Capstone Project 2 Proposal: Wine Review

Background:

Somm is a documentary movie on how four sommeliers attempt to pass the prestigious Master Sommelier exam, an exam that tests how good one is in identifying the type of wine.
The dataset given here looks to predict the wine variety given the description or review of the wine.

Problem:

Given the 9 independent variables describe below, including the description of the wine about how it tastes, smells, and feels, is it possible to figure out the wine's variety? If yes, how accurate is the model in predicting it and how can we improve the model?

Data:

The data is available on Kaggle: https://www.kaggle.com/zynicide/wine-reviews/data
The data contains 10 fields where there are nine independent variables, and one dependent variable.
The nine independent variables include:

- Points: The number of points a person rates the wine on a scale of 1-100
- Title: The title of the wine review
- Description: A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.
- Country: The country that the wine is from
- Province: The province or state that the wine is from
- Region1: The area where the wine is grown in a province or state
- Region2: The more specific area about where the wine is grown, can have missing values
- Winery: The winery that made the wine
- Designation: Vineyard within the winery where the grapes that made the wine are from
- Price: Cost of a bottle of wine

The variable we want to predict: Variety, which is the type of grapes used to make the wine (ie. Pinot Noir).

Problem Solving Approach:

First, I would do some data cleaning which includes analyzing what to do with the missing values, especially under the column Region2, where I supposed there will be a lot of missing data. Second, I would figure out if any variables are highly correlated with one another. If so, I would figure out how to transform those variables into one or for those variables with non-normal distribution, I would transform the variable.
Third, I would do EDA (Exploratory Data Analysis) to better understand how the variables are related to one another.
Fourth, I would transform the categorical variables into more meaningful features that can be used for prediction.
Last, I would try out several machine learning techniques to figure out the model's accuracy on the testing data.

Deliverables:

Deliverables to be expected from this project include lines of codes written down as well as PDF which includes all the data cleaning process, data visualization, and an explanation of the machine learning models used in this project.