# CO3093/CO7093 - Big Data & Predictive Analytics

## CW Assignment

## Classification & Clustering

## Assessment Information

| | |
|---|---|
| **Assessment Number** | 2 |
| **Contribution to overall mark** | 70% |
| **Submission Deadline** | Friday 19 March 2021 at 5:00 pm |

## Assessed Learning Outcomes

This second assessment aims at testing your ability to

- carry out data cleansing and visualization
- develop a classifier and evaluate its performance
- perform appropriate and justified clustering of the data
- communicate your findings on the data

## How to submit

For this assignment, you need to submit the followings:

1. A short report (about 8 pages in pdf including all the graphs) on your findings in exploring the given dataset, a description of your model and its evaluation, a description of your clusters and its justification, as well as any decisions or actions that may be taken following your analyses.
2. The Python source code written in order to complete the tasks set in the paper. You should submit the Python code file, say emt12_solution.py, emt12_solution.ipynb, or if you are in a group, group1_solution.ipynb for your solution to the given problem.
3. A signed coursework cover – this should include the names of all the students involved in the work submitted.

Please put your source code, report and signed coursework cover into a zip file CW2_YouremailID.zip (e.g., CW2_emt12.zip, or CW2_Group1.zip) and then submit your assessment through the module's Blackboard site by the deadline. Note that to submit, you need to click on the Coursework link on Blackboard and then upload your zipped file.

## Problem Statement

Consider this dataset Oscars-demographics.csv, which can be downloaded from Blackboard. The given dataset contains records about the race, religion, age, and other demographic details of all Oscar winners from 1927 to 2014 in various categories such as best actor, best actress, best supporting actor, best supporting actress, and best director.

**Objective:** Using the given dataset, 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 that may make business sense of the movies industry.

## Exploring the data

Your first task is to prepare the data and carry out data munging or cleansing, bearing in mind the question you would like to answer. Namely, what is the impact of country of origin, race, religion, age in winning an Oscar award? Address the following questions:

## 1 Part 1 - Building up a basic predictive model

Load the dataset, and consider the subset of the `dataframe` formed by the following columns:

```
cols = ['birthplace', 'date_of_birth','race_ethnicity', 'year_of_award', 'award']
```

In this section, we will only analyse this subset of the given dataset.

1. **Data cleaning:** Using `pandas`, show the first 3 rows of the subset. Then, display all the distinct values for the column award in the entire subset.

   If you have a closer look at the entire subset, you will see that there are some inconsistencies on the way the birthdays have been recorded and that for some rows, the country of origin is missing. Add a new column `ldob` to your current `dataframe` to record the length of the date of birth for each row; then show the distinct values in the column `ldob`. Write the following functions:

   - Assuming that a year that has two digits is a twentieth century one, write a function that will re-write a given unclean date of birth to a clean one with the format 'Day-Month-Year'.
   - Assuming that any place of birth ending with two characters (e.g., `Los Angeles, Ca`) is in USA, write a function that will add the country of birth to those rows that are missing the country of birth.
   - Use the two functions you have defined to clean the columns `birthplace`, `date_of_birth` and add a new column `award_age` to your `dataframe` to record the age of the individual when she or he received the award. This new column is essentially the difference between the year of award and the year of birth. Using the column `birthplace`, add a new column `country` that records solely the country of origin. This means a birth place such as `'City of New York, USA'` will become only `'USA'`.

- Check the resulting `dataframe` for missing values and treat them as appropriate. Check for duplicates and treat them as deemed necessary.

2. **Data exploration:** Carry out a data exploration using appropriate plots to identify patterns or trends in the data. Note that we have few numerical variables in the current subset of data we are working on. Nonetheless, we need to assess the impact of the predictors (age, race, and country of origin) on the outcome (award). Use graphs to prove or disprove the following hypotheses:

   - Most Oscar winners are from USA.
   - Most Oscar winners are white.
   - Best Directors tend to be older than best Actors or Actresses.

   **Hint:** Check for distinct values in categorical data and their frequencies. If there are too many distinct values (levels), then you may want to reduce the number of levels by grouping some of the detailed levels. This could be the case for the country of origin in this dataset.

3. **Model building**. Note that age is a numerical variable. Discretise the age by using buckets. For example, we can form the following buckets:

   - Bucket 1: `age < 35`
   - Bucket 2: `35 ≤ age < 45`
   - Bucket 3: `45 ≤ age < 55`
   - Bucket 4: `age ≥ 55`

   Update the `dataframe` accordingly and build up a model that predicts the award type based on age, race, and country of origin. Split the data into a training and test sets, build the model and show the confusion matrix. Evaluate your model and discuss its performance.

## 2 Part 2 - Improved 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.

## Marking Criteria

The following areas are assessed:

1. Cleansing, visualizing, and understanding the data **[35 marks]**
2. Building up and evaluating the predictive model **[15 marks]**
3. Building up and justification of your clusters **[15 marks]**
4. Coding style **[15 marks]**
5. Writing the report (up to six pages) interpreting the results. **[20 marks]**

Indicative weights on the assessed learning outcomes are given above. The following is a guide for the marking:

- **First++ (≥ 90 marks)**: As in **First+** plus a classification model with excellent performance, excellent justification and visualisation of the clusters and a report of professional standards.

- **First+ (≥ 80 marks)**: As in **First** plus a comprehensive coverage of data cleansing techniques leading to a classification model of high performance and a well-structured and maintainable code usefully using functions.

- **First (≥ 70 marks)**: As in **Second Upper** plus well-justified models by the data exploration and a concise and well-structured report containing any decisions that may be recommended.

- **Second Upper (60 to 69 marks)**: A good coverage of data cleansing techniques exploring the dataset, a good visualisation of the clusters, a predictive model with an appreciable accuracy with a rationale behind it, a working code and a well-structured report on the results obtained from the dataset.

- **Second Lower (50 to 59 marks)**: Some techniques used for data cleansing are overlooked, a predictive model partially justified with an appreciable accuracy, a working clustering, a partially commented code with very few functions, and a narrative of the findings about the dataset with few deficiencies.

- **Third (40 to 49 marks)**: Essential data cleansing techniques are covered, a predictive model is given with some justification, a working but basic block code with no clustering, and a written report describing some of the work done.

- **Fail (≤ 39 marks)**: Not satisfy the pass criteria and will still get some marks in most cases.

- **None-submission**: A mark of 0 will be awarded.

**N.B. Make yourself available for presenting your work after submission, meaning from the week starting on 22 March 20121.**

Last Updated 21 February 2021 by Emmanuel Tadjouddine