**1. Student bi-weekly performance summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Adm. No. | Name | No. of hours present | Progress1 | Remarks |
| 1. 2007476 | Johnnie | 8 | A | - |
| 2. 2112790 | Jayden | 8 | A | - |
| 3. 2112802 | Wee Loon | 8 | B | Distracted by CNY |

1 State whether: A=On Schedule B=Ahead Schedule for no. of days C=Behind Schedule for no. of days

**2. Weekly Scrum**

|  |  |
| --- | --- |
| Week No: 13-14 Date: 16/01/2022 | |
| Member Name 1: | **Johnnie (Data Engineer)** |
| Last week’s Progress | * Implemented Dask framework when importing data and used appropriate datatypes to save RAM. * Outlier detection: Apply domain knowledge to univariate outlier detection using interquartile ranges (IQR). * Outlier detection: Apply isolation forest algorithm to remove outliers, some domain knowledge still needed to remove values that do not make sense (e.g. negative speed) * Feature Selection: Used correlation matrix to remove features that I highly correlation (after feature engineering), to prevent multi-collinear issues within data. |
| This week deliverables | * Feature Selection: Applied RFECV using f1-score as metrics to select useful features (without categorical columns) * Feature Selection: Categorical columns are one hot encoded and added back into selected features from RFECV to compare result. * Unsupervised Learning: Trained an unsupervised learning model and interpreted the clusters. |
| Obstacles | * Categorical columns can be challenging to decide whether they are useful since categorical columns will be one hot encoded. (Have to manually add categorical columns and compare after performing RFECV or using permutation importance) |
| Member Name 2: | **Jayden (Data Engineer)** |
| Last week’s Progress | * Product Backlog: Allocated number of man hours our team will spend for each product (task). * Product Backlog: Created a list of products that we need to complete for project organisation. * Created Gantt chart for visualisation of project timeline. * Data Imputation |
| This week deliverables | * Sampling: Implemented multiple oversampling methods and did under sampling. * Sampling Analysis: Analysed objectively and compared different sampling methods against no sampling and eventually pick SMOTE oversampling. * Outlier Removal Comparisons: Trained and tested models on the three different datasets obtained from different outlier removal methods. |
| Obstacles | * Must be as objective as possible and weight out the trade-off between precision, recall and accuracy when deciding on the sampling method * Avoid data leakage by using pipelines |
| Member Name 3: | **Wee Loon (AI Engineer)** |
| Last week’s Progress | * Feature Engineering: Aggregated sensor data to summarize relevant information such as maximum, minimum, mean, standard deviation, kurtosis, and skewness. * Feature Engineering: Generated new columns such as age from driver date of birth. * Feature Engineering: Generated net acceleration from acceleration x, y and z. |
| This week deliverables | * Model Comparison: Compared different models on the dataset and selected the best model. * Model Improvement: Tuned the hyperparameters of the selected model to improve performance. |
| Obstacles | * Difficult to evaluate model as different metrics has to be considered when evaluating the best model. * Difficult to decide range of hyperparameters to test. Depending on the number of searches and how large the parameter space is, some parameters might not be explored enough. |