Student Id: s3835204

Student Name: Shonil Dabreo

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show we I agree to this honor code by typing "Yes": Yes

Classifying people's demographics based on their attitude

To classify people's demographics based on their opinion on starwars movies, we need to first extract the features from the data collected in the survey.
Following are the survey questions i.e. attitude or opinion on starwars movies:

- Have you seen any of the 6 films in the Star Wars franchise?
- Do you consider yourself to be a fan of the Star Wars films franchise?
- Which of the following Star Wars films have you seen? Please select all that apply. (Star Wars: Episode I The Phantom Menace; Star Wars: Episode II Attack of the Clones; Star Wars: Episode III Revenge of the Sith; Star Wars: Episode IV A New Hope; Star Wars: Episode V The Empire Strikes Back; Star Wars: Episode VI Return of the Jedi)
- Please rank the Star Wars films in order of preference with 1 being your favorite film in the franchise and 6 being your least favorite film. (Star Wars: Episode I The Phantom Menace; Star Wars: Episode II Attack of the Clones; Star Wars: Episode III Revenge of the Sith; Star Wars: Episode IV A New Hope; Star Wars: Episode V The Empire Strikes Back; Star Wars: Episode VI Return of the Jedi)
- Please state whether you view the following characters favorably, unfavorably, or are unfamiliar with him/her. (Han Solo, Luke Skywalker, Princess Leia Organa, Anakin Skywalker, Obi Wan Kenobi, Emperor Palpatine, Darth Vader, Lando Calrissian, Boba Fett, C-3P0, R2-D2, Jar Jar Binks, Padme Amidala, Yoda)
- Which character shot first?
- Are you familiar with the Expanded Universe?
- Do you consider yourself to be a fan of the Expanded Universe?
- Do you consider yourself to be a fan of the Star Trek franchise?

We have already cleaned and explored the data in Assignment 1. Based on the analysis, we need to extract features from the above survey questions then select a classification model to train and test the data. The trained data should be then validated to predict the unseen future data i.e. expected input and output.

- **Feature Engineering and Model Selection**

Gender

Features	Values	Encoded (mapped) values
1. Have you seen any of the 6 films in the Star Wars franchise?	▪ Yes ▪ No	▪ 0 ▪ 1
2. Do you consider yourself to be a fan of the Star Wars films franchise?	▪ Yes ▪ No	▪ 0 ▪ 1
3. Please state whether you view the following characters favorably, unfavorably, or are unfamiliar with him/her. (Han Solo, **Luke Skywalker,** Princess Leia Organa, Anakin Skywalker, Obi Wan Kenobi, Emperor Palpatine, Darth Vader, Lando Calrissian, Boba Fett, C-3P0, R2-D2, Jar Jar Binks, Padme Amidala, Yoda)	▪ Very favorably ▪ Somewhat favorably ▪ Neither favorably NOR Unfavorably (Neutral) ▪ Somewhat unfavorably ▪ Very unfavorably ▪ Unfamiliar.	▪ 1 ▪ 2 ▪ 3 ▪ 4 ▪ 5 ▪ 0

I have selected the above mentioned features to predict the values of Gender.

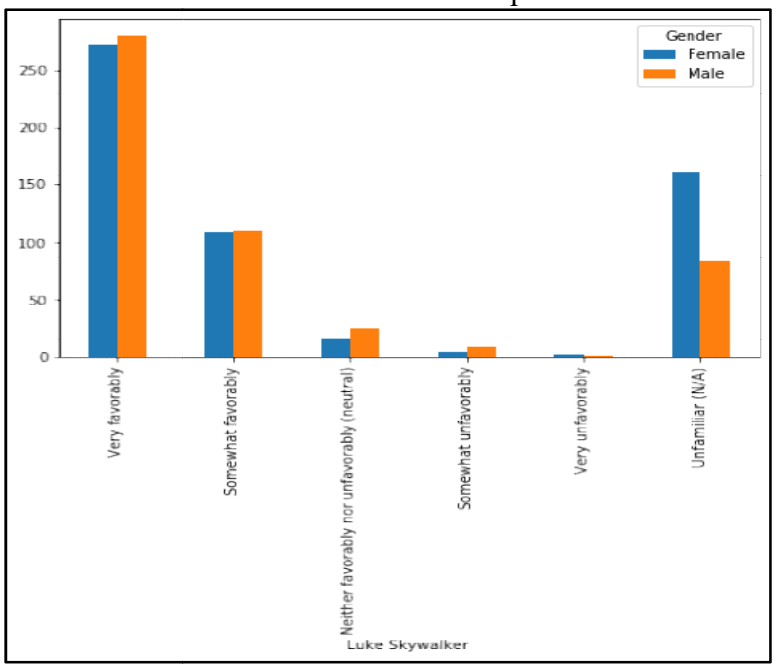

Fig 1

As we can see in the Fig 1, Gender and Luke Skywalker columns are grouped and the total count of the values of 3rd feature are calculated following which the bar graph is plotted. My hypothesis was that, given that Luke Skywalker is a male actor, the actor would have more 'Very Favorably' ratings from the female population. Gender (Target Feature) and Luke Skywalker (3rd Feature) was used to find out how actors would have ratings given by different people of the population. 1st feature is used to see the comparison between population that has seen films. 3rd feature is used to see the comparison between population (i.e. fans of starwars films franchise).
All the mentioned features are of category datatype. 'KNN classification model' is used as the 3rd feature has more than 2 values (i.e. several separated classes).

Age

Features	Values	Encoded (mapped) values
1. Have you seen any of the 6 films in the Star Wars franchise?	▪ Yes ▪ No	▪ 0 ▪ 1
2. Do you consider yourself to be a fan of the Star Wars films franchise?	▪ Yes ▪ No	▪ 0 ▪ 1

I have selected the above mentioned features for predicting Age groups

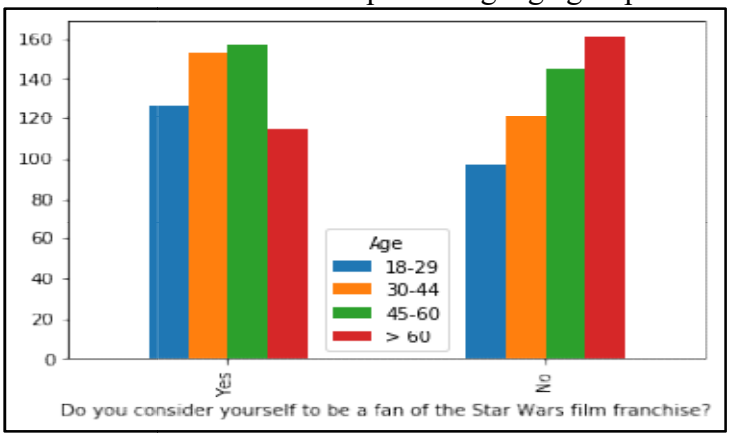

Fig 2

As we can in the Fig 2, Age and the Do you consider yourself to be a fan of the Star Wars films franchise? (2nd feature) columns are grouped and the total count of the values of 2nd feature is calculated following which the bar graph is plotted. My hypothesis was that, as most of the old people don't watch films they usually won't be the fan of the star wars films. Also, this type of films are generally for kids or teens, therefore, assuming there would be more fans in Age ranging from 18 to 29. Based on the results, the hypothesis about the old age group was true. 1st feature is used to see the comparison between age groups (i.e. fans of starwars films franchise).
All the mentioned features are of category datatype. 'Decision tree' is used to classify as the features includes 2 values.

- **Training the Model**

The labels are encoded with a dictionary of values and their corresponding numerical values. This is done to save time for the KNN classifier in calculating the distance between the points. Also, these encoded values are easy to interpret while feeding the data into a model.
The dataset is divided into 75% train data and 25% test data.

The value of k is an odd number 5. For KNN classifier, classification report is generated to see the accuracy of the model.

For maximizing the purity (i.e. homogenized groups), gini index is used in decision tree. Also, the information gain of gini index is higher than entropy information gain. Confusion matrix is used to gauge the accuracy of the model. The numbers on the diagonal of the confusion matrix correspond to correct predictions and the other values as the total no of errors.

- **Model Validation and selection**

K-folds cross validation is selected as it divides the dataset into K parts and uses each part one time as a test dataset while using rest of the data as a training dataset. This process is repeated until every k parts serves as a train set. To have better results the value of k is between 5-10.
For each process the accuracy score is calculated.
Then the average of all the scores is calculated to get the overall accuracy of the model.

- **Applying the trained model to unseen future data**

We expect the model to predict the same values (Output) while applying the trained model to unseen data (Input). The result would be an accuracy rate closed to approximation value between over fitting or under fitting.

Features	**Possible values**
1. Gender	▪ Yes ▪ No
2. Age	▪ Yes ▪ No