DEPARTMENT OF
SOFTWARE TECHNOLOGY

STINTSY

Machine Project

Major Details

Groupings:	At most 4 members in a group
Deadline:	November 29, 2024 (Friday) 6:00 PM
Demo Schedule:	December 2 to 6, 2024 (Week 14)
Percentage:	30%
Submission guidelines:	Submit the `zip` file to AnimoSpace
Filename format:	`STINTSY-Project-<Section>-Group<#>.zip`

Deliverables

`Zip` file containing:

- Jupyter Notebook file – `ipynb` file
- Other Python 3 files – `py` files
- Dataset files – `csv` files

Specifications

You are tasked to go through the process of selecting a dataset, describing the dataset, performing exploratory data analysis, data preprocessing and cleaning, model training, hyperparameter tuning, model selection, and extracting insights from the data.

The project is to be submitted as a Jupyter Notebook and, optionally, some Python 3 source files. The notebook should be a self-explanatory document containing a report of the entire process undertaken to come up with the generated insights from the raw dataset. It should contain markup cells explaining the processes undertaken in the project, as well as code cells showing all the code that was performed. Please make sure that the codes could be successfully run sequentially to replicate the processes done in the project. Offshoots (a new task different from the original task because the new task seems interesting) are also encouraged, but make sure that the original task has already been completed.

As part of your task for this project, you are required to set-up a consultation with your instructor at least once during the term. **The consultation should be done, at the latest, one week before**

the deadline of the project. During the consultation, you need to discuss your partial work. Completion of this requirement is part of the grading scheme for this project.

Outline for the Notebook

Section 1. Introduction to the problem/task and dataset

Each group should select one real-world dataset from the list of datasets provided for the project. Each dataset is accompanied with a description file, which also contains detailed description of each feature.

The target task (i.e., classification or regression) should be properly stated as well.

Section 2. Description of the dataset

In this section of the notebook, you must fulfill the following:

- State a brief description of the dataset.
- Provide a description of the collection process executed to build the dataset. Discuss the implications of the data collection method on the generated conclusions and insights. Note that you may need to look at relevant sources related to the dataset to acquire necessary information for this part of the project.
- Describe the structure of the dataset file.
  - What does each row and column represent?
  - How many instances are there in the dataset?
  - How many features are there in the dataset?
  - If the dataset is composed of different files that you will combine in the succeeding steps, describe the structure and the contents of each file.
- Discuss the features in each dataset file. What does each feature represent? All features, even those which are not used for the study, should be described to the reader. The purpose of each feature in the dataset should be clear to the reader of the notebook without having to go through an external link.

Section 3. List of requirements

List all the Python libraries and modules that you used.

Section 4. Data preprocessing and cleaning

Perform necessary steps before using the data. In this section of the notebook, please take note of the following:

- If needed, perform preprocessing techniques to transform the data to the appropriate representation. This may include binning, log transformations, conversion to one-hot encoding, normalization, standardization, interpolation, truncation, and feature engineering, among others. There should be a correct and proper justification for the use of each preprocessing technique used in the project.

- Make sure that the data is clean, especially features that are used in the project. This may include checking for misrepresentations, checking the data type, dealing with missing data, dealing with duplicate data, and dealing with outliers, among others. There should be a correct and proper justification for the application (or non-application) of each data cleaning method used in the project. Clean only the variables utilized in the study.

Section 5. Exploratory data analysis

Perform exploratory data analysis comprehensively to gain a good understanding of your dataset. In this section of the notebook, you must present relevant numerical summaries and visualizations. Make sure that each code is accompanied by a brief explanation. The whole process should be supported with verbose textual descriptions of your procedures and findings.

Section 6. Initial model training

Use machine learning models to accomplish your chosen task (i.e., classification or regression) for the dataset. In this section of the notebook, please take note of the following:

- The project should train and evaluate at least 3 different kinds of machine learning models. The models should not be multiple variations of the same model, e.g., three neural network models with different number of neurons.
- Each model should be appropriate in accomplishing the chosen task for the dataset. There should be a clear and correct justification on the use of each machine learning model.
- Make sure that the values of the hyperparameters of each model are mentioned. At the minimum, the optimizer, the learning rate, and the learning rate schedule should be discussed per model.
- The report should show that the models are not overfitting nor underfitting.

Section 7. Error analysis

Perform error analysis on the output of all models used in the project. In this section of the notebook, you should:

- Report and properly interpret the initial performance of all models using appropriate evaluation metrics.
- Identify difficult classes and/or instances. For classification tasks, these are classes and/or instances that are difficult to classify. Hint: You may use confusion matrix for this. For regression tasks, these are instances that produces high error.

Section 8. Improving model performance

Perform grid search or random search to tune the hyperparameters of each model. You should also tune each model to reduce the error in difficult classes and/or instances. In this section of the notebook, please take note of the following:

- Make sure to elaborately explain the method of hyperparameter tuning.

- Explicitly mention the different hyperparameters and their range of values. Show the corresponding performance of each configuration.
- Report the performance of all models using appropriate evaluation metrics and visualizations.
- Properly interpret the result based on relevant evaluation metrics.

Section 9. Model performance summary

Present a summary of all model configurations. In this section of the notebook, do the following:

- Discuss each algorithm and the best set of values for its hyperparameters. Identify the best model configuration and discuss its advantage over other configurations.
- Discuss how tuning each model helped in reducing its error in difficult classes and/or instances.

Section 10. Insights and conclusions

Clearly state your insights and conclusions from training a model on the data. Why did some models produce better results? Summarize your conclusions to explain the performance of the models. Discuss recommendations to improve the performance of the model.

Section 11. References

Cite relevant references that you used in your project. All references must be cited, including:

- **Scholarly Articles**
  - Cite in APA format, and put a description of how you used it for your work.
- **Online references, blogs, articles that helped you come up with your project**
  - Put the website, blog, or article title, link, and how you incorporated it into your work.
- **Artificial Intelligence (AI) Tools**
  - Put the model used (e.g., ChatGPT, Gemini), the complete transcript of your conversations with the model (including your prompts and its responses), and a description of how you used it for your work.

<u>Final Project Presentation</u>

Here are some guidelines regarding the final project presentation:

- Each group is given 45 minutes: 25 minutes to present, and 20 minutes for Q&A.
- Presentations will be done either online or face-to-face.
- Open all the necessary files before your allotted presentation time slot. Do not wait until the presentation itself to load anything.
- All members should be present and should discuss a part in the final project presentation.
- Kindly read the rubrics to check different requirements and expectations on the project presentation.

Working With Groupmates

For this project, you are encouraged to work in groups of at most 4 members. Make sure that each member of the group has approximately the same amount of contribution for the project. Problems with groupmates must be discussed internally within the group, and if needed, with the lecturer.

Use of Artificial Intelligence Tools

You are allowed to use AI tools to assist you in the creation of your work, under the following conditions:

1. You must declare the use of such tools following the prescribed format in the *References* section of your submission (see Section 11 under the outline for the Notebook).
2. You must **not** use any code written by AI directly in your submission.
3. You must validate any AI response through your own understanding of the concepts or through your own research.
4. You must be able to articulate the thought processes, rationales, and implementation details of your work, and through this you must be able to show that human agency was maintained even if AI was used in augmenting the process.

Use of AI outside of these parameters is considered academic dishonesty.

Deliverables

Submit a `zip` file containing the source code files via AnimoSpace. All exploratory data analysis, machine learning, and core algorithms should be performed using Python 3 code and integrated into the Jupyter Notebook. Other code that you used for the project other than those in the Notebook should also be included in the submission of the project.

Academic Honesty Policy

Honesty policy applies. Please take note that you are NOT allowed to borrow and/or copy-and-paste – in full or in part – any existing related program code or solutions from the internet or other sources (such as printed materials like books, or source codes by other people that are not online). You should develop your own codes and solutions from scratch by yourselves.

The student handbook states that (Sec. 5.2.4.2):

"Faculty members have the right to demand the presentation of a student's ID, to give a grade of 0.0, and to deny admission to class of any student caught cheating under Sec. 5.3.1.1 to Sec. 5.3.1.1.6. The student should immediately be informed of his/her grade and barred from further attending his/her classes."

The student handbook also states that (Sec. 10.3):

A student caught cheating, as defined in Sec. 5.3.1.1., shall be penalized with a grade of 0.0 in the requirement or in the course, at the discretion of the faculty member, without prejudice to an administrative sanction. In cases of alleged cheating, the faculty member should report the incident to the Student Discipline Formation Office (SDFO).

RUBRIC FOR GRADING

<table>
<tr><th>Criteria</th><th colspan="12">Ratings</th><th>Points</th></tr>
<tr><td><b>Description of the dataset and the task</b></td><td colspan="4"><b>COMPLETE
5 pts</b>
An overview or description of the data is provided, including how it was collected, and its implications on the types of conclusions that could be made from the data. A description of the variables, observations, and/or structure of the data is provided.
The target task is well introduced and clearly defined.</td><td colspan="4"><b>INCOMPLETE
2 pts</b>
An overview or description of the data is provided but lacks details. A description of variables, observations, and/or structure is present but is missing for some aspects of the dataset.
The task is not clearly defined.</td><td colspan="4"><b>NO MARKS
0 pt</b>
No overview or description of the data is provided.
No description of variables, observations, and/or structure is provided.
The task is not defined.</td><td>5 pts</td></tr>
<tr><td><b>Exploratory data analysis</b></td><td colspan="3"><b>COMPLETE
5 pts</b>
The data is sufficiently explored to get a grasp of the distribution and the content of the data. Appropriate summaries and visualizations are presented. Insights into how the EDA can help the model training is mentioned.</td><td colspan="3"><b>INCOMPLETE
3 pts</b>
Exploratory data analysis is not sufficiently performed. Summaries and visualizations are presented but have minor issues in terms of methods chosen.</td><td colspan="3"><b>INCOMPLETE
1 pt</b>
Exploratory data analysis is rudimentary. Inappropriate methods of summarizing and visualizing data are frequently chosen.</td><td colspan="3"><b>NO MARKS
0 pt</b>
No exploratory data analysis is attempted.</td><td>5 pts</td></tr>
<tr><td><b>Knowledge about exploratory data analysis</b></td><td colspan="4"><b>COMPLETE
5 pts</b>
The group was able to discuss the exploratory data analysis process correctly. All questions about this section were answered correctly and sufficiently.</td><td colspan="4"><b>INCOMPLETE
2 pts</b>
Some parts of the exploratory data analysis process were not correctly discussed. Some questions about this section were answered correctly and sufficiently.</td><td colspan="4"><b>NO MARKS
0 pt</b>
The group was not able to discuss the exploratory data analysis process at all. Questions about this section were not answered correctly.</td><td>5 pts</td></tr>
</table>

<table>
<tr><td><b>Data pre-processing and cleaning</b></td><td colspan="3"><b>COMPLETE</b>
<b>5 pts</b>
The necessary steps for pre-processing and cleaning are performed, including explanations for every step for each feature. If no preprocessing or cleaning is done, there should be a justification on why it is not needed.</td><td colspan="3"><b>INCOMPLETE</b>
<b>3 pts</b>
Pre-processing and cleaning steps are performed but lacks explanation. Or, pre-processing and cleaning are insufficiently performed for less than half or half of the number of features.</td><td colspan="3"><b>INCOMPLETE</b>
<b>1 pt</b>
Pre-processing steps do not match the ML model chosen. Or, pre-processing and cleaning are insufficiently performed for more than half of the number of features.</td><td colspan="3"><b>NO MARKS</b>
<b>0 pt</b>
No pre-processing and cleaning are done, and no justification was provided as to why it was not done, or the justification is weak or incorrect.</td><td>5 pts</td></tr>
<tr><td><b>Knowledge about data pre-processing and cleaning</b></td><td colspan="4"><b>COMPLETE</b>
<b>5 pts</b>
The group was able to discuss the data pre-processing and cleaning process correctly. All questions about this section were answered correctly and sufficiently.</td><td colspan="4"><b>INCOMPLETE</b>
<b>2 pts</b>
Some parts of the data pre-processing and cleaning process were not correctly discussed. Some questions about this section were answered correctly and sufficiently.</td><td colspan="4"><b>NO MARKS</b>
<b>0 pt</b>
The group was not able to discuss the data pre-processing and cleaning process at all. Questions about this section were not answered correctly.</td><td>5 pts</td></tr>
<tr><td><b>Model training</b></td><td colspan="3"><b>COMPLETE</b>
<b>5 pts</b>
Appropriate models are used to accomplish the machine learning task. Justification of choosing the models is discussed.</td><td colspan="3"><b>INCOMPLETE</b>
<b>3 pts</b>
A lot of various models are used without proper justification. Or some of the models are not appropriate for the task.</td><td colspan="3"><b>INCOMPLETE</b>
<b>1 pt</b>
Only one model is generated. Or all of the models are not appropriate for the task.</td><td colspan="3"><b>NO MARKS</b>
<b>0 pt</b>
No model training is performed.</td><td>5 pts</td></tr>
<tr><td><b>Correctness of model training</b></td><td colspan="3"><b>COMPLETE</b>
<b>10 pts</b>
All models are trained correctly. The report shows that all models are not overfitting nor underfitting.</td><td colspan="3"><b>INCOMPLETE</b>
<b>7 pts</b>
At least two models are trained correctly. The report shows that at least two models are not overfitting nor underfitting.</td><td colspan="3"><b>INCOMPLETE</b>
<b>3 pts</b>
At least one model is trained correctly. The report shows that the model is not overfitting nor underfitting.</td><td colspan="3"><b>NO MARKS</b>
<b>0 pt</b>
The report shows no evidence proving that the models are not overfitting nor underfitting.</td><td>10 pts</td></tr>
<tr><td><b>Knowledge about model training</b></td><td colspan="4"><b>COMPLETE</b>
<b>5 pts</b>
The group was able to discuss the model training process correctly. All</td><td colspan="4"><b>INCOMPLETE</b>
<b>2 pts</b>
Some parts of the model training process were not correctly discussed. Some questions about</td><td colspan="4"><b>NO MARKS</b>
<b>0 pt</b>
The group was not able to discuss the model training process at all.</td><td>5 pts</td></tr>
</table>

	questions about this section were answered correctly and sufficiently.	this section were answered correctly and sufficiently.	Questions about this section were not answered correctly.		
Error analysis	**COMPLETE** **10 pts** Comprehensive error analysis is made based on the result of all models. Difficult classes and/or instances are correctly identified.	**INCOMPLETE** **7 pts** Error analysis is made based on the result of some models. There is an effort to identify difficult classes and/or instances.	**INCOMPLETE** **3 pts** Error analysis is made based on the result of one model. Or difficult classes and/or instances are not correctly identified.	**NO MARKS** **0 pt** No error analysis is performed.	10 pts
Knowledge about error analysis	**COMPLETE** **5 pts** The group was able to discuss the error analysis process correctly. All questions about this section were answered correctly and sufficiently.	**INCOMPLETE** **2 pts** Some parts of the error analysis process were not correctly discussed. Some questions about this section were answered correctly and sufficiently.	**NO MARKS** **0 pt** The group was not able to discuss the error analysis process at all. Questions about this section were not answered correctly.		5 pts
Improving model performance	**COMPLETE** **5 pts** Hyperparameter tuning and adjustments are performed to improve model performance. The study exhausts improvements that can be done to all models.	**INCOMPLETE** **3 pts** Hyperparameter tuning and adjustments are performed exhaustively but without proper justification or analysis. Or improvements to the models are not exhausted.	**INCOMPLETE** **1 pts** Hyperparameter tuning and adjustments are performed, but no efforts to further improve the model are done.	**NO MARKS** **0 pt** Hyperparameter tuning and adjustments are not performed.	5 pts
Knowledge about improving model performance	**COMPLETE** **5 pts** The group was able to discuss the process of improving model performance correctly. All questions about this section were answered correctly and sufficiently.	**INCOMPLETE** **2 pts** Some parts of the process of improving model performance were not correctly discussed. Some questions about this section were answered correctly and sufficiently.	**NO MARKS** **0 pt** The group was not able to discuss the process of improving model performance at all. Questions about this section were not answered correctly.		5 pts
Model performance summary	**COMPLETE** **5 pts** Multiple appropriate evaluation metrics and visualizations are used to report the performance of all	**INCOMPLETE** **2 pts** Incorrect evaluation metric or visualization is used to report the performance of at least one models.	**NO MARKS** **0 pt** No evaluation metric nor visualization is used to report the performance of the models.		5 pts

	models. Results are correctly interpreted.	Or, results are incorrectly interpreted.		
Knowledge about model performance summary	**COMPLETE** **5 pts** The group was able to discuss the summary of model performance correctly. All questions about this section were answered correctly and sufficiently.	**INCOMPLETE** **2 pts** Some parts of the summary of model performance were not correctly discussed. Some questions about this section were answered correctly and sufficiently.	**NO MARKS** **0 pt** The group was not able to discuss the summary of model performance at all. Questions about this section were not answered correctly.	5 pts
Notebook	**COMPLETE** **5 pts** The report discusses all steps in the machine learning process.	**INCOMPLETE** **2 pts** The report discusses some steps in the machine learning process.	**NO MARKS** **0 pt** No steps are discussed in the report.	5 pts
Consultation	**COMPLETE** **5 pts** The group consulted with their instructor regarding this project at least once.		**NO MARKS** **0 pt** The group did not consult with their instructor about this project.	5 pts
Presentation manner	**COMPLETE** **5 pts** The presenter seldomly looks at notes. The presenter displays a relaxed, self-confident nature about self, with no mistakes.	**INCOMPLETE** **2 pts** The presenter looks at his notes most of the time. The presenter displays mild tension; has trouble recovering from mistakes.	**NO MARKS** **0 pt** The presenter reads the entire report from his notes. The presenter displays tension and nervousness; has trouble recovering from mistakes.	5 pts
Presentation organization	**COMPLETE** **5 pts** Information is presented in a logical and interesting sequence which the audience can follow.	**INCOMPLETE** **2 pts** Audience has difficulty following the presentation because the presenter jumps around different topics.	**NO MARKS** **0 pt** Audience cannot understand the presentation because there is no logical sequence of information.	5 pts
			Total points:	100

Note: Each member of the group is expected to have a good understanding of the group's submission, even the parts that were not directly delegated to them. Failure to answer the questions during the demo, in a such a way that suggests that one or more group members did not sufficiently understand the work that was delivered, will result in a grade of 0 for those members for the entire project.