STUDENT ACADEMIC PERFORMANCE PREDICTION AND REASONING

**Table of Contents**

1. Project Details
   1. Objective
   2. Resources
2. Project Selection Motivation
   1. Machine Learning
   2. Interpretable AI
3. Basic Theory
   1. Exploratory Data Analysis
      1. Univariate Analysis
      2. Multi-variate analysis
      3. Missing values
   2. Machine Learning Models
      1. Naive Bayes
      2. Logistic Regression
      3. Decision Tree
      4. Random Forest
      5. Multi-layer Perceptron (sklearn, keras)
   3. Interpretable AI
      1. Lime (Local Interpretable Model-agnostic Explanations)
4. Project Workflow
   1. Architecture
   2. Exploratory Data Analysis
      1. Univariate Analysis
      2. Multi-Variate Analysis
      3. Missing Values
   3. Model Building and Testing
      1. Pre-processing
      2. Model Training and Testing
   4. Machine Learning Interpretability (MLI)
   5. Improvement (Manual Analysis)
   6. Final Inferences
5. Conclusion
6. ***Project Details***

1.1. Objective:

To predict the performance of a student based on historical data and identify the reason for said performance. We also look into the identification of parameters that can influence the performance.

1.2. Resources

Language Used: Python

Libraries: numpy, pandas, sklearn, keras, tensorflow, lime

1. ***Project Selection Motivation***

2.1. Machine Learning:

Machine Learning has taken over the world by reducing the limitations of rule-based systems. In this project, we look into various fields of Machine Learning which includes data manipulation, encoding, model building, model validation and model testing. Since this topic has taken over my interest, I chose this to be an essential part of my project.

2.2. Interpretable AI

Today, when a model is created, the general way of testing a model is using metrics like accuracy, precision, recall, m.s.e., etc. But these metrics don’t provide the behaviour of a model. That’s where Interpretable/Explainable AI comes in. This has the responsibility to understand why a model responds in a certain way and can also help validate model understanding. Let’s call this as MLU (Machine Learning Understanding).

1. ***Basic Theory***

3.1. Exploratory Data Analysis

3.1.1 Univariate Analysis

This deals with the variables at an individual level, where we observe the variations, distribution and effectiveness of the variable.

3.1.2 Multi-variate Analysis

This deals with the variables at a group level, where we observe the relationship, distribution and effect of one variable with others. It can be 1-1, 1-many type relationship

3.1.3 Missing Values

This is the segment where we deal with missing values. In our project, there were no missing values but, we have the code set up to take care of the issue if occurred.

3.2. Machine Learning Models

3.1.1 Naive Bayes

This algorithm is based on Bayes’ Theorem with an assumption of independence among predictors. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

3.2.2 Logistic Regression

This algorithm is a supervised classification algorithm where the target variable/label can take only discrete values for a given set of features/input.

3.2.3 Decision Tree

Decision tree builds classification in the form of a tree structure. It breaks down the dataset into subsets and incrementally develops the tree.

3.2.4 Random Forest

This algorithm performs classification, consisting of multiple decision trees. It uses bagging and feature randomness while creation of the tree and tries to make an uncorrelated tree collection whose net prediction is more accurate than the individual tree.

3.2.5 Multi-layer Perceptron

This model algorithm is a feedforward artificial neural network (ANN). It has layers and nodes of neurons which store non-linear patterns and provide enhanced results over big data.

3.3. Interpretable AI

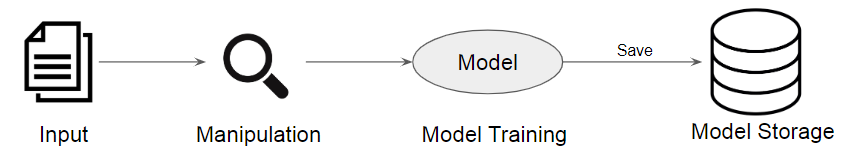
3.3.1 Lime (Local Interpretable Model-Agnostic Explanations)

This segment deals with the understanding and explanation of a model.

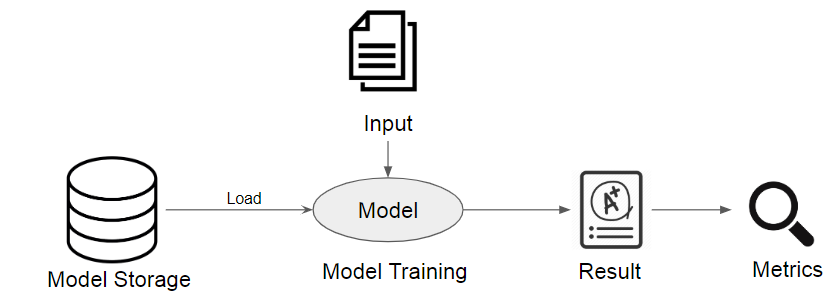
Lime is an explanation technique that explains the prediction of any classifier in an interpretable fashion by learning the model locally around the prediction.

1. ***Project Workflow***

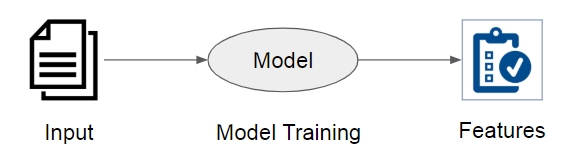
4.1. Architecture



Model Training Pipeline



Model Testing Pipeline



MLI Pipeline

**Architecture Brief:**

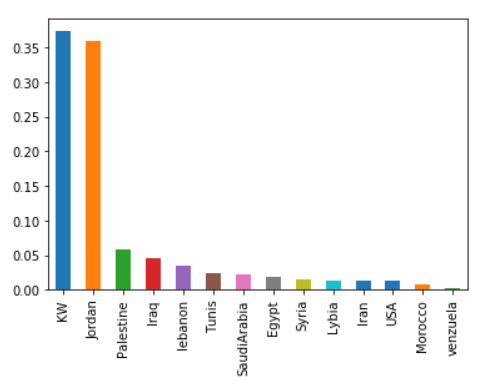
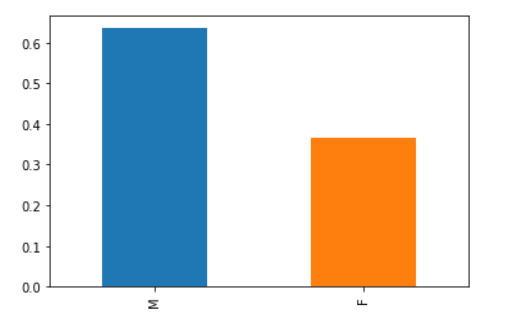
* Exploratory Data Analysis is performed on the dataset to identify the patterns and quality of the data.
* Data is then used to build a model after pre-processing (Categorical and Numerical).
* Manipulation techniques are used to check the model accuracy.
* MLI is performed on the testing data in order to identify the features which prove the understanding of the model.
* We try to identify the patterns from MLI on how the student performance can be increased.

4.2. Exploratory Data Analysis

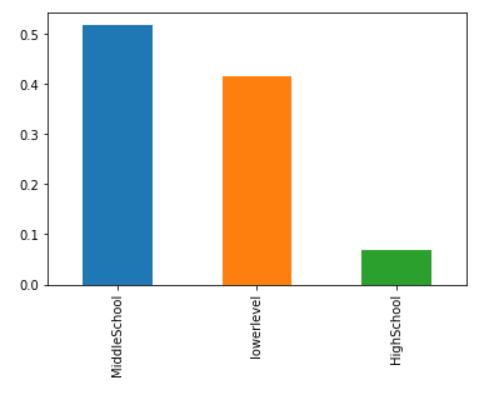
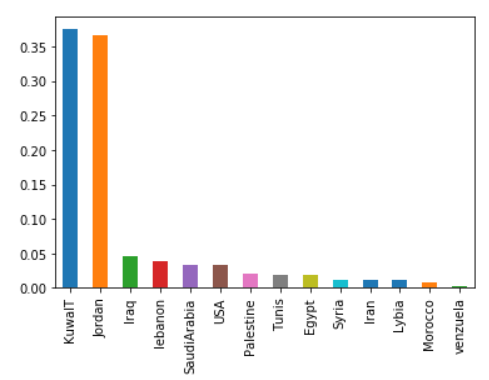
4.2.1 Univariate Analysis

Variables Used:

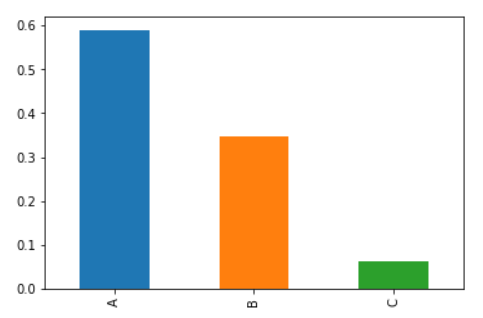
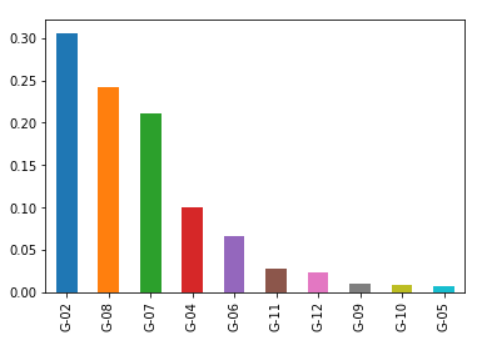
1. **Gender** - student's gender (nominal: 'Male' or 'Female’)



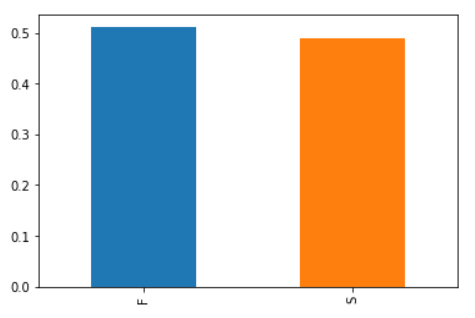
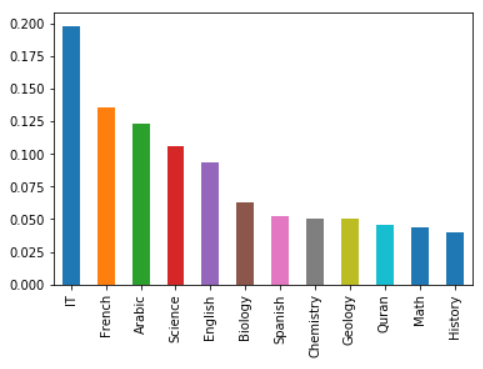
1. **Nationality**- student's nationality (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’ Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’)
2. **Place of birth** - student's Place of birth (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’ Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’)



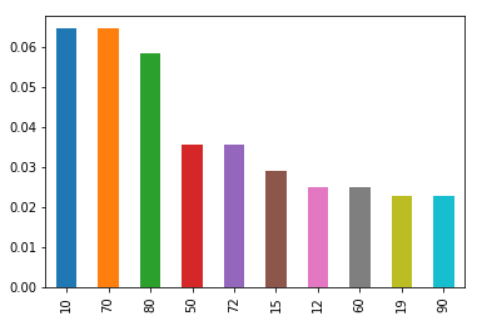
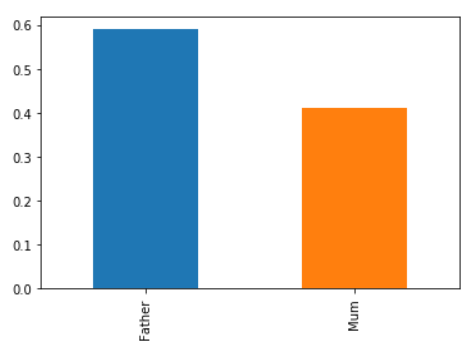
1. **Educational Stages** - educational level student belongs (nominal: ‘lowerlevel’,’MiddleSchool’,’HighSchool’)
2. **Grade Levels** - grade student belongs (nominal: ‘G-01’, ‘G-02’, ‘G-03’, ‘G-04’, ‘G-05’, ‘G-06’, ‘G-07’, ‘G-08’, ‘G-09’, ‘G-10’, ‘G-11’, ‘G-12 ‘)



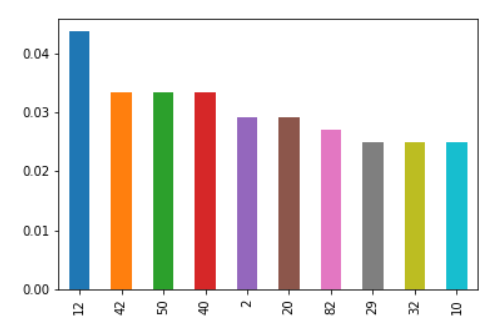
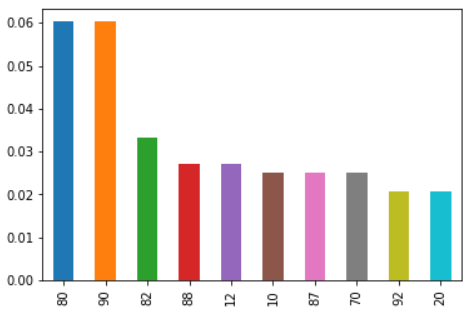
1. **Section ID** - classroom student belongs (nominal:’A’,’B’,’C’)
2. **Topic** - course topic (nominal:’ English’,’ Spanish’, ‘French’,’ Arabic’,’ IT’,’ Math’,’ Chemistry’, ‘Biology’, ‘Science’,’ History’,’ Quran’,’ Geology’)



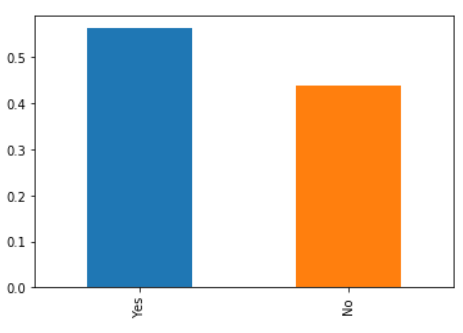
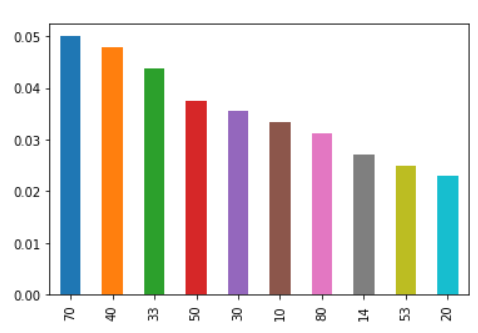
1. **Semester** - school year semester (nominal:’ First’,’ Second’)
2. **Parent relationship** with student (nominal:’mom’,’father’)



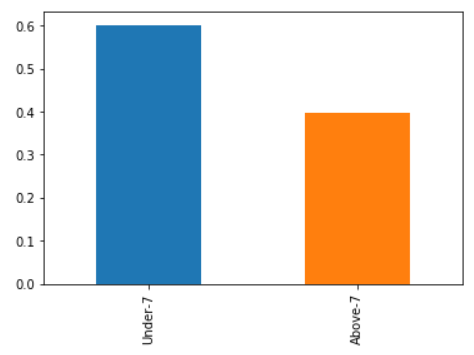
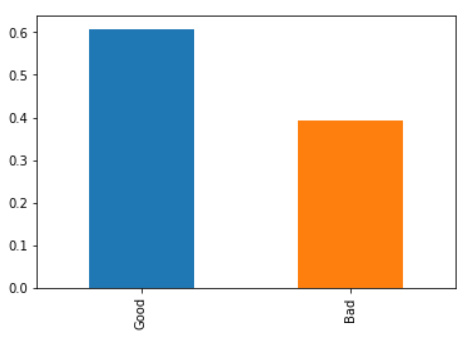
1. **Raised hand** - how many times the student raises his/her hand on classroom (numeric:0-100)
2. **Visited resources** - how many times the student visits a course content(numeric:0-100)



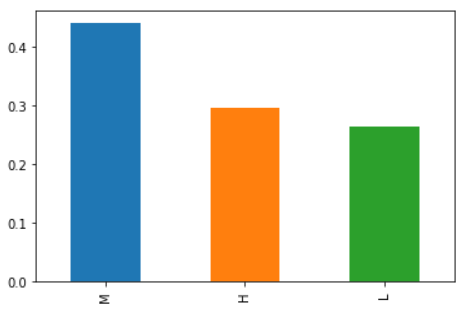
1. **Viewing announcements** - how many times the student checks the new announcements(numeric:0-100)
2. **Discussion groups** - how many times the student participate on discussion groups (numeric:0-100)



1. **Parent Answering Survey** - parent answered the surveys which are provided from school or not (nominal:’Yes’,’No’)
2. **Parent School Satisfaction** - the Degree of parent satisfaction from school(nominal:’Yes’,’No’)



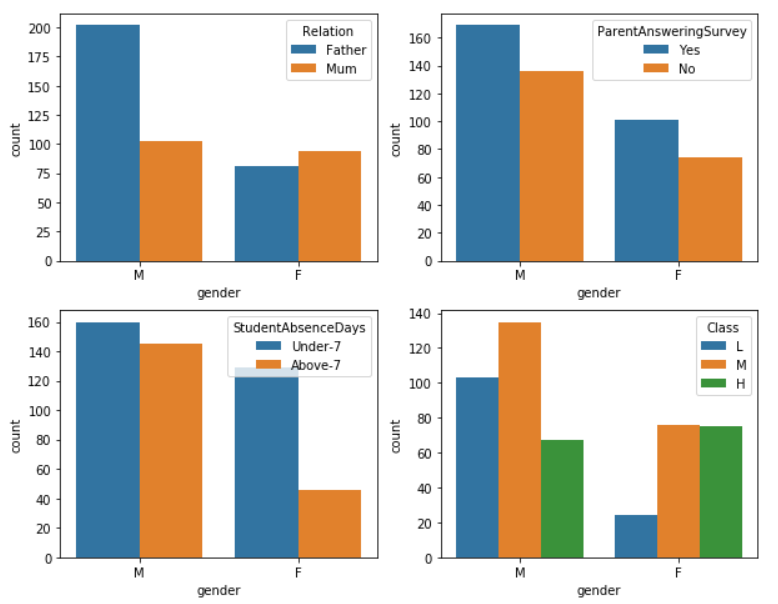
1. **Student Absence Days** - the number of absence days for each student (nominal: above-7, under-7)
2. **Class** - The class/dependent variable



4.2.2 Multi-variate Analysis

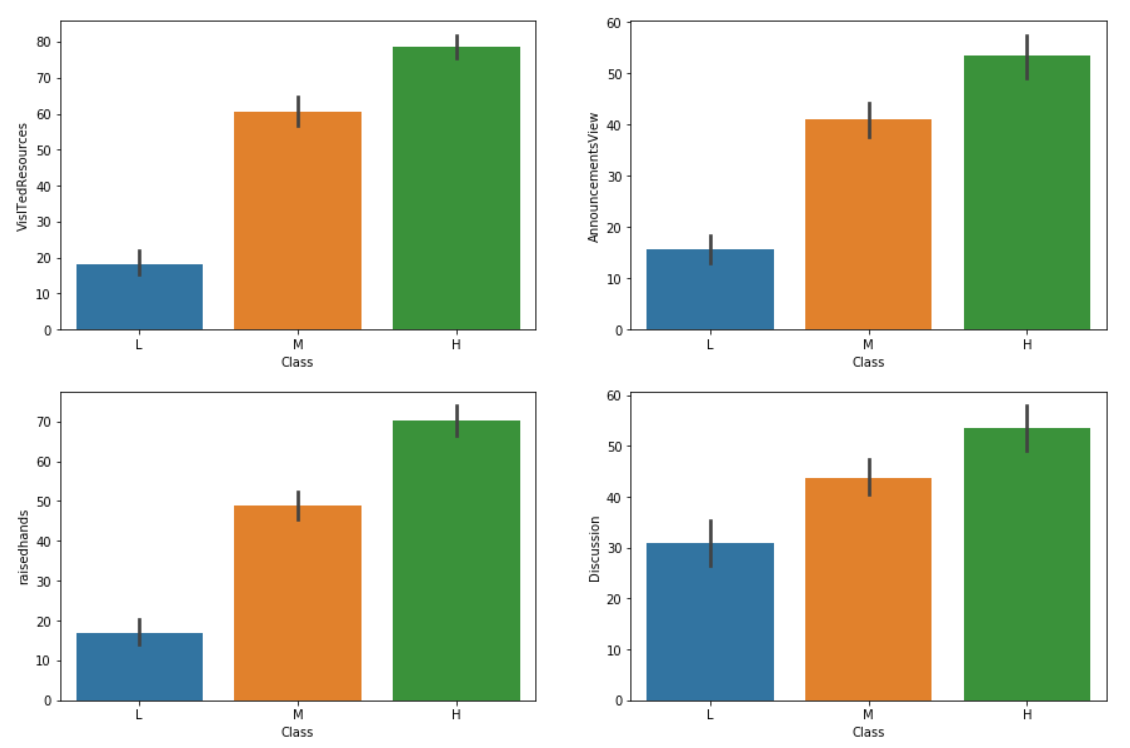
This deals with the variables at a group level, where we observe the relationship, distribution and effect of one variable with others. It can be 1-1, 1-many type relationship.

The relation between (Gender vs Relation, ParentAnsweringSurvey, StudentAbsenceDays, Class)



We observe:

1. Males have higher count for relation with father compared to females.
2. Males have higher count for Parents answering survey compared to females.
3. Males are absent in school for much higher duration compared to females. (Females are focused!)
4. Males in average score lower compared to girls.



We observe:

The students who participated in more of (Discussion, raisedhands, AnnouncementViews, RaisedHands) got a higher grade.

More observations directly on the Jupyter-notebook

4.2.3 Missing Values

This is the segment where we deal with missing values. In our project, there were no missing values but, we have the code set up to take care of the issue if occurred.

4.3. Model Building

4.3.1 Pre-processing

Categorical Data:

These were converted to encodings and also, dummy variables were tested, but later removed due to higher model accuracy in encodings.

Numerical Data:

These were normalized at a mean level.

4.3.2 Model Training and Testing

In this step, we train the model based on the following algorithms along with the Accuracy, Precision, Recall and F-score (On Testing Data):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No. | Model Type | Accuracy | Precision | Recall | F1-Score |
| 1 | Naive Bayes | 0.58 | 0.60 | 0.58 | 0.59 |
| 2 | Logistic Regression | 0.78 | 0.80 | 0.78 | 0.78 |
| 3 | Decision Tree | 0.72 | 0.76 | 0.72 | 0.73 |
| 4 | Random Forest | 0.79 | 0.84 | 0.79 | 0.79 |
| 5 | SVM | 0.79 | 0.81 | 0.79 | 0.79 |
| 6 | Multi-layer Perceptron | 0.42 | 0.25 | 0.42 | 0.31 |

4.4. Machine Learning Interpretability (LIME)

In this step, we identify the features where the model and training data is used to identify the reason a classification was provided by the model for a given input from test data.

Example:

For a student who scored ‘H’ i.e. High, these are the features that model understood:

('0.00 < StudentAbsenceDays <= 1.00', -0.16438850448986833),

('raisedhands > 72.00', -0.12369068516327807),

('AnnouncementsView > 52.25', -0.1156616501437044),

('NationalITy <= 3.00', -0.09413064481102569),

('ParentAnsweringSurvey <= 0.00', 0.07574953587129417)

From this, we can say that the student who :

* Raises his hands more than 72 times
* Views announcement regularly more than 52.25 times

Has a higher chance of getting ‘H’.

For a student who scored ‘M’ i.e. High, these are the features that model understood:

('StudentAbsenceDays <= 0.00', 0.1683702600913632),

('AnnouncementsView <= 14.00', 0.10384104704630431),

('NationalITy <= 3.00', -0.10004602741150374),

('ParentAnsweringSurvey <= 0.00', 0.07531229765791045),

('PlaceofBirth <= 3.00', 0.07183647522952041)

From this, we can say that the student who :

* Place of Birth is 0: Egypt, 1: Iran, 2: Iraq, 3: Jordan
* Views announcement regularly but less than 14 times

Has a higher chance of getting ‘M’.

For a student who scored ‘L’ i.e. High, these are the features that model understood:

('StudentAbsenceDays <= 0.00', 0.1541718062372064),

('raisedhands <= 15.00', 0.14985311199938714),

('VisITedResources <= 18.75', 0.11207606481585251),

('AnnouncementsView <= 14.00', 0.10716960113076818),

('ParentAnsweringSurvey <= 0.00', 0.06801177362323842)

From this, we can say that the student who :

* Raises his hands less than 15 times
* Views announcement regularly but less than 14 times
* Visited a course regularly but less than 18.75 times.

Has a higher chance of getting ‘L’.

4.5. Improvement (Manual Analysis)

1. First, we incorrectly classified data
2. We run the MLI over the remaining data and store it as a file.
3. Next, we separate the ‘H’, ‘M’, ‘L’ into separate groups to find the features that affected them. Similar to the mentioned analysis in the previous section.
4. We then take these features and compare them within the segment and across the segment.
5. Finally, the results are shown as to how can a student improve his grade.

**For ‘H’ the major features found are:**

1. ('0.00 < StudentAbsenceDays <= 1.00', 24),
2. ('raisedhands > 72.00', 17),
3. ('AnnouncementsView > 52.25', 16),
4. ('0.00 < ParentAnsweringSurvey <= 1.00', 16),
5. ('NationalITy <= 3.00', 15),
6. ('VisITedResources > 84.00', 7),
7. ('PlaceofBirth > 4.00', 6),
8. ('ParentAnsweringSurvey <= 0.00', 4),
9. ('NationalITy > 4.00', 4),
10. ('42.00 < raisedhands <= 72.00', 3),
11. ('62.50 < VisITedResources <= 84.00', 2),
12. ('PlaceofBirth <= 3.00', 2),
13. ('AnnouncementsView <= 14.00', 1),
14. ('0.00 < gender <= 1.00', 1),
15. ('StageID <= 1.00', 1),
16. ('0.00 < Relation <= 1.00', 1)

**For ‘M’ the major features found are:**

1. ('0.00 < ParentAnsweringSurvey <= 1.00', 15),
2. ('0.00 < StudentAbsenceDays <= 1.00', 13),
3. ('NationalITy <= 3.00', 10),
4. ('StudentAbsenceDays <= 0.00', 9),
5. ('ParentAnsweringSurvey <= 0.00', 7),
6. ('PlaceofBirth <= 3.00', 7),
7. ('AnnouncementsView > 52.25', 7),
8. ('StageID <= 1.00', 7),
9. ('42.00 < raisedhands <= 72.00', 6),
10. ('AnnouncementsView <= 14.00', 5),
11. ('0.00 < gender <= 1.00', 5),
12. ('raisedhands > 72.00', 3),
13. ('NationalITy > 4.00', 3),
14. ('gender <= 0.00', 3),
15. ('VisITedResources <= 18.75', 2),
16. ('15.00 < raisedhands <= 42.00', 2),
17. ('62.50 < VisITedResources <= 84.00', 2),
18. ('VisITedResources > 84.00', 1),
19. ('Discussion <= 20.00', 1),
20. ('PlaceofBirth > 4.00', 1),
21. ('1.00 < StageID <= 2.00', 1)

**For ‘L’ the major features found are:**

1. ('StudentAbsenceDays <= 0.00', 11),
2. ('VisITedResources <= 18.75', 10),
3. ('AnnouncementsView <= 14.00', 8),
4. ('raisedhands <= 15.00', 7),
5. ('NationalITy <= 3.00', 6),
6. ('ParentAnsweringSurvey <= 0.00', 5),
7. ('NationalITy > 4.00', 3),
8. ('PlaceofBirth > 4.00', 1),
9. ('PlaceofBirth <= 3.00', 1),
10. ('0.00 < ParentAnsweringSurvey <= 1.00', 1),
11. ('42.00 < raisedhands <= 72.00', 1),
12. ('0.00 < gender <= 1.00', 1)

**The relation between the ‘H’ and ‘L’ score are the features below:**

1. {'0.00 < Relation <= 1.00',
2. '0.00 < StudentAbsenceDays <= 1.00',
3. '62.50 < VisITedResources <= 84.00',
4. 'AnnouncementsView > 52.25',
5. 'StageID <= 1.00',
6. 'VisITedResources > 84.00',
7. 'raisedhands > 72.00'}

**The relation between the ‘M’ and ‘L’ score are the features below:**

1. '0.00 < StudentAbsenceDays <= 1.00',
2. '1.00 < StageID <= 2.00',
3. '15.00 < raisedhands <= 42.00',
4. '62.50 < VisITedResources <= 84.00',
5. 'AnnouncementsView > 52.25',
6. 'Discussion <= 20.00',
7. 'StageID <= 1.00',
8. 'VisITedResources > 84.00',
9. 'gender <= 0.00',

4.6. Final Inferences:

1. ‘L’ score -> ‘H’ score
   1. Student needs to reduce Absent Days
   2. Student needs to increase his Resource/Course visit to greater than 84 times.
   3. Student needs to view announcements more than 52.25 times.
   4. Student needs to raise hands by more than 72 times.
2. ‘L’ score -> ‘M’ score
   1. Student needs to reduce Absent Days
   2. Student needs to bring his Resource/Course visit to between 62.50 and 84 times.
   3. Student needs to view announcements more than 52.25 times.
   4. Student needs to raise hands between 15 to 42 times.
3. ***Conclusion***

Student Performance is a vital factor for a students future and with our project, we tried understanding the factors that lead to lower or higher performance. Using the final inferences, the students can improve their score (other factors remain constant). We have been able to perform exploratory analysis, train a machine learning model, validate it using the validation and test data, found the features affecting the output using MLI and used them to derive the inferences mentioned in the previous section.

Future endeavours include using a more complex dataset and automating the Improvement Pipeline to avoid human intervention.