INTRO :

* Présentation du jeu de données : 119 sportifs issus de 8 sports différents : Baseball, Basketball, Bowling,Rowing, Rugby, Savate, Swimming and Synchronised Swimming.
* In order to reduce the risk of injuries among the athletes and their downtime, the CREPS coaches had developed a scored protocol test called TM2S. Fourteen exercises compose the TM2S, split into 4 categories :

• Lower mobility, scored on 8 points;  
• Upper mobility, scored on 4 points;  
• Lower stability, scored on 6 points;  
• Upper stability, scored on 2 points.

* Certain exercises have to be performed for the right side and the left side, which leads to a global score of 20 points. The TM2S is binary scored; points are attributed by the fitness coach depending on the athlete’s ability to complete the exercises.
* Each athlete had completed the TM2S test at the beginning of the 2020-2021 season.
* Among the 14 TM2S variables, we had small groups of exercises and 3 asymmetries. We remove the asymmetry variables from the dataset because athletes without asymmetries are more often injured than others. This is contrary to the medical literature ; therefore, the asymmetry variables are ill defined.
* To avoid an excessive number of variables, we group some of them into 4 quantitative ones : “Lower Mobility Sum” (LMS), “Upper Mobility Sum” (UMS), “Pelvic  
  Stability Sum” (PSS), “Scapular Sum and Core Stability” (SSCS).
* Finally, we also consider other qualitative variables that might have an influence on injuries, such as the Sport, Past injuries, Current pain, and the fact that an athlete is new / old in the CREPS structure.

MAIN GOALS OF THE CREPS PROJECT :

* The main goal was to construct a statistical model to understand the underlying mechanisms of injury in high-level athletes.
* We wanted to extend this model to predict targeted injuries based on 3 body areas: Lower Limbs (LL), Upper Limbs (UL), Torso and Spine (T&S).
* Find the variables of importance to explain injuries
* Estimate the unavailability of athletes in case of injury for each body area
* Provide CREPS staff with an interpretable decision support tool

PRELIMINARY DATA PROCESSING :

* First, we reorganize the entire database: we merge the different CREPS files and put them in a format interpretable by R.
* Due to the poor size of the database, we choose to transform response variables into binary ones (Injured / Not injured). We also defined 3 body areas to have more precise predictions :

• Lower Limbs injuries (LL), which groups the entire area extending from the top of the thigh  
to the end of the feet;  
• Upper Limbs injuries (UL), which groups all injuries located at the level of the arm (from the  
collarbone to the fingertips);  
• Torso and Spine injuries (TS), regarding torso, bust, lower back and spine injuries.

* In order to determine if the variables previously described are relevant to predict athletes’ injuries, we use the Chi-square test of independence. Indeed, if an explanatory variable is independent of an injury, it is useless in the model.

MODELS USED AND EXPLANATION :

Once the data processing step is done, we choose to use different statistical models and machine learning algorithms to predict athletes’ injuries. First, we highlighted that the dataset contains a significant number of variables. This is a problem for 2 reasons :

* in terms of computation time: the more variables we have, the longer the algorithm will take to compute
* from a prediction perspective: with fewer variables, we increase the bias but decrease the variance, which leads to better predictions. Moreover, a simpler model will be more interpretable.

Therefore, we select variables thanks to a LASSO penalized regression : we constrain the l1 norm of the estimator to have a model more understandable.

We also use an alternative method to reduce dimension with the clustering of variables. We group variables which are strongly related to each other, and thus create homogenous clusters. Where LASSO cancels parameters (which means that we do not take them into account in our final model), ClustOfVar arrange variables into meaningful structures and finally select one synthetic metavariable from each group. Then, each variable of each group contributes some information, so ClustOfVar build models more complete than LASSO.

To compare these 2 methods, we combine them with a logistic model, which is the most commonly used for the prediction of a binary variable. Predictions are better with LASSO than with ClustOfVar, so we decide to combine only LASSO with other machine learning algorithms.

In addition to the LASSO logistic regression, we choose to use 3 other algorithms to predict the binary variable Injury :

* Classification trees, which are an easy-to-understand tool to predict injuries. Since one of our goal is to provide CREPS coaches with a decision support tool, trees seem to be the better choice. They present the information in a visual way, are easy-to-read and thanks to them, coaches can determine quickly if an athlete has a high risk of injury. However, trees are instable, have a lack of robustness (results highly depend on the train and test sets) and overfit the dataset.
* We try to improve the disadvantages of classification trees by reducing the variance thanks to random forests. It also improves predictions thanks to the bootstrap procedure. If the predictions should be better than the trees one, the results would not be visually available. For this reason, we first presented trees to our CREPS supervisors to explain them how these type of methods works. Therefore, the were able to understand better the random forest algorithm.
* Finally, we use a simpler method which is well adapted to classification problem and easy to implement: the Naïve Bayes classifier. The fundamental hypothesis of this method is that the effect of an attribute on a given class is independent of the values of the other attributes. It is a strong hypothesis since most of our variables are not independent. However, Naives Bayes has proven its efficiency on hard-to-use datasets, so it seems adapted to our case.

COMPARISON CRITERIA

* In binary classification, the most widely used comparison criterion is the ROC curve, and more particularly the AUC (area under the curve). The larger the AUC, the better the model. A ROC curve represents the distribution of individuals in the test sample who are well classified compared to those who are bad classified. A model that predicts at random has an AUC of 0.5, and its ROC curve corresponds to the red diagonal.
* However, the AUC is highly dependent on the test sample. For this reason, we implement a Monte Carlo cross-validation. The data set is separated 200 times into training and test samples. We learn the model on the training sample and we compute the AUC on the test sample.
* as a second comparison criterion, we created 3 groups of predictions:
  + significantly small predictions, which correspond to "not injured"
  + significantly large predictions, which correspond to "injured"
  + and finally, the intermediate predictions where it is considered that the prediction is not decided
* The closer the prediction is to 0.5, the more uncertain the algorithm. But an algorithm that says "I don't know if the athlete will be injured" is totally useless. We have therefore penalized the methods with a lot of "uncertains"

MODELS SELECTION

* First, we compared the 4 main algorithms in terms of AUC. CART trees are clearly bad compared to others so we decided to abandon this algorithm.
* On the right picture, we display the number of “uncertains” for random forest, linear model and naive bayes, for the prediction of LL injuries. In this example, the linear model is clearly the best.
* The first three boxplots represent the correct predictions for LL Injury and the others represent the correct predictions for No LL injury.

R INTERFACE FOR COACHES

Our main objective was to build a statistical model to predict injuries; however, this model must be reusable a posteriori by CREPS coaches. The purpose of the project is not really the report, but the creation of a usable tool for the future.

Thus, the most optimal way we found to execute code and present the results clearly was to create an R interface. Then we thought about the qualities that our interface should have

* First, the CREPS tutors should not have to code to use the interface
* Second, its use must be intuitive and the results must be clear
* Then, each athlete is different so the interface must fit all profiles
* Moreover, the interface should change injury probabilities in real time as the user changes the athlete's characteristics.
* Last but not least, as the interface is intended to be used in the future, its accuracy must improve as the excel files are filled. This is why these excel files must be easy to fill in for coaches.

I will now quickly show you what the interface looks like and how to use it.

On the left side, coaches can enter the values of their athlete. They can choose the sport, adjust the sliders etc.

On the right there is a summary of the characteristics of the athlete, the probability of being injured at each area of the body and the downtimes for an injury in this area with a pie chart for all athletes and one specifically for athletes from the chosen sport.

Finally, let’s show the interactivity of the interface. Look at the probability of being injured at UL, it is 0.68. If we say that our athlete has never been injured at UL, this probability drops to 0.3

Thus it is possible to determine the impact of the various variables on the risk of injury.

CONCLUSION

Our main goal was to predict athletes’ injuries thanks to the TM2S score and other significant  
variables. With machine learning algorithms and statistical models, we were able to predict each  
type of injury correctly between 60 and 80% of the time, which is very satisfactory.

We also wanted to provide coaches with an easily understandable tool that can be used without particular mathematical or computational skills. When we presented the Shiny interface to our supervisors, we taught them how to position the different cursors and to interpret the resulting probabilities. Hopefully, they will be able to use the Shiny interface on new athletes.

We managed to create a decision support tool ; thanks to the probabilities computed and the empirical observations of the coaches, the coaches will be able to make decisions to decrease the risk of injury of an athlete (such as corrective routines, a diminution of  
training volume or an adaptation of some motions).

Finally, we want to discuss the difficulties we had to overcome. First of all, the size of the dataset  
is very limited compared to the number of variables so it was hard to  
find efficient models to predict injuries. Moreover, most of the variables we studied are qualitative, and are harder to exploit than quantitative ones.

Another problem we faced concerns the duration of downtime. Our tutors wanted us to predict it ; however, a basic statistical analysis showed that it was not possible. For example, some athletes had several injuries on the same part of the body, but  
their unavailability was not the same for each injury. In this case, we cannot predict the value of  
the duration of downtime.

One last point concerns the bias naturally present in the TM2S scores and exercises :

* Exercises may not be independent two at a time
* For the scores, we used the results obtained at the beginning of the 2020-2021 season; corrective routines are then not taken into account