Projects 2017/2019

Dec 2018 - Present

- Classification_skin_lesion_images_Py: classification of 7 types of dermatoscopic images. Under-sampling, Random Forests, Neural Network and Convolutional Neural Network in Python.

  Details: 10 015 images + metadata

March/April 2019

- Classification_DeceptiveReviews_Py: two supervised binary classification problems, deceptive/truthful and positive/negative, are analyzed. Logistic, Support Vector Machine and Naïve Bayes are applied with personalized pipelines.

  Details: 1600 reviews. https://www.kaggle.com/rtatman/deceptive-opinion-spam-corpus

Feb 2018 – June 2018

- Pattern_workabsenteeism_R: PCA, MCA, Clustering. The performed clustering techniques were hierarchical (Ward, Average, Single, Complete and Centroid linkages) and partitional (K-means and K-medoids) with several numbers of clusters. The chosen distances were Spearman and Kendal.

  Details: 667 instances (34 employees), 20 features.

- Regression_GeneticProgramming_Java: no information about the dataset was provided during the project. After we knew that the problem consists in the prediction of human oral bioavailability of a new drug as a function of its molecular descriptors. In the data, there are other features not related to this problem.

  Genetic Programming with tree-based individuals. Implement crossovers, mutations, selections and elitism. Restrictions such as population size of 300 individuals and 300 number of generations.

  Details: 282 instances and 242 features (36 problem un-related).

- Marketing_Compaign_Classification_SAS: produce the highest profit for the next campaign, defining if the customer will accept the offer or not. Restriction: maximum of 9 features in the final model.

  Details: 1450 instances to train and 1855 to predict, 30 features.

- USACancerDeathRate_R: linear models to predict the cancer death rate per county with geolocation features. Assess all the theoretical assumptions (with appropriate tests when it was possible). Generalized Linear Model (logit and probit) to predict a

dichotomous target variable defined as follows: if the death rate is bigger than the second quartile then the variable assumes value 1, otherwise 0.

Details: 3047 instances and 34 features.

Sept 2017 – Jan 2018

- MonaLisa_withTriangles_GeneticAlgorithms_Java: reproduce Mona Lisa. Each solution has 100 triangles (each triangle represents a raster), 2000 generations and population size 25. Euclidian distance between solution and target (Mona Lisa image). Tuning initialization, selection, elitism, crossovers and mutations.

- Breast_Cancer_Diagnosis_Winconsin_R_Py: classification between benign and malignant breast cancer using clustering algorithms (K-means and K-medoids) in R and decision trees in Python. Principal components analysis.

  Details: 569 instances, 30 features

- Customer_Segmentation_SAS: segmentation of clients. High Performance (HP) Clustering using ABC with Global Peak Criterion and Hierarchical Clustering (HC) Principal Component Analysis. Self-Organizing Map (Kohonen net).

  Details: 10 000 instances, 28 features

Sept 2017 – June 2018

- DataWarehousing_BusinessInteligence_Projects: in the data warehousing part, it was built a data warehouse using SQL Management Studio and the ETL process using SQL Server Data Tools. In the business intelligence part, it was built a dashboard and reports in PowerBI. It is possible to access the reports in here: https://app.powerbi.com/view?r=eyJrIjoiMDcwMzc3MjItYmI2OS00ZTg3LWJiMTQtOTMzMmYwYzA4NjEyIiwidCI6ImU0YmQ2OWZmLWU2ZjctNGMyZS1iMjQ3LTQxYjU0YmEyNDkwZSIsImMiOjh9.