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’ and ‘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 **‘clean\_regions\_olympic\_data.csv’** 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 predicting the model.

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

**Summer & Winter Models**

Following steps have been performed

* Label Encoder was used for the aforementioned reasons
* Data was split into summer and winter dataframes
* Train and test split was done for each dataframe
* Random Forest & Decision Tree Classifier used on each dataframe
* Using make\_pipeline() performed Imputation, Standardscaler, on season models.
* Did Cross Validation on winter and summer data with accuracy score
* Classification report was used to assess model performance

**Issues**

With both models and for both dataframes the recall score (metric used for evaluation) was low.

Text

Description automatically generated with medium confidence

Text

Description automatically generated

**Resolution**

A technique to oversample the minority (is\_medal) class and undersample the majority (not is\_medal) class was used to improve recall score. Significant improvements were obtained.

A picture containing table

Description automatically generated

A picture containing table

Description automatically generated

# **Reasons for using SMOTE & Undersampling**

* The combination of oversampling the minority (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 classification 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.