CIS 6610 - Machine Learning (Spring 2024)
Project 2 - Due Date: Sunday, 3/24/2024 11:59 pm

This project asks you to implement a machine learning system to solve a classification/regression problem.

You will use the **NBA Rookie Stats** dataset to predict if a player will last over 5 years or not. The details of the dataset and the questions you will try to answer are as follows.

1 The Data

You are given the NBA Rookie Stats dataset (provided at data.world). This dataset is accessible on our Canvas course site.

This dataset totals 21 columns and 1340 rows. The 21 features are play name (Name), games played (GP), minutes played (MIN), points per game (PPG), field goals made (FGM), field goal attempts (FGA), field goal percent (FG%), three points made (3PM), three point attempts (3PA), three point percent (3P%), free throws made (FTM), free throw attempts (FTA), free throw percent (FT%), offensive rebounds (OREB), defensive rebounds (DREB), rebounds (REB), assists (AST), steals (STL), blocks (BLK), turnovers (TOV), and target (TAR).

Each row in the table represents a player's *rookie statistics*, stats of that player's first season.

Out of these 21 attributes, the last attribute is the class attribute for which your system will predict about. It is a Boolean attribute, where "o" means the career length of the player is less than 5 years, and "1" greater than or equal to 5 years. The other 20 attributes are the features your models may consider. Out of these 20, there is 1 text attribute and 19 numerical attributes.

2 The Tasks

You are to explore the following classification/regression models to predict the target value and report the comparison of their performances based F_1 scores.

1. K-nearest neighbors [1]
2. Random forests [2]
3. Logistic regression [3]
4. Artificial neural networks [4]

3 The Questions

Here are the questions you need to address eventually in the project report.

1. When you prepare the data for training the models, did you discover any attribute to remove or any new attribute to add? If you did, discuss the choices.
2. Normalizing (a.k.a., scaling) features is desirable for distance-based models, e.g., k-nearest neighbors. Did you try feature normalization for some of the models? If so, talk about if any improvement.
3. Regularization is a common practice to battle overfitting. How is varying the penalty parameter in logistic regression affect the performance F_1 score on testing? (The logistic regression penalty parameter may be 'none', 'l1', 'l2' or 'elasticnet'.)
4. These models have hyperparameters. When training, experiment using GridSearch to select hyperparameters for your models. What are the best hyperparameters among those you tried?
5. Which model you experimented with gives the best F_1 score on testing?

4 The Requirements

1. Your project should be in Python 3 and you are free to use any Python package to help you develop your programs.
2. When separate the dataset for training and testing, use 80% randomly selected for training and the rest for testing. During training, use 10-fold cross validation to pick the best model learned. Finally, it will be tested on the testing set.
3. Your project should provide the solutions and answers to the aforementioned tasks and questions.

5 The Deliverables

Zip the following to [your-last-name]_Project2.zip and submit to Canvas.

1. A directory that contains all your Python programs that are may be .py or .ipynb.
2. A README file that contains instructions to run your Python programs.
3. A PDF report that describes your results on these two parts. In particular, it must address the aforementioned questions.

[1] sklearn.neighbors.KNeighborsClassifier
[2] sklearn.ensemble.RandomForestClassifier
[3] sklearn.linear_model.LogisticRegression
[4] sklearn.neural_network.MLPClassifier