A picture containing clipart

Description automatically generated

Olympic\_Medal\_Predictions

Final Group Project

Maimoona I Elahi |Data Analytics & Visualization | January 23-29, 2022

# Assigned Task for the Final Group Project

Using Machine Learning build a prediction model of winning a medal based on user input of biometric data.

# Machine Learning Prediction Model Process



The preliminary dataset downloaded from <https://www.kaggle.com/datasets> as

1. **‘athlete\_events.csv’** It includes columns that describe medal or no medal for participants. Using only 2016 Olympics dataset and created a prediction model.
2. Used **‘cleaned\_olympic\_data.csv’(1964-2016)** and missing data in csv for medal column had NA where there was no medal won, replaced it with “no medal” instead and imputed NA values from other columns ‘age’, ‘height’ and ‘weight’ by using Iterative Imputer.
3. Using make\_pipeline() method with OneHotEncoder created different prediction models like ‘RF’, ‘LR’, ‘LDA’, ‘DTC’ ‘KNN’, ‘GNB’ and ‘SVM’ to analyze model performances on 200 rows.

**Issues**

OneHotEncoder took forever to run even for limited number of rows.

**Resolution**

Team discussion and independent research I changed the OneHotEncoder to LabelEncoder and it runs without delay in time and best fit for prediction the model.

1. Using **‘clean\_regions\_olympic\_data.csv’** created two prediction models based on season ‘summer’ and ‘winter’

**Summer & Winter Models**

Following steps have been performed

* Label Encoder
* Train, test & split data on ‘is\_medal’
* Random Forest & Decision Tree Classifier used
* Using make\_pipeline() performed Imputation, Standardscaler, on season models.
* Did Cross Validation on winter and summer data with accuracy score
* Print the classification report

**Issues**

With RF and DTC models recall score came low for both summer and winter models.

**Resolution**

Used

# that describe medal or no medal for participants

# · Begin to build out prediction model with Machine Learning

Refine prediction model with Machine Learning



# Reasons for using SMOTE & Undersampling

* The combination of oversamppling the minortiy(Medal) class and undersampling the majority(no medal) class were used to improve the performance of our model.
* The increase in performance classifier is evident from the results of classfication reports.
* This is also proven by the following research. <https://arxiv.org/abs/1106.1813>
* The decision to use Decision Tree classifier was based on the original results. Decision Tree Classifier performed better on both the winter and Summer Datasets prior to using SMOTE and Undersampling.

# Reasons for not using a multinomial appoach

* Although using a multinomail approach might improve the classifier results. However, for this dataset it is suspected that the improvement in performance would not be as significant as can be achieved using SMOTE and Undersampling.
* The reason for this is the fact that the majority class is still no medal and there are almost 27 times more no medals than the medal classes(Gold, Silver and Bron

In [ ]:



​