Department of Computer Science
The University of Hong Kong
Final Year Project
Final Report

The next generation eLearning platform for computing education

Supervisor: Dr. Chui Chun Kit

Group Members: Cai Tung San 3035090470
Chan Wai Lun 3035089835
Cheung Siu Wai 2013588661
Wong Ching Quen 3035098161

Date of Submission: 16th April, 2017
Table of Content

**Introduction** 3
- Brief Project Background 3
- Project Objectives 3
- Project Requirements 4

**Project Background** 5
- A detailed description of the project background 5
- Current available counterparts 7
- Shortcomings of current available products 7
- Incentive for doing the project 8
- Desirability of the project 8

**Project Methodology** 10
- Design of the platform 10
  - Frontend 11
    - User interface & User interaction 11
    - Coding feedback from the platform 11
  - Backend 11
    - Data storage 11
    - Code compilation 11
    - Code grading 11
    - Data analytics 12
      - Coding feedback generation and other applications 12
- Implementation details 12
  - Programming languages/tools used 12
    - Frontend 12
    - Backend 14

**Data Analytics / Machine Learning Techniques** 17
- Data preprocessing 18
- Data Extraction 18
- Data Formatting 18

**Data modeling** 19
- Submission Vectors 19
- Delta Vectors 19
- Correlation tuples 20
- Penalty function 20
- Correlation matrices 21

**Learning** 22
K-means 22
DBSCAN 22
Agglomerative clustering 23
Evaluation 25
Prediction 25

Experiments and Results 26
Detailed result of experiments 26
Experiment procedure 26
Visualization of algorithm result 26
Elbow method 28
Silhouette plot 30
Other details of K-means 31
Visualization of correlation matrix 31

Conclusion and Future Works 33
findings and conclusion of the project 33
Difficulties encountered 36
Time issue 36
Data size 36
Future improvement 37
Optimization 37
Storage 37
Algorithm 37
Code analyzing 38
Conclusion 38

Personal Reflection 39
1. **Introduction**

1.1. **Brief Project Background**

Currently in the aspect of computing education, there exists few, if not none, e-learning platform that provide specific support to programming learning. At present, schools and lecturers choose general e-learning platforms as supporting tools for conducting computing education courses in the same manner as other non-computing education courses. These chosen general e-learning platforms are not designed specifically for computing education. As a result, these platforms do not provide specific aids to the learning of students who take part in the computing course. With a view to tackling this problem, the project team would like to set up a new platform that is designed for students taking part in computing education courses. The platform will feature tools like collaborative tools and auto-generated hint for students. The features are implemented with data analytics and machine learning techniques. Students are expected to be able to access more support in programming learning through this platform, and hence gaining technical and practical knowledge in programming more effectively in the self-learning manner. This final report will serves as an informative document regarding the methodology, details about experiments and results conducted in the development of the platform and some proposed improvements for future development work.

1.2. **Project Objectives**

The platform is assumed to be an auxiliary tool for the course COMP2123 Programming Technologies and Tools. The platform should allow students learning programming in self-learning manner with support from peers and the platform’s feedback, instead of teachers. The platform should be able to facilitate communication between students in order to allow peer support.
1.3. **Project Requirements**

Table 1 shows the prerequisites of the project, which are essential to the completion of the project.

<table>
<thead>
<tr>
<th>Medium</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server</td>
<td>1. Allow hosting of the platform</td>
</tr>
<tr>
<td></td>
<td>2. Conduct computational task like data analytics</td>
</tr>
<tr>
<td>Web Development framework</td>
<td>1. Serve as development tools to the platform</td>
</tr>
<tr>
<td>Data analytics / Machine learning</td>
<td>1. Conduct Data analytics / Machine learning tasks</td>
</tr>
<tr>
<td>Tools</td>
<td>2. Facilitate experiments involved in development</td>
</tr>
<tr>
<td>Coding data</td>
<td>1. Serve as raw materials for analysis</td>
</tr>
</tbody>
</table>

Table 1. Project requirements
2. **Project Background**

2.1. **A detailed description of the project background**

With the advancement in computation resource, data analytics and machine learning have become more efficient. Relevant studies and different applications of data analytics and machine learning have become more popular. However, there seems to be few, if not none, applications of machine learning in the computing education field. This project, while serving the purpose of assisting in coding education of students as stated above; also acts as a proof of concept that machine learning can be applied to the computing education. With the data collected from platform (including but not limited to, source code written by students, timestamp of submissions), we can discover or deduce additional information about the coding habits of students. One of the few examples is given a certain programming task to students, from students’ submissions we can observe which test cases students tends to solve first. For some test cases that are always being solved by students in early stage of their development, we can reasonably conclude these test cases are associated with low difficulty, given that we collect and observe enough submission data. Another example is we can discover association between test cases. If several test cases are frequently solved by same submission, we may conclude that these test cases belongs to the same test case group. In reality, different test case group often related to different programming concepts while test cases in same test case group shares similar concepts. To illustrate, consider a scenario that there is two test cases group in one programming task, which is shown in Table 2. The programming task is given an integer N, output “even” if it is even number; output “odd” if it is odd number. Students’ solutions are also expected to be able to perform basic checking in the input. For inputs that is not integer, submission should output “invalid”.

<table>
<thead>
<tr>
<th>Test Case ID</th>
<th>Test Case Input</th>
<th>Test Case Expected output</th>
<th>Test Case group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>even</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>even</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>odd</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>193</td>
<td>odd</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>12.93</td>
<td>invalid</td>
<td>B</td>
</tr>
<tr>
<td>6</td>
<td>Invalid input</td>
<td>invalid</td>
<td>B</td>
</tr>
</tbody>
</table>

Table 2. Example of test case groups

For this naive example, it is obvious that the first 4 test cases are valid input to the program and the last 2 test cases are not. Intuitively, the first 4 examples should be grouped as 1 test case group (test case group A) and the last 2 should belong to another test case group (test case group B). This simple examples clearly shows that test case group A are related to the concepts of completing basic requirement of the task; and test case group B contains test cases that are related to the idea of error handling. In reality, the test case grouping is not as obvious as the example is. In more complicated case, it may require expert knowledge in software testing to correctly classify test cases into test case groups. The platform is expected to automate this classification process, based on submission data from students and machine learning techniques. After the classification is completed, the platform is able to utilize the classification information to provide feedback to students. Following the same examples, a common scenario is a student complete the basic requirement of the question (i.e. passed the first four test cases) but he has not performed the input checking (i.e. have not passed the last the two test cases).
The platform will suggest the student to try the last two test cases. In this case, the student is reminded to perform input format handling for the question, which is a different concept that should be learned by students through completing this question.

2.2. **Current available counterparts**

Currently, there is no e-learning platform targeting computing education. Computing education courses adopts general e-learning platforms like Moodle, iClass. These e-learning platforms are not designed specifically for students learning programming and have a focus on assignment submissions and course material management. For the usage in computing education course, students uploaded their source code files after they completed the question. The e-learning platform simply store the the files. To mark the submission, instructor has to download the files and assess them manually. After that, feedback is given to student through the platform. Automation of the marking process may be able to allow instructors to save time for designing and reviewing future teaching materials, which can improve the learning experience of students. The platform we developed aims to help students learning coding in more efficient manner, through facilitating peer support by various tools and the feedback generated by the platform automatically.

2.3. **Shortcomings of current available products**

As discussed above, the general e-learning platform emphasized on assignment submissions and course material management, with little, if not none, specific support to coding learning. To directly tackle this problem, we design the platform to be able to provide support to coding learning as mentioned previously, while maintaining the functionality provided by those general e-learning platform. The platform will also be able to allow instructors and
students to submit assignments and manage course materials in traditional way. In other words, the new platform can be said to be a superset of the traditional e-learning platform, with additional functions for supporting coding learning.

2.4. **Incentive for doing the project**

This project is proposed by the project supervisor, Dr CK Chui. Dr Chui is very keen on computing education field and contributes a lot to this field. Our team shares the same vision as Dr Chui and wants to contribute to education also. We wholeheartedly believe the project, after completion, can truly benefit students who are new to learning programming. The reason behind is when we are also new to coding in the past, we often encounter difficulties and would like to seek assistance from various sources. We believe the platform can provide the help students needed.

On the other hand, this project matches our team's interest and skills. After years of studies, we gained various programming and CS knowledge and skills. This project can be a good summarize and demonstration of what our team has learned in our studies. This project, involves data analytics / machine learning and web programming, which are all popular skills required in CS / IT field. We believe the project can be beneficial to the education field, while being able to demonstrate our skills and knowledge.

2.5. **Desirability of the project**

As stated above, with the rising trend of coding education, more students are expected to learn coding and it is common that they require support. The support from instructors may not be sufficient as the number of experienced instructors is limited. As we observed in COMP2123 class, despite there are more TAs and student helpers attending the class to help students, sometimes a student may need to wait certain amount of time before they are acknowledged. This
phenomenon may come from the fact that it requires patience and time for instructors to thoroughly explain certain concepts to students, especially those abstract ones. Through the project, we hope to solve this issue by providing certain amount of help to students. As the platform allows many students to access it, it is scalable and can be applied to MOOC (Massive Open Online Course). Students can learn remotely at home through the platform. At the same time, with more students participating in the platform, more coding data can be collected and the data analytics can be more accurate since we have more details about the distribution on the submission data. This forms a virtuous cycle that the more data we collect, the platform can produce more accurate feedback. And, if the platform are able to provide feedback with higher quality, it can attracts more users. In conclusion, the platform’s performance will keep improving.
3. **Project Methodology**

3.1. **Design of the platform**

The platform is separated into two parts, the frontend and backend. The frontend is a web system that is responsible for providing user interface to users and handling user interaction. The objective of the backend included data storage, code compilation, code grading, and also data analytics. Figure 1 shows the architecture of the platform.

![Figure 1. Architecture of the platform](image)

As illustrated in figure 1, the frontend of the platform mainly resided in the middle of the diagram; while the backend is located in the right. The collaborative tools are included in the frontend while data...
storage, code compilation / grading and data analytics are handled in the backend. More information about the two parts will be given below.

3.1.1. **Frontend**

3.1.1.1. **User interface & User interaction**

The frontend part is a web system that provide user interface and handle user interaction. For example, as shown in Figure 1, lecturer can use the frontend to input test cases and post assignments. Students can view the assignments and use the collaborative tools provided by the platform through the user interface. The collaborative tools will be explained in later section.

3.1.1.2. **Coding feedback from the platform**

The coding feedback generated by the platform backend will be returned to user through frontend.

3.1.2. **Backend**

3.1.2.1. **Data storage**

Various data is stored in the platform, including assignments, chat history and forum posts. The submission data stored like passing of test cases and timestamp will be used in data analytics.

3.1.2.2. **Code compilation**

The platform can perform code compilation. The user input can be piped to the compiled code and the output of the compiled program is returned to the user.

3.1.2.3. **Code grading**

Code grading can also be achieved by the platform. By utilizing the compilation function, with test cases specified by instructor as input to the program and comparing the expected output of the test case and the
output of the student’s program, the submission of student can be verified against all test cases in grading. Hence, grading can be achieved by examining the passing of test cases.

3.1.2.4. **Data analytics**

The data analytics program extract data stored in database. After performing data processing, which will be discussed in later section, the data will be used as input to machine learning algorithm, the result will be stored and used for other usage, including generating feedback to students.

3.1.2.5. **Coding feedback generation and other applications**

Recommendation can be generated by the platform, through the analysis on the submission data. There are other potential applications to the data stored in the platform, which will be discussed in later section

3.2. **Implementation details**

3.2.1. **Programming languages/tools used**

3.2.1.1. **Frontend**

3.2.1.1.1. **User interface & User interaction**

For general user interface and handling user interaction, the platform adopts Laravel as the development framework. Laravel is a PHP framework with a mature online user community. This is important to the development of the project as the community is able to provide interactive and timely support for Laravel developers. Common problems encountered in the development is likely to be encountered by other members of the
community before and an existing solution can be expected. The development time and difficulty can be reduced as a consequence.

Another advantage of Laravel is the abundance of available libraries. Laravel provides more than 9000 packages which is commonly used in web development. These functions can be utilized in development of the project and the development progress can be highly speeded up.

3.2.1.1.2. Code Editor

The platform embedded external code editor instead of developing one by the team. The main reason behind this choice is the time issue. It requires a lot of time to develop a code editor. We prefer to utilize our time in building other functionalities of the platform. In addition, the external code editors are well-written and contain a lot of functionalities. For the project, Firepad, Ace and Firebase are selected for constructing the code editor.

Firepad is a real-time collaborative code editing tool, which allows multiple users access and modifying the same source code file. Ace is a code editor which provide common functions exist in other IDEs and code editors. Some of the functions of Ace are syntax highlighting, code completion and multiple cursors. For integration into the platform, detailed instructions are provided by the original developers of the tools. Thus, it can be concluded that the workload and time needed for the integration is minimized.
The aforementioned auto-generated code feedback will be given to students via message box in the code editor.

Firebase serves as a native data logging system which records all input of students when they use the code editor. The code editing history is logged in json-like format every time when the source code is changed.

3.2.1.1.3. Collaborative tools

3.2.1.1.3.1. Forum

Forum, as the one of the basic and most common tool for communication and idea exchanged, is built using external library. Similar to code editor, it requires a lot of time and effort to build a forum function from scratch. Therefore, the platform adopts external library again, taking the same advantages of well-written external library. The library adopted is named laravel-forum, developed by Riari.

3.2.1.1.3.2. Drawing Board

Following the same reasoning of using external libraries for code editor and forum, external libraries are planned to be imported for developing drawing board. Unfortunately, due to time issue, the drawing board is not completed. The time issue will be discussed in later section.

3.2.1.1.3.3. Video presentation

Similar to the situation of drawing board, video presentation is not completed due to time issue.
3.2.1.2. **Backend**

3.2.1.2.1. **Server hosting**

The team adopted web hosting service provided by DigitalOcean. The team used free credits provided for students for free hosting. Upon subscription, our server’s IP address is given to us. We can then setup the Ubuntu environment. To use the server, the team use SSH to remotely launch Bash commands. That is also the main reason we choose the service of DigitalOcean. It is relatively easy to use for our team as our members has experience in Bash.

3.2.1.2.2. **Data storage**

The data collected in the platform in stored in the database as shown in the architecture diagram. The database tool we adopted is MySQL. One of the main reasons we adopted MySQL is we are relatively familiar with its relational database design. We have accumulated experience working with relational database in our studies (including the course COMP3322 and COMP3278) and internship experiences. We have considered MongoDB as another database candidate for the projects. However, due to the facts that we are more familiar with the MySQL and extra learning curve may be induced if we choose to develop with MongoDB, we finally used MySQL as our database model.

In addition, another advantage of MySQL is we find existing tools like XAMPP that allow us to setup the environment easily. This facilitated the
development and testing process, allowing us to allocate our time in other parts of the project.

3.2.1.2.3. Code compilation

The code compilation is simply achieved by calling bash command `g++` in PHP code. As stated previously, the server allows us to interact in Bash manner through SSH. By calling the `g++` command, the object code is compiled, in the same manner like developing c++ program in local linux environment.

3.2.1.2.4. Code grading

After code compilation, the object code is obtained. It is worth mention that if the source code from students fail to compile, the platform simply return the error message from the compiler to the student. If the code can successfully compile, test cases (in text format) are inputted into the object code through Bash and the output is stored in text format. The output is then returned to the PHP and passed to students. By repeating the process for every assigned test cases for the question, we obtained a series of output of the program. The outputs are then checked against expected output, which is provided with the test case when being inputted by instructors / TAs and stored in the database. By comparing output of student’s code and the expected output, we can mark the student’s code as 1(pass) or 0(fail) for every test cases. The marked result is then stored in database, for the usage of data analytics, which will be discussed in details in later section.

3.2.1.2.5. Coding feedback generation
As stated above, for the case that the source code fail to compile, the error message is returned to students. If the code passed the compiler, the output against test cases are returned to students. This process can facilitate students in testing their program during development as students do not need to repeat the input process every time when they are testing. If a student struggle to pass test cases, he can request for hints, which is presented via message box implemented through external libraries bootstrap-notify.js and jquery-popup-overlay.js. The content is generated after data analytics, which will be discussed in later section in details.

3.3. Data Analytics / Machine Learning Techniques

The platform heavily relies on data analytics and machine learning to provide coding feedback automatically. In the following section, details about data analytics and machine learning will be discussed. Figure 2 shows the general workflow of data analytics and machine learning.

![General workflow of data analytics and machine learning](image)

Figure 2. General workflow of data analytics and machine learning
The general workflow of data analytics / machine learning involves 4 stages: preprocessing, learning, evaluation and prediction. This workflow serves as a rough guideline for the data analytics task in the project. In the project, the data will undergo data preprocessing, data modeling, learning, evaluation, and finally prediction. Python is the programming language used in the data analytics. Python is selected due to its user-friendliness, abundance of data analytics and machine learning libraries (Numpy, Scikit-learn, Pandas). These advantages help reduce the development time of the data analytics / machine learning part of the platform. In addition, due to the fact that Python serves as a general purpose programming language, it is not difficult to connect the Python program to the other parts of the platform. The server only needs to call the Python code through launching system commands in Bash. With suitable parameters, the correct output will be provided by the Python program and can be read and returned to the user by the web system.

3.3.1. **Data preprocessing**

3.3.1.1. **Data Extraction**

To conduct data analytics, data is needed. As stated before, data is stored in a database. To extract data from the database, a Python library pymysql is used. This library serves as a connection between the database and the data analytics program. SQL statements can be issued using pymysql and the result can be returned to the Python program used in data analytics.

3.3.1.2. **Data Formatting**

The data extracted cannot be used in analytics directly. The machine learning algorithms require matrices as input. Raw data is required to undergo
formatting and modeling to ensure correct inputs are used. This follows the “garbage-in-garbage-out” principle and ensures that the output of the machine learning program is accurate.

3.3.2. **Data modeling**

As mentioned above, the data needs to be formatted to be suitable input to the machine learning program. The expected input format of the machine learning algorithms is matrix. The data structure is numpy array (numpy.ndarray). The data extracted will be constructed into this form.

3.3.2.1. **Submission Vectors**

For each submissions from every students, it will be formed into submission vectors. To illustrate, if student passed #0, #2, #4 out of the 10 test cases associated with this question, the submission vector will be [1 0 1 0 1 0 0 0 0 0]. The submission vector will be further processed to be used in data analytics.

3.3.2.2. **Delta Vectors**

For every pair of submissions, we can compute the difference between them, which is named delta vector. The delta is computed using element-wise XOR operator. For example, The delta vector computed from [0 1 1 0] and [1 0 1 0] is [1 0 1 0]. Noted that by computing delta vector, different submissions with changes in source code are considered the same submissions if they passed the same set of test cases. As a result, if the students simply format the code like adding comments for improving readability, it will not be treated as a new submission as it has no effect on the submission vector. Therefore, the delta vector is also not affected and the result of data analytics will not be changed.
3.3.2.3. **Correlation tuples**

The delta vectors calculated are used for generating correlation tuples. The intuition of correlation tuple is the linkage between test cases. To illustrate, given a delta vector \[1 \ 0 \ 0 \ 1 \ 1\], correlation tuples \((0, 3), (0, 4), (3,4)\) are generated. The correlation tuples are used for updating correlation matrices with penalty function, which both will be explained in later sections. Given a penalty function \(P(x) = 0.8^x - 2\), where \(x\) is the number of “1” entries of the delta vectors, for the vector \([1 \ \ 1 \ \ 1 \ \ 0 \ \ 0]\), the output of the penalty function is \(0.8^{3-2} = 0.8\). Then for every correlation tuples, which indicates the indices of entries to be updated in the correlation matrices, the output of the penalty function (0.8) is used. In other words, for the correlation tuple (0,3), the 0th row and 3rd column of the correlation matrices are updated with 0.8.

3.3.2.4. **Penalty function**

Penalty function serves the purpose of regulating correlation between test cases. The final candidate selected for the penalty function is \(0.95^x - 2\), where \(x\) is the number of “1” entries in the delta vector. The output of the penalty function is named penalty coefficient. For every delta vectors, element-wise multiplication of the penalty coefficient is applied. One simple example is given delta vector \([1 \ \ 0 \ \ 1 \ \ 0 \ \ 1]\), output of penalty function is \(0.95^{3-2} = 0.95\). The final form of the delta vector is \([0.95 \ \ 0 \ \ 0.95 \ \ 0 \ \ 0.95]\). The intuition behind penalty function is a measure for regulating correlation between test cases. By imposing higher penalty coefficient, the correlation between test cases are increased. The choice of the
penalty coefficient is based on empirical result, which will be discussed in later section.

3.3.2.5. Correlation matrices

After obtaining delta vectors from submissions, correlation matrix can be calculated. The entries in the matrix range from 0 to 1, implying low degree of correlation and high degree of correlation correspondingly. Each row in the correlation matrix is initialized to identity matrix with dimension equal to number of test cases. The choice of identity matrix is based on the assumption that each test cases are correlated with themselves and different test cases are not related at first. The test cases will have correlation only after being analysed by the system. The update of the correlation matrix will be based on the clustering result.

The whole process of the data modelling is summarized and visualized in Figure 3.

Figure 3. The data modelling process
3.3.3. Learning

Upon accepting data for input, learning algorithms has to be decided for classification. For the project, unlabelled data is provided. In other words, the test cases are not associated with prior knowledge that which class it belongs to. As mentioned before, to manually label the test cases, it can be complicated and may require expert knowledge in software testing, which is limited. As a result, unsupervised learning model is adopted. We considered several candidate algorithms, including K-means, DBSCAN, Agglomerative clustering, which will be explained below. They are all provided as api from the Scikit-learn library, therefore implementation of the algorithms is not required, only suitable parameters are required for the input of the algorithms. Details will be provided in later section.

3.3.3.1. K-means

Among the clustering algorithms, K-means has a relatively efficient time complexity, $O(tkn)$, where $t$ is number of iterations; $k$ is number of clusters; and $n$ is number of data points. However, K-means algorithm requires the parameter $k$ (number of clusters) to be specified in advance. Extra experiments are required to be conducted to find optimal set of parameters. The algorithm also has the disadvantage of being unable to handle outliers. Dataset with outliers can have classification result with low accuracy since the algorithm attempts to put outliers in clusters.

3.3.3.2. DBSCAN

DBSCAN stands for Density-based spatial clustering of applications with noise. DBSCAN is associated with two parameters, $\varepsilon$ and minPts. For a unvisited point $p$, if there are at least minPts points with
distance not greater than $\varepsilon$, $p$ is concluded as core point. In other words, $p$ will become a core point if there are $minPts$ points inside the circle area with radius $\varepsilon$ and centered at $p$. After all points have been processed and all core points have been found, the core points and the points within their radii will form groups. If two groups’ core points are within the radius of their counterparts, the two groups will be merged. Hence, Clusters are formed and outliers are separated since outliers will not be able to form cluster. The advantage of DBSCAN is the number of clusters is not required as parameters and robust to outliers. It can also form arbitrary shape of cluster, which may be favoured depends on situation. However, it required large amount of experiments to find out suitable values for the two parameters for good clustering result.

### 3.3.3.3. Agglomerative clustering

Agglomerative clustering assigns each data point to its own cluster first. By using a evaluation metrics like single link, clusters are merged together. The process continues until termination condition is reached. The evaluation metrics determines whether clusters should be merged together. Take single link as example, if there is two data points in two clusters which has distance smaller than a certain threshold, the two clusters are merged. Eventually all clusters will be merged to a single cluster which contains all data points. For desired clustering result, termination condition is introduced. Common termination condition includes number of clusters in current result smaller than a threshold. A variant to this clustering algorithm is divisive clustering, which reverse
the process. Instead of merging clusters, a single big cluster is divided until termination condition is reached. This clustering algorithm can determine number of clusters with the help of termination condition. However, the algorithm requires $O(n^2)$ time and hence does not scale up well.

The algorithms have their merits and demerits. Table 3 provides summarization of the algorithms and compares them.

<table>
<thead>
<tr>
<th></th>
<th>K-means</th>
<th>DBSCAN</th>
<th>Agglomerative clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantage(s)</td>
<td>● Relatively efficient</td>
<td>● Can determine number of clusters</td>
<td>● Manual tagging is not required</td>
</tr>
<tr>
<td></td>
<td>● Suitable for general clustering task</td>
<td>● Robust to outliers</td>
<td>● Relatively efficient</td>
</tr>
<tr>
<td>Disadvantage(s)</td>
<td>● Sensitive to outliers</td>
<td>● Require manual tuning of parameters</td>
<td>● Do not scale up well due to high time complexity</td>
</tr>
<tr>
<td></td>
<td>● Need to specify number of clusters manually</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Summarization of difference between clustering algorithms

Sample data is used as input to the algorithms and see the effect of the algorithms. The details will be stated in experiment section.
3.3.4. Evaluation

After the learning phrase, the algorithm obtain a set of parameters which extracts best knowledge from the input data. The algorithm is said to be trained at this phrase. The trained algorithms should undergoes evaluation to ensure its performance is up to standard for the application. If the trained algorithm is not satisfactory, the parameters need to be tuned. The details of the experiment process will be stated in later experiment section, including the evaluation metrics used, parameters to optimize and choosing of algorithms.

3.3.5. Prediction

The final model after evaluated could then be used for prediction. New data including test cases’ rates comes from students could be inputted to the model. The analysis will attempt to classify the test cases and offering feedback back to the code editor. These data can also be used for further tuning of the model and algorithms in the evaluation part.
4. **Experiments and Results**

4.1. **Detailed result of experiments**

Upon series of experiments, the system adopts K-means as the final clustering algorithm. The parameters chosen for the algorithm included number of cluster groups is determined by using elbow method. For every question, the elbow method is used to find optimal number of cluster groups in an automatic manner. The result is verified by another analysis method called silhouette analysis.

4.2. **Experiment procedure**

In the following section, some of the experimental procedures mentioned above will be described thoroughly.

4.2.1. **Visualization of algorithm result**

At the start of the project, to better understand the clustering algorithm candidates, sample data provided by Scikit-learn is used to simulate the clustering process. The sample data used is the Iris flower data set. Figure 4 shows the difference between results of K-means, DBSCAN, Agglomerative clustering.
The code that is used to plot the graph is adopted from COMP3314 Machine Learning Course. The result implies that K-means tends to give result with highest accuracy as it misclassifies the least number of data points. This conclusion matches the general comment from our supervisor Dr Chui, that in most cases, K-means is a suitable candidate for clustering task. K-means is then determined to be the candidate for the data analytics task for the platform. However, this result is not finalized until further verification on submission data is completed. It only helps us to narrow down the scope of trial and error for experiment.

4.2.2. Elbow method

As mentioned above, the K-means requires some parameters to be set manually. The most prominent one is the number of cluster groups K. K has to be predetermined as the algorithm needs to initialized K points as centroids and keep
updating them in learning process. These K centroids are finalized when they stop differing from result of previous iteration. For determining K, we adopted elbow method. Elbow method plots distortions against all possibilities of K. Distortion is the average distance of every data point to the centroid of the clustering group. For a good clustering, the distortion should be low. The possibilities of K ranges from 1 to N, which correspond to the case that every data point belongs to the same group and every data points gain their own cluster group. For K=N case, the distortion will be minimal as every data point is the centroid of their cluster group, and the distortion will be zero in this case. However, it is not a desired result as it does not provide grouping information. Elbow method is designed to find the optimal value of K. Figure 5 shows one of the Elbow plots in the experiment process.
The plot is generated when using penalty function of $0.95^{-2}$. From the plot, intuitively the elbow point ($n=4$) is the solution we are seeking. The elbow point strikes a balance between minimizing distortions and loss of grouping information. To compute the the elbow point instead of plotting the graph and spot it manually every time, second derivative is computed and the point associated with the maximum absolute value is the elbow point. To compute the second derivative efficiently, approximation is used, for an array $a[1..n]$, the second derivative for the $i$-th element is approximated by $A[i+1] + A[i-1] - 2 * A[i]$. For the 2 end points ($n=1$ and $n = N$), they are disregarded as elbow point. In addition if there is only 2 cases, which corresponds to the 2 end points, the program report no elbow point is found. For general situation, this case is rare and we can often find the elbow point. The elbow point found is used as the value of $K$ for K-means.

### 4.2.3. Silhouette plot

The silhouette analysis mainly measure 2 quantities, coherence in between cluster and separation between clusters. High coherence and high separation implies good clustering. After computing silhouette coefficient for every data point as Figure 6, average silhouette value in clusters can be found and plotted as silhouette plot. Similar silhouette distribution is a sign for good clustering result.
4.2.4. Other details of K-means

The K-means algorithm used for the platform is also set up with initialization with K-means++. The maximum number of iterations is limited to 300.

4.2.5. Visualization of correlation matrix

To ensure the test case classification is running correctly, correlation matrix is visualized using heat map. Figure 7 shows one of the heat map in experiments.
Figure 7. Sample heat map

The entries with brighter colors indicate value with higher value; darker entries corresponds to lower value. From this heat map, we can easily see that correlation matrix is symmetric, which is correct since the correlation between test cases are symmetric. We can also see there is 4 groups, indicated by the brightest entries on the diagonal. This matches our earlier plot of elbow method that n=4 is the optimal grouping.
5. **Conclusion and Future Works**

5.1. **findings and conclusion of the project**

To simulate the real situation, we set up a virtual course with only one assignment that is compute the check digit of HKID. The method of calculating the check digit is stated in [Chinese Wikipedia](https://zh.wikipedia.org). Students are expected to perform basic input format checking. We made up 10 test cases, which are shown in Figure 8.

![Test cases](image)

**Figure 8. Test cases**

The first three test cases (testcase_id = 1, 2, 3) are basic requirement. The next three (testcase_id = 4, 5, 6) are special cases for HKID requirements. The next two (testcase_id = 7, 8) are error cases. The last two are special cases that the HKID has 2 starting alphabets (testcase_id = 9, 10). Figure 9 shows the output of the algorithm, which is stored in database.
The algorithm gives correct result. The test case group are then ranked based on their relative position in submission of students. For example consider test case group 4, says student #1 passed the test cases in the group in 4th submission out of his total 5 submissions; student #2 passed the test cases 8th submission out of her total 10 submissions; and student #3 passed the test cases in the group in 2nd submission out of his total 3 submissions. We can observe that students pass the test cases in test case group 4 in relatively late submissions. We can then reasonably conclude that test case group is a difficult test case group. Mathematically by computing the relative position of the passing, then taking average for every student entries, we obtain the difficulty rating of test case group. Following the example, Student #1’s relative submission is $\frac{4}{5} = 0.8$; Student #2’s relative submission is $\frac{8}{10} = 0.8$; Student #3’s relative submission is $\frac{2}{3} = 0.67$. The rating for test group 4 will then be $0.8 + 0.8 + 0.67 / 3 = 0.76$. For our previous example, the test case rating are computed and stored in database, which is shown in Figure 10.
From the figure we can see the ordering of difficulty of the test case group is 3, 1, 2, 4. This implies test case group 3 is the easiest, and test case group 4 is the hardest. This conclusion also matches our expectation as we expect student will complete basic requirement first, then complete error checking. By ranking test cases, we can extract the test case group that a student has not passed with the minimum rating, which is the easiest test cases that the student has not passed. When student require hints, we remind him to try these test cases first. If student still need further help, we extract the previously stored code that other students completed to passed that test case and show to the struggling student. We believe students can learn step-by-step this way. Figure 11 are two screenshots of the hint we provide to students.

<table>
<thead>
<tr>
<th>id</th>
<th>avg_rate</th>
<th>course_id</th>
<th>assignment_id</th>
<th>title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.40</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7.45</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.10</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8.80</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
5.2. **Difficulties encountered**

The two main difficulties encountered in the project is time issue and data size.

5.2.1. **Time issue**

The platform has a lot of features to implement. With only time of two semesters (around 6 months), we can only partially finish the features. Among all proposed features, we focused on core functions related to coding. Those functions related to coding included code editor, code compilation, data analytics, providing coding feedback. We successfully complete these functions on time. However, we can only partially complete other features. Among proposed collaborative tools, we only implemented forum, which can serve basic communication.

5.2.2. **Data size**

For data analytics task, amount of data is crucial to the result. Given more input data, intuitively we have more information about the distribution of the data, in this project the submissions of students. Due to aforementioned time issue, and lack of data source, we have a relatively small data size, only 1
sample question is featured (HKID). Although we believe it is an appropriate demonstration of the platform, in future we hope to have a more sophisticated setup to further prove the capability of the platform.

5.3. **Future improvement**

In following section, some improvement will be proposed to the final product.

5.3.1. **Optimization**

5.3.1.1. **Storage**

For ease of development, every submissions of students are stored in the database. This implies that in every submission, the whole source code of student is stored to database. With native logging of Firebase, we believe it is redundant to store source code in every submissions as we can reconstruct the the source code by using the code change in Firebase.

5.3.1.2. **Algorithm**

For best accuracy, the platform now reconstruct the correlation matrix every time and reperform clustering when there is new test cases. For performance, it is proposed that we can sample a portion of test cases and do clustering on them instead of using all of them. This idea is inspired by stochastic gradient descent, which is designed to deal with large amount of data. Aside from this, it is also suggested that the classification of new test cases can based on existing result instead of reclustering. By using existing result, we can treat the test case grouping as labelled data, and use them to train supervised learning model like SVM and use the model for classification of new test cases.
5.3.2. Code analyzing

At the start of the project we have discussed the idea of analyzing student’s codes to provide feedback, which involves techniques from NLP (Natural Language Processing). However, as suggest by Dr Chui, this may requires a lot of time to understand and apply to the project. Due to the aforementioned time issue, we do not adopt this idea. However, it is believed that by analyzing student’s code, we can provide the most accurate feedback.

5.4. Conclusion

In conclusion, the platform is an e-learning platform that aims to provide specialized coding support for computing students. By applying data analytics and machine learning techniques, the platform is able to analyze students’ submissions and use the result to generate feedback to students. Regarding the technical side, the platform is mainly built with Laravel and other web system tools and the data analytics part are mainly built using python and clustering algorithm. We hope that the platform can contribute to the computing education field.
6. **Personal Reflection**

In this project, I mainly focus on the data analytics part. I learned much from doing this project and gained a invaluable chance for practising what I have learned in my 4 years of CS study. Concretely, I have touched linux programming, data analytics / machine learning, data modelling, algorithms, scientific computing in this project, which covered most of my study, if not all. I believe this project can serves as a summarization of what I have learned in my studies.

However, the platform would not be successfully built without the help of many individuals. First of all, I would like to give my sincere thank to the project's supervisor, Dr Chui, for proposing this project and providing many useful suggestions in the whole process.

On the other hand, I would like to thank my teammates, Alan, Dennis and Timmy for doing this project with me. There are numerous difficulties encountered in the project, but they are often solved by our discussion and cooperation. The platform is impossible to be completed without any one of us.

In addition, I would like to acknowledge all professors in CS department for providing CS knowledge throughout our studies. This knowledge are fundamentals to the completion of the platform.