Methodology

The first phase is to collect the data that we are interested in collecting for pre-processing and to apply classification and Regression methods. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.

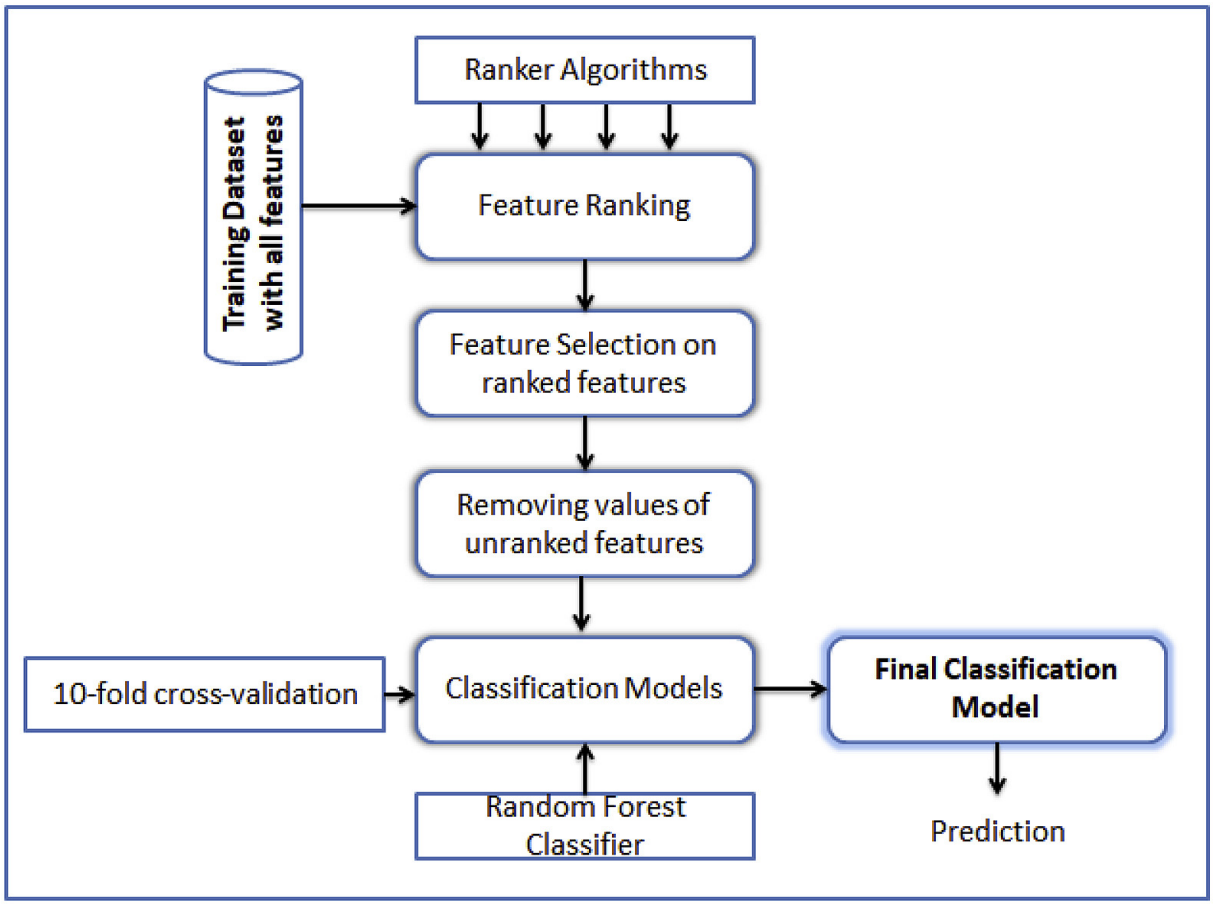
Real world data is often incomplete, inconsistent, and lacking certain to contain many errors. Data preprocessing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing.

For pre-processing we have used a standardization method to pre-process the UCI dataset. This step is very important because the quality and quantity of data that you gather will directly determine how good your predictive model can be.

We first check whether all features are important for the classification task. This is done using several feature ranking algorithms. Then based on the ranking, a subset of the top-ranked features are selected.

Finally, the Random Forest algorithm is applied on the selected features to train and construct the final model. We apply 10-fold cross validation during the model training. Notably, cross validation is a method to evaluate a predictive model by partitioning the original sample into a training set to train the model, and a validation/test set to evaluate it.

In 10-fold cross validation, the original samples are randomly partitioned into 10 equal sized subsamples, and among these subsamples a single subsample is retained as the validation data for testing the model, while the remaining 9 subsamples are used as training data.


Ranker Algorithms
Training Dataset with all features
Feature Ranking
Feature Selection on ranked features
Removing values of unranked features
10-fold cross-validation
Classification Models
Final Classification Model
Random Forest Classifier
Prediction