WEEKLY PROGRESS REPORT

Name: Sahil Sanjay Pokharkar

Domain: Data Science and Machine Learning

Date of Submission: 28/01/2026

Project Title: Prediction of Crop Production in India

Organization: Upskill Campus

Week Ending: Week 03

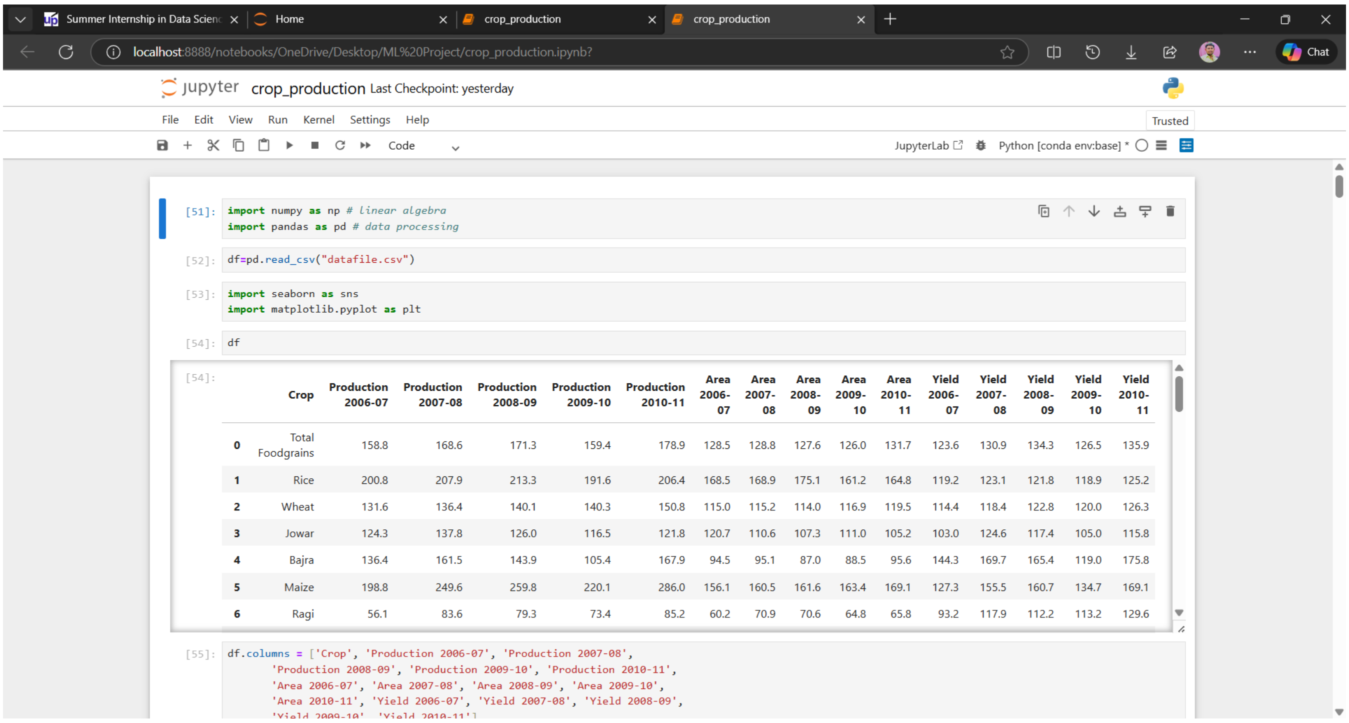
ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Upskill Campus and my mentors for their continuous guidance and support throughout the third week of my internship. Their valuable suggestions helped me understand machine learning concepts and apply them effectively to real-world data. I am also thankful to my faculty members, family, and friends for their encouragement and motivation during this learning phase.

Table Of Content

Chapter 1: Overview

The third week of the internship focused on **feature engineering, data encoding, and implementation of initial machine learning models** for predicting crop production. The objective was to transform the pre-processed dataset into a machine-learning-ready format and build baseline models to evaluate prediction performance.



Summer Internship in Data Scienc
Home
crop_production
crop_production
localhost:8888/notebooks/OneDrive/Desktop/ML%20Project/crop_production.ipynb?
Chat
jupyter
crop_production Last Checkpoint: yesterday
File
Edit
View
Run
Kernel
Settings
Help
Trusted
Code
JupyterLab
Python [conda env:base] *

[51]:
```
import numpy as np # linear algebra
import pandas as pd # data processing
```

[52]:
```
df=pd.read_csv("datafile.csv")
```

[53]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

[54]:
```
df
```

[54]:

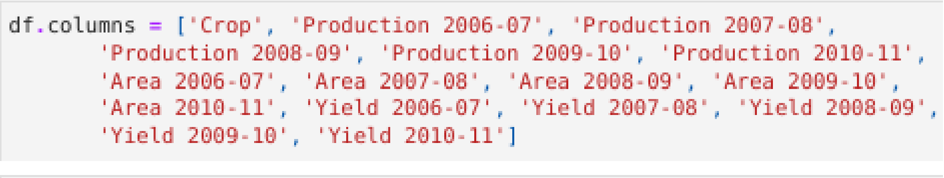
	Crop	Production 2006-07	Production 2007-08	Production 2008-09	Production 2009-10	Production 2010-11	Area 2006-07	Area 2007-08	Area 2008-09	Area 2009-10	Area 2010-11	Yield 2006-07	Yield 2007-08	Yield 2008-09	Yield 2009-10	Yield 2010-11
0	Total Foodgrains	158.8	168.6	171.3	159.4	178.9	128.5	128.8	127.6	126.0	131.7	123.6	130.9	134.3	126.5	135.9
1	Rice	200.8	207.9	213.3	191.6	206.4	168.5	168.9	175.1	161.2	164.8	119.2	123.1	121.8	118.9	125.2
2	Wheat	131.6	136.4	140.1	140.3	150.8	115.0	115.2	114.0	116.9	119.5	114.4	118.4	122.8	120.0	126.3
3	Jowar	124.3	137.8	126.0	116.5	121.8	120.7	110.6	107.3	111.0	105.2	103.0	124.6	117.4	105.0	115.8
4	Bajra	136.4	161.5	143.9	105.4	167.9	94.5	95.1	87.0	88.5	95.6	144.3	169.7	165.4	119.0	175.8
5	Maize	198.8	249.6	259.8	220.1	286.0	156.1	160.5	161.6	163.4	169.1	127.3	155.5	160.7	134.7	169.1
6	Ragi	56.1	83.6	79.3	73.4	85.2	60.2	70.9	70.6	64.8	65.8	93.2	117.9	112.2	113.2	129.6

[55]:
```
df.columns = ['Crop', 'Production 2006-07', 'Production 2007-08',
       'Production 2008-09', 'Production 2009-10', 'Production 2010-11',
       'Area 2006-07', 'Area 2007-08', 'Area 2008-09', 'Area 2009-10',
       'Area 2010-11', 'Yield 2006-07', 'Yield 2007-08', 'Yield 2008-09',
       'Yield 2009-10', 'Yield 2010-11']
```

Chapter 2: Tasks Completed & Achievements

1. Feature Engineering and Encoding

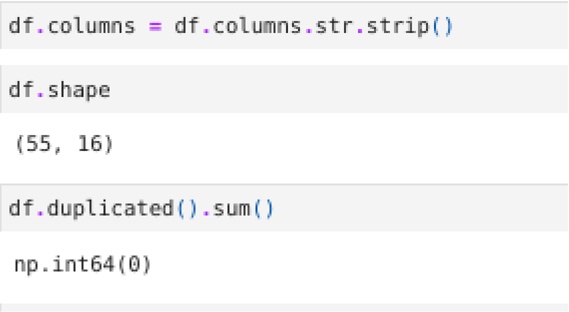
- Performed encoding of categorical variables such as State, District, Crop, and Season using suitable encoding techniques.
- Normalized and scaled numerical features like Area to improve model performance.
- Finalized the feature set required for training machine learning models.



```
df.columns = ['Crop', 'Production 2006-07', 'Production 2007-08',
       'Production 2008-09', 'Production 2009-10', 'Production 2010-11',
       'Area 2006-07', 'Area 2007-08', 'Area 2008-09', 'Area 2009-10',
       'Area 2010-11', 'Yield 2006-07', 'Yield 2007-08', 'Yield 2008-09',
       'Yield 2009-10', 'Yield 2010-11']
```

2. Dataset Splitting

- Split the dataset into training and testing sets to evaluate model generalization.
- Ensured appropriate data distribution between training and testing data.



```
df.columns = df.columns.str.strip()
```

```
df.shape
```

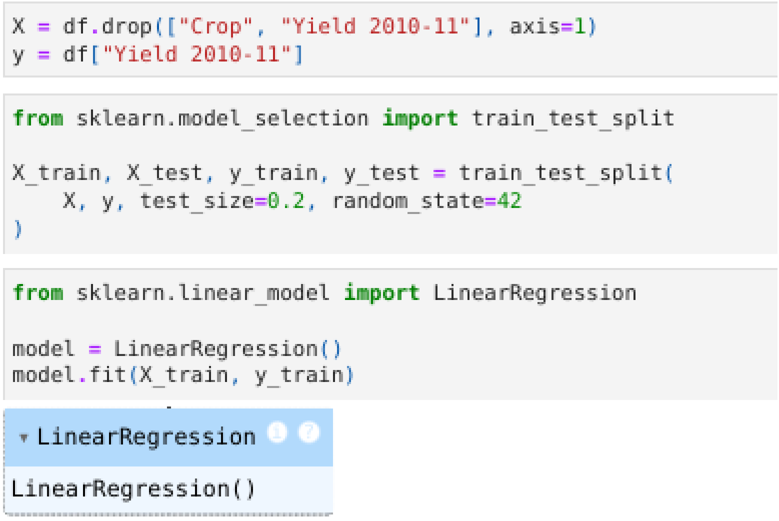
```
(55, 16)
```

```
df.duplicated().sum()
```

```
np.int64(0)
```

3. Machine Learning Model Implementation

- Implemented baseline machine learning models such as:
- Linear Regression
- Decision Tree Regressor
- Trained models on the processed dataset to predict crop production values.



```
X = df.drop(["Crop", "Yield 2010-11"], axis=1)
y = df["Yield 2010-11"]
```

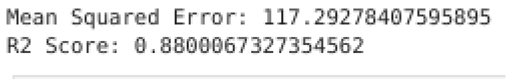
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

```
from sklearn.linear_model import LinearRegression

model = LinearRegression()
model.fit(X_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

4. Model Evaluation

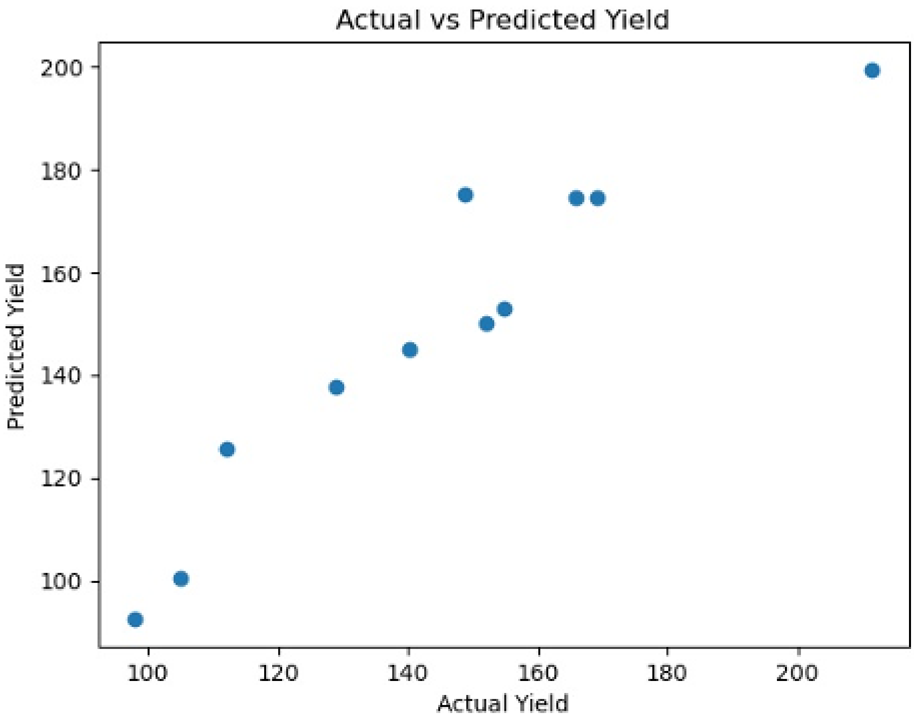
- Evaluated model performance using metrics such as:
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R² Score
- Compared model results to identify the better-performing algorithm.



```
Mean Squared Error: 117.29278407595895
R2 Score: 0.8800067327354562
```

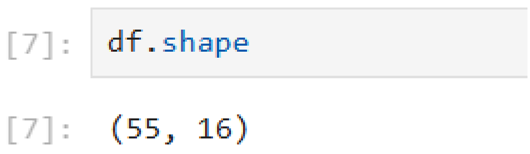
5. Performance Analysis

- Analysed overfitting and underfitting issues in initial models.
- Observed how feature selection and scaling impacted prediction accuracy.

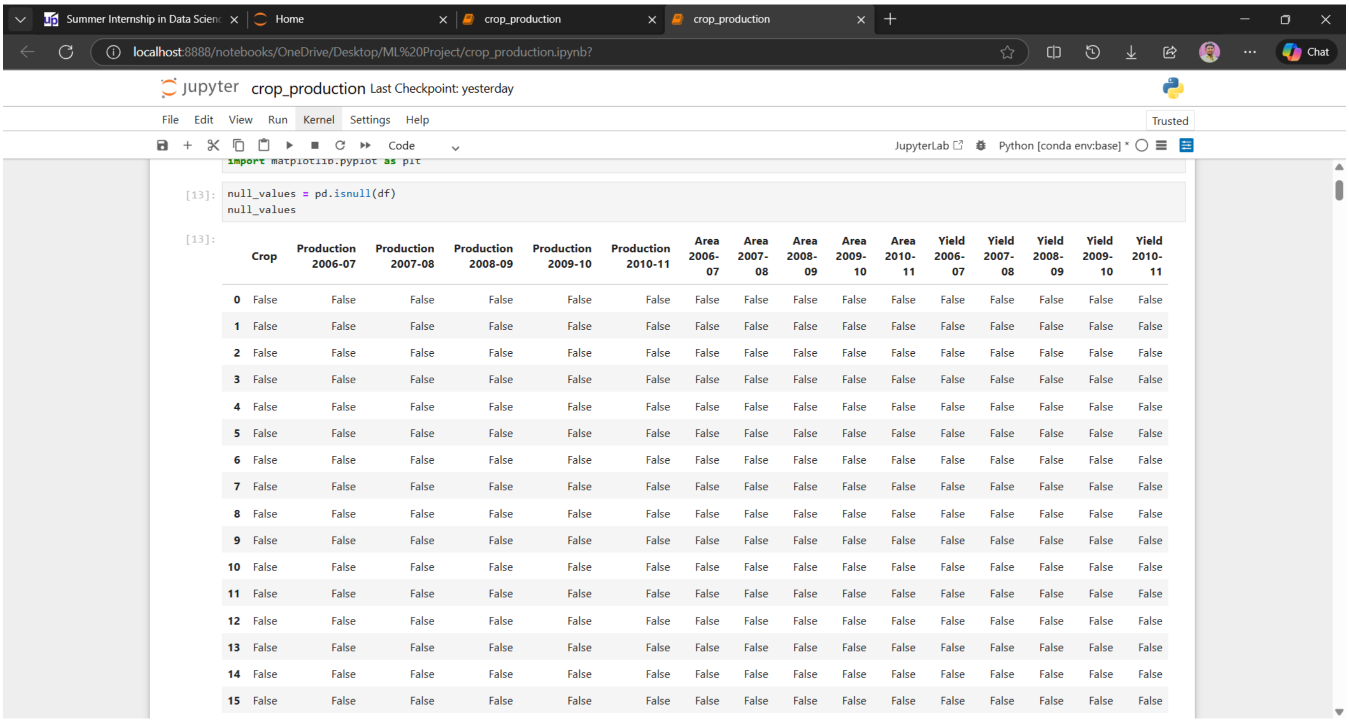


Chapter 3: Challenges

- **High Cardinality in Categorical Features:** Encoding state and district columns increased feature dimensionality.
- **Overfitting in Decision Tree Model:** Required careful tuning of model parameters.
- **Computational Complexity:** Training models on large datasets increased processing time.
- **Metric Interpretation:** Understanding regression evaluation metrics required additional practice.



```
[7]: df.shape

[7]: (55, 16)
```


Summer Internship in Data Scienc
Home
crop_production
crop_production
localhost:8888/notebooks/OneDrive/Desktop/ML%20Project/crop_production.ipynb?
Chat
jupyter crop_production Last Checkpoint: yesterday
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python [conda env:base] *

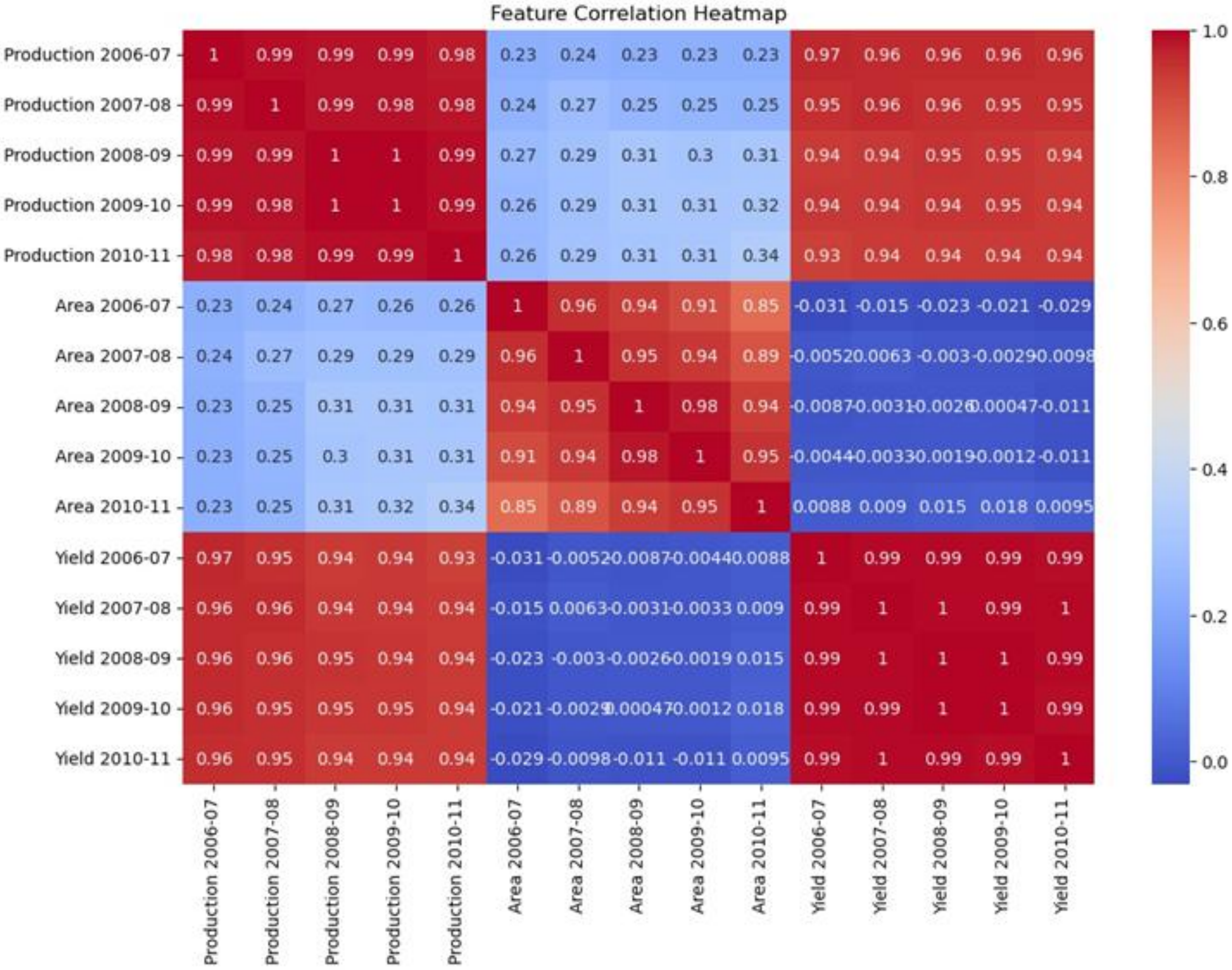
```
import matplotlib.pyplot as plt

[13]: null_values = pd.isnull(df)
      null_values
```

[13]:

	Crop	Production 2006-07	Production 2007-08	Production 2008-09	Production 2009-10	Production 2010-11	Area 2006-07	Area 2007-08	Area 2008-09	Area 2009-10	Area 2010-11	Yield 2006-07	Yield 2007-08	Yield 2008-09	Yield 2009-10	Yield 2010-11
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
5	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
6	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
7	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
8	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
9	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
10	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
11	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
12	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
13	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
14	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
15	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False

Chapter 4: Lessons Learned

- Gained hands-on experience in converting real-world data into machine-learning-ready format.
- Learned how different algorithms behave on the same dataset.
- Developed a strong understanding of regression evaluation metrics.
- Understood the importance of feature scaling and encoding in predictive modelling.
- Improved debugging and optimization skills during model training.



Chapter 5: Learning Resources

- Scikit-learn official documentation
- Online tutorials on regression algorithms
- Upskill Campus learning materials and mentor guidance
- Blogs and articles on crop yield prediction using machine learning

Chapter 6: Next Week's Goals

- Implement advanced models such as **Random Forest and Gradient Boosting**.
- Perform hyperparameter tuning to improve model performance.
- Apply cross-validation techniques.
- Compare all models and select the best performing one.
- Prepare the model for final deployment and reporting.

Chapter 7: Additional Comments

Week 03 was a crucial phase in transitioning from data analysis to machine learning implementation. Building and evaluating predictive models provided valuable practical exposure to real-world data science workflows. This week significantly strengthened my confidence in applying machine learning techniques to solve agricultural prediction problems.