**I.** **Assessment Requirements**

This coursework is an individual assignment. You must implement it independently.

In this coursework, you are asked to analyse the "2020 Stack Overflow Annual Developer Survey" dataset which is available at <https://insights.stackoverflow.com/survey>. The data were collected by Stack Overflow, which is a question and answer site for professional and enthusiast programmers. In 2020, there were 65,000 responses from over 180 countries and dependent territories. This survey examines all aspects of programmers’ experience from career satisfaction to opinions on open source software. This information could be useful for you as program developers. For an overview of 2020 Stack Overflow’s Developer Survey, you may refer to <https://insights.stackoverflow.com/survey/2020>.

The following tasks are required in the coursework:

1. To understand the data set. You must implement this task by exploratory data analysis through Python programming.
2. To describe characteristics of low-income developers and high-income developers. You must implement this task by cluster analysis through Python programming. Note: Low income is defined as salary up to the median of all developer’s annual salaries, while high-income as salary more than the median of all developer’s annual salaries. Income is measured by the column **ConvertedComp** which represents Salary converted to annual USD salaries.
3. To build machine learning models for predicting whether a developer is in high-income based on survey data. You must implement this task by classification through Python programming.

There are over 60 columns in the data set. Obviously, some attributes (e.g. country or education) have a bigger impact on the income than others. Therefore, you will decide by yourself which features should be adopted in data analysis tasks. Good selection of features may improve the performance of your machine learning models.

You will present the coursework in the form of a technical report containing the five sections listed in Table 1. A report template is provided at the end of this specification. Please follow the template.

Table 1 Structure of report, weights, and recommended pages

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Section** | | **Weighting** | | **Recommended Pages per Section** | |
| 1. Introduction | | 0.1 | | 1-2 | |
| 2. Data Understanding and Exploratory Data Analysis | | 0.2 | | 7-10 | |
| 3. Cluster Analysis | 0.1 | | 2-4 | |
| 4. Machine Learning Methods and their Implementation | 0.3 | | 7-10 | |
| 5. Evaluation Machine Learning Models | 0.2 | | 4-5 | |
| 6. Discussions and conclusions | 0.1 | | 1 | |

A report with 20-30 pages is recommended. The report in total, however, must not exceed 30 pages (excluding title page, contents page, references, and appendices) with the font Calibri and size 11 or 12 in the main text. A penalty of a single grade will be incurred if you exceed the 30-page limit. You may put extra information in appendices which is not counted in 30-page limit.

You are asked to write the report with the provided report template at the end of the template. It is recommended to cite and list referees using **Harvard Referencing style** (see <https://www.ntu.ac.uk/m/library/referencing-made-easy>). However, other (author, year) styles like APA are also accepted.

By the submission deadline, you are expected to submit both **your report (in MS Word or PDF format) and the Python source code (in plain text format)** to NOW Dropbox under the `report’ folder.

Your report will be assessed according to the assessment criteria provided in Section II.

The remainder of this specification provides you with detailed requirements for each area of content – you should read it very carefully.

**1. Introduction**

* Describe the CRISP-DM methodology for data analytics projects (Wirth and Hipp, 2000) and discuss its applications and importance with appropriate reference to the literature.
* State the data analytic task in your coursework and explain how you are applying CRISP-DM to your coursework.
* State the insight you intend to gain in the coursework.

### 2. Data Understanding, Data Pre-processing, Exploratory Data Analysis

* Describe the Survey Data Set (Stack Overflow Developer Survey, 2020)
* Briefly describe data attributes with attribute name, description and data type use descriptive statistics and exploratory data analysis (Larose, and Larose, 2015). Note: it is not necessary include all results (e.g., tables and figures) in the main text. Only select most important results in the main text and leave others in the appendix.
* Describe the characteristics of the data set, such as (though not limited to) the number of instances, possibly duplicate or conflicting instances, missing values, or erroneous values, outliers.
* If any duplicate or conflicting instances, missing values, outliers/erroneous values, outliers exist, describe the process of cleaning these data.
* Conduct the exploratory data analysis on the data set, for example (though not limited to), identify outliers using histogram or box-plot, or scatter-plot; visualise the percentage of classes using pie-plot; explore the relation between features and target variable using crosstab and staked bar plot and so on.

### 3. Cluster Analysis

* Split the data set into two subsets, labelled low-income (= up to the median of all developer’s annual salaries) or high-income (= more than the median of all developer’s annual salaries). Note: income is given in the column **ConvertedComp** which represents Salary converted to annual USD salaries. Perform cluster analysis of these two subsets separately using some clustering methods (such as k-Means and hierarchical clustering).
* If applicable to a machine leaning method, describe the process of data transformation and normalization used in that method.
* Implement cluster analysis on each subset. Describe parameter setting, initialisation, stopping criterion and discuss your choice of cluster number.
* Describe characteristics of low-income developer clusters and high-income developer clusters found in cluster analysis.

**4. Machine Learning for Classification and their Implementation**

* Describe the workflow of machine learning for classification using a flow-chart(s).
* State and describe classification methods that are used in your coursework. At least three classifiers should be chosen for the classification problem. The methods may be chosen from those taught in this module, such as k-Nearest Neighbour, Decision Trees, Logistic Regression, Artificial Neural networks. It is also allowed to choose methods that are not taught in this module.
* Describe parameter setting, data transformation and normalisation in the methods that you have chosen for the task.
* If applicable to a machine leaning method, describe the process of data transformation and normalization used in that method.
* Build and implement machine learning models and tune hyper-parameters in these models for good performance. You may implement these models using Sklearn modules. It is also allowed to use any other Python libraries that are not taught in this module.
* Implement ensemble learning of combining your classifiers together. Describe the ensemble method(s) that you are using.

**5. Evaluation Machine Learning Models**

* Evaluate and compare the performance of the machine learning models (both base and ensemble models). You should at least use one or more of the performance metrics (as appropriate), such as accuracy, confusion matrix, recall and precision, or ROC curve.
* Explain your results. Generate tables to list the results or figures to visualize the results.
* Review the performance of different models (base and ensemble models). You may critically review which model performed best and which hyper-parameter settings were most effective? Provide necessary explanations.

**6. Discussions and Conclusions**

* Summarise your task and your findings in the data analysis on this survey data set.
* Describe what kind of insight that you have gained from the module “machine learning for data analytics”.
* Explain whether and how well has the module developed your understanding of machine learning for data analytics?

Finally, it must be pointed out that there exist a few notebooks on Kaggle and GutHub which are working on Stack Overflow Annual Developer Survey 2020. It is allowed for you to study these notebooks, but you must implement own coding in your coursework. You have one chance to check the similarity between your coursework and others through submitting your report and source code to Draft report and Draft code respectively on NOW Dropbox. Turnitin similarity score for report must be below 30%, while the score for your code below 50%. Coursework which violates the thresholds will be downgraded.

**References**

1. Larose, D.T. and Larose, C.D., 2015. *Data mining and predictive analytics*. John Wiley & Sons.
2. Stack Overflow Developer Survey, 2020. *Survey Data Set*, viewed 1 November 2020, <https://insights.stackoverflow.com/survey>

Wirth, R. and Hipp, J., 2000. *CRISP-DM: Towards a standard process model for data mining*. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). London, UK: Springer-Verlag.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class/**  **Grade/**  **Assessment Criteria** | **First**  **Low 13 | Mid 14| High 15**  **\*Exceptional Distinction 16** | **Upper Second**  **Low 10 | Mid 11| High 12** | **Lower Second**  **Low 7 | Mid 8| High 9** | **Third**  **Low 4 | Mid 5|High 6** | **Fail**  **Low 1 | Mid 2 | Marginal 3**  **\*Zero 0** |
| Section 1. Introduction  Weighting is 0.1 | An excellent description of the CRISP-DM methodology and its applications.  An excellent description of the machine learning task in the coursework.  A deep understanding of the problem area and related industry.  An excellent use of sources that evidence independent study and, in some cases, content that is not taught.  Excellent justification and intended insight are given.  \*Writing and content are nearly perfect | A good description of account is of the CRISP-DM methodology and its applications with appropriate reference to the literature but may miss some details.  A good description of the machine learning task in the course work.  A good understanding of the problem area and related industry.  Good justification and intended insight are given. | A reasonable description of the CRISP-DM methodology and its applications but may miss many details and but lacks appropriate reference to the literature. A reasonable description of the machine learning task in the coursework.  A reasonable understanding of the problem area and related industry but without depth.  Reasonable justification and intended insight given. | Some description of the CRISP-DM methodology and its applications without details.  Some description of the machine learning task in the course work.  Little understanding of the problem area Insufficient justification may not be based on the literature is provided.  Some justification and intended insight mentioned. | No meaningful description of the CRISP-DM methodology and its applications.  No information about the machine learning task in the coursework.  No literature is provided and/or consideration of intended outcomes.  \* This section is missing |
| Section 2. Data Understanding, Data Pre-processing, Exploratory Data Analysis  Weighting is 0.2 | An excellent explanation of the data using statistical summary with details.  An excellent description of the relevant data pre-processing with excellent critical consideration and justification.  An excellent exploratory data analysis with excellent use of graphics and detailed explanation. In some cases, show independent study of content that is not taught.  \*Writing and analysis are nearly perfect | A good explanation of the data using statistical summary but may miss some details.  A good description of the relevant data pre-processing with some critical consideration and justification.  A good exploratory data analysis with appropriate use of graphics but may miss some details. mainly using the module content/notes/resources. | A reasonable explanation of the data but may lack some statistical summary.  A reasonable description of the relevant data pre-processing but with insufficient critical consideration and justification.  A reasonable exploratory data analysis but with insufficient use of graphics and miss some details. | Some explanation of the data but may lack statistical summary.  Some description of the relevant data pre-processing but without critical consideration and justification  Some exploratory data analysis but without use of graphics and miss some details. | No meaningful explanation of the data and no statistical summary.  No meaningful description of the relevant data pre-processing and no critical consideration and justification.  No meaningful exploratory data analysis.  \* This section is missing |
| Section 3. Cluster Analysis  Weighting is 0.1 | An excellent cluster analysis with detailed explanations of clusters.  In some cases, show independent study of content that is not taught.  \*Writing and analysis are nearly perfect | A good cluster analysis but may missing some detail.  The analysis is clear and detailed, but mainly using the module content/notes/resources. | A reasonable cluster analysis and miss some details and explanations. | Some cluster analysis but lack explanations of results. | No meaningful cluster analysis.  \* This section is missing |
| Section 3. Machine Learning for Classification and their Implementation  Weighting is 0.3 | An excellent description of machine learning workflow with excellent flow chat.  An excellent explanation and justification of the selected machine learning methods (including ensemble) and if applicable, with excellent use of mathematical formulas or diagrams.  An excellent description of data transformation and normalization with details.  An excellent use of hyper-tuning methods with detailed explanation.  In some cases, show independent study of content that is not taught.  The performance of the built machine learning models is on the top within the class.  \*Writing and implementation are nearly perfect | A good description of machine learning workflow with flow chat but may miss some details.  A good explanation and justification of the selected machine learning methods (including ensemble) but miss some details, and if applicable, may without use of mathematical formulas or diagrams.  A good description of data transformation and normalization but may miss some detail.  A good use of hyper-tuning methods but not with detailed explanation.  The performance of the built machine learning models is good. mainly using the module content/notes/resources. | A reasonable description of machine learning workflow but may without flow chat and miss some details.  A reasonable explanation and justification of the selected machine learning methods (including ensemble) but may miss many details and if applicable, without use of mathematical formulas or diagrams.  A reasonable description of data transformation and normalization but may miss many details.  A reasonable use of hyper-tuning methods but without explanation. The performance of the built machine learning models is acceptable. | Some description of machine learning workflow but may be incomplete.  Some explanation and justification of the selected machine learning methods (including ensemble) but without much details.  Some description of data transformation and normalization but without much details.  Some use of hyper-tuning methods but without explanation. The performance of the built machine learning models is poor. | No meaningful description of machine learning workflow.  No meaningful explanation and justification of the selected machine learning methods.  No meaningful description of data transformation and normalization.  No use of hyper-tuning methods.  No implementation using Python language.  \* This section is missing |
| Section 4. Evaluation Machine Learning Models  Weighting is 0.2 | Excellent and critical evaluation and comparison of the performance of different machine learning models through use of different performance metrics.  Excellent explanation of the performance metrics.  Excellent explanation of results using excellent tables or visualisation.  \*Writing and evaluation are nearly perfect | Good evaluation and comparison of the performance of different machine learning models through use of different performance metrics but may miss some critical evaluation.  Good explanation of the performance metrics but may miss some detail.  Good explanation with tables or visualisation. | Reasonable evaluation and comparison of the performance of machine learning models but use only use one or two performance metrics and may miss some critical evaluation.  Reasonable explanation of the performance metrics but without details.  Reasonable explanation of results but may without tables or visualisation. | Some evaluation and comparison of the performance of machine learning models but use only one performance metrics and may miss critical evaluation.  Some explanation of the performance metrics.  Some explanation of results but may without tables or visualisation. | No meaningful evaluation and comparison of the performance of machine learning models.  No meaningful explanation of the performance metrics.  No meaningful explanation of results.  \* This section is missing |
| Section 5. Discussions and Conclusions  Weighting is 0.1 | An excellent, clear and concise summary of the work and findings is given. An erudite account of the insight gained is provided together with clear indications as to how to advance the work further.  \*Writing and summary are nearly perfect | A good comprehensive summary and concise discussion is provided that clearly articulates the purpose and findings of the analytics project.  A good insight into how their data analytic skills have been developed by the assessment and how they could be further developed. | A reasonable summary and discussion is provided that clearly articulates the purpose and findings of the analytics project.  A meaningful insight into how their data analytic skills have been developed by the assessment and how they could be further developed. | Some summary of the work and findings is given.  Some insight gained is provided together and some indications as to how to advance the work further. | No summary of the work and findings is given.  No insight gained is provided together and no indications as to how to advance the work further.  \* This section is missing |