Markdown business project

Uzor Chike

2024-04-21

### Importing and Visualizing the Dataset

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ forcats 1.0.0 ✔ readr 2.1.5  
## ✔ ggplot2 3.5.0 ✔ stringr 1.5.1  
## ✔ lubridate 1.9.3 ✔ tibble 3.2.1  
## ✔ purrr 1.0.2 ✔ tidyr 1.3.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggplot2)  
library(caret)

## Warning: package 'caret' was built under R version 4.3.3

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

data <- read.csv('Crop\_recommendation (2).csv')  
df <- data.frame(data)  
head(df)

## N P K temperature humidity ph rainfall label  
## 1 90 42 43 20.87974 82.00274 6.502985 202.9355 rice  
## 2 85 58 41 21.77046 80.31964 7.038096 226.6555 rice  
## 3 60 55 44 23.00446 82.32076 7.840207 263.9642 rice  
## 4 74 35 40 26.49110 80.15836 6.980401 242.8640 rice  
## 5 78 42 42 20.13017 81.60487 7.628473 262.7173 rice  
## 6 69 37 42 23.05805 83.37012 7.073454 251.0550 rice

### Data Cleaning

The first step in data preprocessing involves cleaning the dataset to address any missing values, outliers, or inconsistencies. This ensures that the data is of high quality and suitable for analysis.

sapply(df, function(x) sum(is.na(x)))

## N P K temperature humidity ph   
## 0 0 0 0 0 0   
## rainfall label   
## 0 0

glimpse(df)

## Rows: 8,800  
## Columns: 8  
## $ N <int> 90, 85, 60, 74, 78, 69, 69, 94, 89, 68, 91, 90, 78, 93, 94…  
## $ P <int> 42, 58, 55, 35, 42, 37, 55, 53, 54, 58, 53, 46, 58, 56, 50…  
## $ K <int> 43, 41, 44, 40, 42, 42, 38, 40, 38, 38, 40, 42, 44, 36, 37…  
## $ temperature <dbl> 20.87974, 21.77046, 23.00446, 26.49110, 20.13017, 23.05805…  
## $ humidity <dbl> 82.00274, 80.31964, 82.32076, 80.15836, 81.60487, 83.37012…  
## $ ph <dbl> 6.502985, 7.038096, 7.840207, 6.980401, 7.628473, 7.073454…  
## $ rainfall <dbl> 202.9355, 226.6555, 263.9642, 242.8640, 262.7173, 251.0550…  
## $ label <chr> "rice", "rice", "rice", "rice", "rice", "rice", "rice", "r…

names(df)

## [1] "N" "P" "K" "temperature" "humidity"   
## [6] "ph" "rainfall" "label"

#classes of target variable  
table(df$label)

##   
## banana blackgram chickpea coconut coffee cotton   
## 400 400 400 400 400 400   
## grapes groundnuts jute kidneybeans lentil maize   
## 400 400 400 400 400 400   
## mango mothbeans mungbean muskmelon orange papaya   
## 400 400 400 400 400 400   
## pigeonpeas pomegranate rice watermelon   
## 400 400 400 400

#removal of classes of crops not grown in Nigeria or surrounding regions  
# Create a vector containing the names of classes to remove  
classes\_to\_remove <- c("blackgram", "chickpea", "grapes", "jute", "lentil", "mothbeans", "mungbean", "muskmelon", "pomegranate")  
# Create a logical index to identify rows with the specified classes in the target variable  
logical\_index <- !(df$label %in% classes\_to\_remove)  
  
# Subset your data using the logical index  
new\_data <- df[logical\_index, ]  
df <- new\_data