

PROJECT DOCUMENT



**Student Name:** Shraddha Kalaskar

**Batch:** AI Elite 16



**Project Name:** Crop Yield Recommendation System

**Domain:** Agriculture

**Type of ML:** Supervised ML

**Type of Problem:** Classification

**Project Methodology:** CRISP-ML (Q)



**Phase I: Business and Data Understanding**

1. **Business Understanding:**Business Objective: Maximize crop yield and minimize resource waste (water, fertilizer, or pesticides)

Constraints: **Availability of Data, Best Model selection.**   
  
Success Criteria:

1. *ML success criteria*: with average 99% accuracy(estimation)
2. Business Success criteria: -
3. Economy Success criteria: -
4. **Data Understanding:**

|  |  |  |
| --- | --- | --- |
| **S No** | **Feature Name** | **Data Type** |
| 1 | N (Nitrogen) | integer |
| 2 | P (Phosphorous) | integer |
| 3 | K (Potassium) | integer |
| 4 | temperature | Continuous |
| 5 | humidity | continuous |
| 6 | pH | continuous |
| 7 | rainfall | continuous |
| 8 | label | Categorical |

**Phase 2: Data Preparation  
  
a) Exploratory Data Analysis:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Type** | **Feature Names** | **Observation** |
| 1 | Missing Values | - | - |
| 2 | Duplicates |  | - |
| 3 | Outliers |  | - |
| 4 | Distributions | N | Right skewed |
|  |  | P | Bimodel dist. |
|  |  | K | Bimodel dist. |
|  |  | temperature | Normal distribution |
|  |  | humidity | Left skewed |
|  |  | pH | Normal dist. |
|  |  | rainfall | Right skewed |
| 5 | Noisy data | - | - |
|  |  |  |  |

**b) Data Cleaning/wrangling:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S no** | **Type of Cleaning** | **Technique** | **Feature Name** | **Reason** |
| 1 | Missing value | - | - | - |
| 2 | Encoding | Label Encoding | Label | nominal |
| 3 | Scaling | Standard Scaling | All feature | Outlier |
|  |  | Minmax Scaling | All feature | Normalize |

1. **Feature Selection:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Removed Feature Name** | **Reason** | **Test Performed** |
| 1 | - | All are imp. | - |
|  |  |  |  |

**Phase 3: Model Building:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Type of Problem** | **Approach** | **Algorithm Name** |
| 1 | Classification | Distance based | KNN |
|  |  | Rules based | Decision Tree |
|  |  | Probability based | Naive Bayes |
|  |  | Boundary based | Logistic Regression |
|  |  | Boundary Based | Support Vector Machine |
|  |  | Ensemble | Random Forest |
|  |  | Ensemble | xgboost |
|  |  | Ensemble | AdaBoost |
|  |  | Ensemble | Gradient Boosting |

**Phase 4: Model Evaluation:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Algorithm Name** | **Metric Score** | **Hyper Parameters** |
| 1 | KNN | 98.18% | K =3 |
| 2 | Decision Tree | 98.86% | Max\_depth = 5 |
| 3 | Naïve Bayes | 99.31% | Alpha = 4 |
| 4 | Logistic Regression | 97.25% | Penalty = l2 |
| 5 | Support Vector Machine | 98.40% | kernel=rbf , c = 1.0 |
| 6 | Random Forest | 99.54% | n\_estimators=100  criterion=gini |
| 7 | xgboost | 98.63% | - |
| 8 | AdaBoost | 19% | n\_estimators = 50 |
| 9 | Gradient Boosting | 97.5% | loss = log\_loss  n\_estimators=100 |

**Phase 5: Model Deployment:  
  
Deployment Platform:**  Streamlit

**Link/URL:**  [app · Streamlit](http://localhost:8501/)