

## COMP615 – Foundations of Data Science

## Lab 4: Feature Selection

This Lab focus is to apply two feature selection methods suitable for categorical variables: Chi-Square Method and the ANOVA F-test Method. Both methods use the `SelectKBest()` class from the [Scikit-learn](#) module. The feature selection scores of all features are computed and the outcome is presented based on these scores.

**The Dataset**

In this lab, you will work with two datasets that you have been introduced to during this course: The ‘*Car Evaluation Dataset*’ and the ‘*Breast cancer Wisconsin Dataset*’.

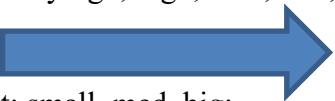
**Download Files**

To start with, download lab 4 files: The two datasets (`car_evaluation.csv` and `breast_cancer.csv`) and the Python code file (`COMP615_Lab4_FeatureSelection.ipynb`).

**1. Application of the Chi-Square method on the ‘*car evaluation dataset*’**

The Car Evaluation dataset contains examples with the structural information removed. The target attribute (*class*) directly relates to the six input attributes namely:

- **buying**: veryhigh, high, med, low;
- **maint (maintenance)**: veryhigh, high, med, low;
- **doors**: 2, 3, 4, 5more;
- **persons**: 2, 4, more;
- **lug\_boot** (luggage boot: small, med, big;
- **safety**: low, med, high.



Both **doors** and **persons** will also be treated as categorical variables

**Class** values are categorised as:

- unacc (unacceptable),
- acc(acceptable),
- good,
- vgood (very good)

**Task 1: Create a Data Frame and Explore the Data**

Create the data frame (*mycar*), check the dimension, and preview the data frame. How many features are presented in your dataset? What is the number of instances?

The data frame column names are not informative. Using the metadata, add meaningful names to the columns. Then review the data frame and check for the possibility of missing values. Are there any missing values in your dataset?

## Task 2: Perform Exploratory Analysis

**2.1** Since all the features are categorical, contingency tables and bar charts can be used for the analysis. Starting with the *Class* feature, explain the distribution of the car condition.

**2.2** Break down the features by the *Class*. The imbalance noticed in the distribution of the *Class* features becomes obvious here. *Two-Passenger* and *Low-Safety* categories did not have any observations in the ‘*acc*’, ‘*good*’, or ‘*vgood*’ classes. So, irrespective of the other features, cars with a passenger capacity of two, and cars with poor safety are always in the ‘*unacc*’ class category.

**Q1:** Perform a similar ‘breakdown’ analysis for the rest of the features. Provide the plots and explain your findings.

## Task 3: Data Preparation

Subset your data into ‘input’ and ‘output’ attributes. Perform feature transformation to ‘Encode’ categorical variables. This can be done through ordinal encoding. Finally, split the data into test (0.3) and train sets.

## Task 4: Chi-Square Method

The Chi-square method works by computing the chi-squared test value which measures the independence between categorical variables. Having three or more levels for the predictor values, we choose this method to find the best sets of features available in the car dataset.

Obtain Feature scores using Chi-square and visualize the results using the bar plot. Explain your findings.

**Q2:** Provide the plots and explain your findings.

## 2. Application of ANOVA F-scores Method on the Breast Cancer Dataset

**Task 5: Data Exploration** Perform initial data exploration to learn about the dataset (as in lab 10). Investigate the possibility of having an Imbalanced Target (Output).

**Task 6: Data Preparation** Set the Input (X) and Target (y) variables, and perform encoding on categorical data values. Finally, split the data into test (0.3) and train sets.

## Task 7: ANOVA F-test Method

Perform the ANOVA F-test to compute scores for each feature and to check how well it differentiates between the classification categories, i.e., how well the feature can discriminate between the classes. The features are ranked based on their ANOVA F-scores.

Obtain feature scores using ANOVA F-scores and visualize the results using the bar plot. Explain your findings.

**Q3:** Summarise the ANOVA F-scores results in a 30 by 2 table with the first column holding the scores and the second column specifying the features’ names. Your table should be sorted based on F-scores in descending order. Present the first 10 rows of your table.

**Q4:** Using the outcome of Q3, choose the top 6 features. Explain the distribution of their Class/features using proper visualization (e.g. factegrid or paired plots)

**TODO Tasks for Submission (1%)**

**Submit Q1:Q4 Answers**

Please convert your Jupyter Notebook to PDF format before submitting it, ensuring that the output of your code for each task is included.