UNIVERSITY OF RIJEKA
FACULTY OF ENGINEERING
SEIP LABORATORY

Data Mining - Data Preprocessing

Samir Jugo, Nikola Lacković, Jovana Brestovac – feature selection
Mateo Hrastnik, Dejan Ćućić, Igor Opačak – over/under sampling

Summary

•Data mining

•Feature selection – get raw data from csv file, feature selection algorithms on data, store results in new csv file.

•Over/Under sampling – get reduced data from csv file, over/under sampling methods on data, stores results from sampling in new csv file.

•Project Goals – reducing data metrics, sampling at specified percentage

• Project Requirements:

Python (scipy, scikit-learn, six, python-dateutil, pandas, numpy), CSV

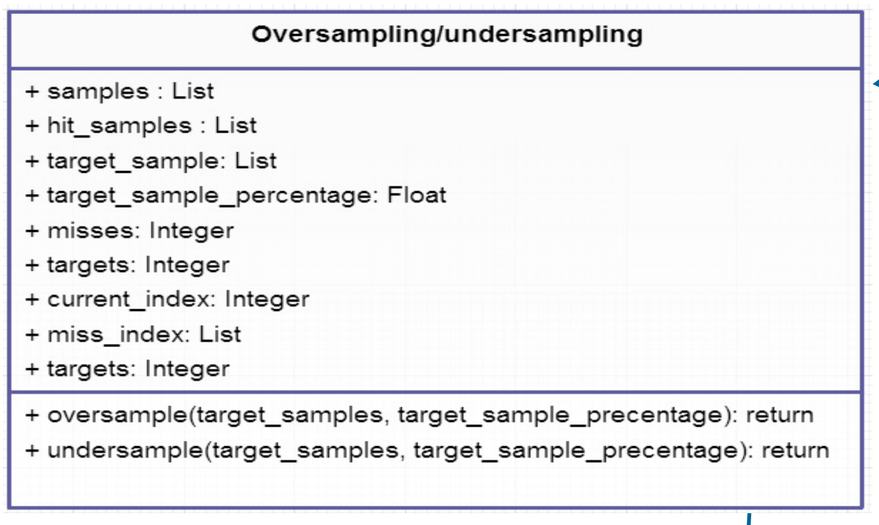
Developed Component

Input / Output for Feature selection

• Input: Raw data from CSV(PDE_R3.csv)

•Output: Data with reduced metrics

Class diagram of Over/Under sampling



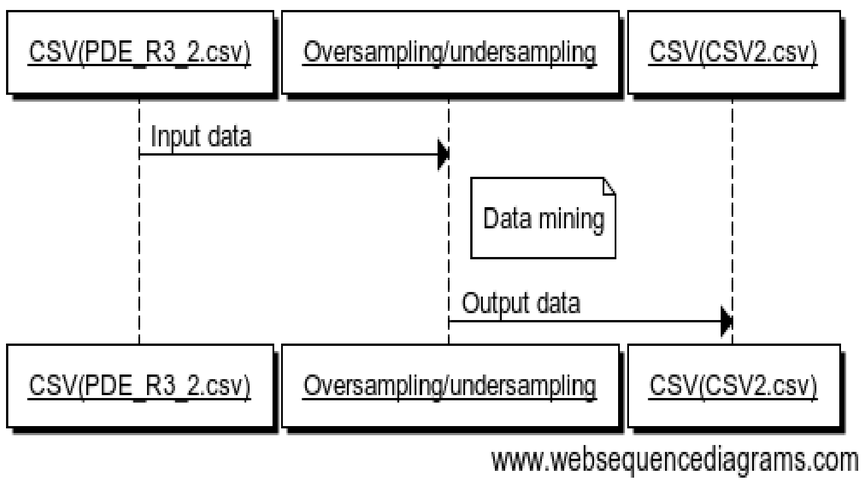
•Class used in project

Input / Output for Over/Under sampling

• Input: Reduced data from CSV(PDE_R3_2.csv)

•Output: Uniform data distributions of two classes of files

Sequence diagram for Over/Under sampling

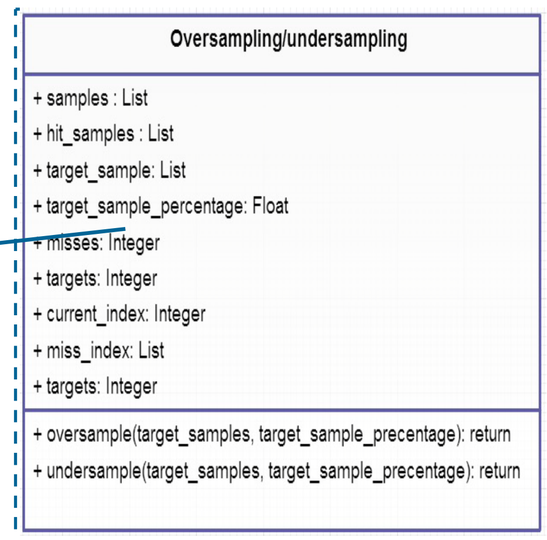


Overall Architecture

Feature selection - functions

Ispis(): Function for printing raw data from csv file
IspisFS(featureIDs): Function for printing data with reduced metrics
UnivariateFS(featureIDs): Function for univariate feature selection
Tree_basedFS(featureIDs): Function for tree-based feature selection
L1_treeFS(featureIDs): Function for L1 + tree-based feature selection
FeatureSelection(featureIDs): Function for feature selection menu
PoljeUDat(featureIDs): Function for creating new csv file with data with reduced metrics
main(): Main function - menu

• Architecture components: CSV files, Feature selection algorithms



• Architecture components: CSV files, Over/Under sampling commands

Conclusion

• Experienced problems

• Learned Concepts:
  • Machine learning – data mining
  • Functionality of feature selection algorithms
  • Functionality of over and under sampling algorithms

• Future work improvements