**DSCI-549: Introduction to Computational Thinking and Data Science**

Vu Truong Si – 6031936649

**Project part 1**

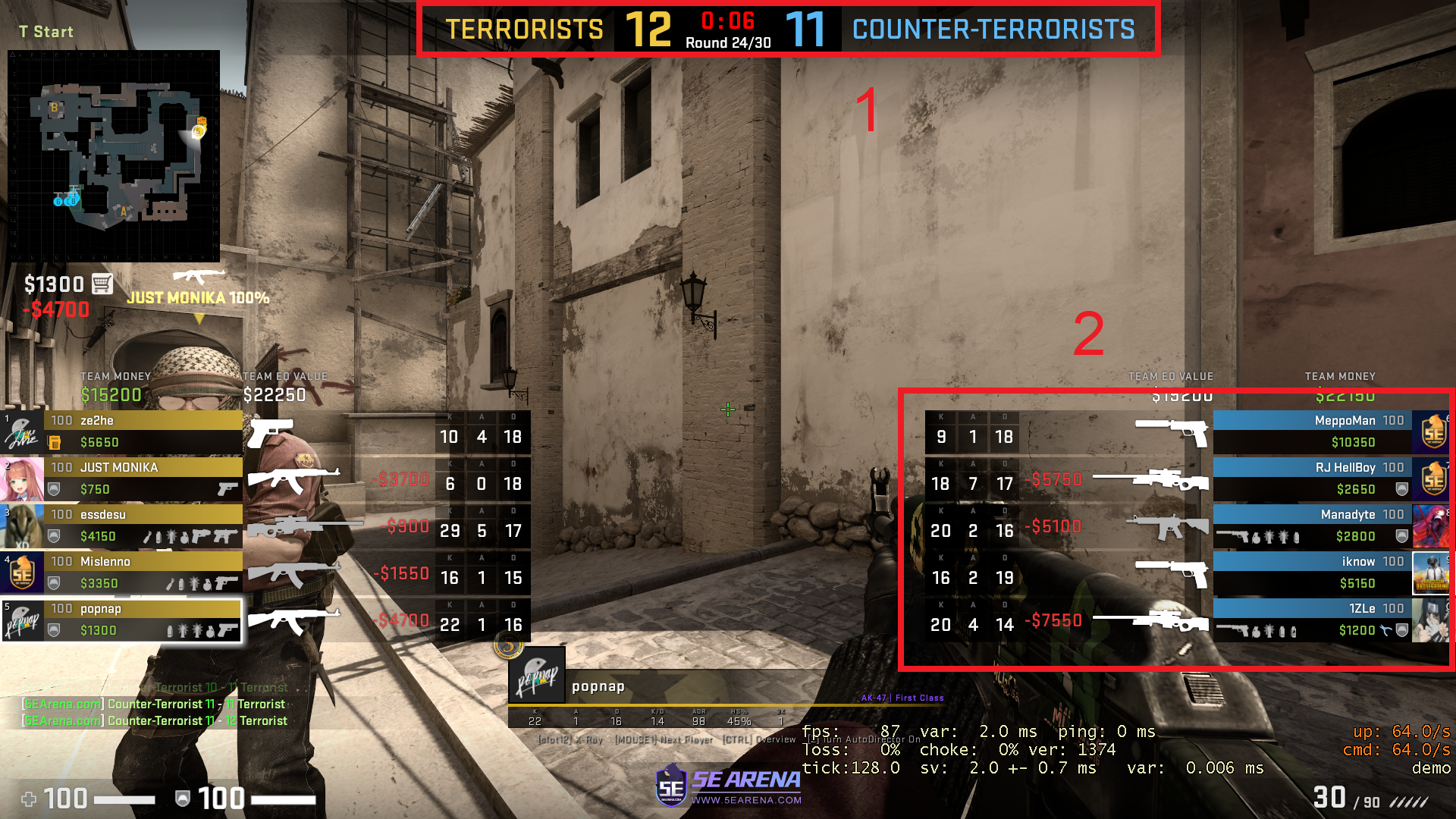
Part 1: Data (50 points)

1. Define an interesting problem that could be answered through data analysis. Describe the problem to a non-informed audience (10 points)

**Counter-Strike: Global Offensive (CS:GO)** is a multiplayer online shooting game developed by Valve. There are multiple game modes in CS:GO, however, the most prominent and widely played mode is called Competitive, in which two teams, also known as Terrorists **(T)** and Counter-Terrorists **(CT)**, of 5 members will play against each other for a maximum of 30 rounds. The Terrorists’ objective each round is to plant a bomb or eliminate all the Counter-Terrorists, while the Counter-Terrorists’ goal is to defuse the planted bomb or eliminate all the Terrorists before the bomb is planted.

During gameplay, there are various factors that will help each team gain their advantages in each round. For instance, if a player has armor, the intake damage will be reduced, hence he will not be eliminated as fast as those who do not have armor. Some other factors are firepower (pistols, rifles, grenades), the number of players alive, or the number of defuse kits. The problem that this project is trying to solve is how to use those factors and analyze them to predict which team will come out on top.

To further clarify the variables, below is a screenshot during a competitive CS:GO match.



In rectangle box 1, we can see the scores of the two teams, round number and timer. Currently, TS is leading 12 to 11 and in 6 seconds, the freeze time will end and the round will officially begin.

In rectangle box 2, we can see the guns, grenades, and money of the CT team. Currently, 3 members of the CT team have a primary weapon, while the other two only have secondary weapons (pistols) because they have not spent their money.

1. Find one or more datasets on the Web that are relevant to these questions, and that are: 1) accessible, 2) released with an open license, and 3) machine processable. For one of the datasets, specify:
   1. A brief description of what the data represents, with a small excerpt of the data as an example (10 points)

The dataset used in this project is “CS:GO Round Winner Classification” by Christian Lillelund. Below is the link to the dataset on Kaggle: <https://www.kaggle.com/christianlillelund/csgo-round-winner-classification>

a. The dataset contains snapshots of around 700 Competitive CS:GO games in 2019 and 2020. The snapshots are recorded 20 seconds between each other until the round ends. There are a total of 122411 snapshots, corresponding to **122411** rows and **97** features (The weapons that the Terrorists have, the amount of armor the Counter-Terrorists have, etc.), corresponding to **97** columns. Below is an example of the first 7 features of some snapshots in the dataset:

# *Table 1: First 7 features of 10 first rows of the data*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **time\_left** | **ct\_score** | **t\_score** | **map** | **bomb\_planted** | **ct\_health** | **t\_health** |
| 175 | 0 | 0 | de\_dust2 | FALSE | 500 | 500 |
| 156.03 | 0 | 0 | de\_dust2 | FALSE | 500 | 500 |
| 96.03 | 0 | 0 | de\_dust2 | FALSE | 391 | 400 |
| 76.03 | 0 | 0 | de\_dust2 | FALSE | 391 | 400 |
| 174.97 | 1 | 0 | de\_dust2 | FALSE | 500 | 500 |
| 114.97 | 1 | 0 | de\_dust2 | FALSE | 500 | 500 |
| 94.97 | 1 | 0 | de\_dust2 | FALSE | 295 | 111 |
| 74.97 | 1 | 0 | de\_dust2 | FALSE | 295 | 111 |
| 54.97 | 1 | 0 | de\_dust2 | FALSE | 249 | 89 |
| 174.97 | 2 | 0 | de\_dust2 | FALSE | 500 | 500 |

Take row 1 for example. Firstly, there are 175 seconds left in the round and the game has just begun since CT score and T score are both 0. The teams are playing in a map named de\_dust2, and looking at the other variables, it looks like that the round has just begun because all the values of the features are default values of the game.

* 1. A discussion on how the dataset satisfies the three criteria above (20 points)

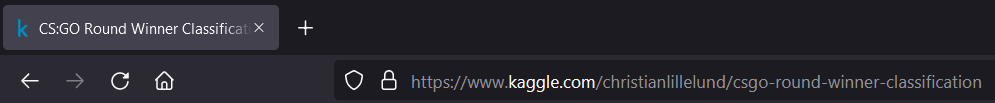
The dataset satisfies the requirements of the question because:

1. It is **accessible** and **downloadable** from Kaggle
2. Its license is **CC0**: Public Domain
3. It can be processed by machine because it is a **Comma Separated Value** (csv) file. The data is tabular that can be read by many software
   1. The **details** for how you can access the data (10 points)

**Use** the technical terms that you have learned in class to describe data.

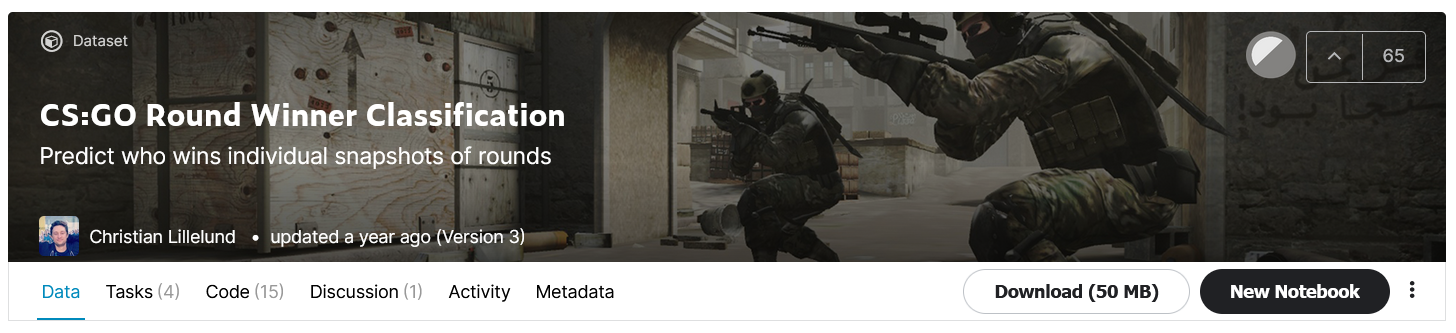
To access the data, follow the following steps:

* Open browser and go to : <https://www.kaggle.com/christianlillelund/csgo-round-winner-classification>



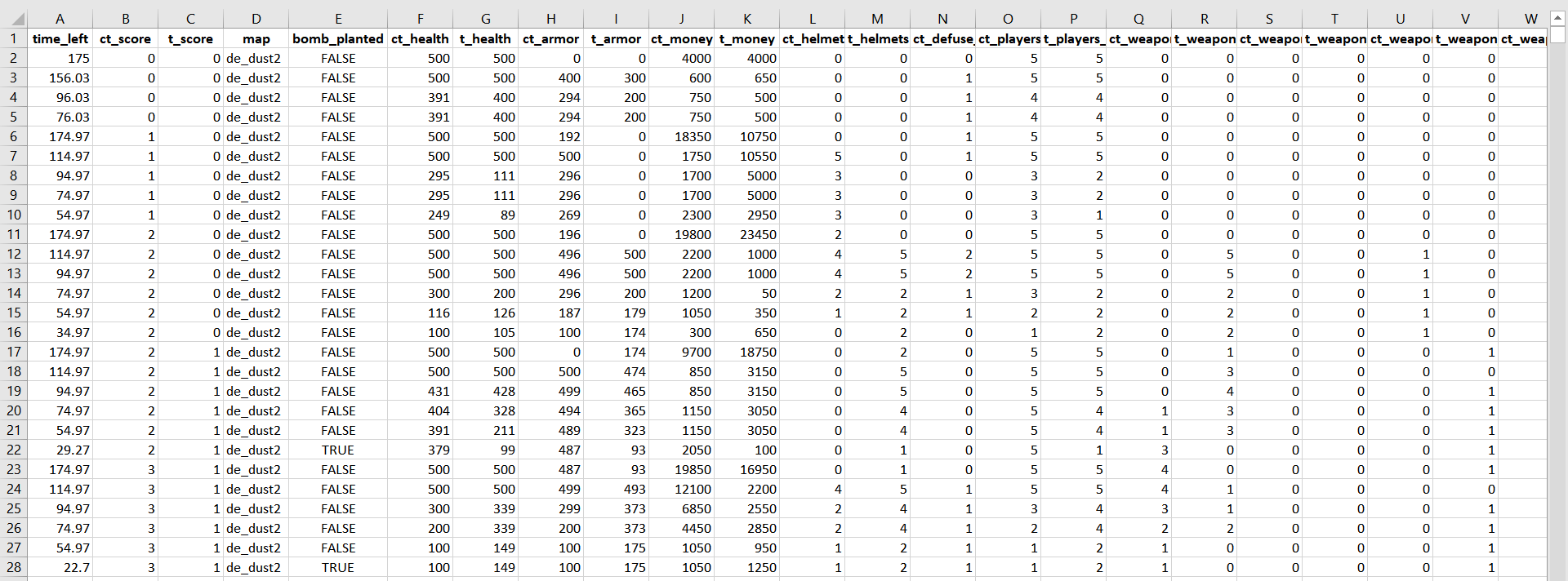
***Figure 1: Browser with dataset link***

* Click the Download (50 MB) button to download the dataset to your computer



***Figure 2: The dataset on Kaggle***

* Open the file with Excel or other software



***Figure 3: The dataset in Microsoft Excel***

The dataset is **tabular** data. Most columns contain numerical values (floats and integers), while some are text values (strings) and one column has boolean values.

Part 2: Hypothesis (30 points)

1. Is the study you want to conduct experimental or observational? Why? How does that affect your conclusions? (10 points)

The study is **observational** because it simply measures or surveys the data without affecting the data itself. In other words, the study does not interfere with the data and only observes what is happening with it. The fact that this is an observational study makes it impossible to establish causal connections and only allows us to associate variables.

1. State the null and alternate hypothesis (10 points)

The project problem is **not** formulated as a Hypothesis test, so there are no main null and alternate hypotheses for it. However, this dataset leads to many sub-questions that can be stated as binary claims, below are some possible hypotheses for the sub-problems:

|  |  |
| --- | --- |
| **H0 (Null hypothesis)** | **H1 (Alternate hypothesis)** |
| There is no statistically significant evidence between the *starting money* of two teams and the *round winner* | There is statistically significant evidence between the *starting money* of two teams and the round winner |
| There is no statistically significant evidence between the *number of defuse kits* of the CT side after the bomb has been planted and the *round winner* | There is statistically significant evidence between the *number of defuse kits* of the CT side after the bomb has been planted and the *round winner* |
| There is no statistically significant evidence between the *map* and the *total win rounds* of a side | There is statistically significant evidence between the *map* and the *total win rounds* of a side |
| There is no statistically significant evidence between the number of *primary weapons* and the *round winner* | There is statistically significant evidence between the number of *primary weapons* and the *round winner* |

1. What are the dependent and independent variables? (10 points)

The **independent** **variables** are variables that affect other variables and are manipulated in a study. Therefore the independent variables in this study are the 96 features (columns) of the data:

['time\_left', 'ct\_score', 't\_score', 'map', 'bomb\_planted', 'ct\_health', 't\_health', 'ct\_armor', 't\_armor', 'ct\_money', 't\_money', 'ct\_helmets', 't\_helmets', 'ct\_defuse\_kits', 'ct\_players\_alive', 't\_players\_alive', 'ct\_weapon\_ak47', 't\_weapon\_ak47', 'ct\_weapon\_aug', 't\_weapon\_aug', 'ct\_weapon\_awp', 't\_weapon\_awp', 'ct\_weapon\_bizon', 't\_weapon\_bizon', 'ct\_weapon\_cz75auto', 't\_weapon\_cz75auto', 'ct\_weapon\_elite', 't\_weapon\_elite', 'ct\_weapon\_famas', 't\_weapon\_famas', 'ct\_weapon\_g3sg1', 't\_weapon\_g3sg1', 'ct\_weapon\_galilar', 't\_weapon\_galilar', 'ct\_weapon\_glock', 't\_weapon\_glock', 'ct\_weapon\_m249', 't\_weapon\_m249', 'ct\_weapon\_m4a1s', 't\_weapon\_m4a1s', 'ct\_weapon\_m4a4', 't\_weapon\_m4a4', 'ct\_weapon\_mac10', 't\_weapon\_mac10', 'ct\_weapon\_mag7', 't\_weapon\_mag7', 'ct\_weapon\_mp5sd', 't\_weapon\_mp5sd', 'ct\_weapon\_mp7', 't\_weapon\_mp7', 'ct\_weapon\_mp9', 't\_weapon\_mp9', 'ct\_weapon\_negev', 't\_weapon\_negev', 'ct\_weapon\_nova', 't\_weapon\_nova', 'ct\_weapon\_p90', 't\_weapon\_p90', 'ct\_weapon\_r8revolver', 't\_weapon\_r8revolver', 'ct\_weapon\_sawedoff', 't\_weapon\_sawedoff', 'ct\_weapon\_scar20', 't\_weapon\_scar20', 'ct\_weapon\_sg553', 't\_weapon\_sg553', 'ct\_weapon\_ssg08', 't\_weapon\_ssg08', 'ct\_weapon\_ump45', 't\_weapon\_ump45', 'ct\_weapon\_xm1014', 't\_weapon\_xm1014', 'ct\_weapon\_deagle', 't\_weapon\_deagle', 'ct\_weapon\_fiveseven', 't\_weapon\_fiveseven', 'ct\_weapon\_usps', 't\_weapon\_usps', 'ct\_weapon\_p250', 't\_weapon\_p250', 'ct\_weapon\_p2000', 't\_weapon\_p2000', 'ct\_weapon\_tec9', 't\_weapon\_tec9', 'ct\_grenade\_hegrenade', 't\_grenade\_hegrenade', 'ct\_grenade\_flashbang', 't\_grenade\_flashbang', 'ct\_grenade\_smokegrenade', 't\_grenade\_smokegrenade', 'ct\_grenade\_incendiarygrenade', 't\_grenade\_incendiarygrenade', 'ct\_grenade\_molotovgrenade', 't\_grenade\_molotovgrenade', 'ct\_grenade\_decoygrenade', 't\_grenade\_decoygrenade']

The **dependent variable** is the outcome variable and is measured in a study. Therefore, in this study, the dependent variable is “round\_winnder”.

Part 3: Data exploration (50 points)

1. List and describe three data exploration steps that you will likely need. (note: visualization is acceptable as three separate steps as long as they allow you to explore three different characteristics of the dataset) (30 points)

* **Null and missing data detection**: This step checks for the null and missing values that can compromise our analysis. For example, Pandas’s functions isnull() or isna() combined with sum() and any() can find out if the dataset contains null or missing values and count them.

Graphical user interface, application

Description automatically generated

***Figure 4: Null values check***

There are currently no null or missing data in the dataset. For more detailed results, sum() is used to count missing values:

Text

Description automatically generated

***Figure 5: Missing values count***

* **Variable identification:** This step explores the data types of each column and decides if the variable is the predictor or the target. We can use Pandas dtype to view all the data types presented in the dataset.

A screenshot of a computer

Description automatically generated with medium confidence

***Figure 6: Datatypes of the features***

* **Univariate analysis:** This step explores the values of each feature in order to have a better understanding of the data. We can use either graphical or non-graphical techniques to perform univariate analysis.

Chart, bar chart, histogram

Description automatically generated

***Figure 7: ct\_score count***

Chart

Description automatically generated

***Figure 8: ct\_weapon\_awp***

Chart, bar chart

Description automatically generated

***Figure 9: round\_winner count***

Besides the three steps above, I would also like to include two more data exploration steps that I think will be useful:

* **Metrics analysis:** This is the phase where we look at the basic metrics of each feature such as mean, standard deviation or maximum value.

Table

Description automatically generated

***Figure 10: Dataset’s metrics***

* **Bivariate analysis:** This step analyzes the relationship between two features.

Chart, bar chart

Description automatically generated

***Figure 11: map vs round\_winner count***

Chart, bar chart

Description automatically generated

***Figure 12: ct\_defuse\_kits vs round\_winner count***

1. For each of the steps, describe how it will help you with pre-processing and analysis. (i.e., if my visualization indicates outliers, then I can remove them). (20 points)

* **Null and missing data detection**: This step is crucial before data pre-processing because it allows us to see whether or not our data has gaps that we need to fill. If there are missing values in our table, various methods can be applied to deal with them such as Deletion or Mean/Mode/Median Imputation.
* **Variables identification:** This step will help decide which features are categorical and whether or not label encoding is required so that the machine learning algorithm will work properly.
* **Univariate analysis:** This step can help see if there are features that have low variance or few unique values so that in pre-processing, relevant features can be selected to speed up the training process and also prevent overfitting.
* **Metrics analysis:** This step help to see if the differences between the metrics of the variables are significant and from that, decides whether or not normalization or standardization is required.
* **Bivariate analysis:** In this step, variables that are not strongly related to the target can be removed.

Part 4: Pre-processing (50 points)

1. List and describe three processing steps that you think are involved in the data science pipeline. (30 points)

Before performing data pre-processing, it is important to split the data into two sets, one for testing and one for training, and then preprocess them separatedly. The reason for this is because the testing set is crucial to validate how the model will work with unseen data, so if the testing set is pre-processed too, the model may produce biased results that seem better than in reality.

**Splitting the data into training and testing** **sets:** This step is important for model evaluation. Basically, the dataset will be split into two parts: a small portion of the data will be used for testing, while the rest will be used for training.

The three pre-processing steps, which are performed on the training set:

* **Missing values treatment:** In this step, a method to handle missing values will be selected. For example, if we choose the deletion method, the row with the missing data simply gets removed, or we can even delete the entire column in some cases. There are many other ways to deal with missing values namely mean, mode, or median imputation.
* **Label encoding:** Because we might encounter categorical variables that are not numerical in the dataset, it is wise to use label encoding to transform the data into numerical values so that the algorithm can handle it. For example, if the value of the “bomb\_planted” column is either True or False, we can convert it into 1 or 0.
* **Feature selection:** Since the dataset has many features, and not everyone is good for analysis, we may want to select the best features only. We can look at the “Select K best” method to select a certain number of features, or we can create a threshold to eliminate features with low variances.

1. For each of the steps, what do you expect to achieve and how does it help your analysis? (20 points)

* **Missing values treatment:** After treating missing values, the dataset is expected to be complete and has no gaps. This will prevent errors from appearing during data analysis. For example, the machine learning algorithm may fail if it encounters rows with missing data.
* **Label encoding:** After this step, categorical values that are non-numerical will be turned into numerical values. This will help the machine learning algorithm to work properly without running into errors when reading categorical data.
* **Feature selection:** After this step, some columns that are considered less important for predicting the results will be removed. This will increase the accuracy when performing data analysis.

Part 5: Analysis (50 points)

1. In your own words, give an overview of the analysis you would need to perform to answer your data science question. (20 points)

**For the main problem of this project, which is to predict the round winner based on the features, the following analysis can be applied:**

This is a classification problem, therefore, classification algorithms will be selected to solve it. Some algorithms that can be applied are Logistics Regression, Decision Tree, Random Forest, Naïve Bayes or K-nearest Neighbours.

In order to determine which model suits best, metrics such as accuracy, precision, recall, and f1-score will be considered. The model with the best scores overall will be selected to work with the dataset.

In this project, Logistic Regression, Decision Tree, and Random Forest will be used as candidates. After pre-processing, the testing data will be fed into the machine learning models, then the models will train and fit the data to produce the predicted labels. During model training, k-fold cross validation be used to validate the results, however, it will not be applied to Random Forest, because the mechanism of Random Forest is already similar to cross validation.

In k-fold cross validation, the data will be split into K parts and the “training-testing” process will run K times. For example, if K = 5, the data will be split into 5 parts. During the first iteration, the first part of the data will be used for validation, while the other 4 will be used for training. During the second iteration, the second part of data will be used for validation while the first and 3-5 parts will be used for training, and so on.

Without the validation scores, an assumption can still be made that Random Forest will perform better in the three because of its ability to handle datasets with large dimensions.

Finally, the predicted results will be visualized alongside with the real results to for overfitting and underfitting investigation. If either case happens, another pre-processing procedure will be performed to reduce overfitting or underfitting.

**For the sub-problems and stated in question 3, the following analysis can be carried out:**

Let us look closer at the hypothesis that *defuse kits* affect the round winner.

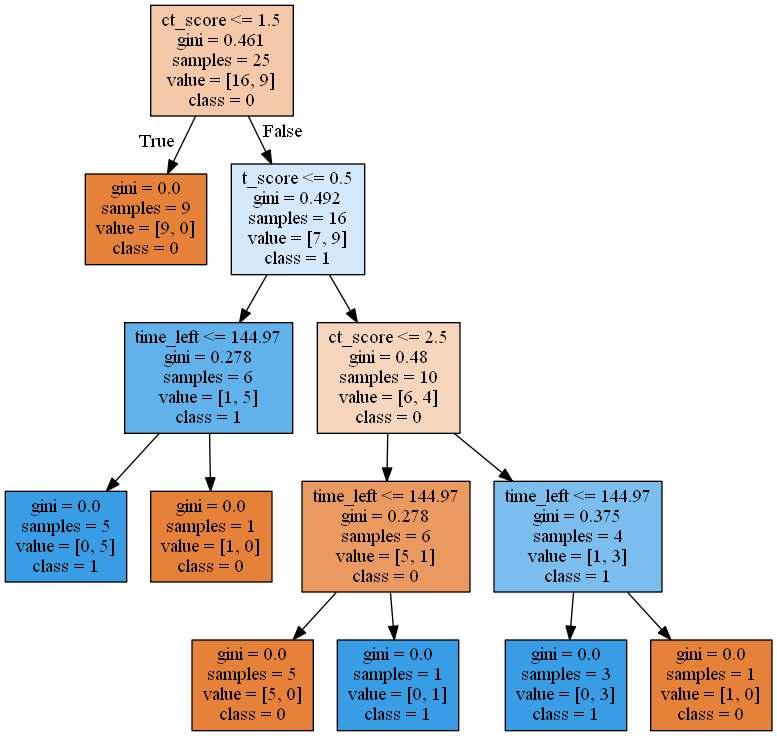
* Null hypothesis: There is no statistically significant evidence between the *number of defuse kits* of the CT side after the bomb has been planted and the *round winner*
* Alternate hypothesis: There is statistically significant evidence between the *number of defuse kits* of the CT side after the bomb has been planted and the *round winner*

Firstly, the training set will be divided into two parts, part 1 with all the rounds where the CT has defuse kits (ct\_defuse\_kits > 0), and part 2 with all the rounds where CT has no defuse kits (ct\_defuse\_kits = 0). The t-test is carried out for the two populations and consider the p-value, if the p-value is smaller than 0.05, the null hypothesis is rejected and if it is greater than 0.05, the alternate hypothesis is rejected.

1. Can you think of an algorithm learned in class that would be best to test your hypothesis? (30 points)
   1. If so, which one? Describe the algorithm

There are various classification algorithms to use in this data science problem, however, in this project, **Random Forest** is chosen.

In order to understand *random forest*, we first need to look at **Decision Tree**. In a decision tree, each node represents a test on a feature. For example, in the small tree below, the first node decides whether or not the feature “ct\_score” is smaller than or equal to 1.5 and then branches to two different nodes. In the “True” branch, the tree decides that the label will be 0, while in the “False” branch, another “test” is made with the feature “t\_score”.



***Figure 12: Decision tree example***

Generally speaking, in *random forest*, multiple decision trees will be formulated and the result of the random forest will be the output that most decision trees make. Let us look at how *random forest* works in detail.

Firstly, new datasets will be constructed based on the original dataset. The number of rows (observations) for each new dataset will be exactly the same as that of the original one. However, the rows themselves are randomly selected with replacement, this means after randomly selecting row K, we will put K back into the pool and it may be selected again in the future. This whole process is known as bootstrapping.

For example:

***Table 2: Original dataset***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | time\_left | ct\_score | t\_score | map | bomb\_planted |
| 1 | 175 | 0 | 0 | de\_dust2 | FALSE |
| 2 | 156.03 | 0 | 0 | de\_dust2 | FALSE |
| 3 | 96.03 | 0 | 0 | de\_dust2 | FALSE |
| 4 | 76.03 | 0 | 0 | de\_dust2 | FALSE |
| 5 | 174.97 | 1 | 0 | de\_dust2 | FALSE |

If the table above is the original dataset, then one new dataset can be:

***Table 3: Newly generated dataset***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | time\_left | ct\_score | t\_score | map | bomb\_planted |
| 4 | 76.03 | 0 | 0 | de\_dust2 | FALSE |
| 2 | 156.03 | 0 | 0 | de\_dust2 | FALSE |
| 3 | 96.03 | 0 | 0 | de\_dust2 | FALSE |
| 4 | 76.03 | 0 | 0 | de\_dust2 | FALSE |
| 3 | 96.03 | 0 | 0 | de\_dust2 | FALSE |

In this case, rows 4 and 3 are selected twice.

Next, only a subset of the features will be used for training, which means Decision Tree 1 may use “time\_left” , “ct\_score”, and “map” as the features, but Decision Tree 2 may use “bomb\_planted”, “map”, and “time\_left”.

After this step, the algorithm will generate N trees, each with its own new dataset and features. The trees will train the data and produce results respectively. Now when a new observation is fed into the random forest model, each tree will give its own prediction and the prediction with the most pickers is the result.

Part 6: Workflow (20 points)

1. Sketch the workflow for your data science problem, omitting the data exploration steps but including pre-processing and analysis. The workflow needs to include the data flow. The analysis step should be one component.

*Diagram

Description automatically generated*

***Figure 13: Workflow***