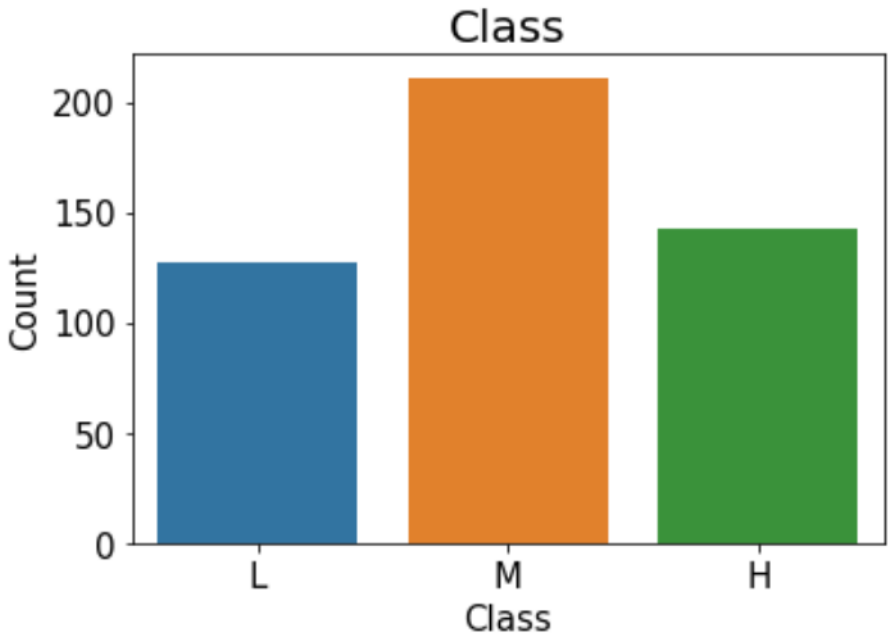
Analyse of Student Performance data

Trang Nghiem

Description of data

```
RangeIndex: 480 entries, 0 to 479
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   gender                    480 non-null    object
 1   NationalITy               480 non-null    object
 2   PlaceofBirth              480 non-null    object
 3   StageID                   480 non-null    object
 4   GradeID                   480 non-null    object
 5   SectionID                 480 non-null    object
 6   Topic                     480 non-null    object
 7   Semester                  480 non-null    object
 8   Relation                  480 non-null    object
 9   raisedhands               480 non-null    int64
 10  VisITedResources          480 non-null    int64
 11  AnnouncementsView         480 non-null    int64
 12  Discussion                480 non-null    int64
 13  ParentAnsweringSurvey     480 non-null    object
 14  ParentschoolSatisfaction  480 non-null    object
 15  StudentAbsenceDays        480 non-null    object
 16  Class                     480 non-null    object
dtypes: int64(4), object(13)
memory usage: 63.9+ KB
```



Student performance is quantified by class :

- L (Low-level) : note 0-69
- M (Middle-level) : note 70-89
- H (High-level) : note 90-100

Steps of analyse & classification

Data Visualization

Correlation
- between features
- between features & class

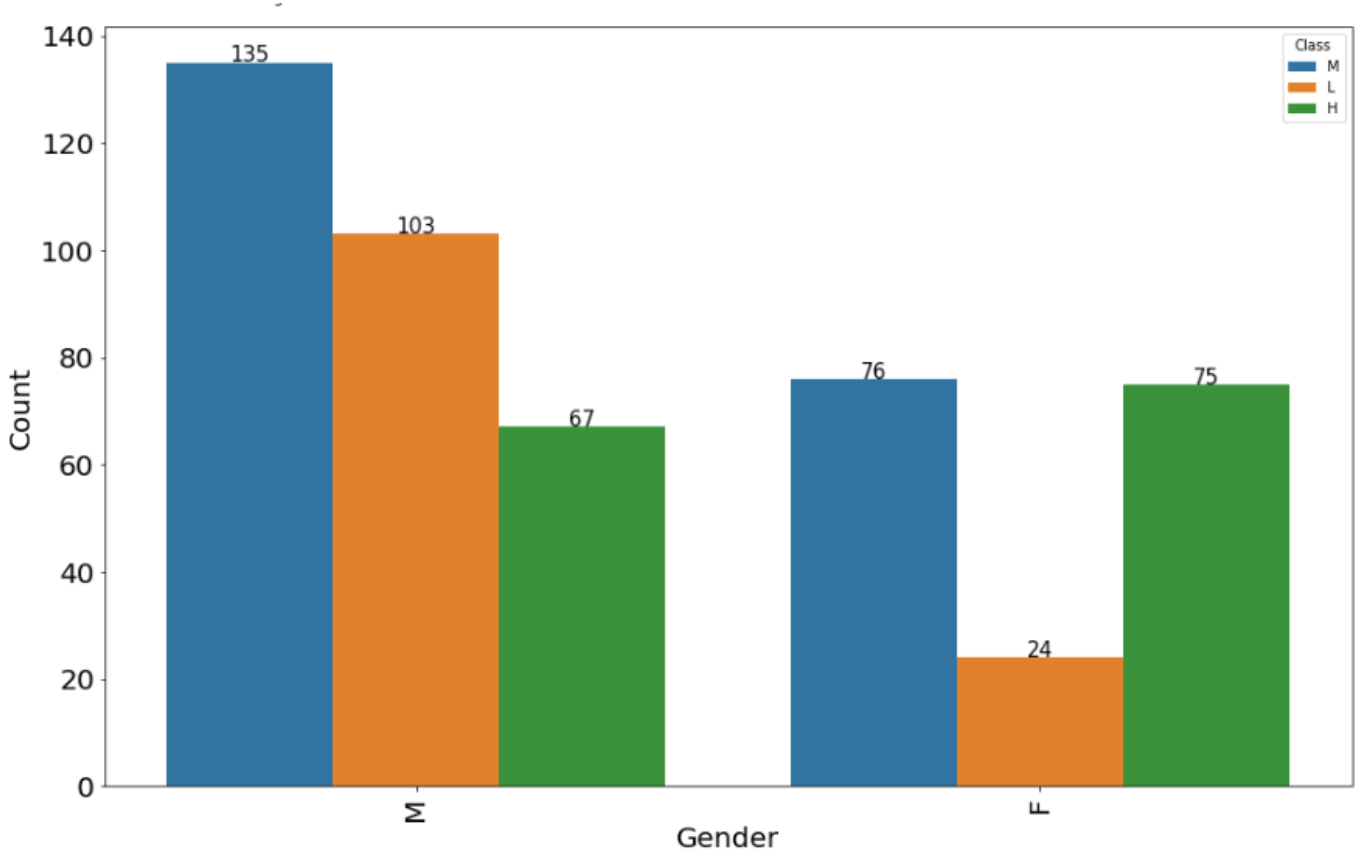
Data - Classification

Data preprocessing

Classification
- Choose of classifiers
- Tune hyperparameters (RandomzedSearch, GridSearch)
- Best model (anlyse score, confusion matrix, feature importance)

Data Visualization

Gender



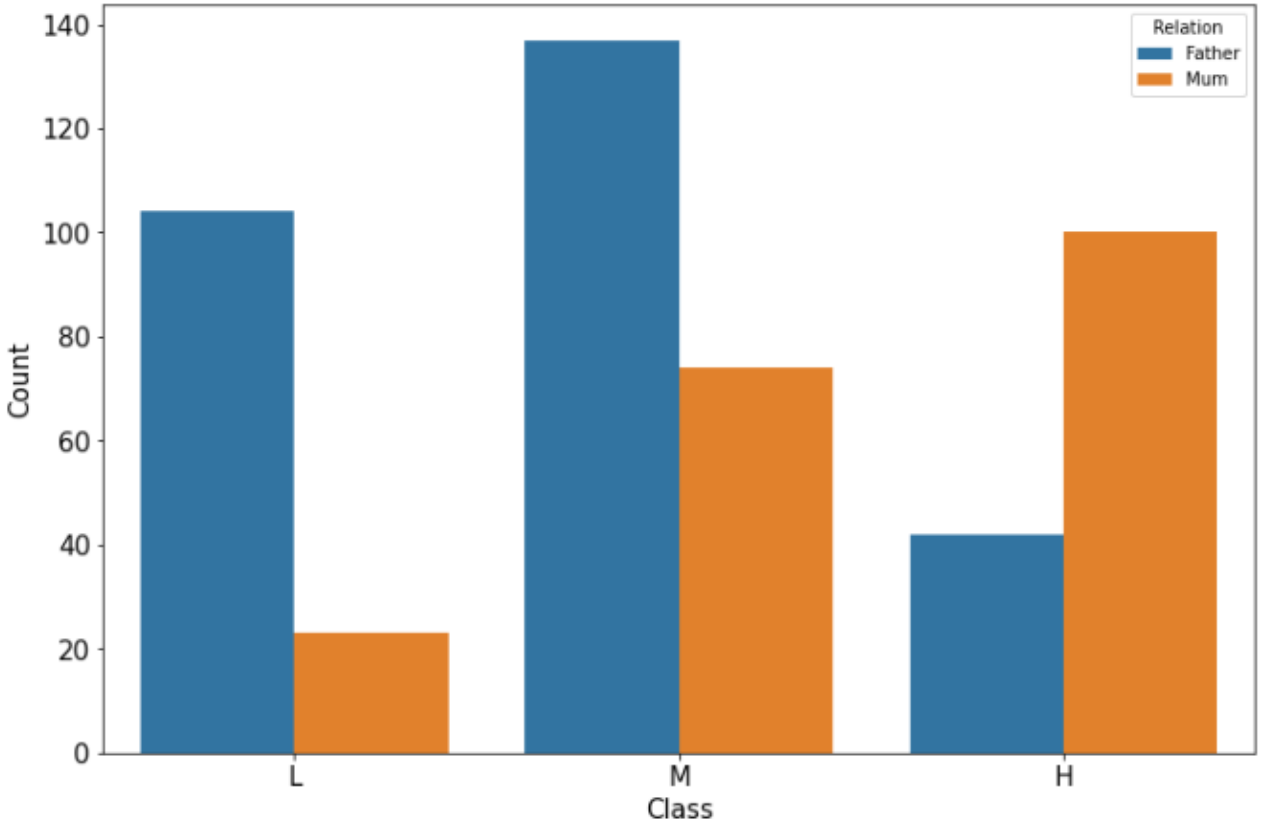
Tendance visuelle

	Positive	Negative
Being male		–
Being female	+	

- Female students : high & middle level
- Male students : low-level note

Data Visualization

Relationship with parents



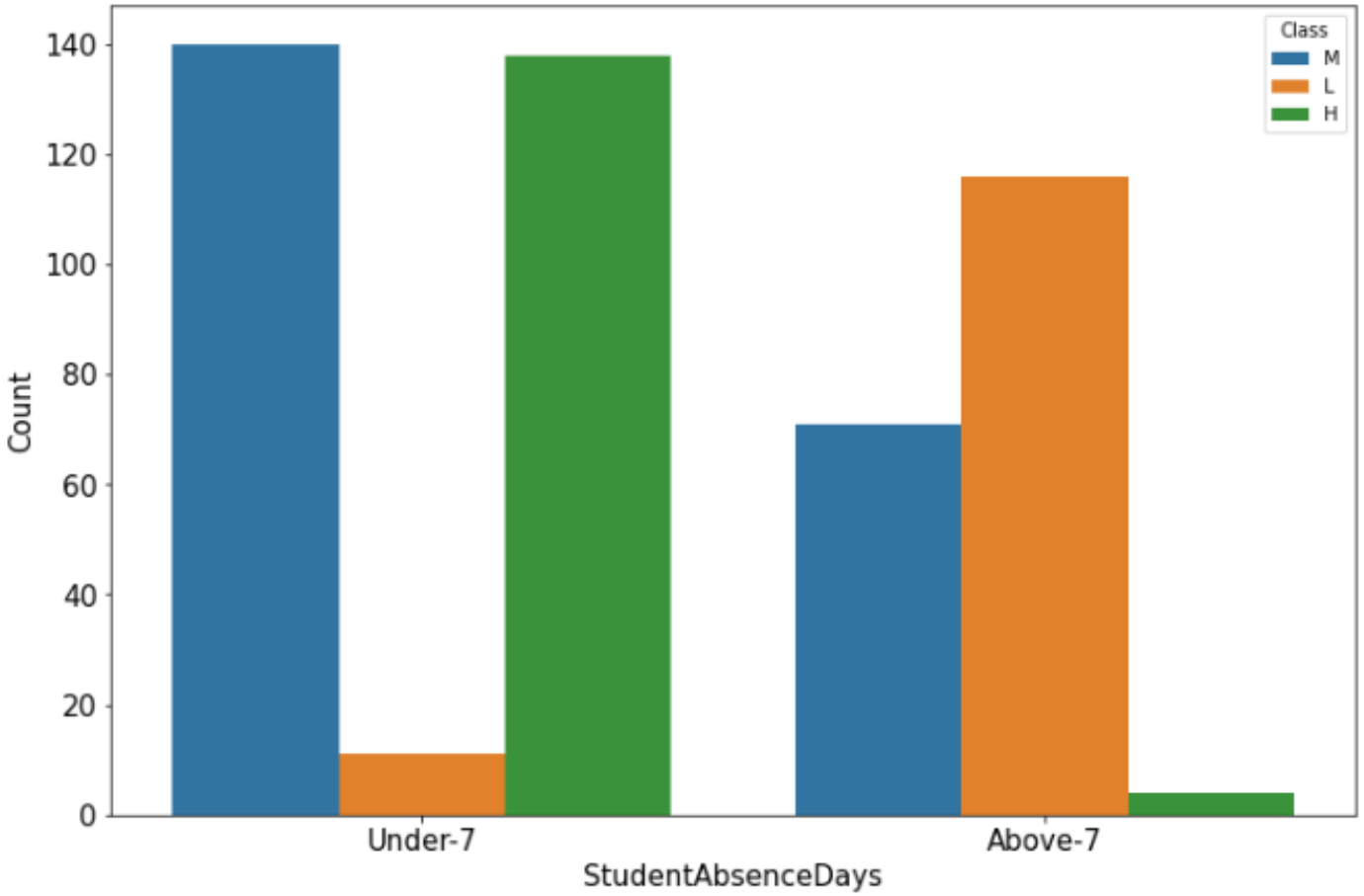
Tendance visuelle

	Positive	Negative
Being male		–
Being female	+	
Relation with mum	+	

Good relationship with mum : better in learning

Data Visualization

Absent Days

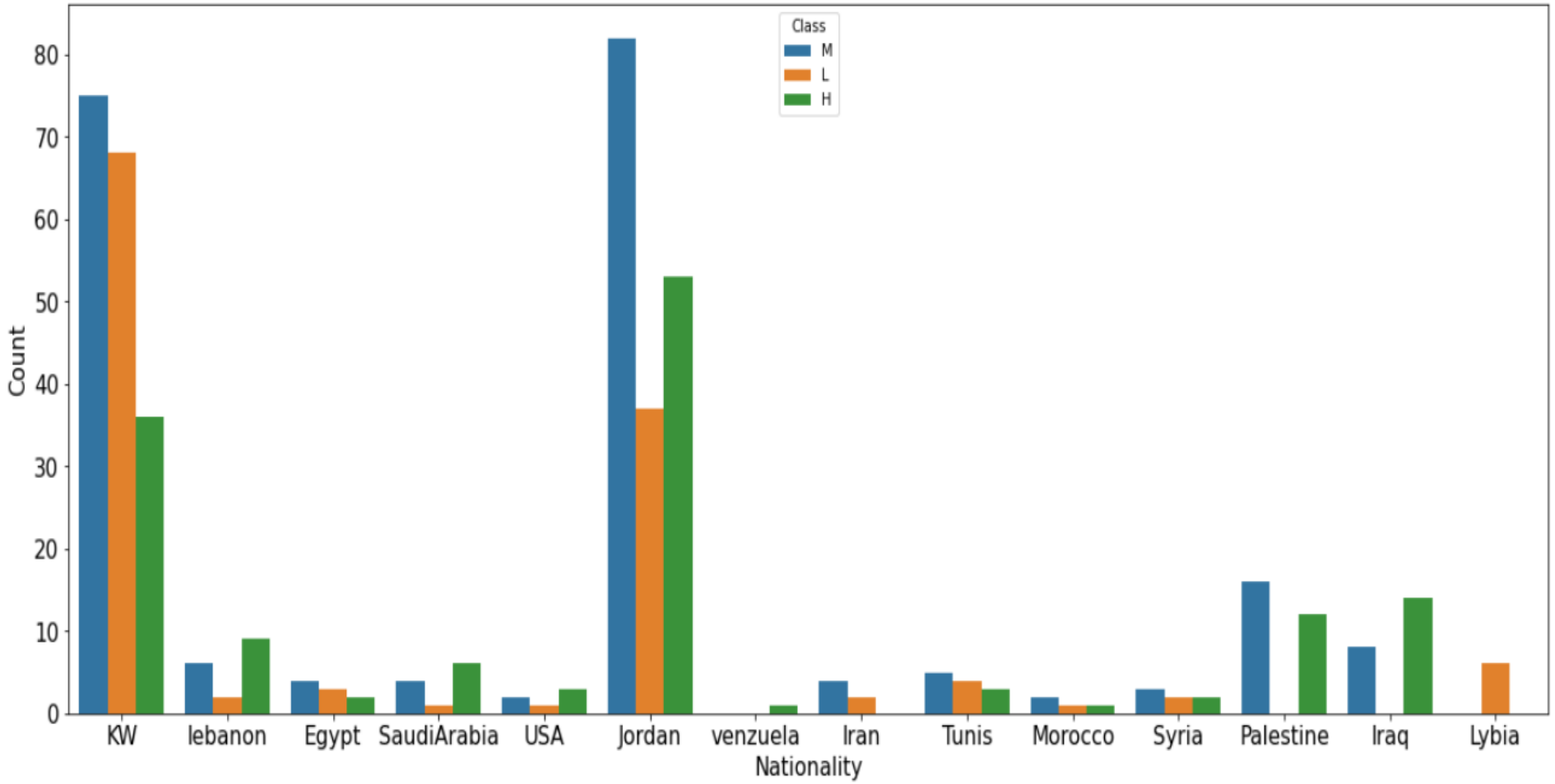


Tendance visuelle

	Positive	Negative
Being male		–
Being female	+	
Relation with mum	+	
Absence Days (under 7)	+	

Data Visualization

Performance vs. Nationality

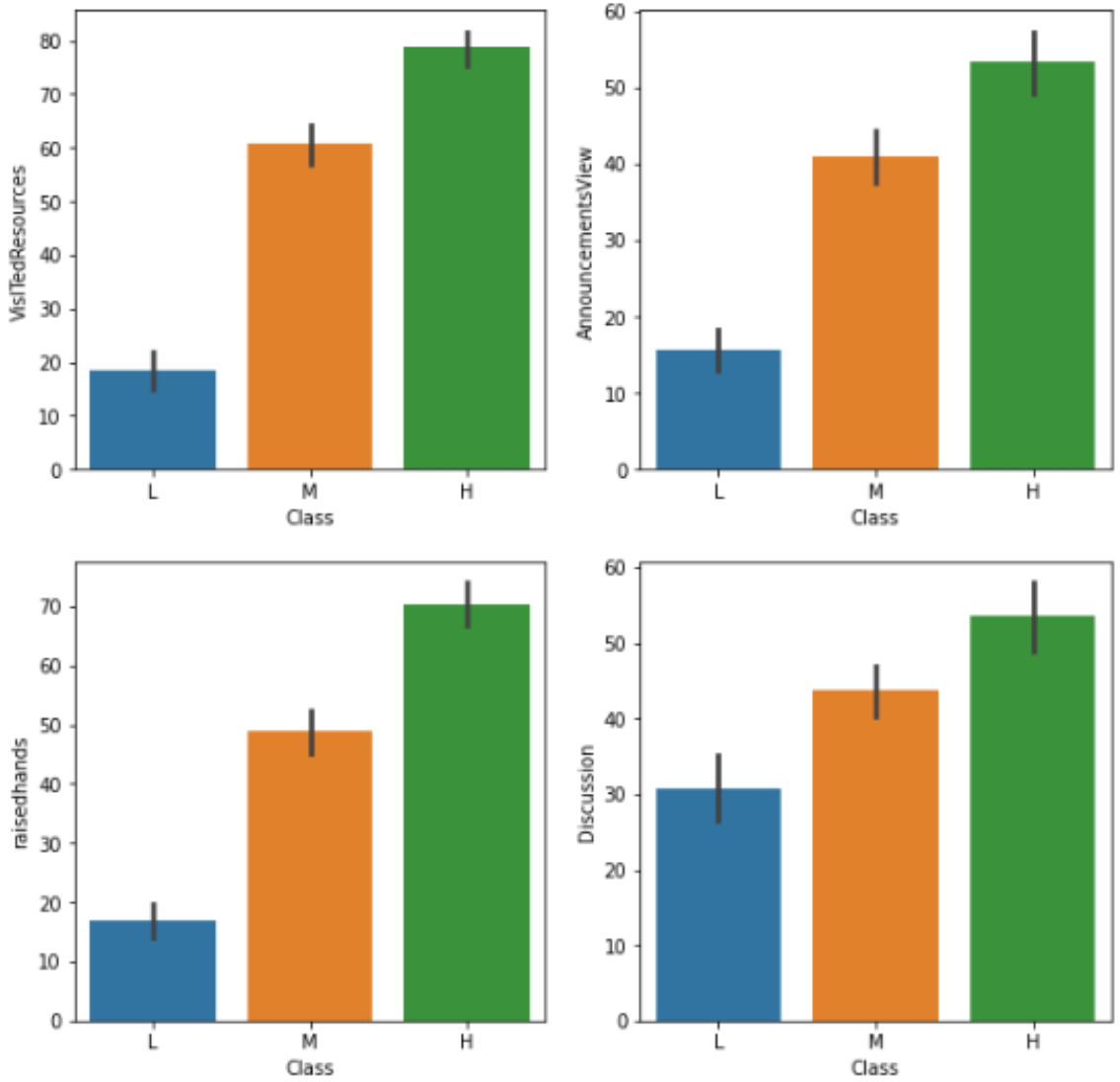


Majority of nationality : Kuwait, Jordan

No explicit relationship

Data Visualization

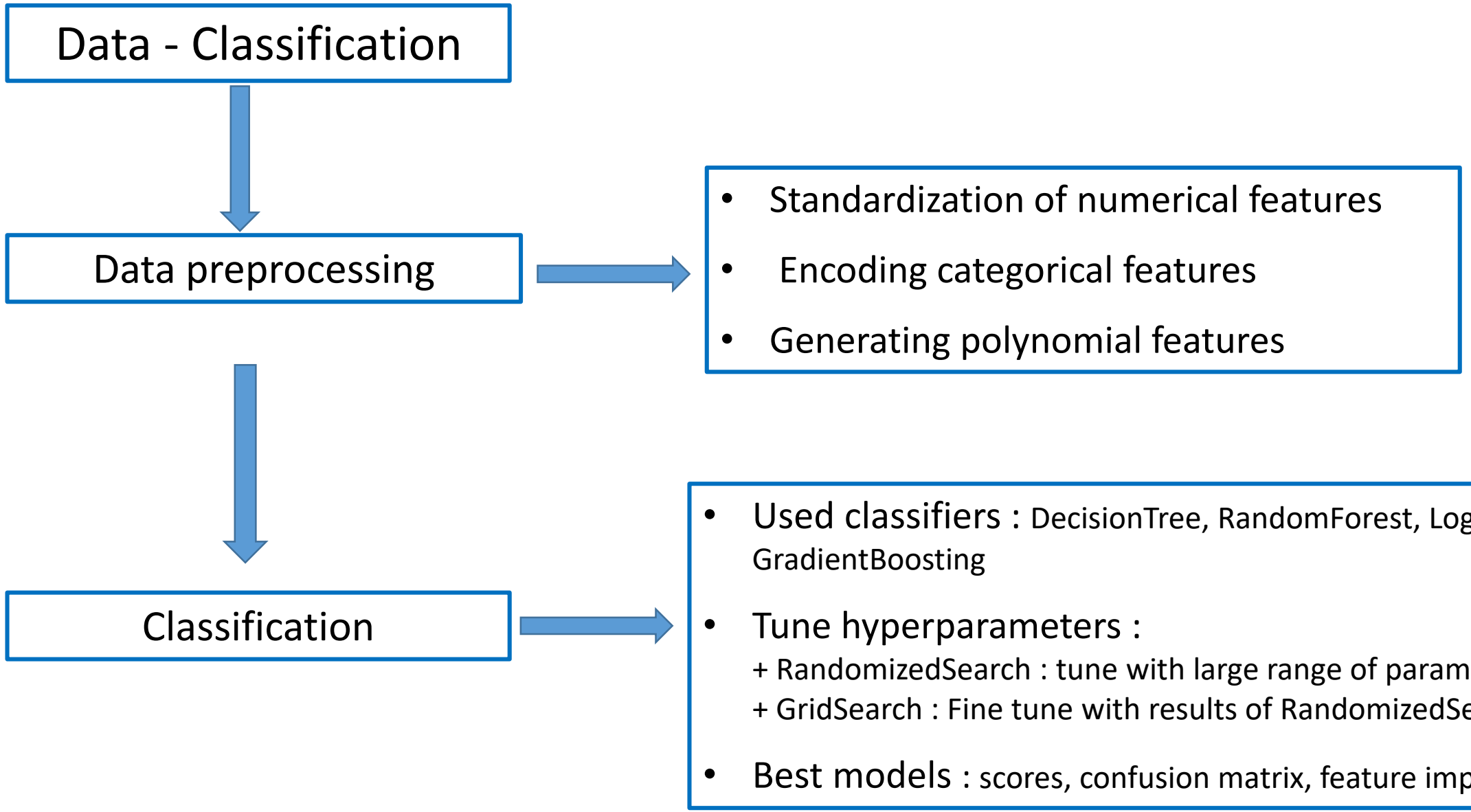
Discussion, Raise Hand, Visited Resource & AnnouncementView



High-level students participe on school's course & activities much more than low-level students

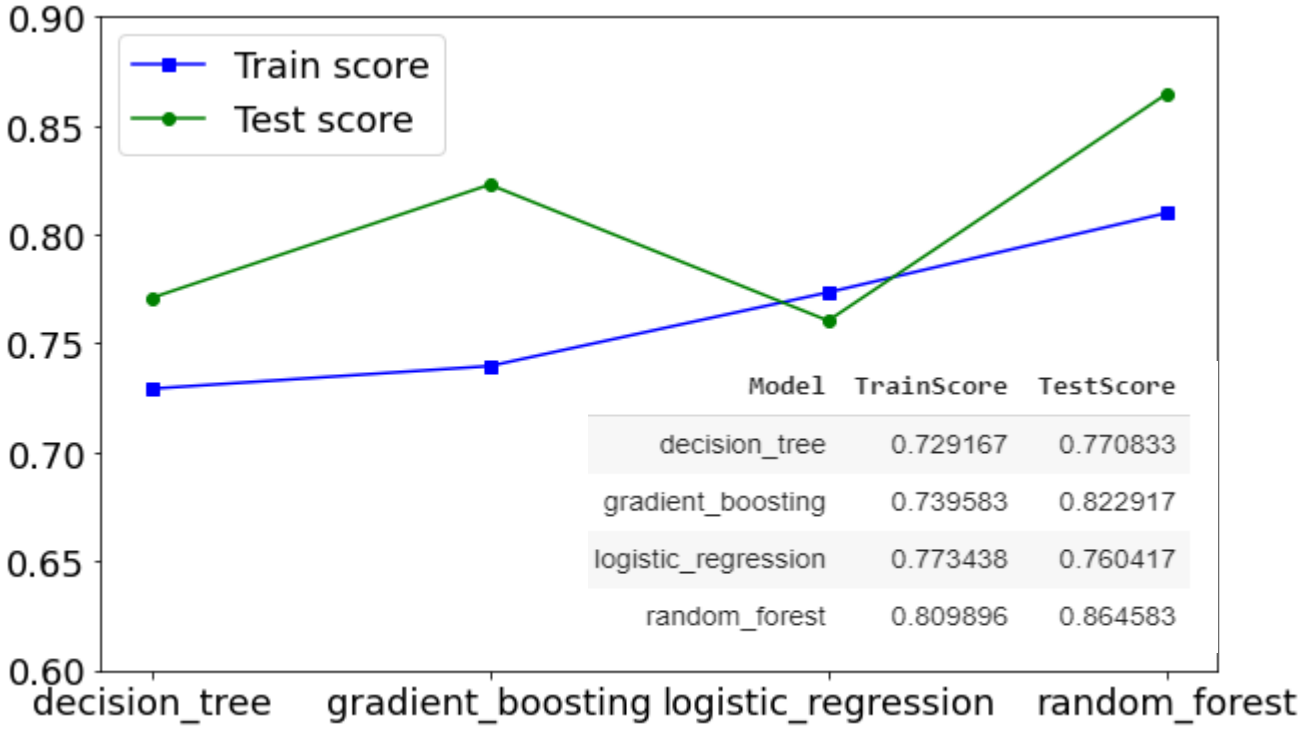
Visual tendency

	Positive	Negative
Being male		–
Being female	+	
Relation with mum	+	
Absence Days (under 7)	+	
Disussion, RaisedHand, Visitesd Ressource, Announcement	+ + + +	


Data - Classification
Data preprocessing
Standardization of numerical features
Encoding categorical features
Generating polynomial features
Classification
Used classifiers : DecisionTree, RandomForest, Logistic Regression, GradientBoosting
Tune hyperparameters :
+ RandomizedSearch : tune with large range of parameters
+ GridSearch : Fine tune with results of RandomizedSearch
Best models : scores, confusion matrix, feature importance

Classification : Results

Score



Model	TrainScore	TestScore
decision_tree	0.729167	0.770833
gradient_boosting	0.739583	0.822917
logistic_regression	0.773438	0.760417
random_forest	0.809896	0.864583

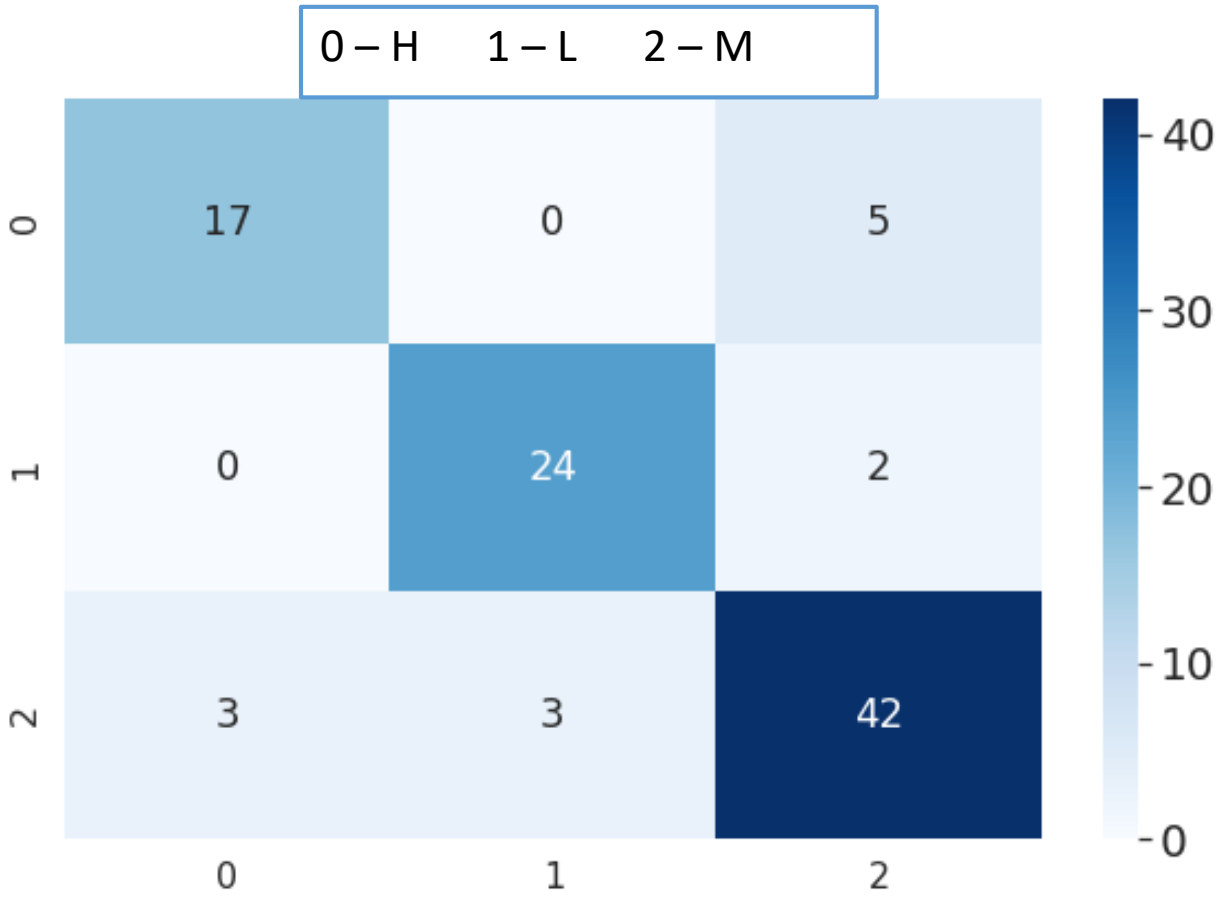
Obtain good score of precision
(logistic regression, random forest)

BUT : underfitting (test_score > train_score)

→ Require : more data

Classification : Results

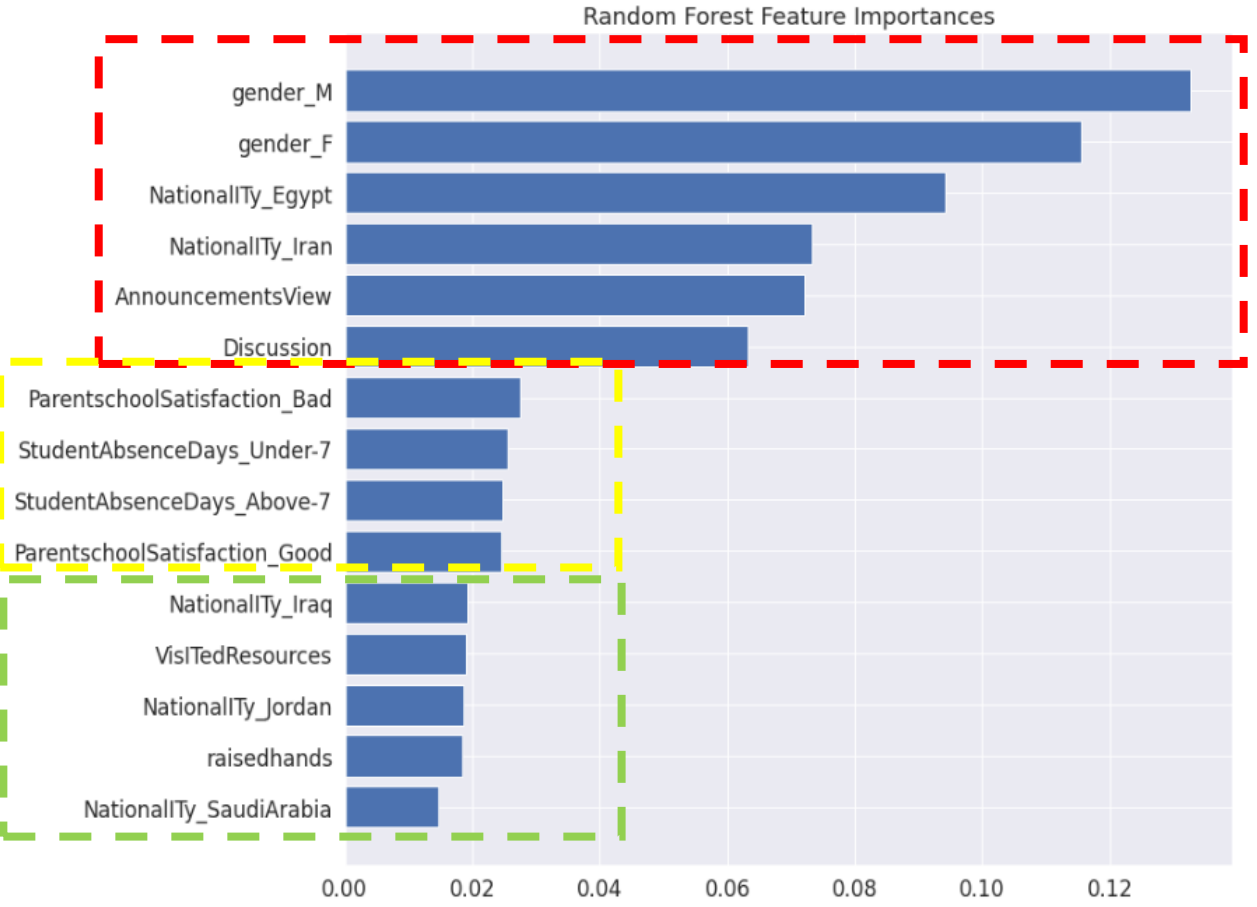
Confusion matrix
Random Forest on test set



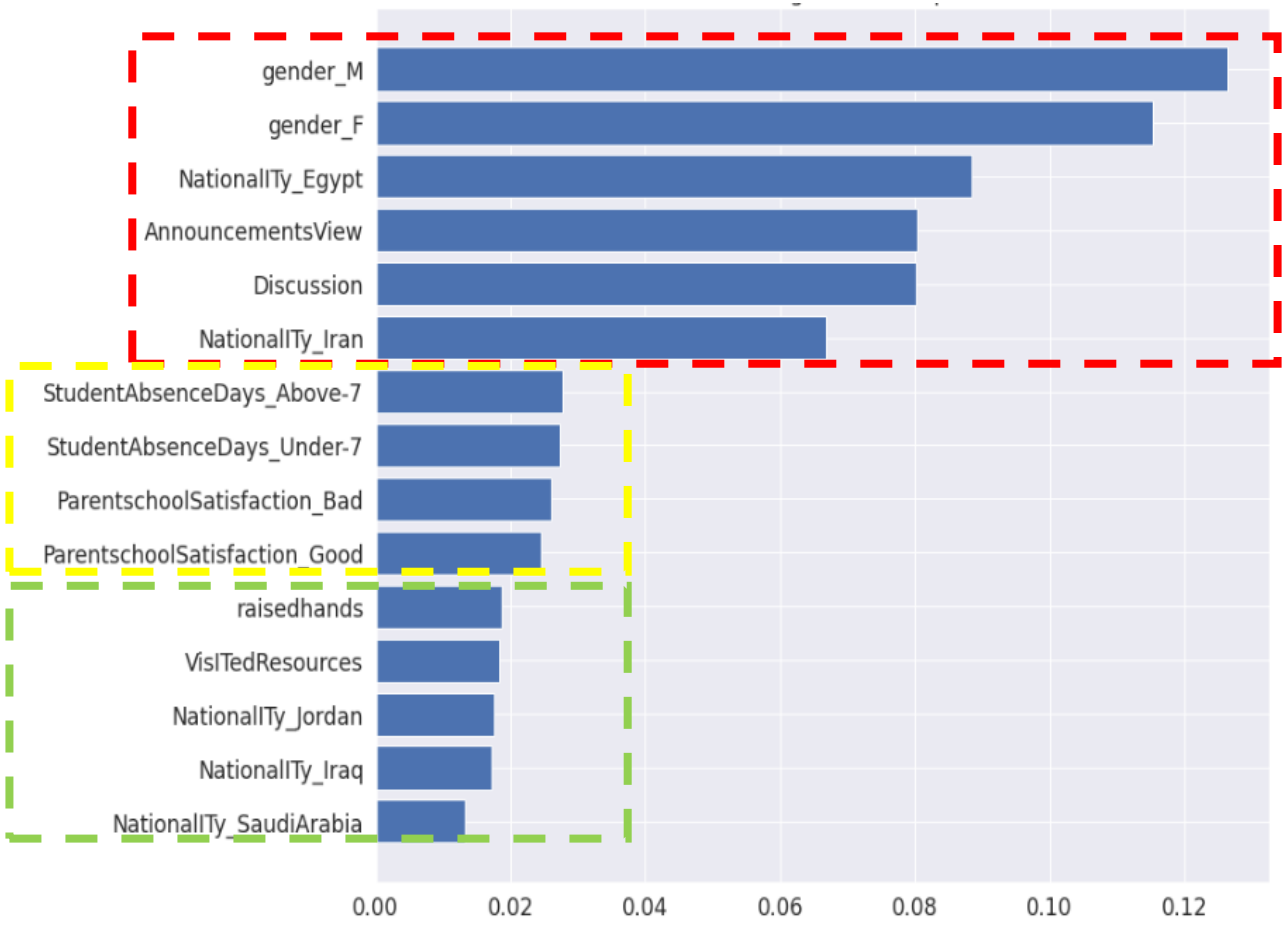
Claifify well all three classes

Classification : Results

Random Forest : Feature importance



Gradient Boosting: Feature importance



Top 6 feature importances are identical in 2 models (Random Forest & Gradient Boosting)

Remarque

- Data preprocessing plays important role in machine learning (for example : OneHotEncoder vs. LabelEncoder)
- Need more data to better accuracy
- Others models to capture better all features (ex : Votting)