**Machine Learning Prediction - based Heart Disease Data Using Logistic Regression**

**Group 10 Members:**

**Ariel Peaceo Gunawan - 2301952874**

**Carlos Martius - 2301922745**

**Wishnu Anindito - 2301899452**

1. **Description:**

Machine Learning Prediction based on heart disease data is a prediction algorithm that uses training datasets that is used to predict whether a person has heart disease or not. This project was carried out by 3 Binusian students named Ariel Peaceo Gunawan, Carlos Martius, and Wishnu Anindito. To work on this project, we sought a dataset from an online community of data scientist and machine learning practice named Kaggle. The name of the dataset we are using is Heart.csv. From the creation of this project, we hope that the accuracy of our algorithm will reach or close to 0,9.

1. **Data: describe the data you will be using in detail (data exploration)**

Here is the medical data we got from Kaggle with the link: https://www.kaggle.com/ronitf/heart-disease-uci. The following is an explanation of each variable:

* age - age in years
* sex - (1 = male; 0 = female)
* cp - chest pain type
* trestbps - resting blood pressure (in mm Hg on admission to the hospital)
* chol - serum cholestoral in mg/dl
* fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
* restecg - resting electrocardiographic results
* thalach - maximum heart rate achieved
* exang - exercise induced angina (1 = yes; 0 = no)
* oldpeak - ST depression induced by exercise relative to rest
* slope - the slope of the peak exercise ST segment
* ca - number of major vessels (0-3) colored by flourosopy
* thal - 3 = normal; 6 = fixed defect; 7 = reversable defect
* target - have disease or not (1=yes, 0=no)

There are 14 types of unique data contained in this data. The data are: Age which has 41 data, Sex has 2 data, Chest pain type (cp) has 4 data, resting blood pressure (trestbps) has 49 data, Cholesterol serum has a total of 49. 152 data, Fasting blood sugar (fbs) which has 2 data, Resting electrocardiographic results (restecg) which has 3 data, Maximum heart rate achieved (thalach) which has 91 data, Exercise induced angina (exang) which has 2 data, ST depression induced by exercise relative to rest (oldpeak) which has 40 data, Slope of the peak exercise ST segment (slope) which has 3 data, Number of major vessels colored by flourosopy (ca) which has a total of 5 data, Thal which has a total of 4 data, and patient status (target) which has a total of 2 data. We did not find null value or NAn on the dataset we are using. Therefore, we do not need to perform data fixing on the dataset used. From the data we used, we managed to find the number of people who have heart disease, which is more than 500 people overall, with the sex being more prone to having heart disease, namely male.

1. **Experimental Design**
   1. Preprocessing

In the preprocessing section, we uploading the dataset into google colabs first so that the dataset can be access to make the machine learning algorithm. After importing the dataset file to google Colabs files, we examine the unique value of each variable. After we find the unique value of each variable, we decide to eliminate the attribute that has a unique value of only 2, because the effect is not too significant (too weak) on the prediction results. With the exception of the *target* variable, we did not eliminate even though the unique value was only 2, because we used it as an indicator variable to determine whether the person has heart disease or not. However, based on our results, we also eliminated sex, fbs, and exang variables because they only had 2 unique values. We performed correlation checking using the heatmap method to check the relationship closeness between all variables to the target variable. As a final step, we separate the dependent variable from the independent variable, so that we can use them to create a logistic regression function.

* 1. Main processing (Machine Learning Algorithm)

In the initial step, we used the initial dataset to be divided into 2 parts, namely train data and test data. We will use the data set that has been divided according to certain needs, namely: training the data to create a logistic regression model, and using test data to predict the results of the logistic regression model. For the next step, we create a logistic regression model using the dataset obtained. And as a final step, we use logistic regression modeling the results obtained from the training data, to test the variables.

1. **Experiment Result and Analysis**
   * 1. Measuring the performance of Logistic regression in terms of:
   * Accuracy

Defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions. The greater the value, the more accurate it is.

* + Mean Absolute Error:

Measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures *accuracy* for continuous variables. The equation is given in the library references. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average. Mean Absolute Error is negatively-oriented scores: Lower values are better.

* + Mean Squared Error:

Measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate. Mean Squared Error is also negatively-oriented scores: Lower values are better.

* + F1 (look for a definition)

A measurement that considers both precision and recall to compute the score. The F1 score can be interpreted as a weighted average of the precision and recall values, where an F1 score reaches its best value at 1 and worst value at 0.

* + 1. Print actual value and predicted value.
    2. Using the confusion matrix function to describe incorrect and correct prediction results, based on factual data:
  + The number of people who are predicted to have no disease and not have a disease in the real world is 111 people.
  + The number of people who are predicted to have the disease and have the disease in the real world is 126 people.
  + The number of people who are predicted not to have the disease, but have the disease in the real world is 23 people.
  + The number of people who are predicted to have the disease, but do not have the disease in the real world is 48 people.
    1. As a final step, we print a classification report, which is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report. The report shows the main classification metrics:
  + Precision

The ability of a classifier not to label an instance positive that is actually negative. For each class it is defined as the ratio of true positives to the sum of true and false positives.

* + Recall

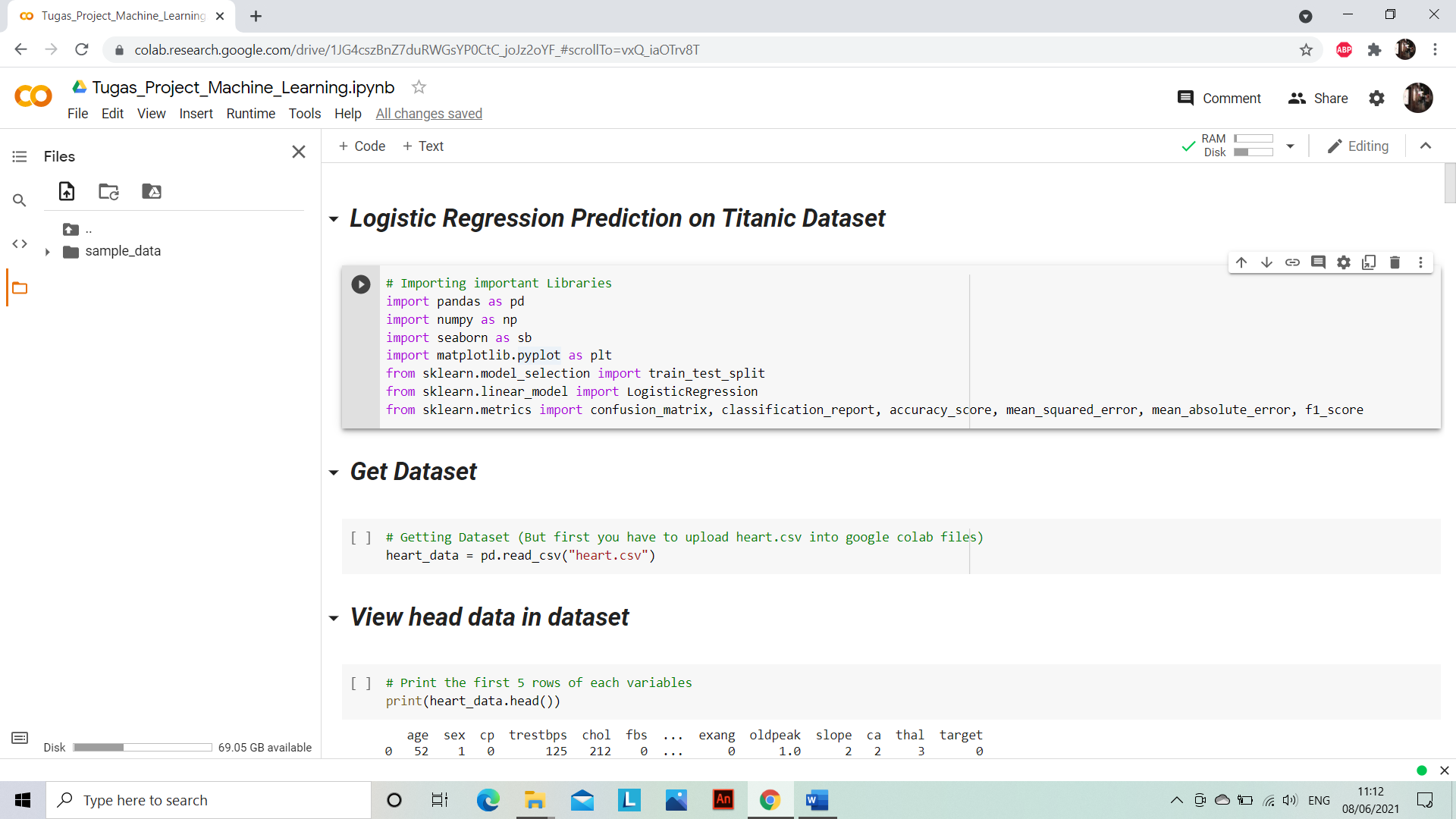
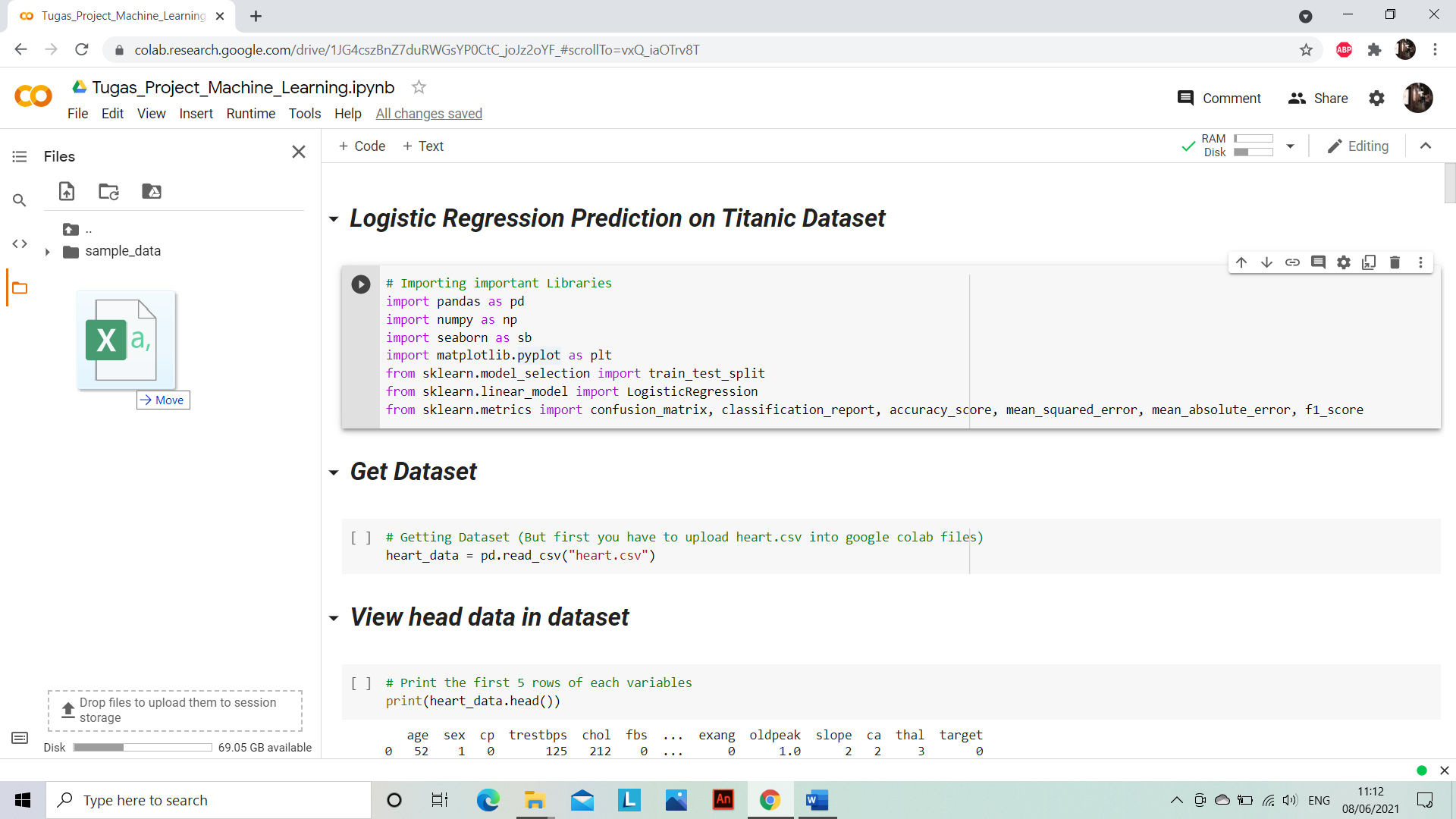
Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

* + - TP / True Positive: when a case was positive and predicted positive.
    - FN / False Negative: when a case was positive but predicted negative.
  + F1

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. Generally speaking, F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.



1. Documentation

Before doing development any Machine Learning algorithm, we must first upload the dataset into google colab files tab by using drag files and press upload file from bottom of text Files.







