PROGRAMMING ASSIGNMENT 5

TAs : Bahar GEZİCİ, Nebi YILMAZ
Due Date : 08.01.2021 (23:00)
Click here to accept your Programming Assignment 5.

1 Introduction

In this experiment, you will implement a program to solve a real wold machine learning problem with a real world data. In this experiment, you will get familiar Python data science libraries such as; Pandas, NumPy, and Matplotlib. At the end, you will build 3 types of machine learning models.

2 BackGround (Technical Debt)

Technical debt (TD) is a metaphor used to describe the lack of software quality in the product. It provides a valuable indicator to track software product quality throughout development and maintenance. It defines the invariance of delayed technical development activities to receive short-term repayments. The consequences related to these skipped activities are accumulated in the software, and it is called as 'debt.' If there is too much technical debt that is accumulated, it causes low software quality, pointing to the initiation of design and code quality problems. Because development will slow down, maintainability of the software will be difficult. To fix these quality problems, it requires extra effort for their mitigation. TD is a measure of the effort needed for fixing these problems in the future and used as a measure of quality. Therefore, the higher value of TD for a software product means more unresolved quality problems included in, and lower overall quality. Since measuring TD directly can be difficult, we propose to analyze whether there is a relation between TD with internal and external metrics.

3 Dataset

In this assignment, we provide you a real world dataset *(Click here to download the dataset.)* that consists of information about 50 open source Android project. In this dataset, you will have 16 different metrics of the each project. We divide these metrics as "external", "internal", and "TD" metrics.

TD metrics that were measured:

Metric1: Code Duplication Ratio (CDR): Density of duplicated line of code
Metric2: Technical Debt (TD):Ratio between the cost to develop the code changed in the new code period and the cost of the issues linked to it

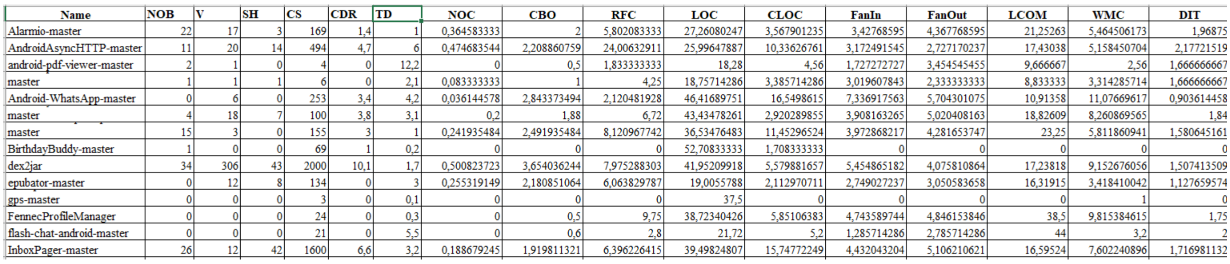
External metrics that were measured:

Metric3: Number of Bugs (NoB): It gives information about the number of bug
Metric4: Vulnerabilities (V): It gives information about the number of vulnerability
Metric5: Security Hotspots (SH): It is about number of security hotspots

Metric6: Code Smells (CS): Total count of code smell issues.

Internal metrics that were measured:

Metric7: Number of Children (NOC): number of subclasses that are linked to a class in the hierarchy
Metric8: Coupling between Object Classes (CBO): number of classes coupled to a given class
Metric9: Lack of Cohesion in Methods (LCOM): Let us assume that class C1 has a set of M1, M2,...Mn methods and that the set is a set of attribute variables used in the Mi method. In this case, LCOM is the number of discrete clusters that are the intersection of these n clusters.
Metric10: Fanin: a measure of how many other classes use the specific class. It indicates the number of classes to be affected if class changes. It can also be expressed as internal coupling.
Metric11: FanOut: Indicates how many classes are coupled on the class being examined. It refers to the external coupling of a class.
Metric12: Response for a Class (RFC): number of different methods that can be executed when an object of that class receives a message (when a method is invoked for that object)
Metric13: Depth of Inheritance Tree (DIT): maximum level of inheritance hierarchy of a class
Metric14: Weighted Method Per Class (WMC): sum of the complexity of all methods of a class
Metric15: Line of Code (LOC): It gives information about the size of the software.
Metric16: Comment Lines of Code (CLOC): It is the number of comment lines written for the program.



Name	NOB	V	SH	CS	CDR	TD	NOC	CBO	RFC	LOC	CLOC	FanIn	FanOut	LCOM	WMC	DIT
Alarmio-master	22	17	3	169	1,4	1	0,364583333	2	5,802083333	27,26080247	3,567901235	3,42768595	4,367768595	21,25263	5,464506173	1,96875
AndroidAsyncHTTP-master	11	20	14	494	4,7	6	0,474683544	2,208860759	24,00632911	25,99647887	10,33626761	3,172491545	2,727170237	17,43038	5,158450704	2,17721519
android-pdf-viewer-master	2	1	0	4	0	12,2	0	0,5	1,833333333	18,28	4,56	1,727272727	3,454545455	9,666667	2,56	1,666666667
master	1	1	1	6	0	2,1	0,083333333	1	4,25	18,75714286	3,385714286	3,019607843	2,333333333	8,833333	3,314285714	1,666666667
Android-WhatsApp-master	0	6	0	253	3,4	4,2	0,036144578	2,843373494	2,120481928	46,41689751	16,5498615	7,336917563	5,704301075	10,91358	11,07669617	0,903614458
master	4	18	7	100	3,8	3,1	0,2	1,88	6,72	43,43478261	2,920289855	3,908163265	5,020408163	18,82609	8,260869565	1,84
master	15	3	0	155	3	1	0,241935484	2,491935484	8,120967742	36,53476483	11,45296524	3,972868217	4,281653747	23,25	5,811860941	1,580645161
BirthdayBuddy-master	1	0	0	69	1	0,2	0	0	0	52,70833333	1,708333333	0	0	0	0	0
dex2jar	34	306	43	2000	10,1	1,7	0,500823723	3,654036244	7,975288303	41,95209918	5,579881657	5,454865182	4,075810864	17,23818	9,152676056	1,507413509
epubator-master	0	12	8	134	0	3	0,255319149	2,180851064	6,063829787	19,0055788	2,112970711	2,749027237	3,050583658	16,31915	3,418410042	1,127659574
gps-master	0	0	0	3	0	0,1	0	0	0	37,5	0	0	0	0	1	0
FennecProfileManager	0	0	0	24	0	0,3	0	0,5	9,75	38,72340426	5,85106383	4,743589744	4,846153846	38,5	9,815384615	1,75
flash-chat-android-master	0	0	0	21	0	5,5	0	0,6	2,8	21,72	5,2	1,285714286	2,785714286	44	3,2	2
InboxPager-master	26	12	42	1600	6,6	3,2	0,188679245	1,919811321	6,396226415	39,49824807	15,74772249	4,432043204	5,106210621	16,59524	7,602240896	1,716981132

Figure 1: A small part of dataset

The aim of this assignment is;

Part1- to find the relation between "TD" metrics with "internal" and "external" metrics by using statistical correlation analysis.
Part2- making a TD estimation by using ML Regression models.

4 Packages

- numpy is the fundamental package for scientific computing with Python.

- matplotlib is a library to plot graphs in Python.
- pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.
- Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.
- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics

5 Problem

5.1 Part1 (Statistical Correlation Analysis)

Correlation analysis is a statistical method used to evaluate the strength of relationship between two quantitative variables. A high correlation means that two or more variables have a strong relationship with each other, while a weak correlation means that the variables are hardly related. We will examine two types of correlation analysis method: Spearman and Pearson. The fundamental difference between the two correlation coefficients is that the Pearson coefficient works with a linear relationship between the two variables whereas the Spearman Coefficient works with monotonic relationships as well. In this part, it is expected you to analyze the correlation of ”TD” with ”internal” and ”external” metrics. Before analyzing the correlation, firstly, you will find the distribution of data. After finding the distribution of data, you will decide which correlation type you will choose; Spearman or Pearson correlation analysis.

REQUIREMENTS FOR PART1

Step1: Show the distribution of 3 metrics that can be evidence for choosing appropriate correlation analysis type (Spearman or Pearson)

Step2: Since we have different data types and ranges in our data set, apply min-max normalization to all the data we have

Step3: Show the correlation matrix of all metrics

Step4: Show p values of correlation tables

Note: The P-value is the probability that you would have found the current result if the correlation coefficient were in fact zero (null hypothesis). If this probability is lower than the conventional 5% (P0.05) the correlation coefficient is called statistically significant.

Step4: Show heatmap of correlation matrix

Step5: Show correlation between External Metrics & TD

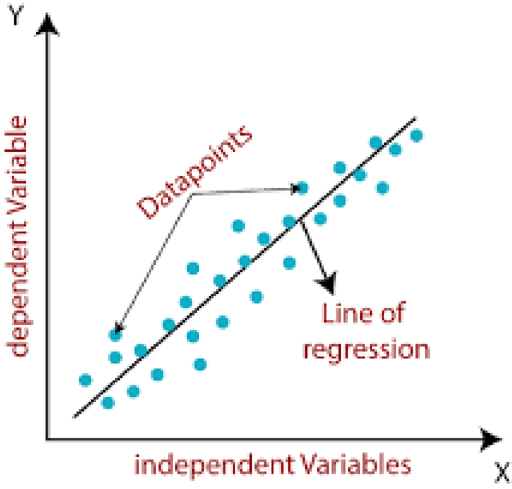
Step6: Show correlation between Internal Metrics & TD

5.2 Part2 (ML Modeling)

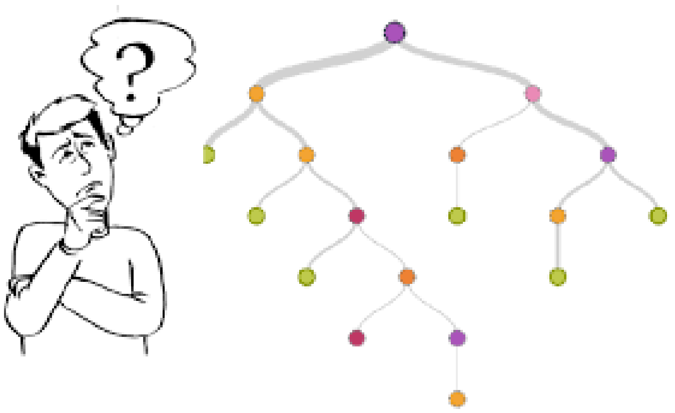
You will use the techniques of a subfield of computer science; which is machine learning. Machine learning can simlpy be defined as building a statistical model from a dataset in order to solve a real world problem. In this part, you will make a Technical Debt Estimation By Using ML Regression Models.

There are types of machine learning. However in this assignment, we expect you using 5 different ML Regression Models as shown below:

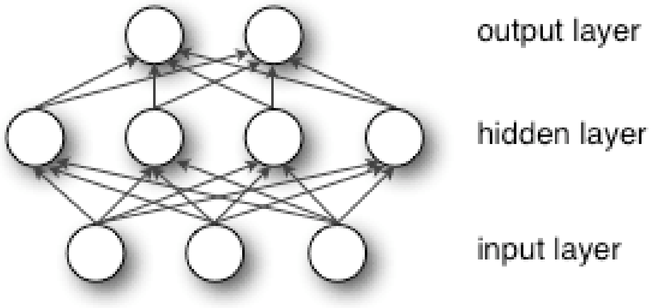
- Linear Regression: is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.



- Support Vector Regression: SVR gives us the flexibility to define how much error is acceptable in our model and will find an appropriate line (or hyperplane in higher dimensions) to fit the data.
- Decision Tree Regression: builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.



- Random Forest Tree Regression: is an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees
- MultiLayer Perceptron Regression: is a supervised algorithm that models neural network.



REQUIREMENTS FOR PART 2

You will read the dataset with Pandas library. Pandas is a data structure and data analysis tool for Python.

Step1: Split Data Into Train and Test Sets: You are expected to split your dataset into 2 sets; train and test sets. The training set is used to train the model. In other words, your model will learn from the training set and tune the weights. For the evaluation purposes; you will use the test set. You will obtain accuracy results on test set. You will use 70% of the data for training set and 30% of the data for test set.

Step2: Define the functions of 5 ML model

```
def linearRegression(X_train,y_train,X_test,y_test):
def svrRegression(X_train,y_train,X_test,y_test):
def decisionTreeRegression(X_train,y_train,X_test,y_test):
def randomDecTreeRegression(X_train,y_train,X_test,y_test):
```

def mlpRegressor(X_train,y_train,X_test,y_test)

Step3: Show machine learning models that estimate Technical Debt using only internal metrics.

Step4: Show machine learning models that estimate Technical Debt using only external metrics.

Step5: Show machine learning models that estimate Technical Debt using all(external and internal) metrics.

Note: You can use the given code for all 2 parts of the assignment. It is important to note that the given code usage is optional.

The aim of this assignment is:
1. Learning to use libraries
2. Learning to understand and analyse a given problem
3. The resulting accuracy percantage is not important. The important part is to understand the problem and implementing it.

6 Grading Policy

Task	**Point**
Part1 (Statistical Correlation Analysis)	50
Part2 (ML Modeling)	50

7 Important Notes

- Do not miss the submission deadline.
- Save all your work until the assignment is graded.
- The assignment must be original, individual work. Duplicate or very similar assignments are both going to be considered as cheating. You can ask your questions via Piazza and you are supposed to be aware of everything discussed on Piazza. You cannot share algorithms or source code. All work must be individual! Assignments will be checked for similarity, and there will be serious consequences if plagiarism is detected.
- *Click here to accept your Programming Assignment 5 for 1 day late submission.* (It will be degraded over 90 points)
- *Click here to accept your Programming Assignment 5 for 2 days late submission.* (It will be degraded over 80 points)
- You must submit your work with the file as stated below:

assignment5.ipynb