Technical Report

**Grades Predictive Modelling Report**

This report presents the procedure of a predictive modelling study aimed at predicting students' grades based on various features. The study involves several steps, including exploratory data analysis, data preparation, and classification using different machine learning algorithms. The following sections provide a summary of each step and its corresponding results.

**1. Exploratory Data Analysis**

The exploratory data analysis (EDA) is conducted to gain insights into the dataset and understand the distribution and relationships between different variables. The analysis includes visualizations such as pie charts, bar plots, and other customized plots.

Some key findings from the EDA include:

* Grades Distribution: The pie chart illustrates the distribution of grades among the students, showing the percentage of students in each grade category.
* Grades by Course ID: The bar plot displays the number of students in each grade category for different course IDs, providing an overview of how grades vary across different courses.
* Distribution of Features: The EDA also explores various features, including age, weekly study hours, scholarship, attendance to classes, listening in classes, taking notes in classes, preparation for midterm exams, fathers' and mothers' education levels, transportation to university, and accommodation type.

**2. Data Preparation**

The data preparation step involves filtering the dataset to exclude certain categories and creating the input matrix (X) and target variable (y) for the predictive modelling task. The classes in the target variable are identified, and the data is split into training and test sets.

**3. Classification Study**

The classification study aims to identify the best model for predicting students' grades. The study involves the use of various machine learning algorithms, including logistic regression, support vector machines, random forest, and others. Different variations of these models are tested, including feature selection using variance threshold.

The performance of each model is evaluated using multiple metrics, such as accuracy, precision, recall, and area under the ROC curve (AUC). The results are visualized using performance plots.

The following subsections summarize the results of the classification study:

**3.1 Standard Estimators**

The performance of standard estimators without feature selection is evaluated using cross-validation. The models are trained on the training set, and their performance is assessed on the validation set. The results are visualized using performance plots.

**3.2 Standard Estimators & Feature Selection (Variance Threshold of 0.10)**

The performance of the standard estimators is evaluated after applying feature selection using a variance threshold of 0.10. The feature selection aims to improve the models' performance by removing features with low variance. The results are presented using performance plots.

**3.3 Standard Estimators & Feature Selection (Variance Threshold of 0.20)**

Like the previous section, the performance of the standard estimators is evaluated after applying feature selection with a higher variance threshold of 0.20. The results are visualized using performance plots.

**3.4 Grid Search for Best Estimators**

A grid search is conducted to find the best hyperparameters for the logistic regression models. The grid search considers different combinations of hyperparameters, such as the variance threshold, regularization strength (C), and multi-class handling (multinomial or one-vs-rest). The best models with optimized hyperparameters are selected based on the accuracy metric.

Overall, the classification study provides insights into the performance of different machine learning models for predicting students' grades. The results highlight the importance of feature selection and hyperparameter optimization in improving the models' accuracy and generalization capabilities.

**4. Test Set Performance**

The performance of the best estimators on the test set was evaluated.

**5. Best Classifiers for Grades Prediction**

**5.1.a Best SGD Classifier**

The best SGD (Stochastic Gradient Descent) classifier was trained and evaluated. The overall performance of the SGD classifier on the test set was assessed. The report includes metrics such as accuracy, precision, and recall. A classification report was generated for the SGD classifier on the test set. It provides detailed information on metrics such as precision and recall for each class. A logodds plot was created to visualize the linear coefficients of the SGD classifier for each attribute and grade. A probabilities plot was created to visualize the linear coefficients of the SGD classifier for each attribute and grade, considering the probability of the predicted class.

**5.1.b Best Extra Tree Classifier**

The best Extra Tree classifier was trained and evaluated. The overall performance of the Extra Tree classifier on the test set was assessed. The report includes metrics such as accuracy, precision, and recall. and F1-score. A classification report was generated for the Extra Tree classifier on the test set. It provides detailed information on metrics such as precision, recall, and F1-score for each class. A feature importance plot was created to visualize the importance of each feature in the Extra Tree classifier's decision-making process.