FINAL PROJECT PROPOSAL

CS418 - INTRODUCTION TO DATA SCIENCE

HASS AVOCADO SALES

Group Info - Martin Bautista - 663597360

Mohmed Hira - 667094300

Sai Bharath Kumar Band - 670244109

Sachin Dattatraya Hegde - 670200929

Problem Statement:

Predict the price of the avocado based on sales and volumes. And classify the avocado into conventional or Organic.

Solution workflow:

- Exploratory data analysis: The dataset is cleaned, visualized and summarized. The irrelevant and redundant variables are identified
- We will remove the irrelevant and relevant variables, and we will standardize the dataset.
- Data Partitioning: We will split the data using holdout method into training data (80%) and Test data (20%).
- We will test the Multiple linear regression model and classification models on combinations of the variables and select the best models based on adjusted R squared value and F1 measure.
- We will use the best regression models to predict the price and classification model to classify the fruit (Organic or Conventional) on the test Dataset.

Data Source:

The dataset we use is the historical data on avocado prices and sales volume in multiple US markets from the year 2015 - 2018.

https://www.kaggle.com/neuromusic/avocado-prices