|  |  |  |  |
| --- | --- | --- | --- |
| LOW LEVEL DESIGN AND IMPLEMENTATION DOCUMENT  PerfectCrop: The right crop for your soil  UE18CS390B – Capstone Project Phase – 2  ***Submitted by:***   |  |  | | --- | --- | | **Srish Srinivasan**  **Akash Kumar Rao**  **Vishruth P Reddy**  **Ishan Agarwal** | **PES1201800051**  **PES1201800089**  **PES1201800102**  **PES1201800291** |   Under the guidance of   |  | | --- | | **Prof. Raghu B A Rao**  Associate Professor  PES University |   **August - December 2021**  **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**  FACULTY OF ENGINEERING  **PES UNIVERSITY**  (Established under Karnataka Act No. 16 of 2013)  100ft Ring Road, Bengaluru – 560 085, Karnataka, India |

TABLE OF CONTENTS

|  |  |
| --- | --- |
| 1. Introduction | 3 |
| 1.1 Overview | 3 |
| 1.2 Purpose | 3 |
| 1.3 Scope | 3 |
| 1. Design Considerations, Assumptions and Dependencies | 3 |
| 3. Algorithm and Pseudocode  4. Implementation and Results | 4 |
| 5. Further Exploration Plans and Timelines | 6 |
| Appendix: Definitions, Acronyms and Abbreviations | 6 |

# Introduction

# Overview

Our crop recommendation system is the project that helps in the prediction of the most suitable crops for a particular climate and soil type. Our model takes in parameters like soil components, weather conditions and indices such as NDVI, ARVI, EVI and GCI as inputs. It provides a list of crops with the crops being ranked in the decreasing order of their yield as the output. Website

* 1. **Purpose**

The purpose of this low-level design document is to throw light upon the methods being implemented to predict the suitable crops, the input parameters and the predicted output.

* 1. **Scope**

This document covers all the details of the design implemented and the models incorporated. The machine learning algorithms, input parameters and the output crops with their yield percentage.

# Design Constraints, Assumptions, and Dependencies

**Design Constraints**

Our model runs on several machine learning algorithms such as K-Nearest Neighbours, Decision Trees, Random Forest Classifier, XG Boost and Artificial Neural Network. The algorithms that provide the best accuracy score are used for prediction of the suitable crops for that particular state.

**Assumptions**

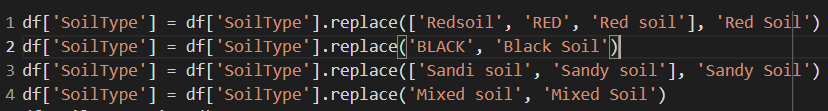
The data obtained is accurate with no altered or forged values.

**Dependencies**

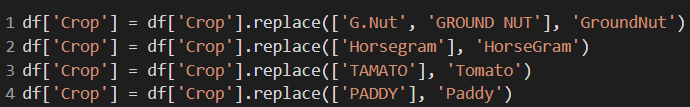
The soil and atmospheric parameters for the area of interest is required. Better analysis can be made with more or all the parameters. The latitude and longitude of that region should be known to calculate the different vegetation indices.

1. **Algorithm and Pseudocode**

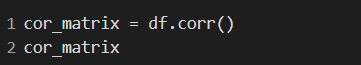
After pre-processing the data, we had all the required columns in hand. There were some soil features that meant the same but had different spellings or words. So we took the similar features and grouped them under a single name as follows:

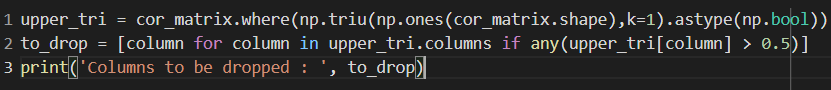


In a similar way, we replaced the crop names that had different spellings or words but meant the same which is as follows:

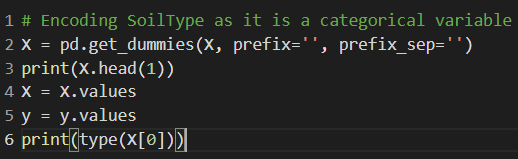


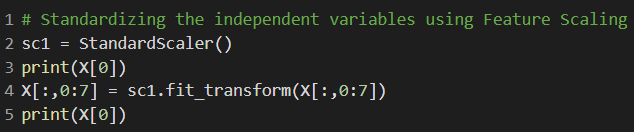
After the finding the crop count and getting rid of the least grown crops with negligible quantity of produce, we construct the correlation matrix to eliminate the columns that have high correlation which is as follows:



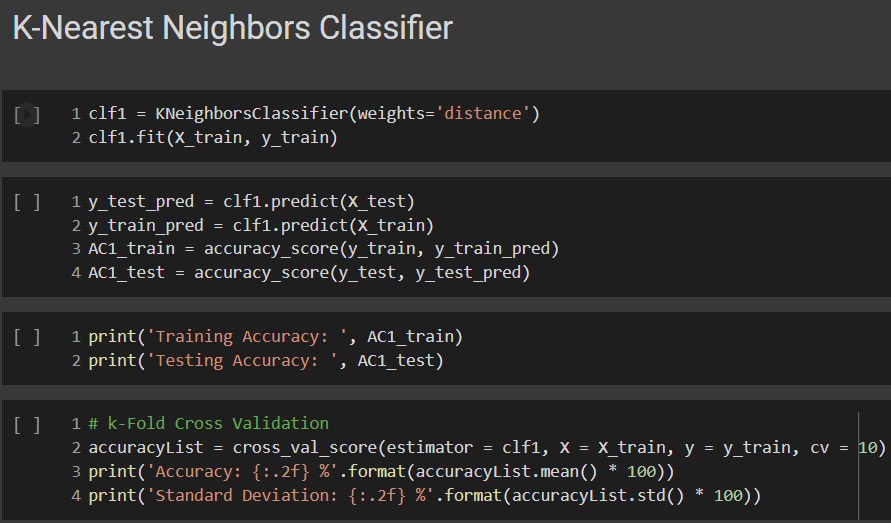


Now we encode the categorical variables and standardize the independent variables using feature scaling which is as follows:

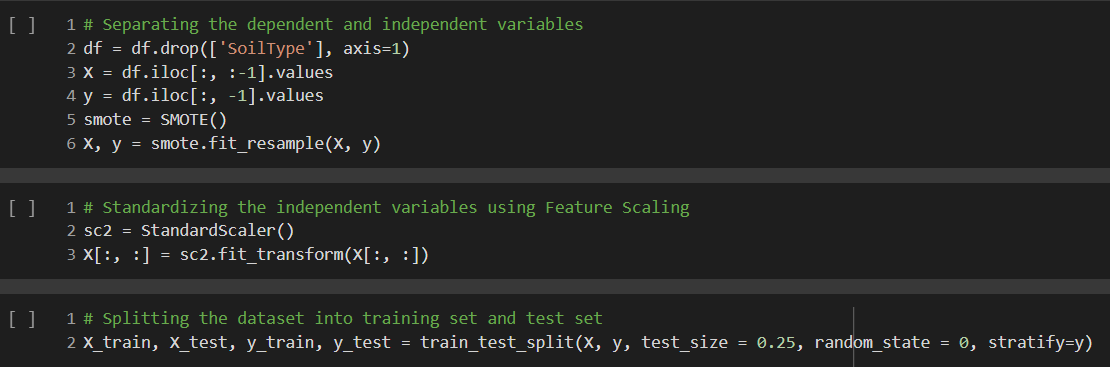




Now we split the data into training and testing sets to perform the analysis. The pseudo code for K-Nearest Neighbours is given below. All the other algorithms follow the same pattern with a change in the choice of the classifier used.



After running the models, we used the technique called SMOTE to give all the crops grown equal weightage so that predictions can be accurate. The SMOTE technique is shown as follows:



**4. Implementation and Results**

Initially, we approached the problem in a very simplistic way. We performed the analysis as soon as we cleaned the data. We didn’t take into consideration the type and amount of crops grown in that region. It was evident that if a crop was grown abundantly in a state, that had the most weightage. Even if the state grew other crops in that region but in small quantity, the majorly grown crops were dominant and hence, were producing inaccurate results. So we used a technique called SMOTE (Synthetic Minority Oversampling Technique) which generates more columns of the minority classes which helps in tackling overfitting issues. This technique enabled us to get better accuracies in prediction of crops. We have also incorporated another feature called the vegetation indices. These indices take in the latitude and longitude values of a the particular area and provides a number called the fertility index which defines how fertile that region is. Using this also as another feature in the dataset, we were able to do justice to the prediction of crops by accounting the fertility of the soil as well.

**5. Further Exploration Plans and Timelines**

We have completed 90% of our project. Small changes such as giving the users an option to input the required parameters as a csv file has to be done. We need to improve the web interface. We are currently working on the research paper and will be completing it soon.

# Appendix: Definitions, Acronyms and Abbreviations

[Provide definition of all terms, acronyms and abbreviations required for interpreting this Low Level Design Document.]

NDVI - Normalized Difference Vegetation Index

ARVI - Atmospheric Resistant Vegetation Index

EVI - Enhanced Vegetation Index

GCI - Green Chlorophyll Index

SMOTE - Synthetic Minority Over-sampling Technique