**Big Data Analytics (CO7093)**

**Group 20**

|  |  |  |
| --- | --- | --- |
| Group Members | | |
| **Student Name** | **Student ID** | **Email** |
| **Falowo Gbolahan** | **209033610** | **gf119@student.le.ac.uk** |
| **Adebowale Adebimpe** | **209054062** | **apa11@student.le.ac.uk** |
| **Yahaya Elmi** | **179040136** | **ye12@student.le.ac.uk** |

University of Leicester

Table of Contents

[Group Members 0](#_Toc67623225)

[INTRODUCTION 2](#_Toc67623226)

[EXPLORING THE DATA 3](#_Toc67623227)

[DATA CLEANING 3](#_Toc67623228)

[METHODOLOGY FOR CLEANING 3](#_Toc67623229)

[DATA EXPLORATION 4](#_Toc67623230)

[HYPOTHESIS: MOST OSCAR WINNERS ARE WHITE. 4](#_Toc67623231)

[HYPOTHESIS: MOST OSCAR WINNERS ARE FROM USA. 5](#_Toc67623232)

[HYPOTHESIS: BEST DIRECTORS TEND TO BE OLDER THAN BEST ACTORS OR ACTRESSES 6](#_Toc67623233)

[MODEL BUILDING 7](#_Toc67623234)

[LOGISTIC REGRESSION 7](#_Toc67623235)

[K-NEAREST NEIGHBOUR 7](#_Toc67623236)

[RANDOM FOREST 8](#_Toc67623237)

[IMPROVING OUR MODEL 9](#_Toc67623238)

[EVALUATING OUR MODEL AND IMPROVING 9](#_Toc67623239)

[USING FEATURE SELECTION TO DETERMINE THE BEST FEATURES TO USE FROM OUR DATASET 9](#_Toc67623240)

[SELECTKBEST 9](#_Toc67623241)

[LOGISTIC REGRESSION 10](#_Toc67623242)

[K-NEAREST NEIGHBOUR 10](#_Toc67623243)

[RANDOM FOREST 11](#_Toc67623244)

[USING K-MEANS ALGORITHM TO CLUSTER 11](#_Toc67623245)

[PRINCIPLE COMPONENT ANALYSIS SHOWING CLUSTER DISTRIBUTIONS OF THE DATASET 12](#_Toc67623246)

[CONCLUSION 13](#_Toc67623247)

[REFERENCES 14](#_Toc67623248)

# INTRODUCTION

Predicative modelling is a mathematical process that utilizes statistics to predict outcomes. In recent years it has become a powerful application of predicative analytics that is widely utilized by many organizations to find and exploit underlying trends found in data. We decided to harness the benefits of predictive modelling and apply this to our dataset. We were provided with a dataset containing related information of Oscar winners during the period of 1927 to 2014. The dataset is split into best actress, best actor, best supporting actress, best supporting actor and best director. We set out to employ different techniques and methodologies to build a predictive model that would be able to predict which person based on their features such as age, race, origin where able to win the different types of awards, using our given dataset. In this report, we present the different methodological approaches to building our respective models along with visualizations of our results/findings.  The discussion underpins justification of the steps we took to achieve our aim, limitations and advantage of methods and recommendations to improve any actions or decisions that we may have undertaken.  Using the given dataset, we developed a predictive model to predict which type of award is won by a person based on a range of features such as country of origin, race, age, etc. and to propose a set of clusters that may make business sense of the movies industry.

# BODY

Using the given dataset, we needed to clean the data, develop a predictive model to predict which type of award is won by a person based on a range of features such as country of origin, race, age, etc. and to propose a set of clusters to properly understand the business of the movies industry.

### DATA CLEANING

Our first task is to prepare the data and carry out data munging or cleansing. In this task, we created a subset of the dataset consisting of the birthplace, date of birth, race ethnicity, year of award and award column.  We found some inconsistencies in the dataset in the date of birth column and the birthplace column.

* In the Date of birth column, we found inconsistent date formatting e.g., 2-Aug-98 and missing day and month of birth.
* In the birthplace column, some birthplace had no country specified.
* We identified Na values and duplicates and expelled them from our dataset accordingly.

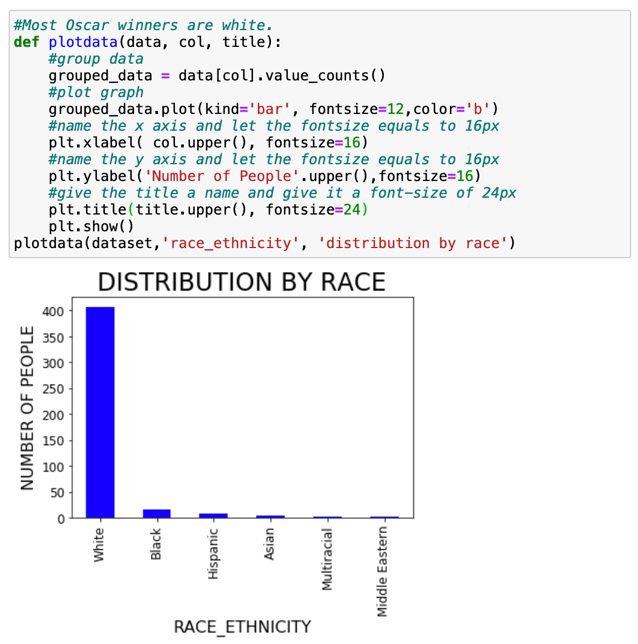
### METHODOLOGY FOR CLEANING

Here we would discuss an overview of our methodologies, intuitions and reasons behind our modes of implementing solution to treat the abnormalities we found in our dataset.

* Firstly, there were errors in the date of birth column and they have now been removed.
* There were also errors present in the birthplace columns and as such we have implementations used in correcting these errors in the notebook submitted.
* The duplicate row in the dataset have been removed from the dataset
* We implemented a for loop that iterates through our column and checks for the rows with incomplete countries and adds the respective country to the new column we created value “USA”. This new column is called “country”.
* We implemented a function that uses the split method to isolate the year of birth and date of birth and updated to the correct format “Day-Month-Year” e.g., 19-Sept-2020
* There has been an addition of columns as instruction which are the country column and the award age column which is going to be used down the line in our analysis.
* A series of our solutions involved the use of function and if else statements in some cases. Furthermore, please find in out notebook alternative methods of implementing the processes for variety.

# DATA EXPLORATION

### HYPOTHESIS: MOST OSCAR WINNERS ARE WHITE.



**Our Findings**

* The visualization above proves to us that the majority of the Oscar winners were white,
* We can also observe that the second highest winners are the blacks.
* We have gone with a bar graph because we are dealing with categorical data.
* It best represents trends and it is easy to observe the differences.
* In this Instance each bar represents an ethnicity.

### HYPOTHESIS: MOST OSCAR WINNERS ARE FROM USA.

Graphical user interface, text

Description automatically generated

Chart, histogram

Description automatically generated

**Our Findings**

* The visualization above proves that most Oscar winners are from USA.
* From the visualization of our findings, we have shown that the hypothesis is valid.
* Initially our graph showed all 33 countries which extrapolated the X axis and minimized the size of the bar chart.
* t also made difficult to read the graph.
* It was decided to replace the 23 countries that had close to 0 Oscar winners with “Others” to better visually represent results.

### HYPOTHESIS: BEST DIRECTORS TEND TO BE OLDER THAN BEST ACTORS OR ACTRESSES

Chart

Description automatically generated

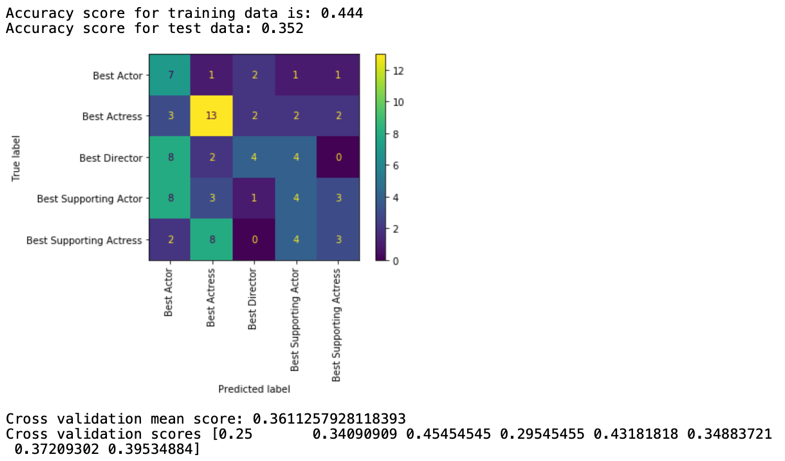
**Our Findings**

* From the visualization above we can prove that the Best Directors tend to be older than best Actors or Actresses.
* From the above bar, we can see that the Best Directors are closer to age 50.
* The Best supporting actress has an award age 40 and slightly above.

# 

# MODEL BUILDING

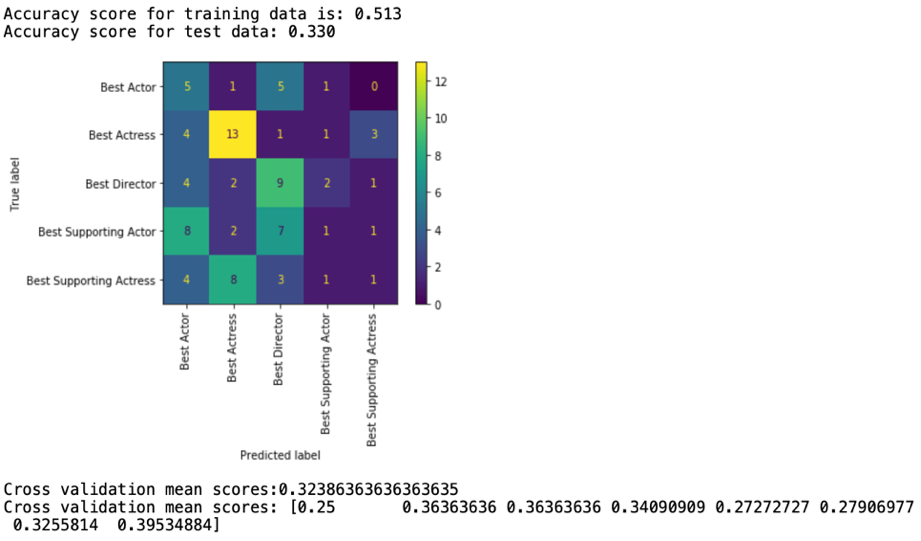
### LOGISTIC REGRESSION



**Our Findings**

* We can see that the Logistic Regression does not have a high accuracy after evaluation.
* The model is not training as well as it should with the default parameters.

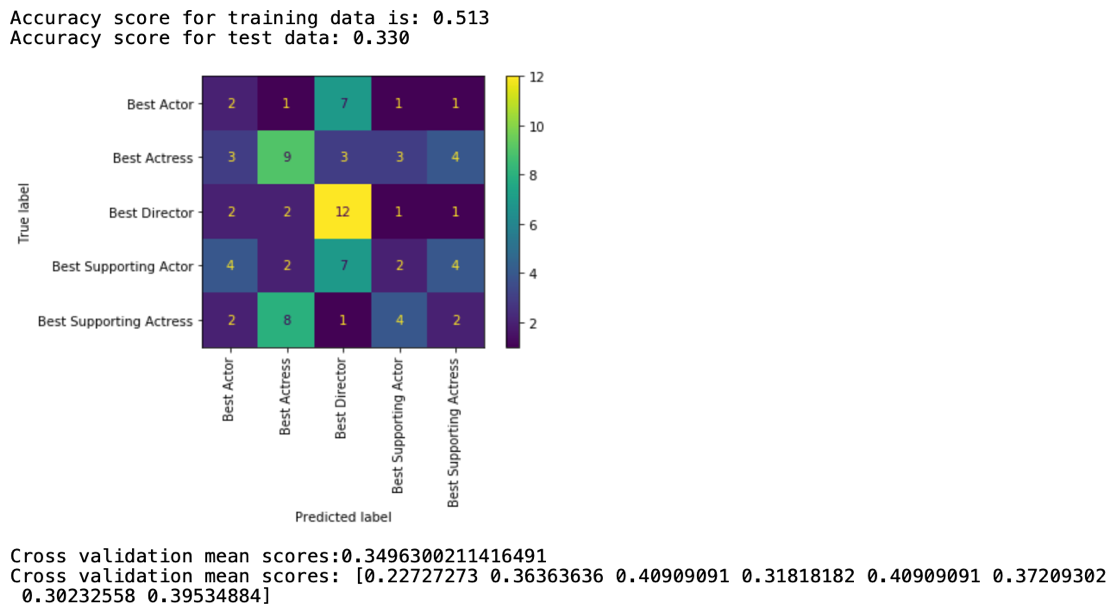
### K-NEAREST NEIGHBOUR



**Our Findings**

* Here we can also see underfitting because the test data does not perform as good as the training data.
* And as such we do not have significant results from this.

### RANDOM FOREST



**Our Findings**

* Here we can say this classifier has similar results to that gotten from logistic regression, although it may be higher slightly considering the performance it is not quite there yet when it comes to giving a solid result.
* Fundamentally we know a way for the model to train better would be to feed it more data which we would do as you go into the report.
* And as such we could suggest that the model is not training as good because of this.

# IMPROVING OUR MODEL

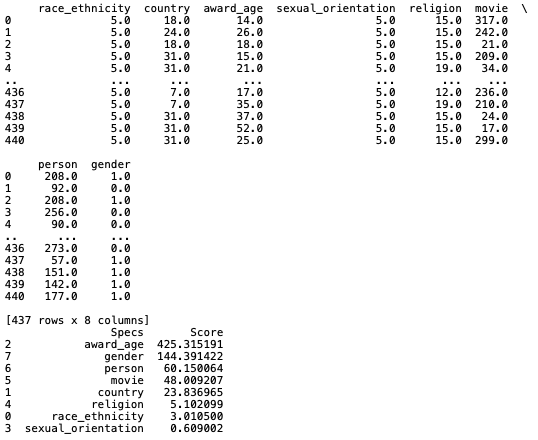
This is an open-ended question and you are free to push your problem-solving skills in order to build up a useful model with higher performance.

1. Consider the entire dataset given in this assignment. Develop an improved predictive model that predicts the award type for a given individual. Make sure your model is validated by using cross-validation. You should aim for a model with a higher predictive accuracy or with results that are easy to explain/interpret.
2. Use the K-Means algorithm to cluster your cleansed dataset and compare the obtained clusters with the distribution found in the data. Justify your clustering and visualise your clusters as appropriate.
3. Include in your report any decisions or actions that may be taken from your improved classification model as well as your obtained clusters on this application.

### EVALUATING OUR MODEL AND IMPROVING

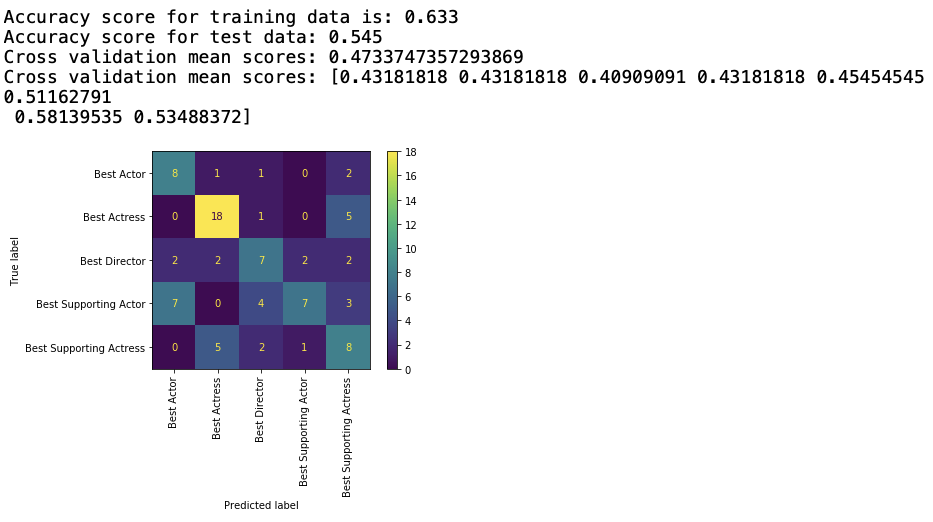
### USING FEATURE SELECTION TO DETERMINE THE BEST FEATURES TO USE FROM OUR DATASET

### SELECTKBEST



Feature selection is one of the core concepts in machine learning which hugely impacts the performance of our model. The benefit for our model is to reduce Over fitting, improve the accuracy and reduce the training times. This was all observed as we tested with more promising columns.

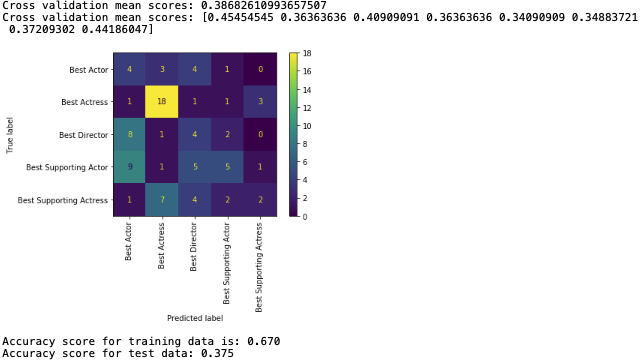
### LOGISTIC REGRESSION



**Our Findings**

* We can observe a little bit of an improvement in the training data.
* Also, we can see a bit of an improvement in the test data even after feeding in more data and tweaking the parameters
* The fluctuations and large gaps between the accuracies during this testing was showing clear signs of overfitting accuracies of 1.00 and such sorts which is not good.

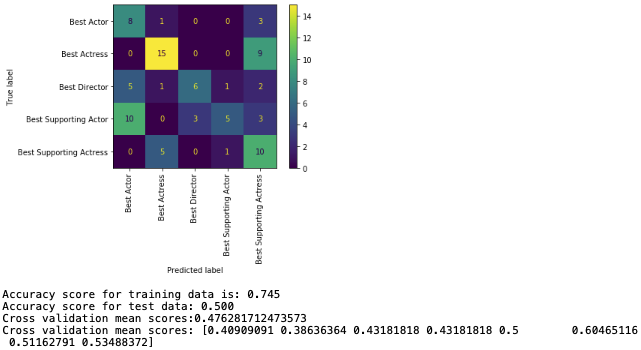
### K-NEAREST NEIGHBOUR



**Our Findings**

* Here we can observe overfitting because the disparity between the training data and the test data is still too large.
* The algorithm still performs poorly on the dataset even after tweaking parameters like the number of neighbors, leaf\_size.
* The algorithm does not provide any significant results.

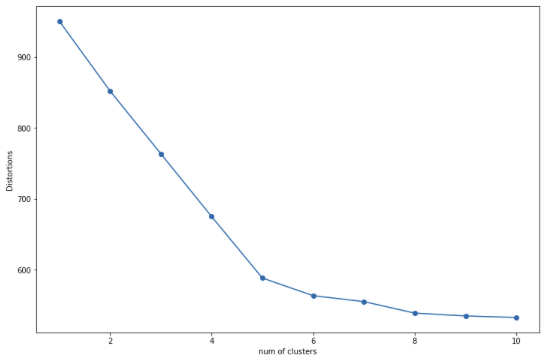
### RANDOM FOREST



**Our Findings**

* Initially the model tested poorly, but after much attempts of tweaking parameters we were able to bridge the gap between the training data and the test data.
* Making changes to the max\_depth and the n\_estimators alongside the min\_samples\_split brought about improvements here and there after which we settled on the set parameters.
* The learning algorithm is not capturing the underlining trend of the data.

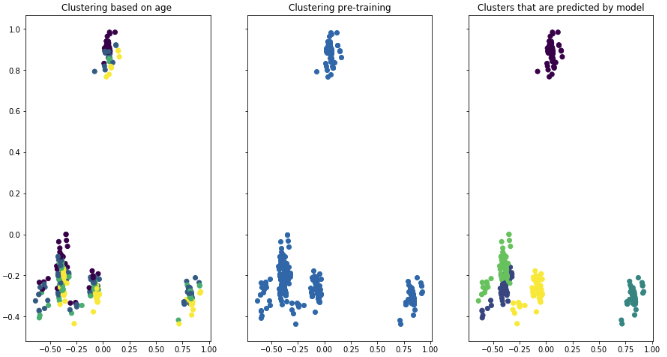
# USING K-MEANS ALGORITHM TO CLUSTER



**Our Findings**

* Here we can observe an elbow of 5.
* Subsequently we use this information for our PCA where we use 5 clusters set.

# PRINCIPLE COMPONENT ANALYSIS SHOWING CLUSTER DISTRIBUTIONS OF THE DATASET



From our visualization you can see we have 5 distinct clusters, It shows that objects that are similar are closely grouped together in the 2-dimensional space. Initially we selected a small data set and observed the clusters were scanty and could see we needed to add more data to observe clear distinctions. In our visualization the colors above represent the various clusters of award distributions Based on the first PCA we can see the clustering is not based on age due to the distribution comparison with the last PCA showing model prediction. We can observe the difference. The first PCA shows messy clustering based on age. We can assume based on the award categories and data it would show a strong influence on the clustering being the dominant feature. In the case where the first PCA doesn't match the last we can say the data is messy.

# CONCLUSION

We were able to successfully employ different techniques and methodologies to build a predictive model that would be able to predict which person based on their features such as age, race, origin where able to win the different types of awards, using our given dataset. We cleaned our data to remove duplicates and treated our missing values by filling them. With this newly cleaned dataset we were able to prove the hypotheses were all correct and visualize our findings to support this. Furthermore, we were able to split our data set for training and thus were able to generate a predictive model and view our confusion matrix. In addition to this we went ahead to improve models by feeding in more data and finding the best set of estimators where necessary and make use of adjustments made to max depth where appropriate to improve our model. We evaluated the models and discussed their performance. We were able to improve the test scores we got from our initial models; this was achieved in our improve model by using web scrapping ' to pull the gender' into our dataset. However, to further improve our test score we could have web scrapped more columns while training our model to improve our accuracy.

# REFERENCES

1. <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>
2. <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>
3. <https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html>
4. <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>
5. <https://medium.datadriveninvestor.com/increase-10-accuracy-with-re-scaling-features-in-k-nearest-neighbors-python-code-677d28032a45>
6. <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>