Moyahabo Rabothata

Meditechy Health Project: Machine Learning Predictive Modeling Approach for Classifying Depression.

25 March 2024

MEDITECHY

INTRODUCTION

The Busara Center in rural Siaya near Lake Victoria in the western Kenya is interested in understanding who is suffering from depression based on the routine survey data conducted in 2015.

- Project Overview
- Exploratory Data Analysis
- Modelling
- Results

MEDITECHY

PROJECT OVERVIEW

1 Project Problem

- Lack of resource for mental health illness
- Requires a different approach for early detection of risk to depression
- Requires a tool to have this detection done in real-time

2 Aim of the Project

To ML predictive modeling techniques to analyze and model survey data to have interpretable and user-friendly ML solution to assist in early detection of risk to depression.

3 Research Goals

- Build ML driven solution
- Create a ML powered application for detecting depression in real time.

4 Research Question

Can ML help the health care systems in early detection of risk to depression?

MEDITECHY

Data collection and EDA

Busara Mental Health dataset contains details of people suffering from depression based on routine survey data of 2015.

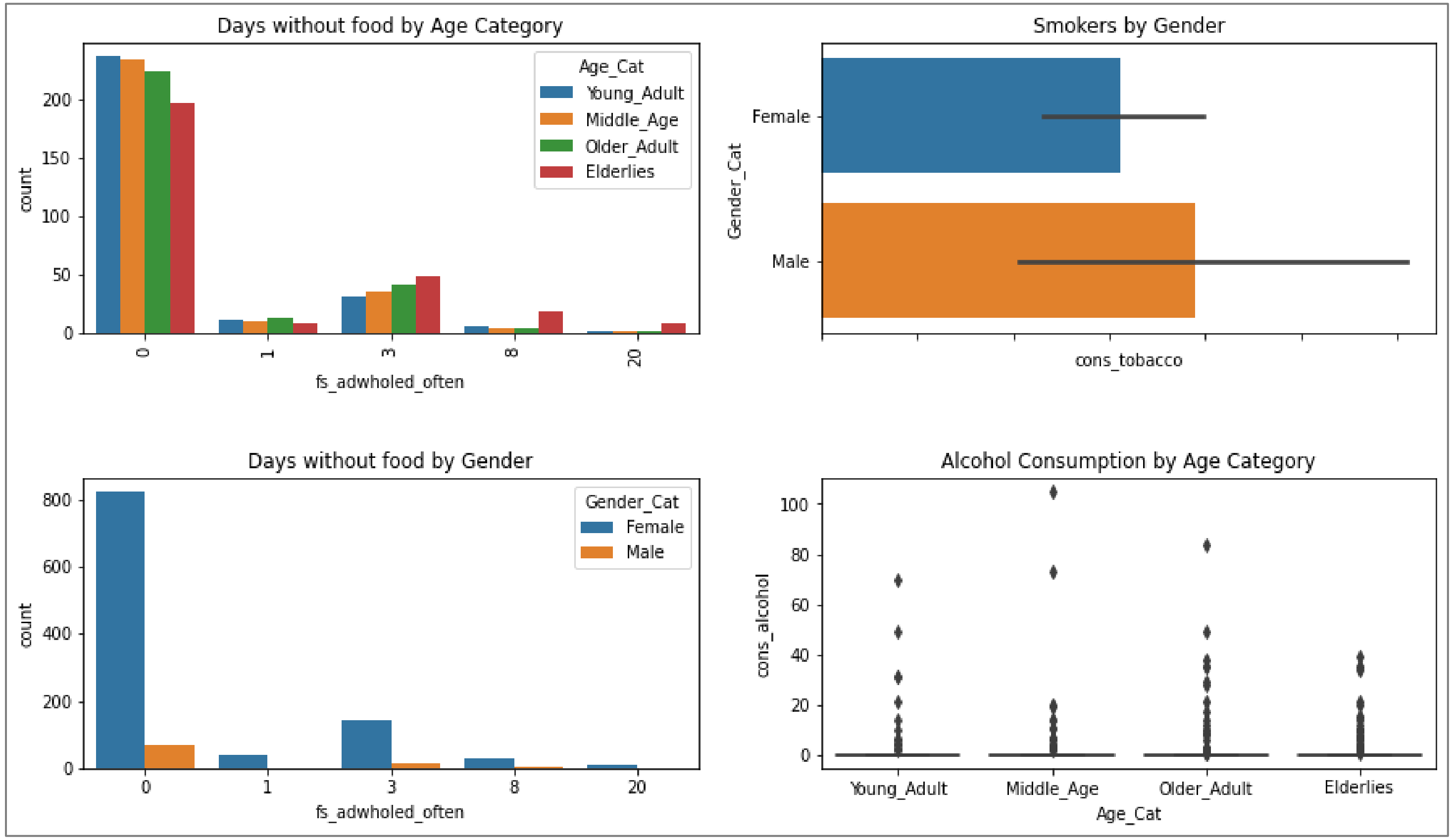
The dataset contains 75 features including information about household composition, economic activity, financial flows and health. In total it has 1 143 records/surveys wot of data.

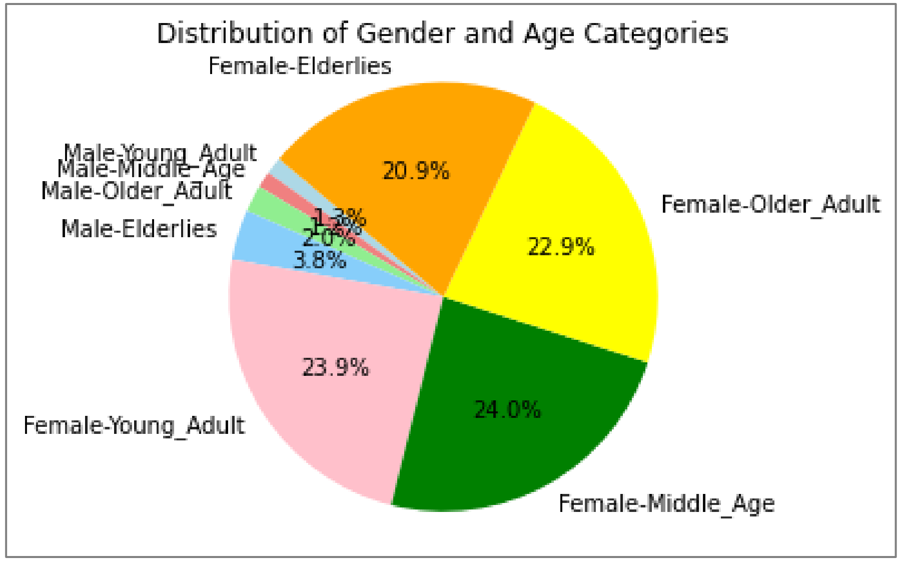
- The stuructured data of 2015 survey was obtained from Busara Mental Health. It consists of 75 features and 1 143 records.
- Checked null values and found that many features have null values this is the issue with survey data if controls on the forms are not put. Rather put an option that indicates that the person taking the survey chose not to answer if the field is not important
- Features ***surveyId***, ***dateofsurvey*** were removed from the data as they do not add any value towards the task at hand
- Exploratory Data Analysis(EDA) to understand and investigate the data to determine to what extend the data can be used to classify patients likely to have depression

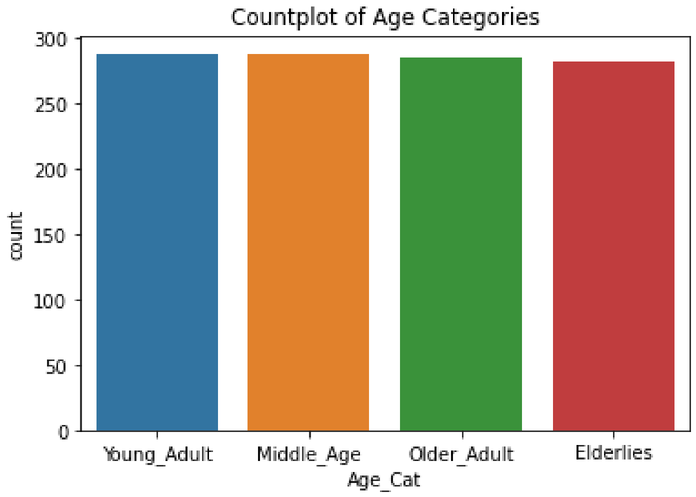
MEDITECHY

EDA VISUALIZATION

- More Females than Males
- Age groups balanced across the data
- Tobacco and Alcohol consumption are some of the major indicators of depression







MEDITECHY

MODELLING

From the gathered results obtained through EDA, we were able to implement the following modelling:

1. ***Support Vector Machine***
2. ***K-Nearest Neighbour***
3. ***Logistic Regression***
4. ***Principal Component Analysis***
5. ***K-Means Clustering***

This then led to predicting the probability of someone taking the survey to be depressed or not.

Support Vector Machine: The default parameters are used with linear kernel. E.g. SVC (kernel = 'linear)

K-Nearest Neighbour: K-NN uses default parameters. E.g. KNeighborsClassifier()

Logistic Regression: Logistic regression with default parameters. E.g. LogisticRegression()

PCA was used to reduce high dimensionality in the dataset since we have many features and less records of surveys. This is done to test curse of high dimensionality in our data

PCA was used to create 3 principal components (3PC) to be fed to K-Means clustering to separate the depression clusters..

MEDITECHY

RESULTS

The results of the modelling allowed us to summarise the classification of depression using Busara Mental Health dataset into as follows:

1. ***SVM is the champion Model***
2. ***83% is the highest accuracy recorded before parameter tuning***
3. ***3-Clusters seen in the data***

Support Vector Machine: The accuracy of the model is 83% which is the highest of the three algorithms experimented

K-Nearest Neighbour: K-NN performed at 82% which makes it the second-best algorithm

Logistic Regression: Logistic regression performed at 81%

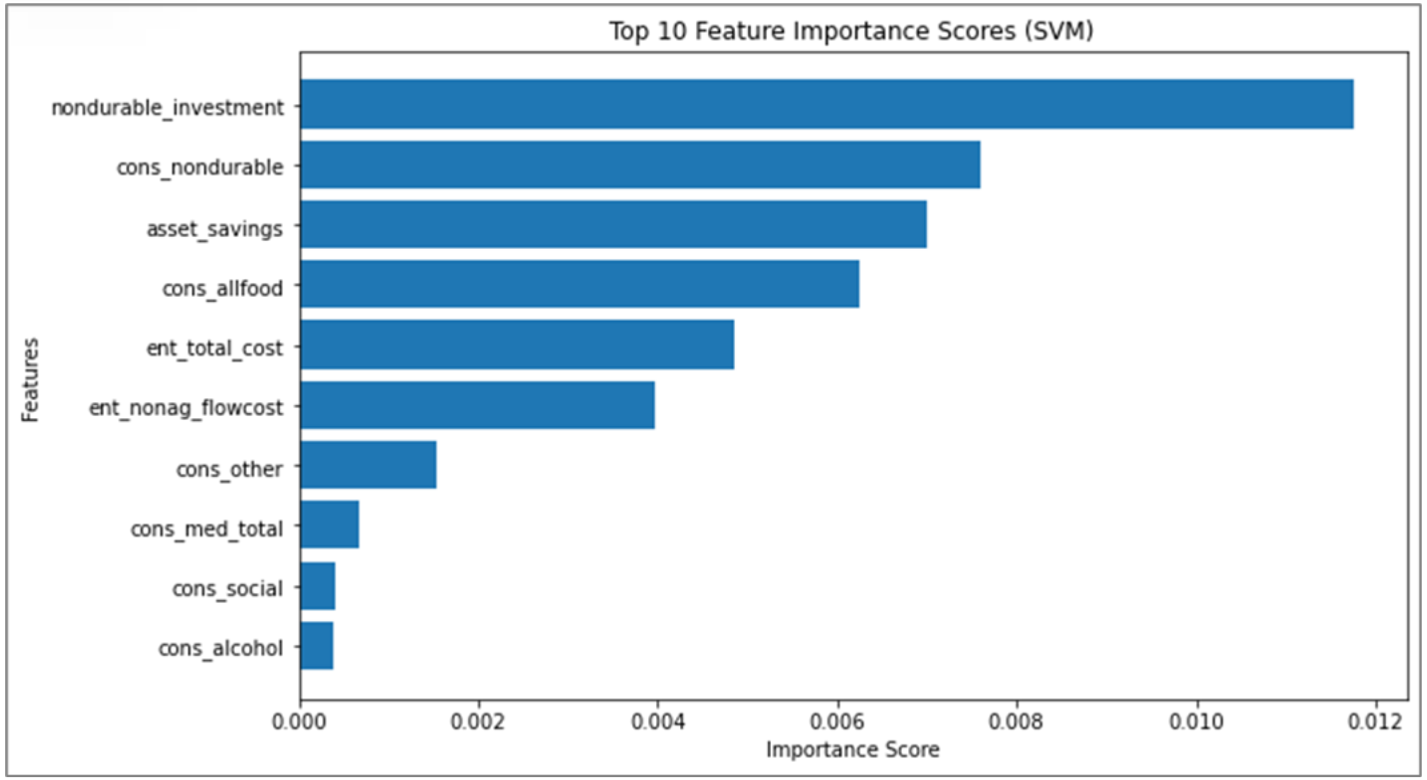
Three distinct clusters are seen which can be used to interpret Low, Moderate and High rate of depression

PCA was used to reduce high dimensionality in the dataset since we have many features and less records of surveys. No significant improvement of the classification algorithms after using PCA was noted

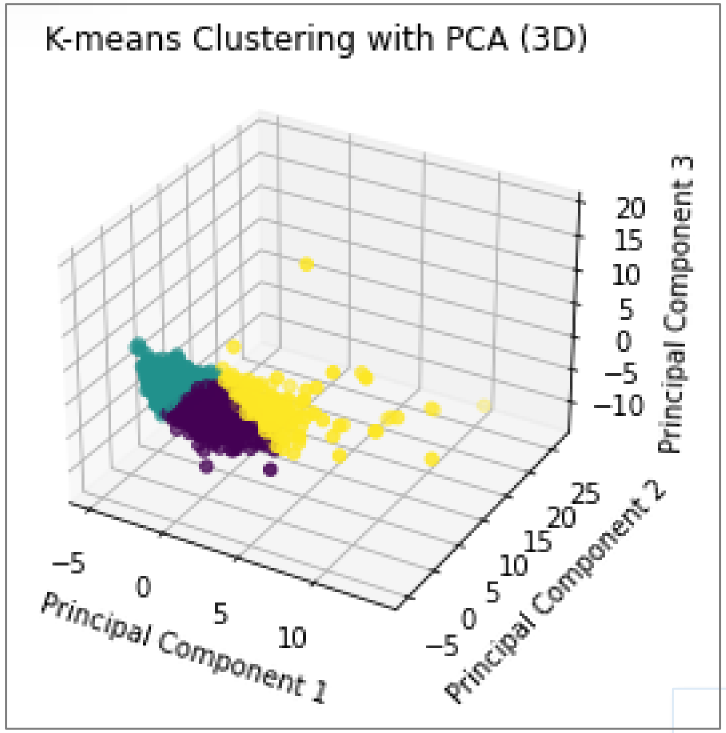
MEDITECHY

MODELLING VISUALIZATION

Best Model Variable Importance



K-Means Clustering K=3



MEDITECHY

THANK YOU !!!