|  |
| --- |
|  |
| **MODULE 1 Statement of Work**  Prachi Jindal 100721507 | Jan 24, 2020 |

Table of Contents

[Executive Summary 2](#_Toc30701392)

[Problem Statement 2](#_Toc30701393)

[Evaluation Metrics 2](#_Toc30701394)

[Analytics Rationale Statement 2](#_Toc30701395)

[Data 3](#_Toc30701396)

[Data Analysis Approach 3](#_Toc30701397)

[Project Plan 4](#_Toc30701398)

# Executive Summary

Starseed Medical Inc is focusing to bring the premium medicinal product in the market to provide easy access to medical cannabis by classifying Starseed strains according to the cannabinoid potency.

The goal of the *Cannabis Growth* project is to establish the best quality strains based on several variables and conditions of potency level so that Starseed Medical Inc. can use the best quality strains to develop and determine high potency cannabis products.

The scope of the Cannabis Growth project is to develop a model that predicts the best type of strain based on several parameters to manufacture premium high potency products.

# Problem Statement

Predict the best quality strain composition and plant size of cannabis strain to manufacture high potency medicinal cannabis products.

The dataset will be analysed to get the best quality strain and all the other quality strain to be treated as other. The best quality strain will be used for medical use.

# Evaluation Metrics

The key metrics for data analysis will be Recall value and Precision. The recall metric will evaluate the percentage of total relevant results correctly classified by the chosen algorithm. The recall value will tell about all the best quality strains when the dataset has different categories of the utility variable. The precision will describe the percentage of results which are relevant. The precision metric will tell the percentage of high quality strains from all the observations.

The key metric for model evaluation will be the confusion matrix. The metric will tell about the performance of our model and that

# Analytics Rationale Statement

To classify and identify the best composition and plant size of the strains which potentiate the effect of THC and CBD to manufacture high potency medicinal cannabis products using those strains and market the best cannabis to potential users.

The dataset has categorical data with five categories ranging from low to best, the model will analyse and identify the best quality strains for manufacturing and marketing which satisfies the scope of this project.

# Data

The dataset was provided by Starseed Medical Inc. which represents raw testing data from Dec 1st to Dec 10th, 2019.

The dataset has 1472 observations in total.

The data provided is raw and has a lot of missing and incomplete values

The dataset has 9 variables/features which determine if the cannabis plant is none, low, average, good or best.

**Variables:**

1. PMCno
2. DBH
3. Ht
4. Surv
5. Vig
6. Ins\_sec
7. Stem\_Fm
8. Crown\_Fm
9. Brnch\_Fm

The utility variable is the categorical data with five possible categories: none, low, average, good and best.

# Data Analysis Approach

**Collect Information & Analyze Data:** A research will be conducted on variables such as the quality of strain genetics, cannabinoids, terpenes, dopamine in order to better understand the data set provided and implement analytics tools for the project. The research will help in analysis of the dataset before developing a procedure and process to deduce the best strain based on the characteristics of the variables of the data given. Data analysis and classification techniques will be considered before model-building.

**Develop Model and Algorithms**

* The model development will begin with selection of variables as inputs to the test model. The data will be analyzed to distinguish various types of strains (none, low, average, good or best).
* The problem here is the classification problem where the output variable is categorical variable with 5 classes (none, low, average, good or best).

**Testing and Evaluation**

* An evaluation will be performed to see whether the data model prediction is accurate.
* The testing of the model and algorithm will ensure that requirements are met by the model.

**Software**

Python will be used as predictive analytics software.

# Project Plan

* The project has been divided into 10 task areas
* The project plan initiates with statement of work and kick off, execution starts with developing model, status of the project is monitored throughout the project and the project closes with final presentation.
* The kickoff is delivered with Statement of Work deliverable. The, most of the time in the project span will be spent in data analysis and model development. Once the model and data gets running, the last phase will be developing deliverables to close the project.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outline Number | Task Name | Duration | Start Date | Finish Date |
| 1 | Statement of Work Kickoff | 4 days | 20-Jan | 24-Jan |
| 2 | Collect Information | 6 days | 24-Jan | 30-Jan |
| 3 | Determine Data Analysis Approach | 5 days | 30- Jan | 3-Feb |
| 4 | Analyze Data | 8 days | 20-Jan | 28-Jan |
| 5 | Data Cleaning | 7 days | 22-Jan | 29-Jan |
| 6 | Develop Algorithm | 4 days | 29-Jan | 1-Feb |
| 7 | Review Algorithm | 4 days | 1-Feb | 5-Feb |
| 8 | Choose software tools | 4 days | 5-Feb | 9-Feb |
| 9 | EDA Action Plan | 4 days | 10-Feb | 14-Feb |
| 10 | Develop Model | 10 days | 14-Feb | 24-Feb |
| 11 | Develop Procedures | 5 days | 15-Feb | 20-Feb |
| 12 | Test Model | 6 days | 24-Feb | 1-Mar |
| 13 | EDA Report | 5 days | 1-Mar | 6-Mar |
| 14 | Train Model | 8 days | 6-Mar | 14-Mar |
| 15 | Validate Model | 10 days | 14-Mar | 24-Mar |
| 16 | Preliminary Report | 4 days | 24-Mar | 27-Mar |
| 17 | Develop Final Model | 8 days | 27-Mar | 3-Apr |
| 18 | Develop Final Report | 10 days | 7-Apr | 17-Apr |