| Academic Year | Module | Assessment Number | Assessment Type |
|---|---|---|---|
| 2025 | Concept and Technology of AI | III | Individual Report |

# Classification Analysis Report


Student Id : 2407947

Student Name. : Shirish Pandit Chhetri

Section : L5CG2

Module Leader. : Siman Giri

Tutor : Sunita Parajuli


Submitted on. : 11 Feb 2025

## Abstract

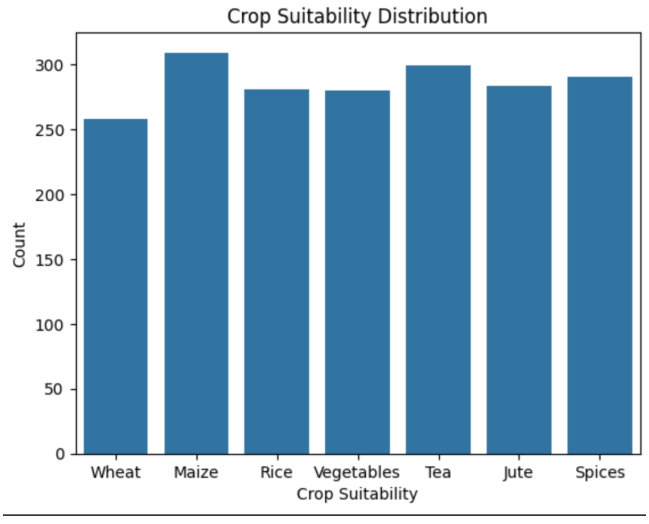
This report details an analysis of a modeling approach toward classifying data with machine learning. Further, this study entails data preprocessing, the modeling process, hyperparameter optimization, conducting feature selections, and evaluating the model with a Logistic Regression. The finalized model is subject to performance metrics to select the one that is most appropriate, and then the results are interpreted to provide insights into the model predictive capability.

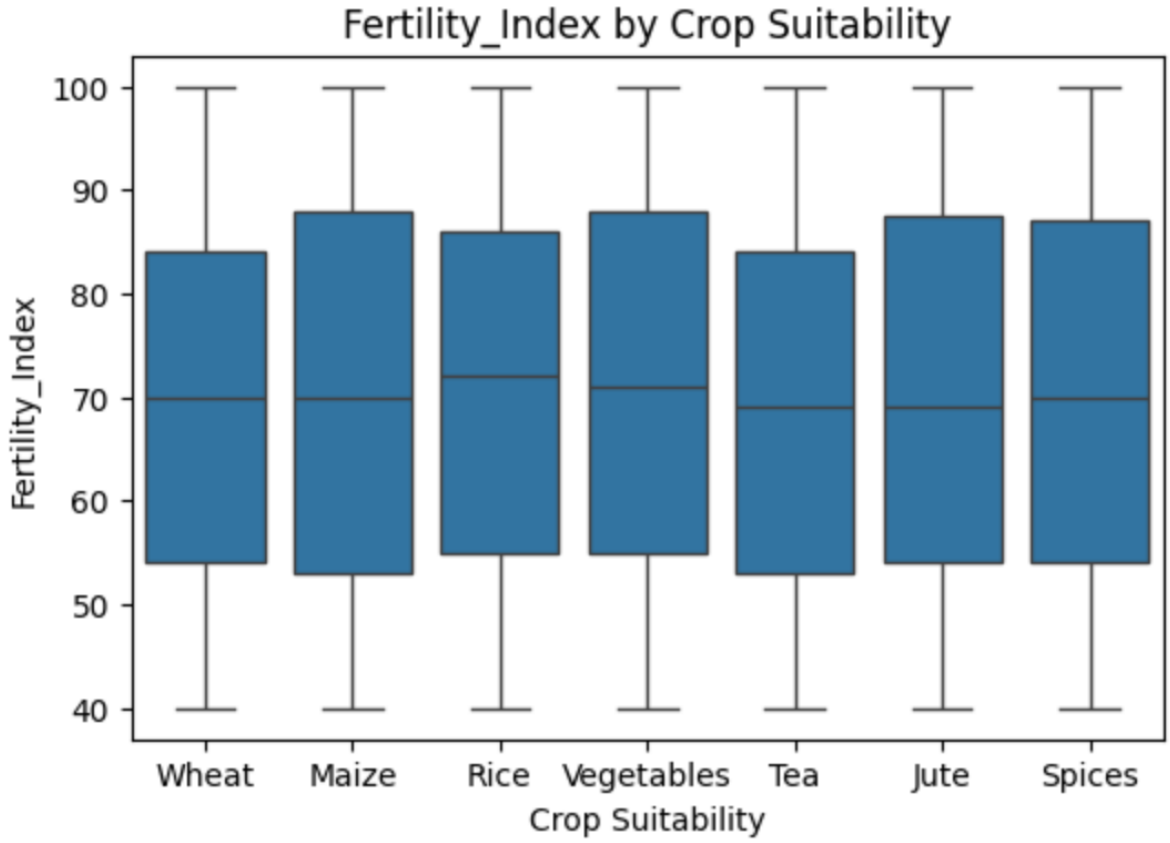
## Introduction

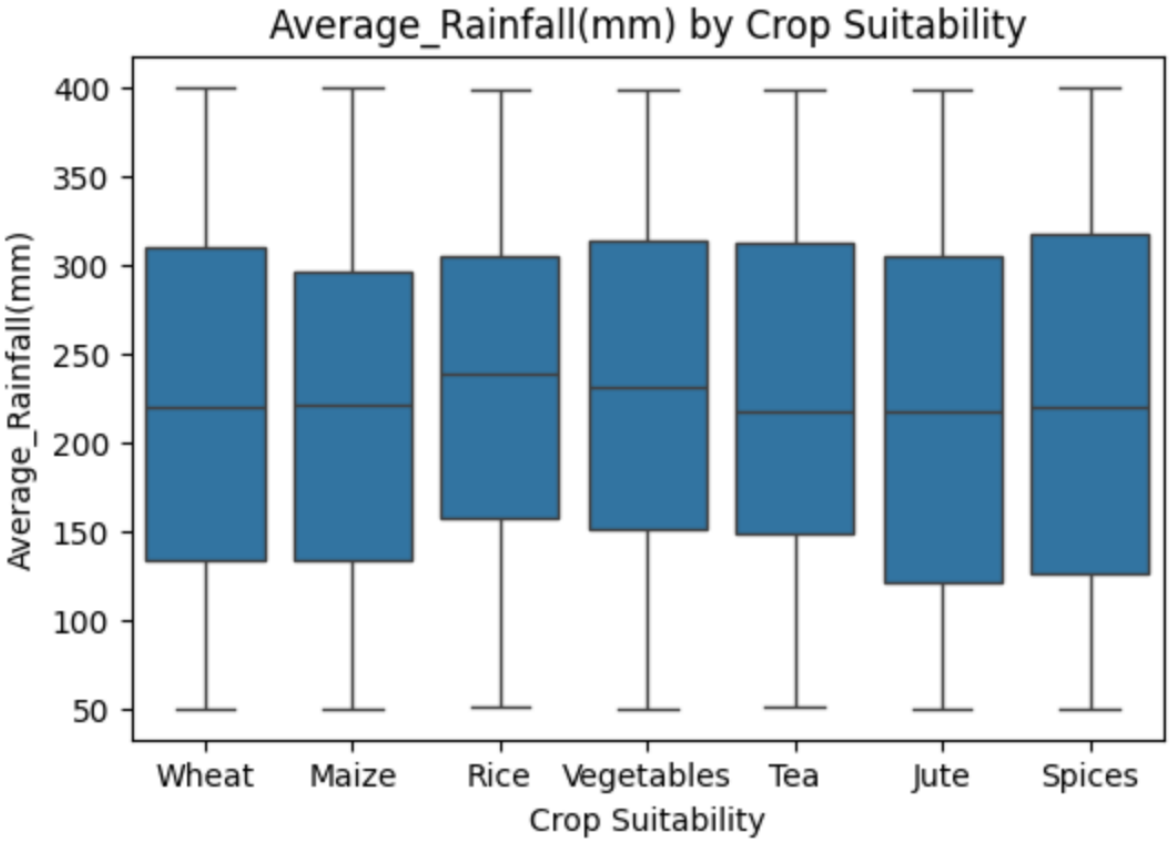
Classification is what primarily emerges as one of the leading problems of machine learning where the problem statement defines predicting a categorical class label for an input based on all its attributes. Efficient classification models can enable enhancement in decision-making across industries, from marketing to finance and healthcare. This study aims to develop a predictive model using logistic regression for classifying data according to various input features.

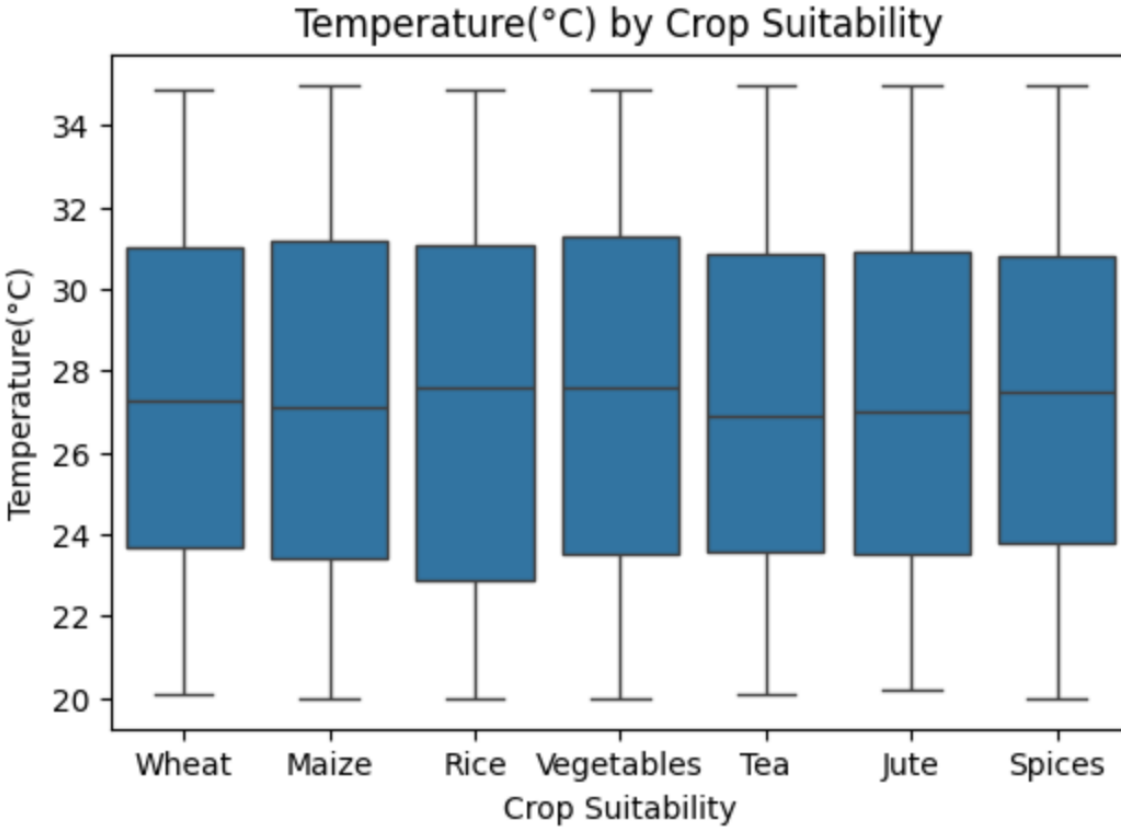
## Data Processing

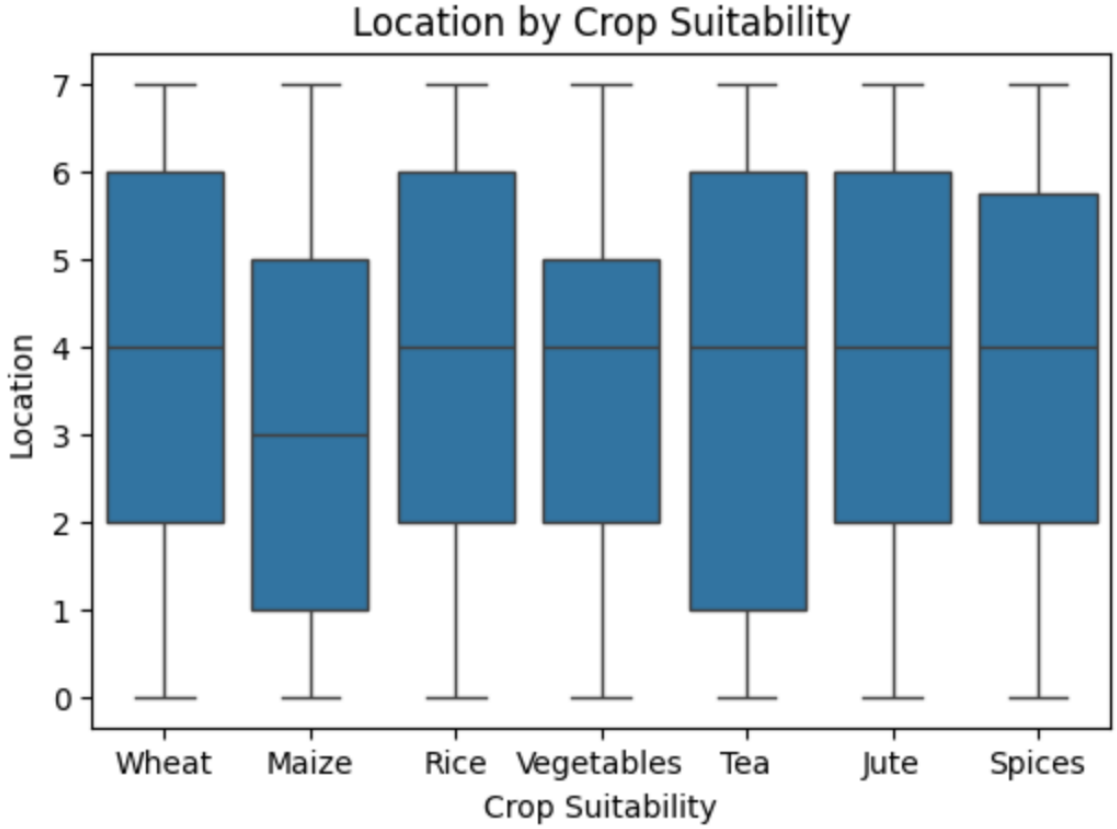
Basic data exploration is carried out to comprehend the structure and features of the loaded dataset used in this study. To guarantee data quality, duplicate entries are eliminated and missing values are located and filled up using the proper techniques. Better model performance is made possible by standardizing the features to have a zero mean and unit variance. The performance of the model is then assessed by separating the data into training and testing sets.

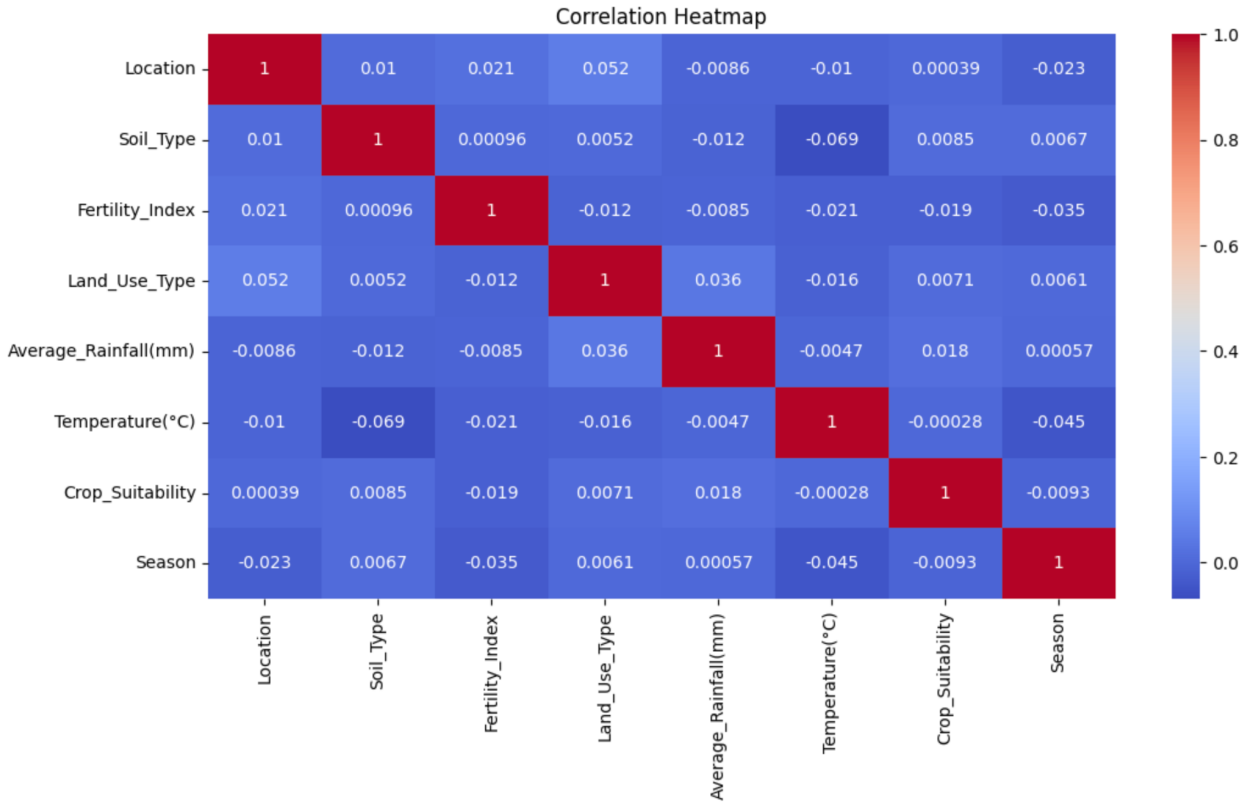



Fertility_Index by Crop Suitability
Fertility_Index
100
90
80
70
60
50
40
Wheat
Maize
Rice
Vegetables
Tea
Jute
Spices
Crop Suitability


Average_Rainfall(mm) by Crop Suitability
Average_Rainfall(mm)
400
350
300
250
200
150
100
50
Wheat
Maize
Rice
Vegetables
Tea
Jute
Spices
Crop Suitability


Temperature(°C) by Crop Suitability
Temperature(°C)
34
32
30
28
26
24
22
20
Wheat
Maize
Rice
Vegetables
Tea
Jute
Spices
Crop Suitability


Location by Crop Suitability
Location
7
6
5
4
3
2
1
0
Wheat
Maize
Rice
Vegetables
Tea
Jute
Spices
Crop Suitability




## Model Building:

Scikit-learn is used to create a model for logistic regression. For binary classification tasks, logistic regression is a linear model that calculates the likelihood that a given input falls into a specific class. Accuracy measures are used to assess the model's first performance after it has been trained on the standardized training data.

## Hyper-parameter Optimization

Improving the productivity of a predictor such as the Logistic Regression model by tuning its hyperparameters is through a varied grid search across hyperparameters including, but not limited to, the solver and regularization parameter (C). Implemented on cross-validation scores, the optimal parameters are derived for the model. In addition, there is being random oversampling to balance the relative distribution between classes in the training dataset, ensuring that the model is not biased toward the majority class.

## Feature Selection

Feature selection intends to find the relevant features that contribute to the classification task. The scoring of the features is done with SelectKBest with ANOVA F-value (f_classif), while the top features are selected accordingly. The selected features were used to retrain the Logistic Regression model, and its performance is evaluated again.

## Conclusion

The final model should be selected based on performance metrics of the Logistic Regression model with optimized hyperparameters and selected features. The model's performance was assessed with test data, and results were reported. The results underscore how well the model performed in classifying the data appropriately.

## Final model

The final model that was chosen is the Logistic Regression model with the best hyperparameters and features chosen. The model had an accuracy of 0.52 on the test data, which signifies the high predictive accuracy and robustness of the model.

## Challenges:

The study had several challenges, such as the treatment of missing values, selection of appropriate hyperparameter values, and identification of the most relevant features. One other critical consideration in the modeling process was the balance between performance and the complexity of the model.

## Future work:

Future work could further broaden the study by exploring other machine learning techniques besides support vector machines and neural networks for improvements in predictive accuracy. Additionally, more features can be engineered using domain knowledge, which might improve the model's performance. The cross-validation techniques could also be enhanced to ensure a sturdy evaluation of the model.

## Model performance:

Therefore, based on the evaluation metrics, accuracies defined the performance of the Logistic Regression model. The best accuracy achieved for the model with selected features and tuned hyperparameters was 0.52, indicating that it is able to correctly classify the data

with minimal errors. These metrics indicate that the model can accurately predict concrete compressive strength with minimal error.

## Interpretation of Results:

The results indicate that the Logistic Regression model is very efficient in classifying the data. The selected features are very significant variables for the classification task, top features defined by SelectKBest. The model can achieve a very high accuracy score, indicating that a significant portion of the input data can be classified correctly.

## Limitations and Suggestions:

Some limitations may, however, be claimed against the promising. The small size of the dataset limits the generalization of the model. Furthermore, a single classification model is studied, while there are other possible models that may yield better results. Future work should adopt larger datasets involving a more extended range of machine learning techniques, with domain knowledge and feature engineering further enhancing the model performance.

## Summary

This investigation proposes a unified procedure for data classification via machine learning techniques. It describes in detail the different facets involved with data processing, model building, hyperparameter optimization, and feature selection; then, upon performance evaluation, the best-performing model is selected. The Logistic Regression model with best-selected features and hyperparameter optimization showed very high accuracy, ensuring

accurate classifications of the data. The study emphasizes the importance of feature selection and hyperparameter tuning in developing reliable predictive models. These findings also offer a foundation for improving the accuracy of models and their applicability in real life.