Assignment 3: Supervised Machine Learning

The goal of the assignment is to help you get familiar with supervised machine learning models in solving problems in real scenario. You will work on three kinds of tasks: regression, multi-category classification, and multi-label classification. For each task, learn and select two different models, and compare their performance on the given dataset. You will write a report to record what you did in the experiments.

Dataset: Student Portuguese Class Performance Data Set, altered version of URL.

Description: This data approach student achievement in secondary education of two Portuguese schools.

Training, validation and testing: 585 students data in assign3_students_train.txt will be used as training and validation data; 64 students data in assign3_students_test.txt will be used as testing data.

Attributes:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 edusupport - student receive extra educational support (nominal: 'school' (extra educational support from school), 'family' (from family), 'paid' (extra paid Portuguese classes) or 'no' (no

extra educational support))
17 activities - extra-curricular activities (binary: yes or no)
18 nursery - attended nursery school (binary: yes or no)
19 higher - wants to take higher education (binary: yes or no)
20 internet - Internet access at home (binary: yes or no)
21 romantic - with a romantic relationship (binary: yes or no)
22 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
23 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
24 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
25 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
26 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
27 health - current health status (numeric: from 1 - very bad to 5 - very good)
28 absences - number of school absences (numeric: from 0 to 93)
29 G3 - final grade (numeric: from 0 to 20, output target)

Task 1: Predict student final grading with regression models.
This is a regression task. In this task, you will predict student final grade with regression model. You can use all or a subset of resting attributes (1-28) as features, and predict the attribute 29. Models can be selected from, but not limited to, Linear Regression, Support Vector Regression, Decision Tree Regression, Nearest Neighbor Regression.

To do this task, you need to:

- choose two regression models,
- figure out the appropriate evaluation metrics,
- tune model parameters based on training data with 10 fold cross-validation,
- report their performance on testing data.

In the report, for each model, you need to write:
(1) model name.
(2) what parameter you have tried, and corresponding performance on training data.
(3) final performance of learned model on the testing data.
(4) how long to train the model on training data.
Also,
(1) what features do you choose to use, and why chose them?
(2) compare two model’s performance on this task

Task 2: Predict the student mother’s job with classification model.
This is a multi-category classification task. In this task, you will predict/classify student mother’s job. You can use all or a subset of resting attributes (1-8, 10-29) as features, and predict the attribute 9. Models can be selected from, but not limited to, Logistic Regression, Support Vector Machine, Decision Tree, Nearest Neighbor, Naive Bayes.

To do this task, you need to:

- choose two classification models,
- figure out the appropriate evaluation metrics,
- tune model parameters based on training data with 10 fold cross-validation,
- report their performance on testing data.

In the report, for each model, you need to write:
(1) model name.
(2) what parameter you have tried, and corresponding performance on training data.
(3) final performance of learned model on the testing data.
(4) how long to train the model on training data.
Also,
(1) what features do you choose to use, and why chose them?
(2) compare two model's performance on this task

Task 3: Predict what kind of extra educational support does the student receive.
This is a multi-label classification task. Student may receive extra educational support from school, or/and family, or/and paid classes, or none. In this task, you will predict this attribute (16) based on other attributes (1-15, 17-29). Models can be selected from, but not limited to, Logistic Regression, Support Vector Machine, Decision Tree, Nearest Neighbor, Naive Bayes. You may want to use One-vs-the-Rest strategy.

To do this task, you need to:

- choose two classification models,
- figure out the appropriate evaluation metrics,
- tune model parameters based on training data with 10 fold cross-validation,
- report their performance on testing data.

In the report, for each model, you need to write:
(1) model name.
(2) what parameter you have tried, and corresponding performance on training data.
(3) final performance of learned model on the testing data.
(4) how long to train the model on training data.
Also,
(1) what features do you choose to use, and why chose them?
(2) compare two model's performance on this task

Tools:
Scikit-learn for Python (https://scikit-learn.org/stable/index.html#), and weka for Java (https://www.cs.waikato.ac.nz/ml/weka/).
If you prefer to use other tools, please get permission from TA before doing assignments.

Coding:

Please feel free to add extra java/python files, functions, attributes to do training, validation, and testing. Your grading will be based on the output of running Assignment3Main. **Assignment3Main** is the main class for running your assignment, which **cannot** be modified. It calls **model_1_run and model_2_run**, where you first implement train the model on training data with your best parameter obtained in validation, and then evaluate and output its performance on testing data.

Grading:

Your submission will be graded based on:

1. Correctness of the implementation on three tasks (35%)
2. A clear report in **PDF** format (30%)
3. A good selection of evaluation metrics for three tasks (10%)
4. A good validation (parameter tuning) procedure (15%)
5. A reasonable selection of feature set (5%)
6. Learned model's performance on testing data (5%)

Submission Requirements

A zipped file package with the naming convention as "pittids_a3". For example, suppose the Pitt id is jud1, then the submission package should be jud1_a3.zip.

The file package should contain:

1. All the scripts/programs you used for this assignment. (**src folder**)
2. A clear report. (**pdf file**)
3. Your output of Assignment3Main. (This should also be included in the **pdf file**.)
4. The tool you used in the assignment. (This should also be included in the **pdf file**.)

Do not upload the assign3_train.txt and assign3_test.txt.