Methodology

In this project, we focused on **predicting agricultural yield** using weather and reservoir features, following a structured data science and machine learning workflow.

Below, we describe in detail the process, tools, and methods used from data preparation to model evaluation.

Data Collection & Understanding

- The primary dataset was provided by our mentor, containing **crop yield data for multiple crops**, daily weather variables (temperature, rainfall), and reservoir data (water level, FRL, live storage) across multiple states in India for several years.
- The data covered multiple crops (e.g., rice, wheat, gram, mustard, etc.) and included thousands of records per crop and state.

Data Preparation & Cleaning

- Cleaned column names and standardized formats.
- Converted date columns to datetime objects and extracted useful temporal features like **year** and **week number**.
- Removed rows with missing critical values to ensure model reliability.
- Merged all crop Datasets and then created separate DataFrames for states.

Feature Engineering

- Created additional features to enrich model inputs:
    - **avg_temp**: mean of yearly max and min temperatures.
    - **Target encoding of crop_name**: replacing categorical crop names with their mean yield in training data.
    - **Interaction features**: experimented with manually multiplying reservoir level and categorical features to capture relationships.
- Explored correlations to identify important features.

Model Selection & Rationale

We focused on classical ML models that are:

- Easy to interpret,
- Robust to nonlinear relationships,
- Require minimal data preprocessing.

Models used:

Linear Regression:

- Used as a baseline model to capture simple linear dependencies between yield and features.

Random Forest Regressor:

- Selected for its ability to handle non-linearities and feature interactions.
- Naturally robust to multicollinearity and outliers.

We trained models separately for each state to account for regional differences.

Data Splitting & Validation Strategy

- Sorted data chronologically by year to avoid data leakage.
- Split each state dataset into:
  - **Training data:** first 80% of years.
  - **Testing data:** remaining 20% of years.
- Avoided random shuffling to maintain temporal order, simulating real-world prediction where future data is unseen.

For additional realism:

- Built separate regressors (e.g., LinearRegression) to predict test set features (avg_temp, rainfall, etc.) **only using year as input**.
- Used these predicted features instead of true test features when predicting yield — mimicking deployment where future weather data isn't known.

Model Training & Evaluation

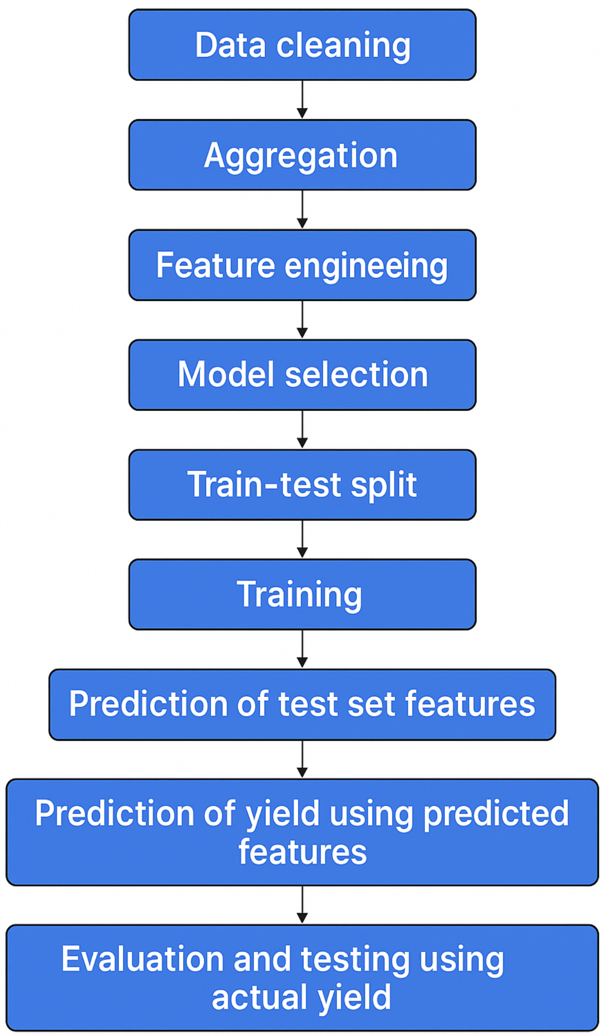
- For Random Forest: tuned number of trees, fixed random_state for reproducibility.
- Experimented with different combinations of feature set to get the best performing features.
- Evaluated models on both training and test data using:
  - RMSE (Root Mean Squared Error)
  - R^2 score (Coefficient of Determination)

This helped assess not only how well models fit the training data but also their generalization capability.

Tools & Libraries

- Python (Jupyter / Colab)
- pandas & numpy (data manipulation)
- sklearn (models and metrics)
- matplotlib & seaborn (visualizations)

Summary Flow-Chart

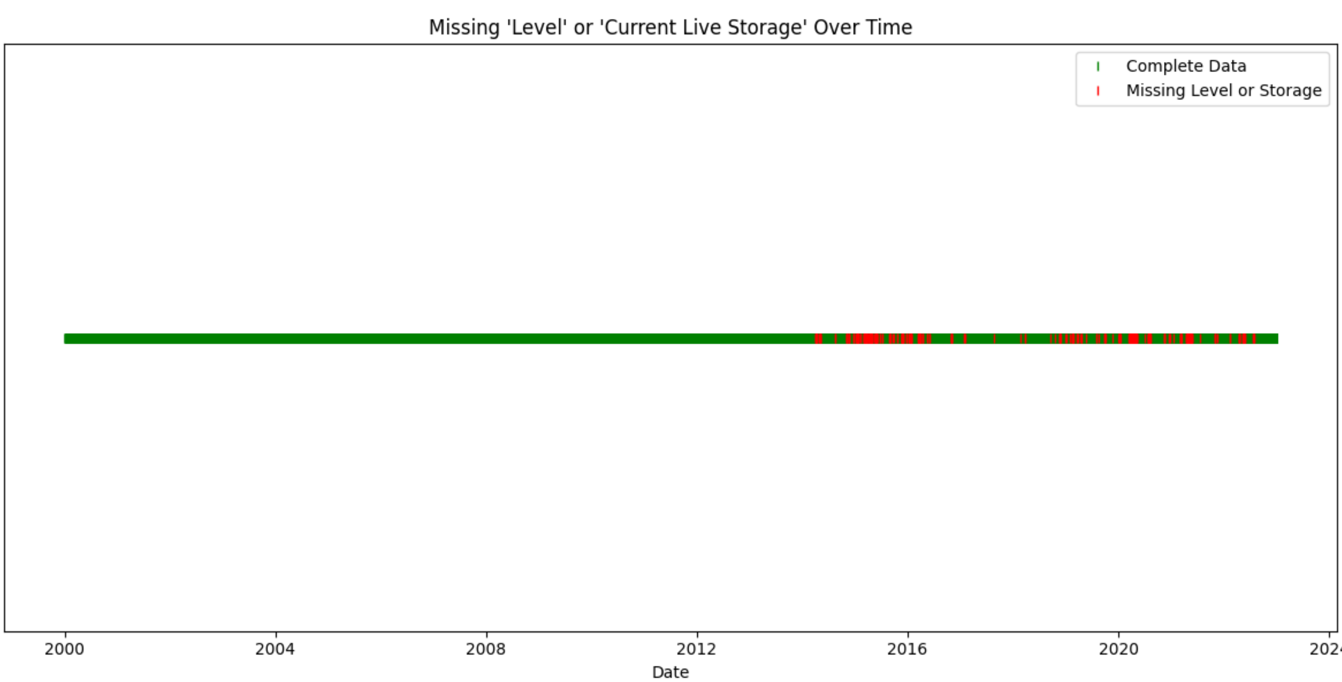


Data Analysis and Results

Missing values and Duplicates(Potato Dataset):

During data analysis, we checked for missing values across all columns. Most features had complete data except for Level and Current Live Storage, which had 382 missing entries each. This is likely due to occasional gaps in reservoir data reporting. Since these features are important for downstream modeling, appropriate imputation strategies (such as using mean, median, or model-based imputation) can be considered in future steps to handle these missing values effectively. No duplicate rows were found, ensuring data consistency.

```
Missing values per column:
 state_name                     0
crop_name                      0
apy_item_interval_start        0
temperature_recorded_date      0
state_temperature_max_val      0
state_temperature_min_val      0
state_rainfall_val             0
yield                          0
FRL                            0
Live Cap FRL                   0
Level                        382
Current Live Storage         382
dtype: int64
Number of duplicate rows: 0
```



Summary Statistics:

The dataset contains 47,848 records covering agricultural and climatic data from 2000 to 2022. Temperature values vary widely, with maximum temperatures ranging from about 11 °C to nearly 48 °C and minimum temperatures from -0.5 °C to around 29 °C, reflecting seasonal and regional diversity. Rainfall shows high skewness, with a median close to zero and a maximum value near 187 mm, indicating many dry days interspersed with occasional heavy rainfall. Yield values range from about 4 to 36, with a mean of roughly 16, suggesting moderate variability across crops and years. Reservoir features like FRL and Level also exhibit broad ranges, highlighting differences in storage capacity and water availability across states and years. Overall, the statistics reveal significant variability, which is essential context for model building and analysis.

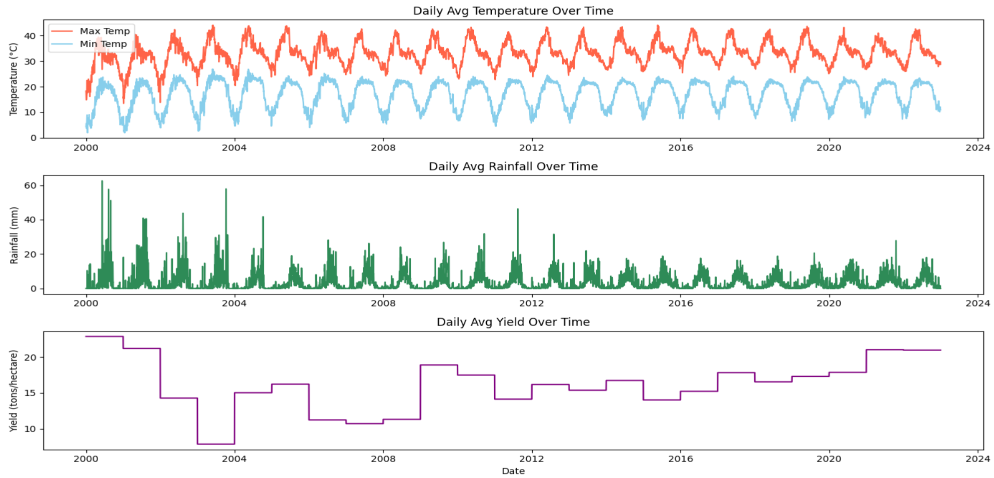
Summary statistics of the dataset:

	apy_item_interval_start	state_temperature_max_val \
count	47848.000000	47848.000000
mean	2014.045645	33.566005
std	5.432290	5.031233
min	2000.000000	11.110000
25%	2010.000000	31.070000
50%	2015.000000	33.420000
75%	2019.000000	36.560000
max	2022.000000	47.510000

	state_temperature_min_val	state_rainfall_val	yield \
count	47848.000000	47848.000000	47848.000000
mean	17.305670	3.262658	16.346552
std	5.572399	6.542797	8.407129
min	-0.500000	0.000000	3.975090
25%	13.420000	0.000000	8.409210
50%	18.490000	0.310000	15.056890
75%	21.790000	3.770000	22.558860
max	28.920000	186.990000	35.916140

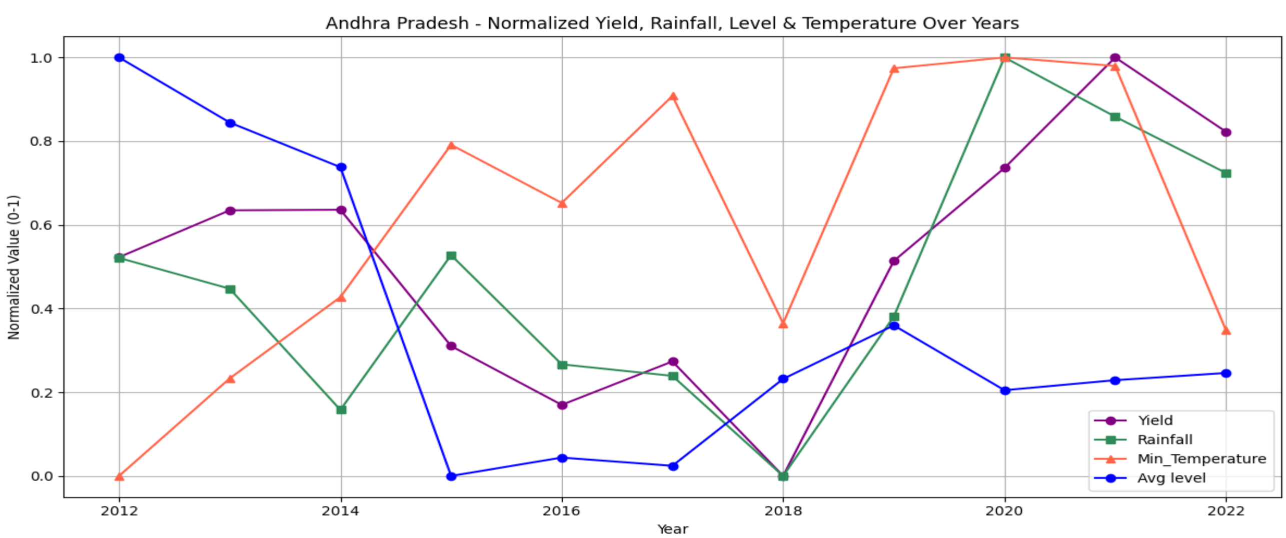
	FRL	Live Cap FRL	Level	Current Live Storage
count	47848.000000	47848.000000	47466.000000	47466.000000
mean	341.616405	1.515693	330.989787	0.752887
std	148.469866	0.867089	148.225816	0.628405
min	135.136667	0.198465	64.960000	0.000000
25%	183.210000	0.523000	177.213125	0.244333
50%	348.926000	1.539062	338.858500	0.576688
75%	461.933333	1.866600	421.389800	1.099721
max	597.731875	2.898000	824.960000	4.327000

Trends:



The above plots illustrate key temporal trends in the dataset from 2000 to 2022.

- The top plot shows clear seasonal cycles in daily average maximum and minimum temperatures, highlighting regular annual fluctuations with higher max temperatures peaking around mid-year and lower min temperatures in cooler months.
- The middle plot displays daily average rainfall, which is highly variable with sharp spikes, capturing the sporadic and monsoon-driven nature of rainfall in India.
- The bottom plot presents daily average yield, aggregated across crops, which shows relatively smoother step-like trends over time, reflecting long-term changes in agricultural productivity. Together, these plots help in understanding the seasonality, variability, and potential factors influencing crop yield over the years.



The above chart shows the normalized trends (scaled between 0 and 1) of yield, rainfall, minimum temperature, and average reservoir level in Andhra Pradesh from 2012 to 2022. We observe fluctuations in yield roughly aligning with changes in rainfall and temperature, while average reservoir levels show a declining trend initially and stabilize later. This visualization helps explore potential relationships and seasonal impacts of climatic and water storage variables on agricultural yield over the years.

Comparative analysis of models

Comparing performance of Linear Regression and Random Forest models tested using actual random test features.

State	LR Train RMSE	LR Train R²	LR Test RMSE	LR Test R²	RF Train RMSE	RF Train R²	RF Test RMSE	RF Test R²
Uttarakhand	1.583138	0.887347	1.646568	0.883942	0.249097	0.997211	0.687708	0.979755
Uttar Pradesh	1.742486	0.963755	1.818623	0.962261	0.401013	0.998080	1.019810	0.988133
Chhattisgarh	0.460810	0.933399	0.466045	0.931195	0.083273	0.997825	0.224716	0.984003
West Bengal	1.771452	0.967269	1.775435	0.967109	0.419865	0.998161	1.078628	0.987860
Andhra Pradesh	1.672503	0.874421	1.698052	0.871731	0.271357	0.996694	0.628160	0.982447
Jharkhand	0.409126	0.969718	0.422939	0.968140	0.095489	0.998350	0.277078	0.986326
Karnataka	0.918453	0.972216	0.908613	0.973086	0.267703	0.997640	0.712759	0.983438
Telangana	1.408565	0.891958	1.447935	0.888519	0.252960	0.996515	0.644023	0.977945
Rajasthan	0.289818	0.919311	0.287629	0.921742	0.044570	0.998092	0.117705	0.986894
Madhya Pradesh	0.440958	0.726663	0.438265	0.730158	0.063497	0.994332	0.167924	0.960385
Gujarat	0.303580	0.851027	0.298526	0.857417	0.054341	0.995227	0.142135	0.967678
Maharashtra	0.167746	0.896734	0.167160	0.899790	0.034631	0.995599	0.091609	0.969903
Tamil Nadu	0.989472	0.989035	0.964188	0.989344	0.192415	0.999585	0.510851	0.997009
Odisha	0.060912	0.980900	0.061269	0.981614	0.010577	0.999424	0.026625	0.996528

Comparing performance of Linear Regression and Random Forest models tested using predicted test features predicted by simple linear regression models.

State	LR Train RMSE	LR Train R²	LR Test RMSE	LR Test R²	RF Train RMSE	RF Train R²	RF Test RMSE	RF Test R²
Uttarakhand	3.050504	0.871694	0.712182	0.958732	0.039830	0.998325	0.295978	0.982849
Chhattisgarh	0.121900	0.949905	0.618529	0.891795	0.002989	0.998771	0.677984	0.881394
West Bengal	3.699386	0.957748	0.878195	0.992971	0.162505	0.998144	1.763023	0.985888
Jharkhand	0.184052	0.971026	0.148013	0.928788	0.009374	0.998524	0.202175	0.902730
Karnataka	0.759944	0.971073	1.295533	0.970228	0.046510	0.998230	1.388033	0.968102
Rajasthan	0.056365	0.942937	0.129117	0.890229	0.000783	0.999208	0.060204	0.948816
Madhya Pradesh	0.152771	0.709419	0.547404	0.445695	0.002670	0.994921	0.192569	0.805003
Gujarat	0.052286	0.914362	0.221269	0.490499	0.001168	0.998087	0.200743	0.537764
Maharashtra	0.020882	0.915420	0.036566	0.882064	0.000723	0.997070	0.144693	0.533329
Tamil Nadu	1.101848	0.986548	0.544978	0.995123	0.034081	0.999584	0.598605	0.994644
Odisha	0.003076	0.987176	0.012952	0.301620	0.000092	0.999618	0.010285	0.445409

Actual test features vs Predicted test features of Rajasthan for 2021

	Actual Value	Predicted Value
avg_temp	28.140000	26.087362
state_rainfall_val	0.140000	1.931910
Level	279.644000	284.743750
FRL	287.252000	287.252000
Current Live Storage	0.451574	0.728505
crop_encoded	1.251720	1.251720

Actual yield vs Predicted yield

Crop-wise across states

	crop_name	State	yield	Predicted Yield
0	gram	Chhattisgarh	0.693304	0.955106
1	gram	Gujarat	1.691451	1.099751
2	gram	Jharkhand	1.202375	0.998426
3	gram	Karnataka	0.671894	0.598654
4	gram	Madhya Pradesh	1.614275	1.151620
5	gram	Maharashtra	1.030654	0.765881
6	gram	Odisha	0.702796	0.618480
7	gram	Rajasthan	1.075413	0.907155
8	gram	Tamil Nadu	0.926280	0.824060
9	gram	Uttarakhand	0.787872	0.827735
10	gram	West Bengal	1.245962	1.033128
11	masoor	Chhattisgarh	0.710361	0.342319
12	masoor	Jharkhand	0.859635	0.736243
13	masoor	Madhya Pradesh	1.057188	0.714615
14	masoor	Odisha	0.470337	0.414680
15	masoor	Rajasthan	1.317518	1.043778
16	masoor	Uttarakhand	0.890343	0.719279
17	masoor	West Bengal	0.865374	0.760473
18	potato	Chhattisgarh	6.718750	5.221069
19	potato	Jharkhand	7.305450	6.567597
20	potato	Karnataka	16.360702	14.811667
21	potato	Tamil Nadu	23.386947	22.737142
22	potato	Uttarakhand	12.218719	11.307472
23	potato	West Bengal	29.878287	31.324976
24	rapeseed &mustard	Chhattisgarh	0.510415	0.494861
25	rapeseed &mustard	Gujarat	1.933926	1.518604
26	rapeseed &mustard	Jharkhand	0.811379	0.816629
27	rapeseed &mustard	Madhya Pradesh	1.550358	1.229001
28	rapeseed &mustard	Maharashtra	0.328988	0.346819
29	rapeseed &mustard	Rajasthan	1.628265	1.440817
30	rapeseed &mustard	Tamil Nadu	0.235114	0.241568
31	rapeseed &mustard	Uttarakhand	0.924425	0.815565
32	rapeseed &mustard	West Bengal	1.219520	0.938837
33	rice	Jharkhand	2.515670	1.501949
34	rice	Karnataka	2.540047	2.506189
35	wheat	Chhattisgarh	1.462655	1.398062
36	wheat	Gujarat	3.195756	2.971460
37	wheat	Jharkhand	2.135435	1.482597
38	wheat	Karnataka	1.225168	0.847122
39	wheat	Madhya Pradesh	3.526995	3.032108
40	wheat	Maharashtra	1.767831	1.260650
41	wheat	Rajasthan	3.892101	3.641733
42	wheat	Uttarakhand	3.006383	2.334568
43	wheat	West Bengal	2.964734	2.353336

Crop-wise across years

	crop_name	year	yield	Predicted Yield
0	gram	2004	0.626860	0.618480
1	gram	2018	1.008188	0.824798
2	gram	2019	1.039189	0.900636
3	gram	2020	1.123991	0.920867
4	gram	2021	1.137153	0.926313
5	gram	2022	1.108544	0.927649
6	masoor	2004	0.397120	0.414680
7	masoor	2018	0.819200	0.708152
8	masoor	2019	0.816920	0.677634
9	masoor	2020	0.914734	0.735954
10	masoor	2021	0.910728	0.722406
11	masoor	2022	1.205086	0.715014
12	potato	2018	13.724845	13.361661
13	potato	2019	15.525360	15.503793
14	potato	2020	16.351722	14.723173
15	potato	2021	17.109432	17.059317
16	potato	2022	18.293028	17.433595
17	rapeseed &mustard	2017	0.342920	0.345954
18	rapeseed &mustard	2018	1.076504	0.865820
19	rapeseed &mustard	2019	1.096448	0.960139
20	rapeseed &mustard	2020	1.157538	0.950945
21	rapeseed &mustard	2021	1.100306	0.952986
22	rapeseed &mustard	2022	1.100691	0.945838
23	rice	2018	2.476320	2.524485
24	rice	2019	2.755020	2.517498
25	rice	2020	2.245620	2.502406
26	rice	2021	2.558790	2.493603
27	rice	2022	2.589446	1.998462
28	wheat	2018	2.157894	1.784618
29	wheat	2019	2.514777	2.165283
30	wheat	2020	2.680861	2.172770
31	wheat	2021	2.601702	2.158983
32	wheat	2022	2.624035	2.148369

Conclusion

In this project, we explored the use of machine learning models—specifically **Random Forest** and **Linear Regression**—to predict crop yield based on features like average temperature, rainfall, reservoir level, and crop-specific encoding.

The dataset covered multiple Indian states over several years, enabling state-level models.

From our analysis:

- **Random Forest and Linear Regression both showed strengths in different states**:
    - For states like *Rajasthan*, *Odisha*, and *Tamil Nadu*, Random Forest achieved lower test RMSE and higher R^2, indicating better predictive performance.
    - In contrast, Linear Regression performed competitively and sometimes better in states like *Gujarat*, *West Bengal*, and *Jharkhand*, where simpler linear relationships between features and yield were more prominent.
- Overall, both models achieved reasonably high R^2 scores on training data (often >0.90), but test performance varied by state, likely due to differences in data distribution, crop diversity, and climatic variability.
- For some larger and more variable states (e.g., Uttar Pradesh, Andhra Pradesh, Telangana), the models struggled, likely due to higher data variability and complex crop-climate interactions.

We also observed:

- **Predicted test features** (using time-based regressions) were reasonably close to actual features, as seen in Rajasthan's comparison table.
- Target encoding of the crop name feature improved model performance by capturing crop-specific yield tendencies.
- The Random Forest model provided feature importance insights, helping identify reservoir variables like level, Current Live Storage had the most impact on yield predictions

Appendices

github.com/TheGiftedExplorer/yield-prediction