

|  |  |
| --- | --- |
| **Student Marks Prediction**  Report  **AI LAB Group Project BSCS 6A**  **Hirdesh Kumar 1812114**  **Malik Rafaquat 1812118**  **Syed Hurrar Hassan Rizvi 1812135 Muhammad Usama 1812137** |  |

|  |  |
| --- | --- |
| **Table Of Contents PgNo.**  [1. Problem statement 3](#_Toc75441963)  [2. Previous work done on this area. 3](#_Toc75441964)  [3. Description of the dataset 3](#_Toc75441965)  [4. Data cleaning techniques to be applied to the dataset. 5](#_Toc75441966)  [5. Description of the AI algorithm you've applied to your dataset. The label and the attributes. 5](#_Toc75441967)  [6. Result 7](#_Toc75441969)  [7. Conclusion 7](#_Toc75441970)  [8. Links to the dataset and any other external help 7](#_Toc75441971) Problem statement The main objective of this project was to assess the performance of students based on various factors such as gender, sex, age, addresss, parent marriage status, father’s eduaction, father’s job, mother’s education, mother’s job, absences, health and other condition. Previous work done on this area. The past work done on this topic is shared on Kaggle, where the people have tried to attempt this problem with algorithms KNN, Linear Regression, Logistic Regression, Naïve Bayes, XGB Classifier, Lasso Regression, and Decision Trees. Description of the dataset **Data Set Information:**  This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school-related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details). |  |
| **Attribute Information:**   1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira) 2. sex - student's sex (binary: 'F' - female or 'M' - male) 3. age - student's age (numeric: from 15 to 22) 4. address - student's home address type (binary: 'U' - urban or 'R' - rural) 5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3) 6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart) 7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education) 8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education) 9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other') 10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')   reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')   1. guardian - student's guardian (nominal: 'mother', 'father' or 'other') 2. traveltime - home to school travel time (numeric: 1 - 1 hour) 3. studytime - weekly study time (numeric: 1 - 10 hours) 4. failures - number of past class failures (numeric: n if 1<=n<3, else 4) 5. schoolsup - extra educational support (binary: yes or no) 6. famsup - family educational support (binary: yes or no) 7. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no) 8. activities - extra-curricular activities (binary: yes or no) 9. nursery - attended nursery school (binary: yes or no) 10. higher - wants to take higher education (binary: yes or no) 11. internet - Internet access at home (binary: yes or no) 12. romantic - with a romantic relationship (binary: yes or no) 13. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent) 14. freetime - free time after school (numeric: from 1 - very low to 5 - very high) 15. goout - going out with friends (numeric: from 1 - very low to 5 - very high) 16. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high) 17. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) 18. health - current health status (numeric: from 1 - very bad to 5 - very good) 19. absences - number of school absences (numeric: from 0 to 93)   These grades are related with the course subject, Math or Portuguese:   1. G1 - first period grade (numeric: from 0 to 20) 2. G2 - second period grade (numeric: from 0 to 20) 3. G3 - final grade (numeric: from 0 to 20, output target)  Data cleaning techniques to be applied to the dataset. We have checked the dataset for null values and we did not encounter them. Qualitative attributes were encoded numerical data. Then each member was given the original dataset and they made duplicate of that dataset and dropped columns where it is necessary for their algorithm. Description of the AI algorithm you've applied to your dataset. The label and the attributes. Lasso Regression - The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.  Linear Regression - Linear regression is the next step up after correlation. It is used when we want to predict the value of a variable based on the value of another variable. The variable we want to predict is called the dependent variable (or sometimes, the outcome variable).  Ridge Regression - Tikhonov Regularization, colloquially known as ridge regression, is the most commonly used regression algorithm to approximate an answer for an equation with no unique solution. This type of problem is very common in machine learning tasks, where the "best" solution must be chosen using limited data.  Decision Tree Classifier - The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data). The understanding level of Decision Trees algorithm is so easy compared with other classification algorithms.  Random Forest Classifier - In the case of random forest, it ensembles multiple decision trees into its final decision. Random forest can be used on both regression tasks (predict continuous outputs, such as price) or classification tasks (predict categorical or discrete outputs). |  |
| Random Forest Regression - It is an ensemble tree-based learning algorithm. The Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object. THE LABEL AND THE ATTRIBUTES Attributes (Features) are used for predicting grade (G3) which will be the label of student Mark prediction dataset. Majority of the models followed this label and attributes.  ● Attributes: School, sex, age, addres, famsize, Pstatus Medu, Fedu,Mjob, FJob, reason, guardian, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, Dalc, health, absences, G1, G2  ● Label: G3  For specifically for Decision Tree Classifier and Random Forest Classifier, GMean was taken from the average of G1, G2, and G3. For Decision Grade was made from the GMean.  ● Attributes: sex, addres, famsize, Pstatus Medu, Fedu,Mjob, FJob, guardian, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, Dalc, health, absences  ● Label: Grade  For Random Forest Classifier  ● Attributes: School, sex, age, addres, famsize, Pstatus Medu, Fedu,Mjob, FJob, reason, guardian, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, Dalc, health, absences  ● Label: GMean Result  |  |  | | --- | --- | | Learning Algorithm | Accuracy % | | Linear Regression | 84.0% | | Lasso Regression | 85.2% | | Ridge Regression | 82.4% | | Decision Tree Classifier | 70.0% | | Random Forest Regression | 87.0% | | Random Forest Classifier | 86.0% |  Conclusion Our most models performed very well as they gave more than 82 percent accuracy except decision trees. The performance of decision Tree classifier in normal, its accuracy can be increased if model is trained more with large dataset. The most accurate learning algorithm is Random Forest Regression, it gives 87.0% accuracy.  The Performance of our model on Random Forest Algorithms is very good as it has achieved more than 80% accuracy. In short, we can assume that with more data, the Random Forest and Lasso and Ridge Algorithm will perform better as their accuracy is more than 80%. The performance of these algorithm can be increased by increasing training dataset size. Links to the dataset and any other external help https://www.kaggle.com/dipam7/student-grade-prediction |  |
|  |  |