# Runners Injury Prediction

## Abstract

Staying injury free is a major factor for success in sports. Our purpose was to use machine learning for the prediction of injuries in runners. I worked with Runners Injury data provided by Kaggle website.

## Design

The project will determine whether a Runner has injured or not. The data provided by Kaggle website for first phase, and I started to data cleansing , explore the dataset, then I visualize the data, after that I build three models to predict then I evaluated each model. Dataset is imbalanced so I will use SMOTE Technique to fix the data imbalance problem.

## Dataset

The data set is expected to have around 42766 rows with 13 columns.

## Algorithms

*Models*

- k-nearest neighbors algorithm (k-NN)
- logistic regression(LR)
- Extreme Gradient Boosting(XGBoost)

*Model Evaluation and Selection*

The entire dataset of 42766 records was split into 70 train /30 test.

Below is the evaluation of each model

- Logistic Regression scores:

- F1 0.86

- Accuracy 62%

- precision

- recall 0.87

- xgboost:

- F1 0.62
- Accuracy 86 %
- precision
- recall 0.63

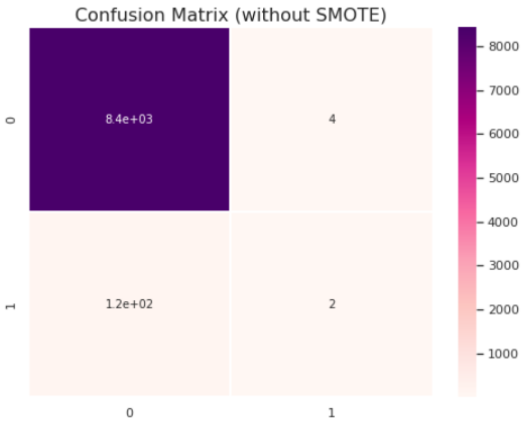
- KNN:

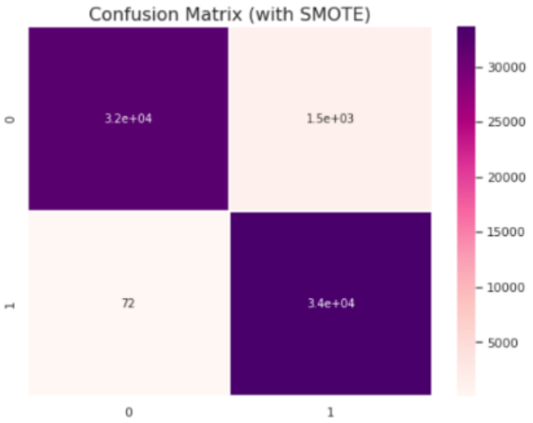
- F1 0.98
- Accuracy 98 %
- precision
- recall 0.96

## Tools

jupyter nootbook with some DS libraries such as pandas and numpy to maipulate the data, matplotlib and seaborn to visulalize the data and scikit learn, Imblearn and XGBoost for model for realted libraries.

## Communication


Confusion Matrix (without SMOTE)


Confusion Matrix (with SMOTE)