1/24/2020

Capstone (DATA 2206)

Module One – Statement of Work (SoW)

Kartik Sojitra
100723768

Table of Contents

Executive Summary

The document illustrates the details of Starseed Medical Inc. which is producing the most high-grade quality cannabis to provide premium medical products to their clients. To accomplish the objective Starseed design system that classifies the best variety of cannabis-based on the high cannabinoid quality starting from green star 1 to blue star 3 potency CBD cannabis strains and classifying high potency THC medical cannabis oil. They are examining how do they classify the most useful CBD cannabis strain and the premium products based on cannabinoid potency grade. To resolve the problem they are using a 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.

Problem statement

Starseed Medical Inc. determine products as the high-grade 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.

Data

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. The dataset contains the independent variable as an actual measurement parameters for the cannabis plant and a dependent variable that classify the different CBD cannabis strains (none, low, average, good, best).

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 we need to add more variables to design a better predictive model and achieve high Recall and f1-score.

The IT and research department of Starseed Medical Inc. has access of all the data during data collection, data processing and data analysis.

Data Analysis Approach

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. To deal with probability and accuracy we are going to use a logistic regression model.

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 validates 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.

Project plan

Tasks	Completion time	Submission date
Analyze data	20-25 days	Jan 24th , 2020
Clean dataset	20 days	Feb 14th , 2020
Review and choose algorithms	20-25 days	Mar 13th , 2020
Train model	7 days	Mar 20th , 2020
Validate model	3 days	Mar 23th , 2020
Develop finalized model	4 days	Mar 27th , 2020
Develop final report	25-30 days	April 17th, 2020

References

Sam Plati_database overview- capstone (Data 2206)

Sam Plati_capstone (Data 2206)_module one_statement of work_week 2