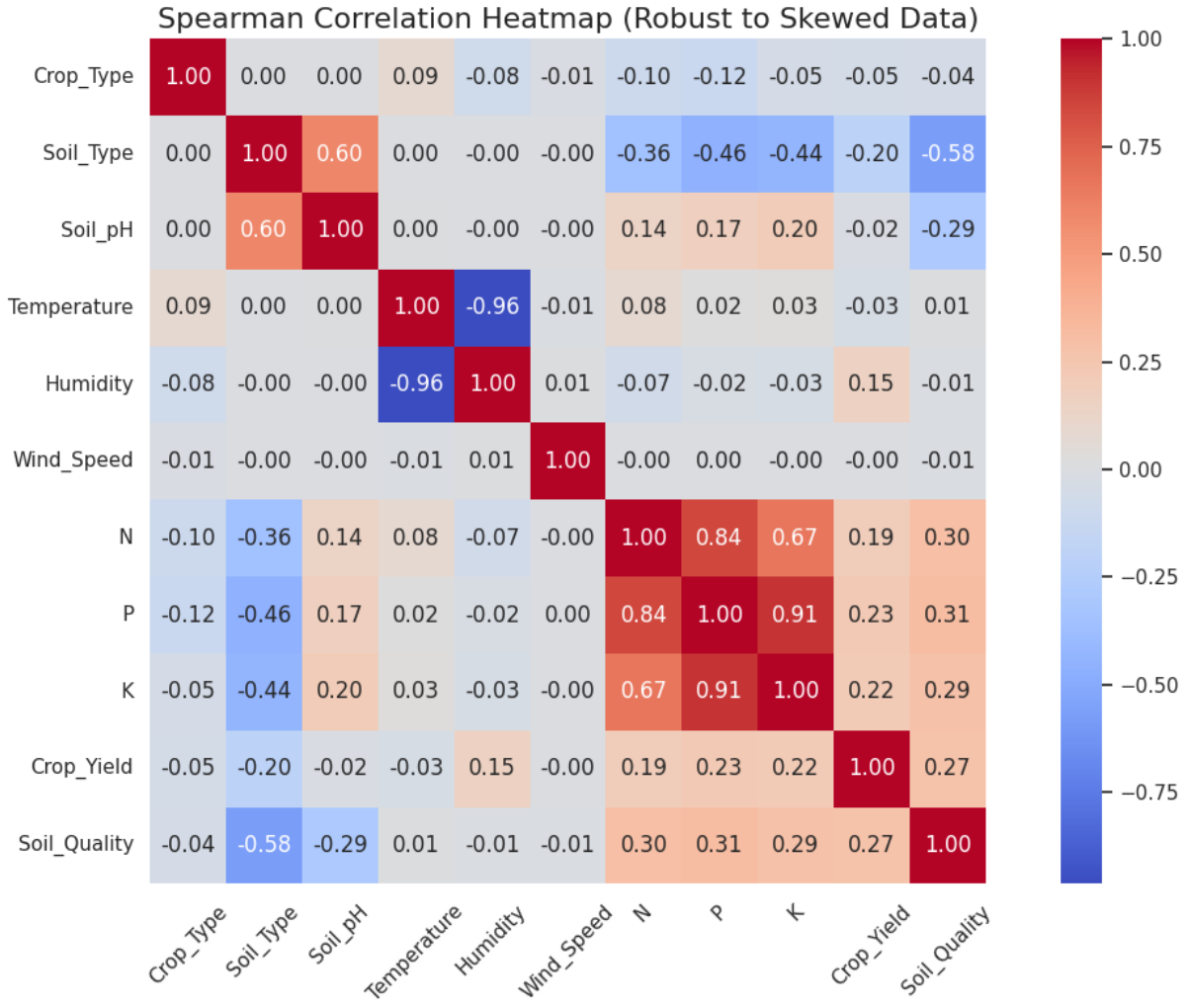
- Dataset provides a daily synthetic record of crop yield and related agro-environmental factors over a 10-year period (2014–2023).
- Contains exactly 36520 samples, offering a solid foundation for machine learning regression tasks.
- Dataset Header:

	Date	Crop_Type	Soil_Type	Soil_pH	Temperature	Humidity	Wind_Speed	N	P	K	Crop_Yield	Soil_Quality
0	2014-01-01	Wheat	Peaty	5.50	9.440599	80.000000	10.956707	60.5	45.0	31.5	0.000000	22.833333
1	2014-01-01	Corn	Loamy	6.50	20.052576	79.947424	8.591577	84.0	66.0	50.0	104.871310	66.666667
2	2014-01-01	Rice	Peaty	5.50	12.143099	80.000000	7.227751	71.5	54.0	38.5	0.000000	27.333333
3	2014-01-01	Barley	Sandy	6.75	19.751848	80.000000	2.682683	50.0	40.0	30.0	58.939796	35.000000
4	2014-01-01	Soybean	Peaty	5.50	16.110395	80.000000	7.696070	49.5	45.0	38.5	32.970413	22.166667
5	2014-01-01	Cotton	Sandy	6.75	14.826739	80.000000	10.366657	55.0	44.0	36.0	29.356115	39.375000

*This dataset offers a **robust, multi-dimensional synthetic simulation of crop yield determinants**, enabling a wide range of agronomic, climatic, and ML research. While it does not replicate all real-world complexities, it is **richly structured and informed by domain logic**, making it an excellent tool for experimentation and discovery.*

EDA and Preprocessing: Correlation Analysis





EDA and Preprocessing: Feature Engineering

Introduced new derived features based on domain knowledge to capture complex relationships.

1. NPK_Ratio:

- Ratio of nitrogen to combined phosphorus and potassium, indicating nutrient balance.
- Helps assess fertilizer efficiency.

$$\text{NPK\_Ratio} = \frac{N}{P + K + \varepsilon}$$

2. Soil_Nutrient_Score:

- Simple average of N, P, and K.
- Provides a consolidated measure of soil fertility.

$$\text{Soil\_Nutrient\_Score} = \frac{N + P + K}{3}$$

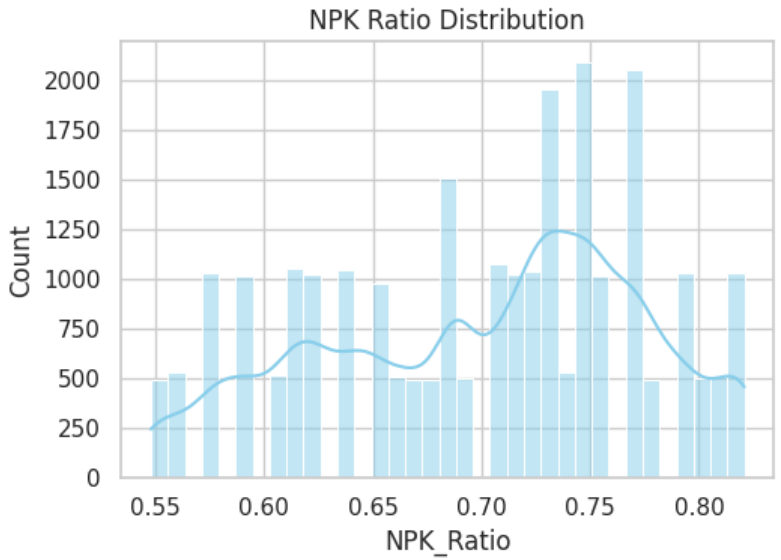
3. Temp_Humidity_Index:

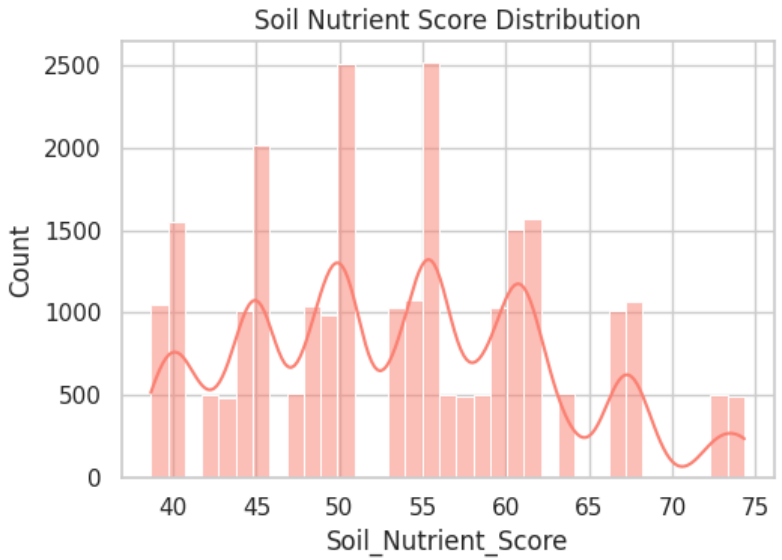
- Product of temperature and humidity.
- Represents combined climatic pressure affecting crop growth.

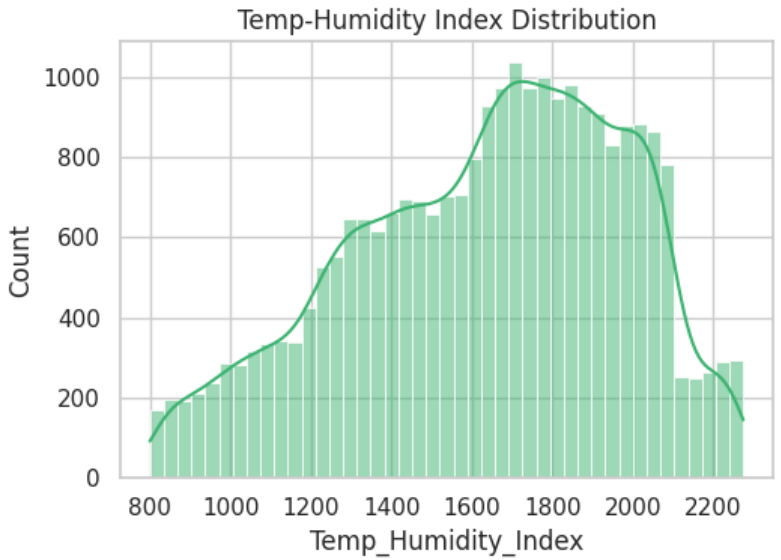
$$\text{Temp\_Humidity\_Index} = \text{Temperature} \times \text{Humidity}$$

These engineered features were included in the final dataset to enhance model learning and potentially uncover hidden interactions.

EDA and Preprocessing: Feature Engineering







Feature	Distribution Insights	Potential Modeling Impact
NPK_Ratio	Fairly uniform with peaks between 0.7 and 0.8.	Indicates a common nutrient balance; good input for models since it's not heavily skewed.
Soil_Nutrient_Score	Appears **multi-modal** due to clustering of N, P, K values (e.g., fertilizer doses).	Suggests multiple fertility zones; may help tree-based models split on those patterns.
Temp_Humidity_Index	Right-skewed with a strong peak around 1800–2000.	Reflects saturation of Humidity at 80; this index compresses variability in climate stress, possibly helping capture subtle effects on yield.

EDA and Preprocessing: Feature Selection

- Recursive Feature Elimination (RFE) was used to rank features by their influence on crop yield prediction.

- **Top predictors identified:**
  - Temperature emerged as the most significant factor.
  - Followed by Soil_Quality, Temp_Humidity_Index, and Crop_Type.
  - Humidity and NPK_Ratio rounded out the top six.

- **Lower-ranked features:**
  - Soil_pH, K, and Soil_Type were among the least influential, suggesting minimal direct impact in this dataset.

- Based on this insight, the dataset was filtered to retain only the top six features along with the target variable for final modeling.
- This targeted feature selection improved model efficiency and maintained predictive power, focusing learning on the most relevant attributes.

Models Implemented

Four tree-based regression models were implemented and systematically benchmarked:

1. **Decision Tree Regressor:**
   - Served as the baseline model to establish initial performance levels.
2. **Random Forest Regressor:**
   - An ensemble technique combining multiple decision trees to reduce variance and improve generalization.
   - Hyperparameters tuned included n_estimators, max_depth, and min_samples_split.
3. **XGBoost Regressor:**
   - A gradient boosting framework that builds trees sequentially to correct previous errors.
   - Hyperparameters tuned included n_estimators, max_depth, learning_rate, subsample, and colsample_bytree.
4. **CatBoost Regressor:**
   - Another boosting algorithm particularly effective with categorical features.
   - Automatically handles categorical data and prevents overfitting through built-in regularization.

All models were trained on the same train-test splits and evaluated using consistent metrics to ensure fair comparison.

Evaluation Metrics

Three standard regression metrics were used to evaluate model performance:

Root Mean Squared Error (RMSE)

- Measures average magnitude of prediction errors.
- Penalizes larger errors more due to squaring.
- Lower values indicate better fit.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Error (MAE)

- Calculates average absolute difference between predicted and actual yields.
- Less sensitive to outliers compared to RMSE.

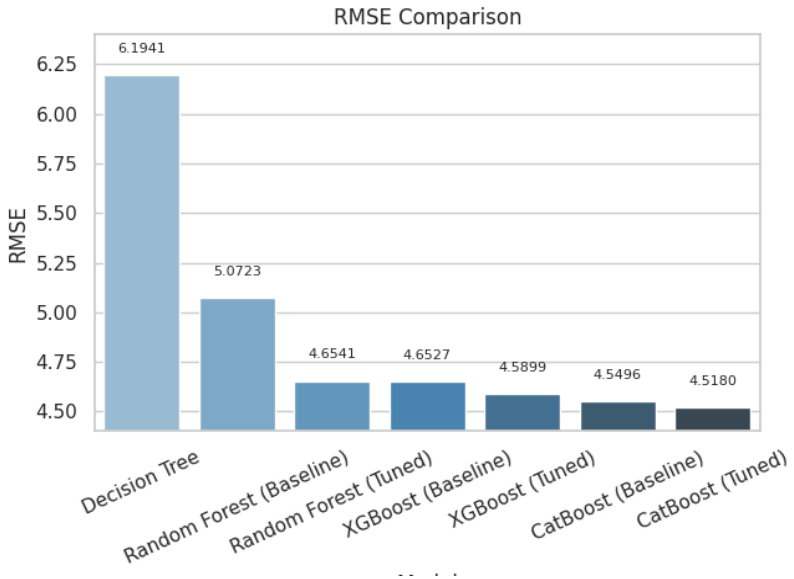
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

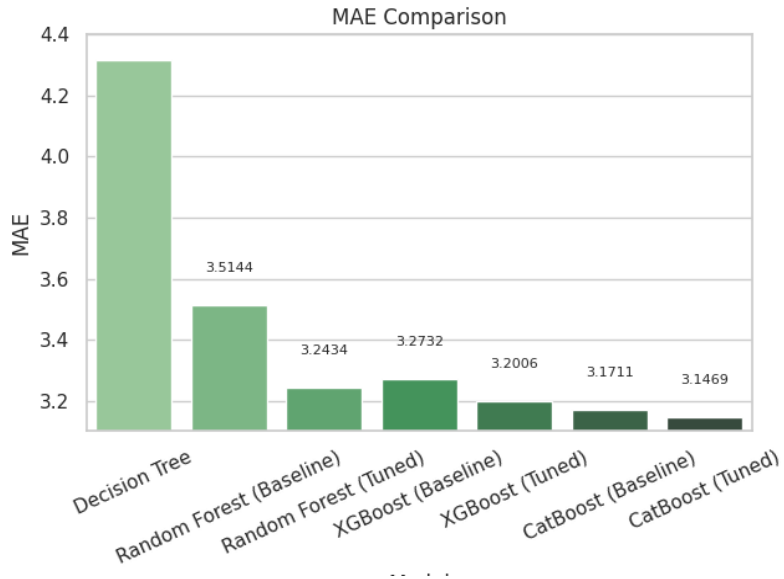
Coefficient of Determination (R^2)

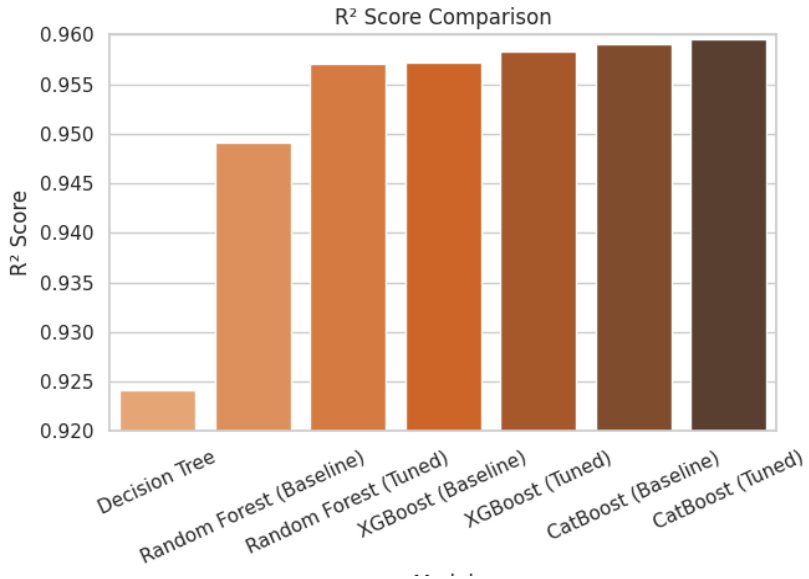
- Indicates proportion of variance in crop yield explained by the model.
- Values closer to 1 imply stronger explanatory power.

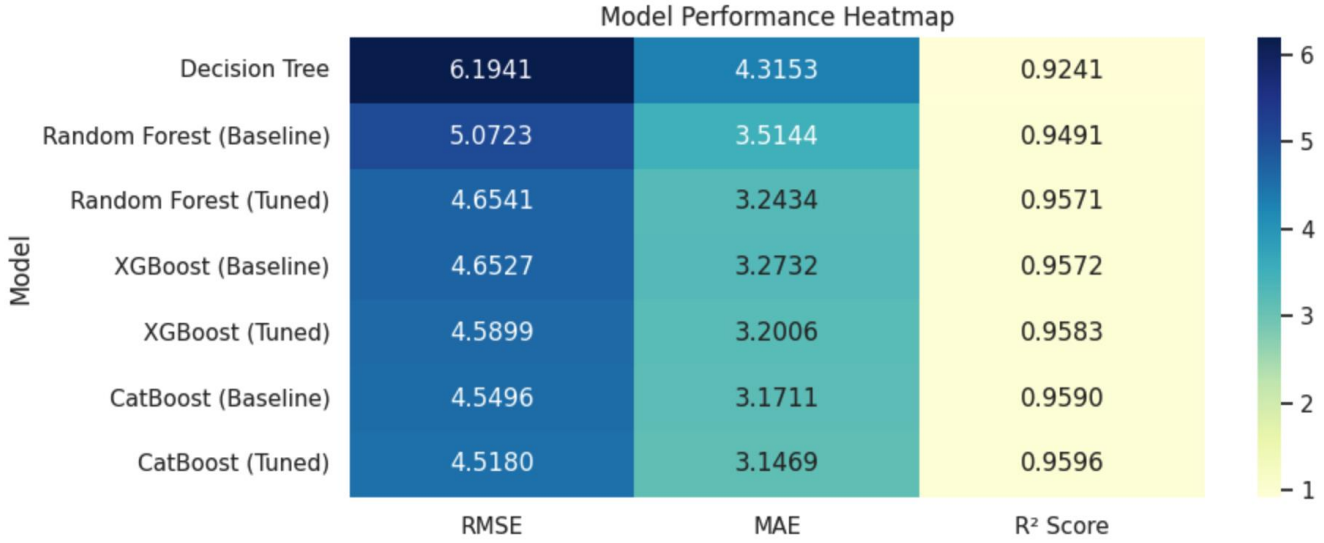
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Results: Comparative Analysis









Results: Impact of Hyperparameter Tuning

Random Forest Performance Before and After Tuning

Metric	Before Tuning	After Tuning	% Improvement
RMSE	5.0723	4.6541	↓ 8.24%
MAE	3.5144	3.2434	↓ 7.71%
R^2 Score	0.9491	0.9571	↑ 0.84%

- Hyperparameter tuning reduced RMSE and MAE while slightly boosting R^2, leading to better generalization and making the tuned Random Forest a strong candidate for deployment.

XGBoost Performance Before and After Tuning

Metric	Before Tuning	After Tuning	% Improvement
RMSE	4.6527	4.5899	↓ 1.35%
MAE	3.2732	3.2006	↓ 2.22%
R^2 Score	0.9572	0.9583	↑ 0.11%

- Tuning XGBoost lowered RMSE and MAE with a modest rise in R^2, enhancing generalization and stability, and confirming its suitability for deployment alongside Random Forest.

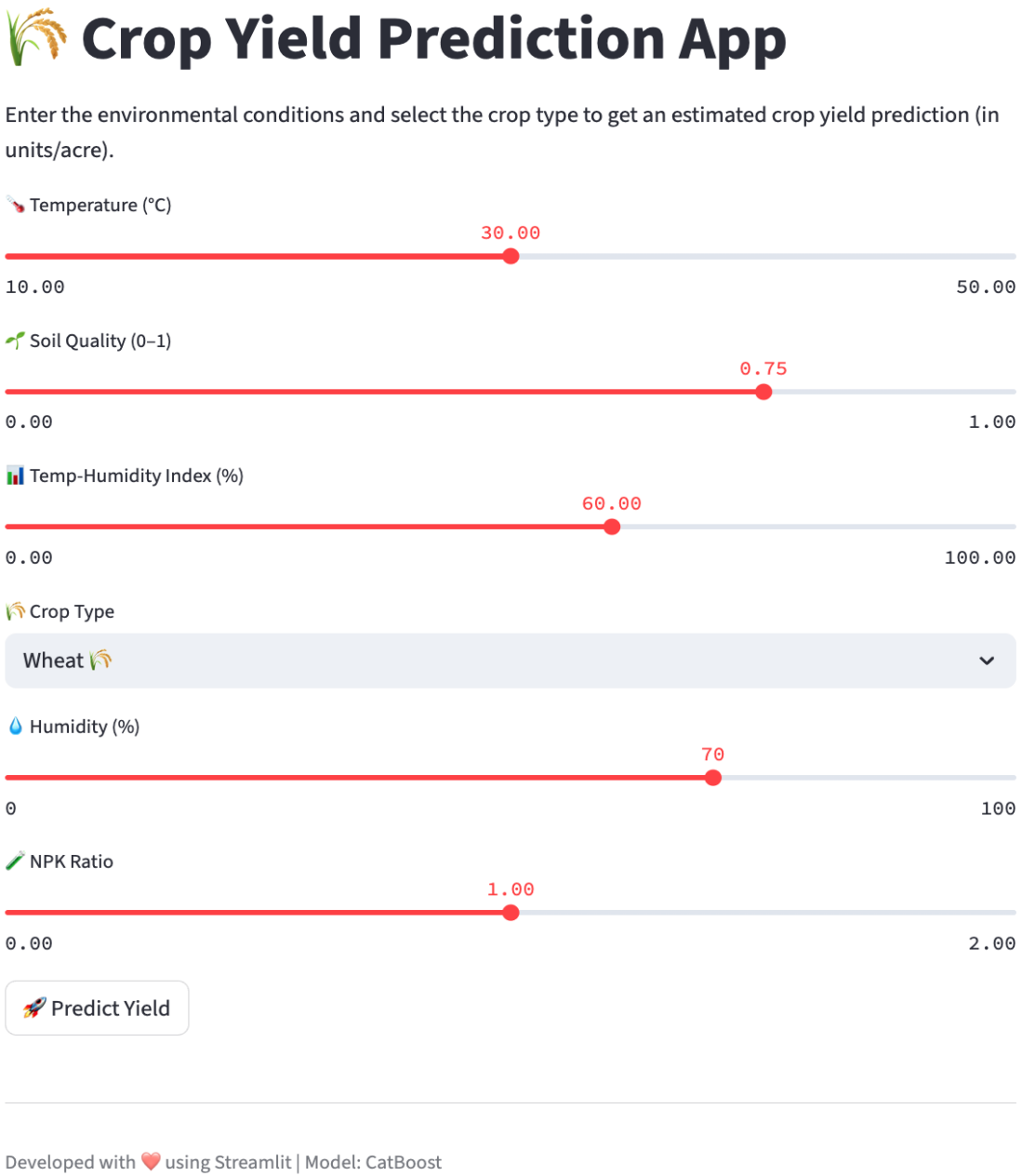
CatBoost Performance Before and After Tuning

Metric	Before Tuning	After Tuning	% Improvement
RMSE	4.5496	4.5180	↓ 0.70%
MAE	3.1711	3.1469	↓ 0.76%
R^2 Score	0.9590	0.9596	↑ 0.06%

- Even with solid baseline performance, fine-tuning CatBoost yielded small but consistent gains across all metrics, showing that parameter optimization further strengthens its deployment readiness.

Deployment

- To extend the practical usability of the machine learning model beyond a research setting, the final CatBoost Regressor was deployed as an interactive web application. This deployment serves to demonstrate how advanced predictive models can be translated into accessible tools for real-world use in agricultural decision-making.



Conclusion

- Developed an extensive crop yield prediction pipeline using a synthetic dataset designed to represent agronomic and climatic patterns over a decade.
- Applied robust data preprocessing and engineered features to handle domain-specific challenges such as multicollinearity and skewness.
- Leveraged non-linear machine learning models, with CatBoost achieving the best performance: RMSE: 4.5180, MAE: 3.1469, R^2: 0.9596
- Demonstrated the model's ability to capture complex feature interactions and maintain stability across a wide input range.
- Transitioned the trained CatBoost model into an interactive Streamlit web app, enabling users to input environmental data and receive instant yield predictions.
- This practical tool bridges the gap from research to real-world agricultural decision-making.

Limitations and Future Work

- **Limitations:**
  - Restricted to the available dataset, which may not cover extreme agro-climatic variations across India.
  - Temporal forecasting (e.g., multi-season or dynamic time-series models like LSTM) not included in this study.
  - Economic or pest-related sudden shocks are outside the predictive scope of these models.
- **Future Work:**
  - Incorporate high-resolution satellite and IoT sensor data for localized predictions.
  - Explore time-series and sequence models (e.g., LSTM, GRU) for multi-season forecasting.
  - Integrate economic, pest, and disease dynamics to handle unexpected yield shocks.
  - Apply interpretability methods (like SHAP) to better understand feature impacts.
  - Develop lightweight, farmer-friendly decision support tools for real-world deployment.

References

1. Hoque, M. J., Islam, M. S., Uddin, J., Samad, M. A., De Abajo, B. S., Vargas, D. L. R., & Ashraf, I. (2024). Incorporating Meteorological Data and Pesticide Information to Forecast Crop Yields Using Machine Learning. *IEEE Access*.

2. Li, Q. C., Xu, S. W., Zhuang, J. Y., Liu, J. J., Yi, Z. H. O. U., & Zhang, Z. X. (2023). Ensemble learning prediction of soybean yields in China based on meteorological data. *Journal of Integrative Agriculture*, *22*(6), 1909-1927.

3. Wang, Y., Zhang, Z., Feng, L., Du, Q., & Runge, T. (2020). Combining multi-source data and machine learning approaches to predict winter wheat yield in the Conterminous United States. *Remote Sensing, 12*(8), 1232.

4. Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and electronics in agriculture, 177, 105709.

5. Sharma, P., Dadheech, P., Aneja, N., & Aneja, S. (2023). Predicting agriculture yields based on machine learning using regression and deep learning. *IEEE Access*.

6. Ashfaq, M., Khan, I., Alzahrani, A., Tariq, M.U., Khan, H., & Ghani, A. (2024). Accurate Wheat Yield Prediction Using Machine Learning and Climate-NDVI Data Fusion. *IEEE Access*, *12*, 40947-40961.

Publications

- Paper 1 (Accepted & Presented):
  - "**Spectral and Environmental Feature Synergies in Crop Yield Prediction: A Model-Wise ML Review**"
  - *Conference: "16th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2025"*
- Paper 2 (Communicated):
  - "**Benchmarking Machine Learning Approaches for Agro-Environmental Yield Forecasting**"
  - *Conference: "2nd International Conference on Signal Processing and Computer Vision (SIPCOV), 2025"*
- Paper 3 (Communicated):
  - **"Feature-Driven Crop Yield Prediction Using Hyperparameter Optimized Ensemble Learners"**
  - *Conference: "1st International Conference on Recent Trends in Computing and Smart Mobility Conference (RCMS), 2025"*

Thank You