Third part: Validation of the results and project review

Introduction

In this part we summarize our project by reviewing the results and the steps we have made.

An analysis of the obtained results versus the goals that had previously been set

One of our goals was to predict, with a high accuracy, if the blue team will win or not only using the important attributes in our dataset. Our second goal was related to describing the data set as a whole by determining global characteristics and dividing examples into groups.

We can say that each of the tasks was completed successfully.

Here we attach the result of our metrics used to guess the result of blue time:

GridSearchCV Classifiers (OrderBy Accuracy)

Classifier	Accuracy	Recall	Precision
K-Nearest Neighbours	72.2 %	70.37 %	72.97 %
Decision Tree	72.1 %	71.5 %	72.29 %
Logistic Regression	72.09 %	72.09 %	72.02 %

The results can be used in practise, for example (although we don't support it) for betting. That is, we can bet which time will win and we can bet around a 72% of accuracy.

A review of the data mining process

What have been done

The data mining task for our dataset is prediction. We use classification to predict the nominal target attribute blueWins.

The steps executed have been these:

- Description of our dataset: Understanding of the dataset (description of the attributes, origin of the dataset, etc.), classify attributes depending on their type (nominal or numerical)
- Exploratory data analysis (EDA): dependencies between variables (correlation), outliers and distribution of the attributes.

- Model creation: We prepare the dataset removing columns that not serve for the model and normalizing values. Once we prepare the dataset, we divide the data set into two subsets: training set and test set.
- Apply Data Modeling algorithm: Logistic regression, K-nearest neighbours and decision tree algorithm.
- Evaluation of the algorithms: Cross-validation calculating these metrics: accuracy, recall and precision.

What could have been possibly done better

If we want to be fussy, when we were doing EDA we could make a comparison between red and blue time for each attribute establishing in the same box histograms and distribution plots to compare better both set of attributes.

Discussion of importance of the results for further exploration

What additional data mining task could be performed for the dataset

Description of the data set (clusterization and segmentation). These techniques describe the data set as a whole by determining global characteristics and dividing examples into groups.

We can apply algorithms like COBWEB that generates a hierarchical division of the dataset in the form of a tree, K-means an Iterative clustering based on distance and finding centroids of the subsets of samples or EM that tries to find the subsets by assuming distributions and estimating probabilities.

Also association rules, but for that we have to transform attributes with a finite set of values to make them nominal attributes.

Results of which steps of the data mining process could be reused

On Data Understanding step and Data preparation, we can reuse it when we want to do another data mining task or use as target value other attribute.