2/14/2020

Capstone (DATA 2206)

Module Two (Part I) – EDA Action Plan

Kartik Sojitra
100723768

Table of Contents

Problem statement

Starseed Medical Inc. determine products as the "best" quality products based on the best cannabinoid quality starting from green star 1 to blue star 3 potency CBD cannabis strains. These products will be used for medical so in the classification model, we will consider Recall and f1-score to resolve this problem.

Analytics Rationale Statement

1) The cannabinoid quality is measured by the classification of the quality of the plant.

2) To achieve the objective, we need to find how can plant provide the best cannabinoid potency to produce the best quality of cannabis-based medical products.

3) Starseed Medical Inc. should register the clients to increase the sale of cannabis-based medical products.

4) We need more than 95% of Recall value and f1-score as we are using the best quality cannabis for medical use.

Identify and Justify Output Variable Class Structure

To resolve the problem, we are using a binomial classification model to distinguish the best potency from low, average, good and best, and used transformation to determine the best cannabinoid quality output by algorithms.

The output variable for this project will be binomial as our output will be whether we are getting the best quality product of cannabis or not based on the "best" cannabinoid quality starting from green star 1 to blue star 3 potency CBD cannabis strains.

We are choosing binomial structure as we are going to identify only "best" quality products from low, average and good and counting the outcomes as either 1 and 0: counting the best as 1 and others (low, average and good) as 0. All the variables given in the dataset considered as independent variables and utility will be the dependent variable.

Action Plan for Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. EDA focuses on checking assumptions required for model fitting and hypothesis testing and handling missing values and making transformations of variables as needed.

The Dataset contains raw data originated from the Starseed Medical Inc. and it represents around 100% of testing data from Dec 1st to 10th, 2019. We are going to analyze cannabis dataset to find the key insights, to analyze dependent and independent variables for various types of data, to detect the outliers, to make assumptions and constrains of dataset, handling missing data using the various method like box plot, scatter plot, correlation matrix and statistical imputation method.

Steps of EDA in Cannabis dataset:

1) Importing the required libraries

2) Loading the data into the data frame

3) checking the types of data

4) Dropping the irrelevant column if necessary

5) Rename the column names if needed

6) Dropping the duplicate rows

7) Counting the number of rows after removing the duplicate rows

8) Dropping the null values

9) Detecting the outliers using box plot

10) Plotting a scatter plot and histogram

Our objective is to determine the best quality of products from cannabinoid potency grades from the plant. Data reveals that the cannabinoid quality of the plant is measured by the size, type, and species of the plant. Based on prescriptive analysis it concludes that no need to add more variables to design a better predictive model and achieve high Recall and f1-score. It is very important to know which and how data is used to build a better model and successful data analysis. It also specifies that what makes a good model, how can we validate and optimized the model and which hyperparameter we used to improve the accuracy of the model. Moreover, we can use visualization to understand the correlation between variables.

Assumptions

1) We have enough data to conduct analysis and it is normally distributed and we will not be getting more data to get high recall and f1-score.
2) We are using the binomial classification method as we are going to identify only “best” quality products from low, average and good and counting the outcomes as either 1 and 0: counting the best as 1 and others (low, average and good) as 0.

3) We are assuming that all columns are necessary, and all independent variables are correlated with the dependent variable (Utility) and not changing any variable name.
4) We are going to use a classification predictive model to map from the independent variable to the dependent variable. We will use the classification algorithm to train the dataset and cluster it.

Constraints

1) We don't have enough information to distinguish the best, low, average and good cannabinoid quality starting from green star 1 to blue star 3 potency CBD cannabis strains.
2) As we are choosing binomial classification as a result, we are not getting very specific results of the analysis.
3) All independent variables might not have correlated with the dependent variable.
4) We are using the classification model to distinguish the best potency so that affects the recall and f1-score.

References

Sam Plati_database overview- capstone (Data 2206)

Sam Plati_capstone (Data 2206)_module two (part - 1)_EDA action plan_week 2