Approach, Architecture, and Decisions

The project is separated into 2 tasks. Linear Regression and Classification. Same dataset was used for both tasks.
Linear Regression folder contains the source code(preprocessing.py, train_models.py, and evaluate.py) in the src folder, dataset, train_evaluate.ipynb and main.py. The models can be trained and evaluated using the train_evaluate.ipynb or the main.py files.
Similarly, **Classification** folder contains the saved models in the models folder, source code(preprocessing.py, train_models.py, and evaluate.py) in the src folder, dataset, eda.ipynb, train_evaluate.ipynb and main.py
As the dataset is the same , eda.ipynb is stored in the classification folder only.

For this task, I have used the Student Performance **Dataset** from UCI Irvine Machine Learning Repository. It contains about 30 features and has about 1045 instances. Column 'G3' contains the final exam scores of the students and has been used as the target variable for the Linear Regression Models. A new column 'Pass/Fail' was created where the students scoring more than 10 out of the total 20 marks were awarded a pass and the rest were assigned a fail. This column was used as the target variable for classification models.

Exploratory Data Analysis was applied to the dataset to understand the data.
There are 33 columns and 1045 rows and thankfully there is no missing value in the dataset.
From the data it can be seen that most of the students scored between 10-12. Also, more than 900 students passed the exam and only about 100 failed.

Looking at the data, there were some interesting questions that were answered.
Question 1- Does the romantic status of the students have anything to do with their exam scores?
Answer - Not really, as the amount of students who failed or passed was almost equal despite their relationship status.

Q2- Does Alcohol consumption on the weekend have anything to do with their scores?
Answer - The students with less consumption of Alcohol on the weekend have performed slightly better than their counterparts.

Q3 - What role does parents' education play in the final scores of the students?
Answer - A well educated mother has a very strong role in the final scores, father's education doesn't seem as important.

Q4 - Does the frequency of how many times a student goes out every week affect their grades?
Answer - Going out about twice a week seems to be working really well for the students. Those going out less than 2 times or more than 2 times a week performed worse.

A lot of other similar questions were looked into to understand the relationships in the data.

Linear Regression with Gradient Descent

I have trained multiple linear regression models:

1. Simple Linear Regression with ‘studytime’ as the only feature.
2. Multiple Linear Regression with multiple features.
3. Multiple Linear Regression with all the features.
4. K-fold cross validation for best r2 score.

‘G3’ column was used as the target variable as it contained the final scores of the students.

I have used a range of learning rates to optimize the models. Mean Absolute Error, r2 score, and the Mean Squared Error were used to evaluate the models. K fold cross validation was used to select the model with the best r2 score. Model with all features used performed the best as it produced an r2 score of 0.84. Cost function and gradient descent were visualized to make sure the best learning rate was selected.

Classification

For classification I have used multiple models:

1. Logistic Regression Gradient Descent
2. Sklearn Logistic Regression
3. Decision Tree Classifier

GridSearchCV was used to tune the hyperparameters of the Sklearn models. K-fold cross validation was used to select the best model of the Gradient Descent. The roc_curve, precision_recall curve, and the confusion matrix were plotted. The models were evaluated using accuracy,precision,recall, and roc_auc scores.

3 different types of training data was created:

1. Original Data
2. SMOTE Sampling
3. RandomUnderSampling

All of the models were trained using all the different training data and produced excellent results as they achieved accuracies of above 85%. Detailed results are in the train_evaluate.ipynb file.