EDM, INC.

EDM PROJECT

WWW.EDM.COM

CONTENT

ABOUT US

- 資工三 110590003 黃政 coding
- 資工三 110590005 蕭耕宏 coding, lead
- 資工三 110590028 黃冠鈞 coding
- 資工三 110590034 楊榮鈞 coding

INTRO OF FIRST DATESET

IBM HR Analytics Employee Attrition & Performance

Goal : predict education level

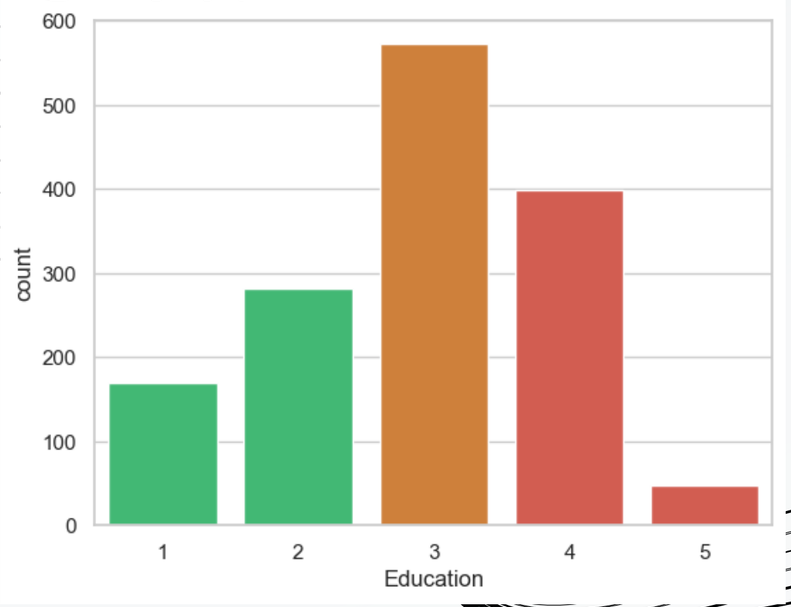
- Below College
- College
- Bachelor
- Master
- Doctor

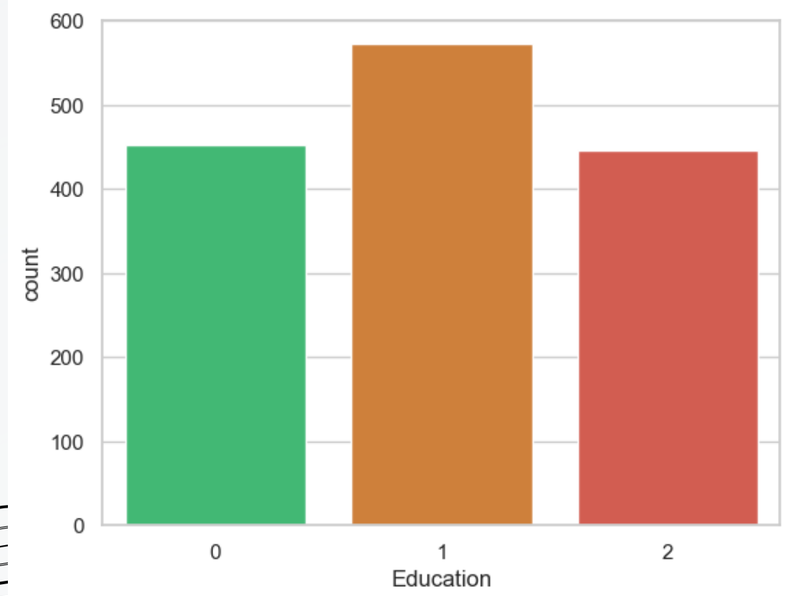
Age, Attrition, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, MonthlyRate, NumCompaniesWorked, Over18, OverTime, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager

INTRO OF FIRST DATESET

IBM HR Analytics Employee Attrition & Performance

simply the prediction task





- Below College 1
- College 2
- Bachelor 3
- Master 4
- Doctor 5

INTRO OF FIRST DATESET

IBM HR Analytics Employee Attrition & Performance

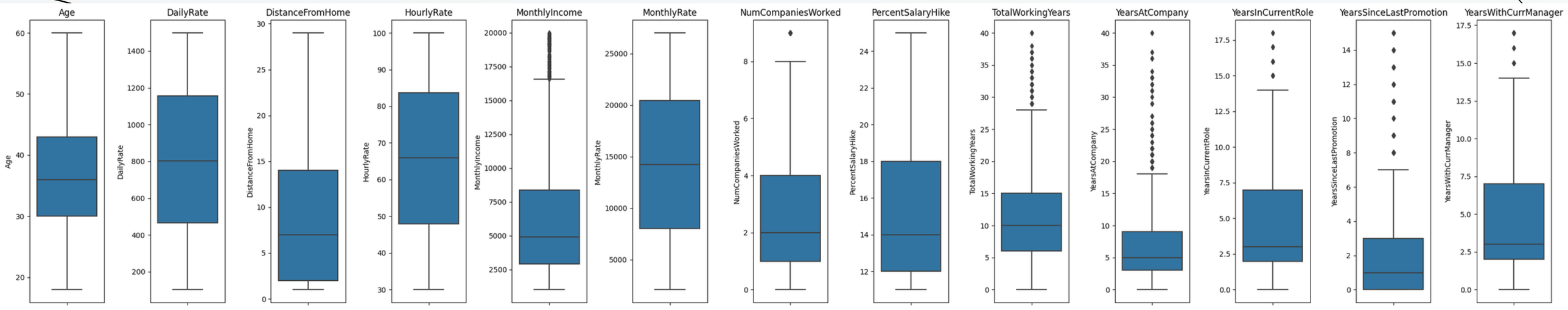
Issues to solve

- some variables contains single value
- some ordinal variables is encoded to integer and some don’t
- some numerical variables have extreme outlier

INTRO OF FIRST DATESET

IBM HR Analytics Employee Attrition & Performance

numerical variables outliers



INTRO OF FIRST DATESET

IBM HR Analytics Employee Attrition & Performance

Solution

- categorical variables : use one-hot encoder
- ordinal variables : use ordinal encoder
- numerical variables : remove outlier and standardize features removing the mean and scaling to unit variance
- use several model to predict
- implement k-fold validation

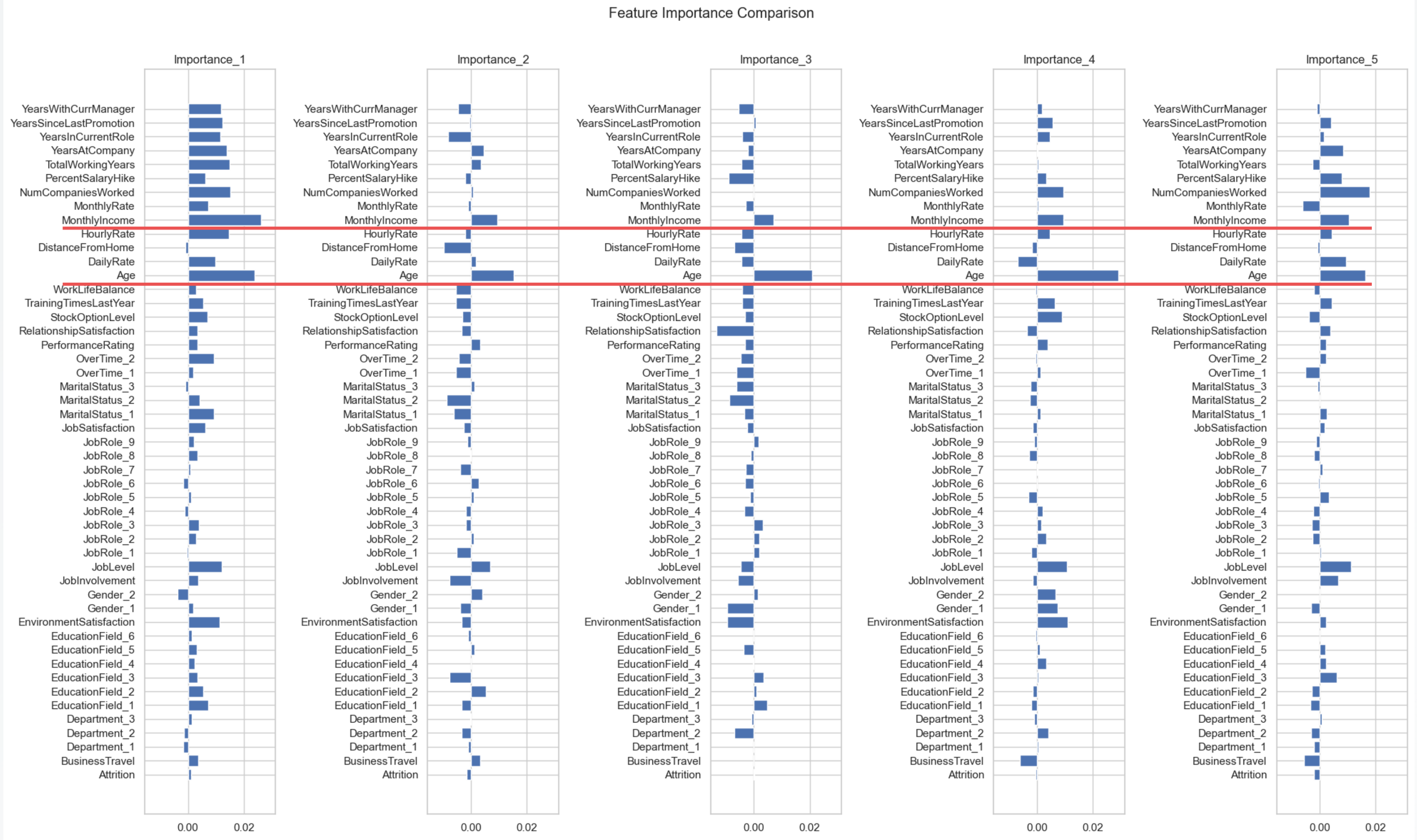
ANALYSIS

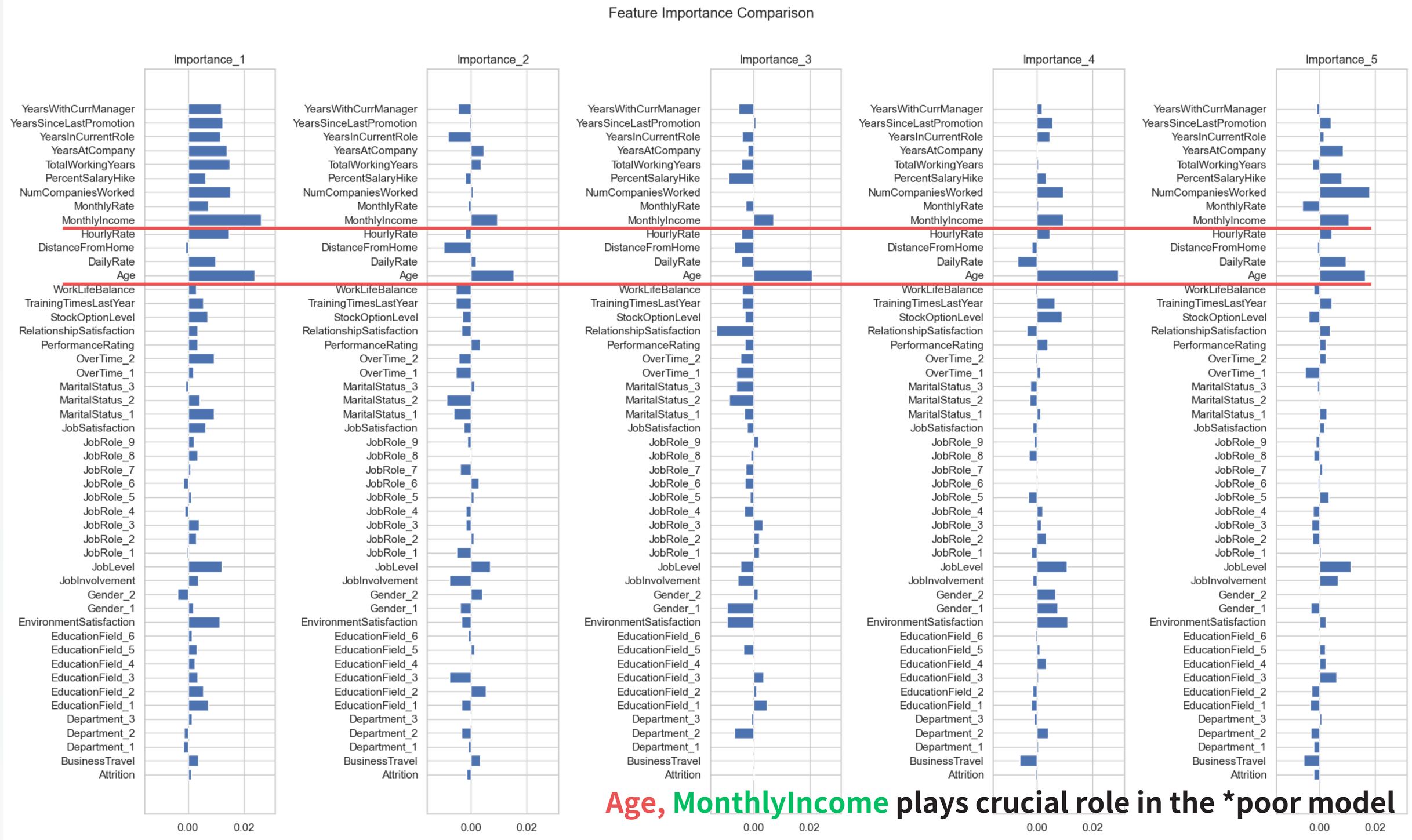
model	acc	precision_score	recall_score	f1_score
GradientBoostingClassifier	0.406	0.4	0.406	0.396
RandomForestClassifier	0.421	0.42	0.421	0.413
SVC	0.386	0.365	0.386	0.385
SVC_rbf	0.388	0.365	0.388	0.377
KNeighborsClassifier	0.354	0.366	0.354	0.348
XGBClassifier	0.404	0.399	0.404	0.399

very poor result

ANALYSIS

Use Permutation feature importance To find features for RFC model


Feature Importance Comparison
Importance_1
Importance_2
Importance_3
Importance_4
Importance_5
YearsWithCurrManager
YearsSinceLastPromotion
YearsInCurrentRole
YearsAtCompany
TotalWorkingYears
PercentSalaryHike
NumCompaniesWorked
MonthlyRate
MonthlyIncome
HourlyRate
DistanceFromHome
DailyRate
Age
WorkLifeBalance
TrainingTimesLastYear
StockOptionLevel
RelationshipSatisfaction
PerformanceRating
OverTime_2
OverTime_1
MaritalStatus_3
MaritalStatus_2
MaritalStatus_1
JobSatisfaction
JobRole_9
JobRole_8
JobRole_7
JobRole_6
JobRole_5
JobRole_4
JobRole_3
JobRole_2
JobRole_1
JobLevel
JobInvolvement
Gender_2
Gender_1
EnvironmentSatisfaction
EducationField_6
EducationField_5
EducationField_4
EducationField_3
EducationField_2
EducationField_1
Department_3
Department_2
Department_1
BusinessTravel
Attrition
0.00
0.02


Feature Importance Comparison
Importance_1
Importance_2
Importance_3
Importance_4
Importance_5
YearsWithCurrManager
YearsSinceLastPromotion
YearsInCurrentRole
YearsAtCompany
TotalWorkingYears
PercentSalaryHike
NumCompaniesWorked
MonthlyRate
MonthlyIncome
HourlyRate
DistanceFromHome
DailyRate
Age
WorkLifeBalance
TrainingTimesLastYear
StockOptionLevel
RelationshipSatisfaction
PerformanceRating
OverTime_2
OverTime_1
MaritalStatus_3
MaritalStatus_2
MaritalStatus_1
JobSatisfaction
JobRole_9
JobRole_8
JobRole_7
JobRole_6
JobRole_5
JobRole_4
JobRole_3
JobRole_2
JobRole_1
JobLevel
JobInvolvement
Gender_2
Gender_1
EnvironmentSatisfaction
EducationField_6
EducationField_5
EducationField_4
EducationField_3
EducationField_2
EducationField_1
Department_3
Department_2
Department_1
BusinessTravel
Attrition
0.00
0.02
Age, MonthlyIncome plays crucial role in the *poor model

CONCLUSION

ADDITIONAL ANALYSIS

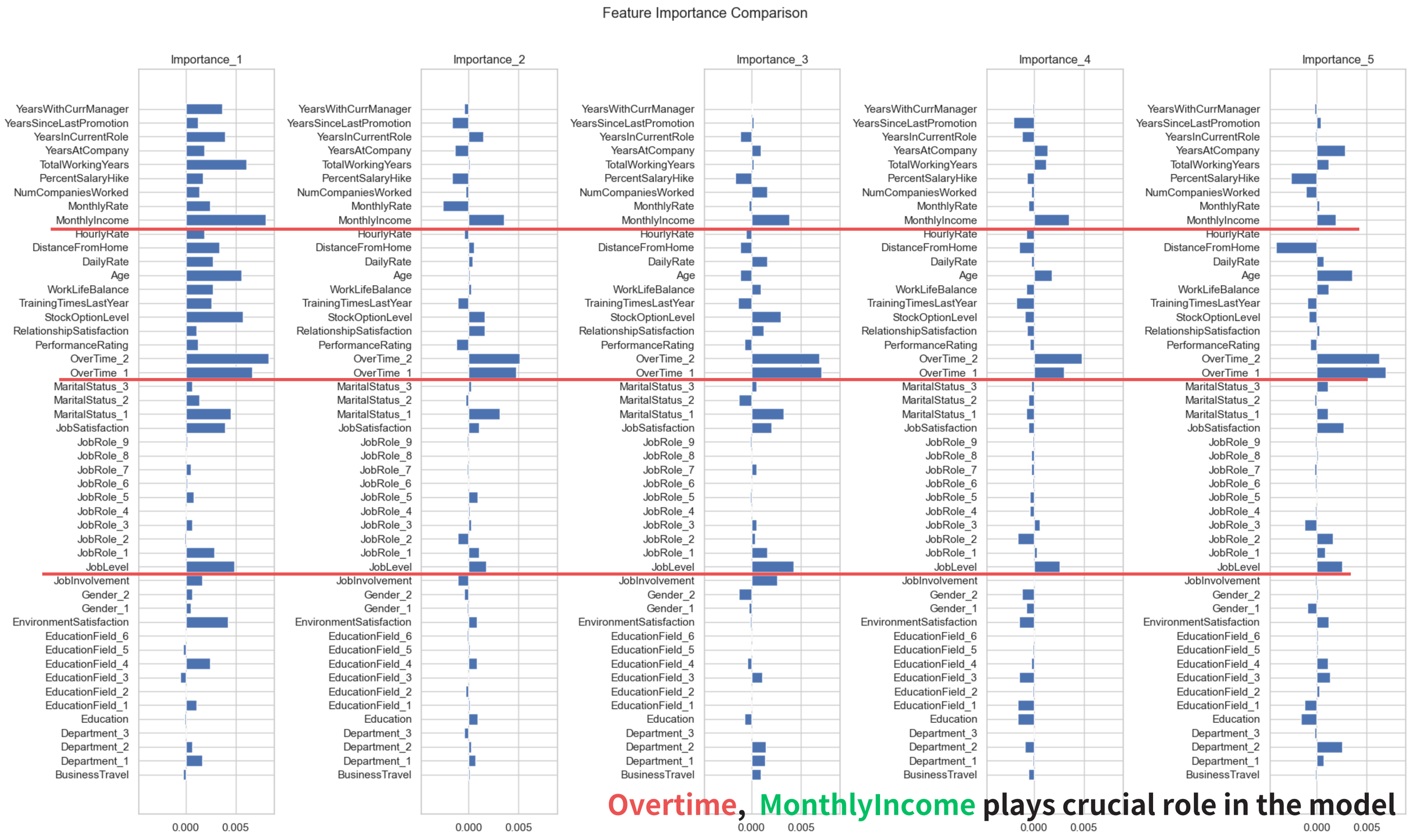
predict Attrition?

model	acc	precision_score	recall_score	f1_score
GradientBoostingClassifier	0.865	0.848	0.865	0.843
RandomForestClassifier	0.859	0.849	0.859	0.82
SVC	0.867	0.873	0.867	0.83
SVC_rbf	0.867	0.873	0.867	0.83
ExtraTreesClassifier	0.859	0.844	0.859	0.826
KNeighborsClassifier	0.844	0.815	0.844	0.797
XGBClassifier	0.868	0.854	0.868	0.848

not bad!!!

ADDITIONAL ANALYSIS

Use Permutation feature importance To find features for RFC model

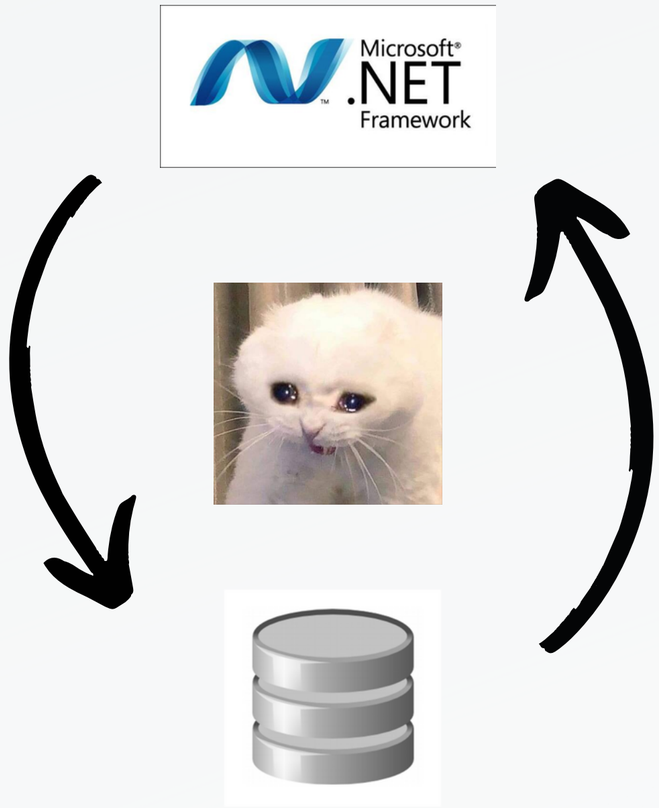

Feature Importance Comparison
Importance_1
Importance_2
Importance_3
Importance_4
Importance_5
YearsWithCurrManager
YearsSinceLastPromotion
YearsInCurrentRole
YearsAtCompany
TotalWorkingYears
PercentSalaryHike
NumCompaniesWorked
MonthlyRate
MonthlyIncome
HourlyRate
DistanceFromHome
DailyRate
Age
WorkLifeBalance
TrainingTimesLastYear
StockOptionLevel
RelationshipSatisfaction
PerformanceRating
OverTime_2
OverTime_1
MaritalStatus_3
MaritalStatus_2
MaritalStatus_1
JobSatisfaction
JobRole_9
JobRole_8
JobRole_7
JobRole_6
JobRole_5
JobRole_4
JobRole_3
JobRole_2
JobRole_1
JobLevel
JobInvolvement
Gender_2
Gender_1
EnvironmentSatisfaction
EducationField_6
EducationField_5
EducationField_4
EducationField_3
EducationField_2
EducationField_1
Education
Department_3
Department_2
Department_1
BusinessTravel
0.000
0.005
Overtime, MonthlyIncome plays crucial role in the model

the result is unexpectedly bad,
so we find another dataset

INTRO OF SECOND DATESET

Student Stress Factors: A Comprehensive Analysis

- Environmental Factors
- Psychological Factors
- Academic Factors
- Social Factor
- Physiological Factors



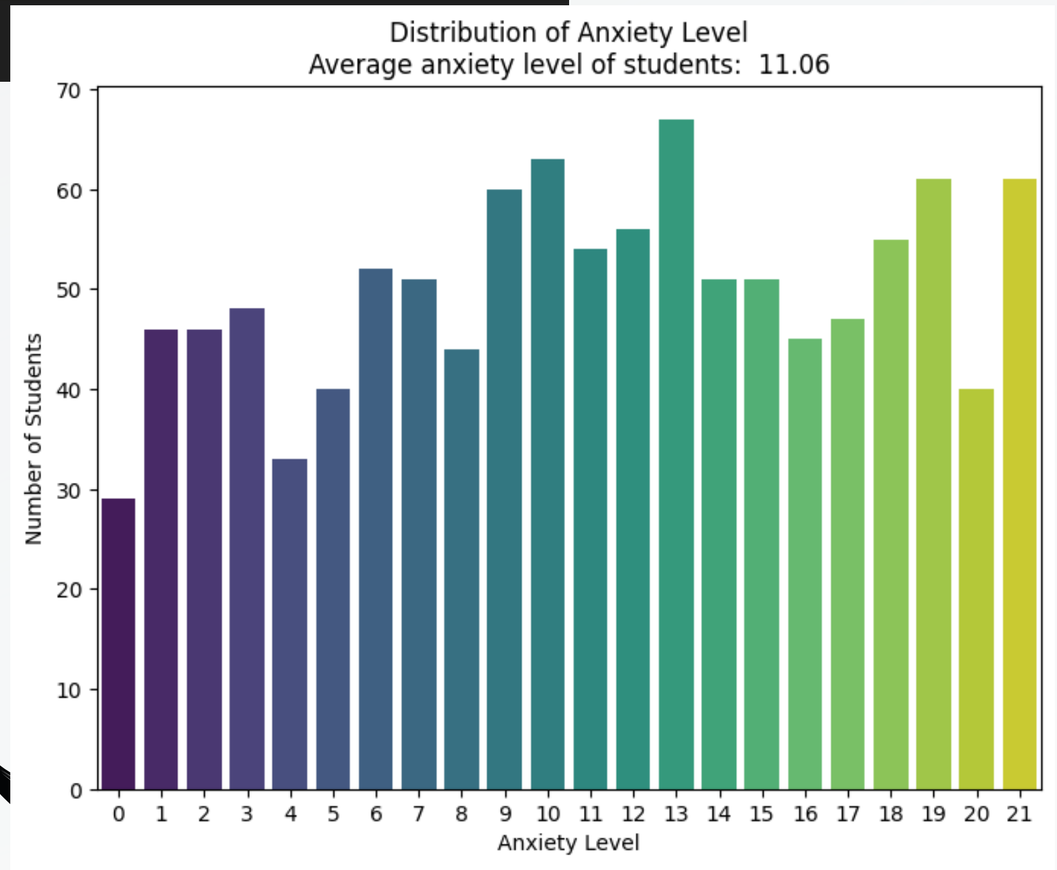
FEATURE DESCRIPTION

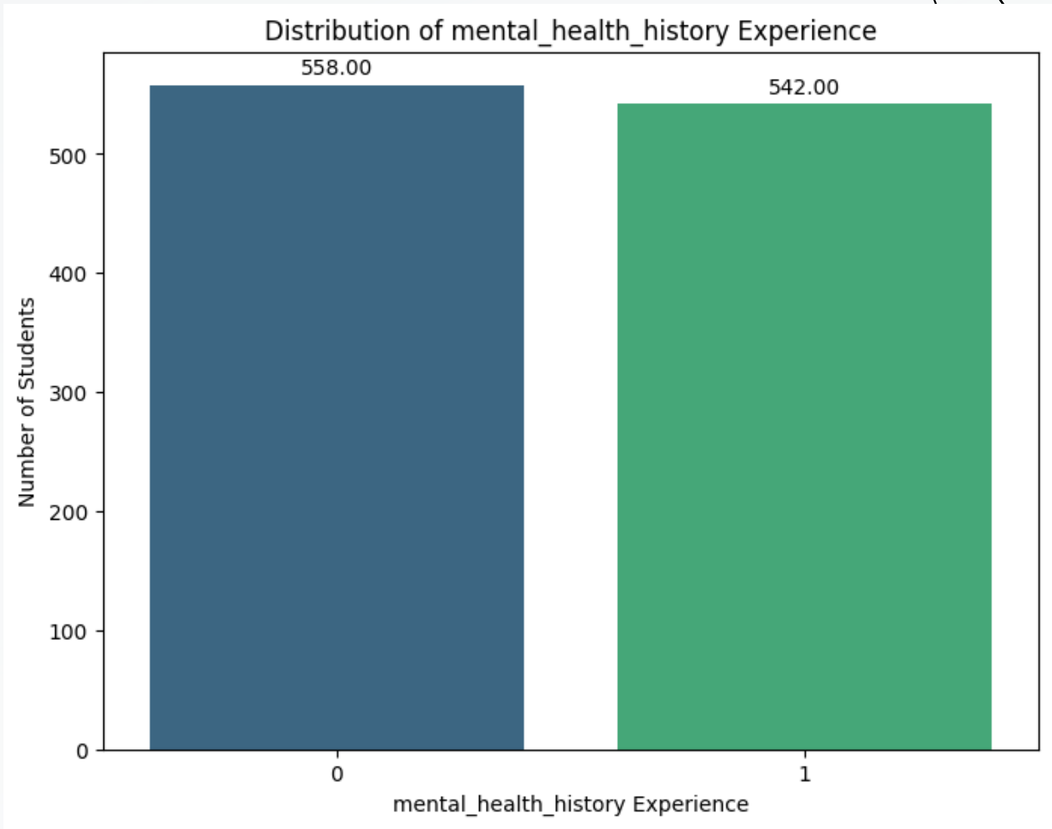
Descriptive Statistics

```
total_students = len(data)
print('total students: ' + str(total_students))
```

✓ 0.0s

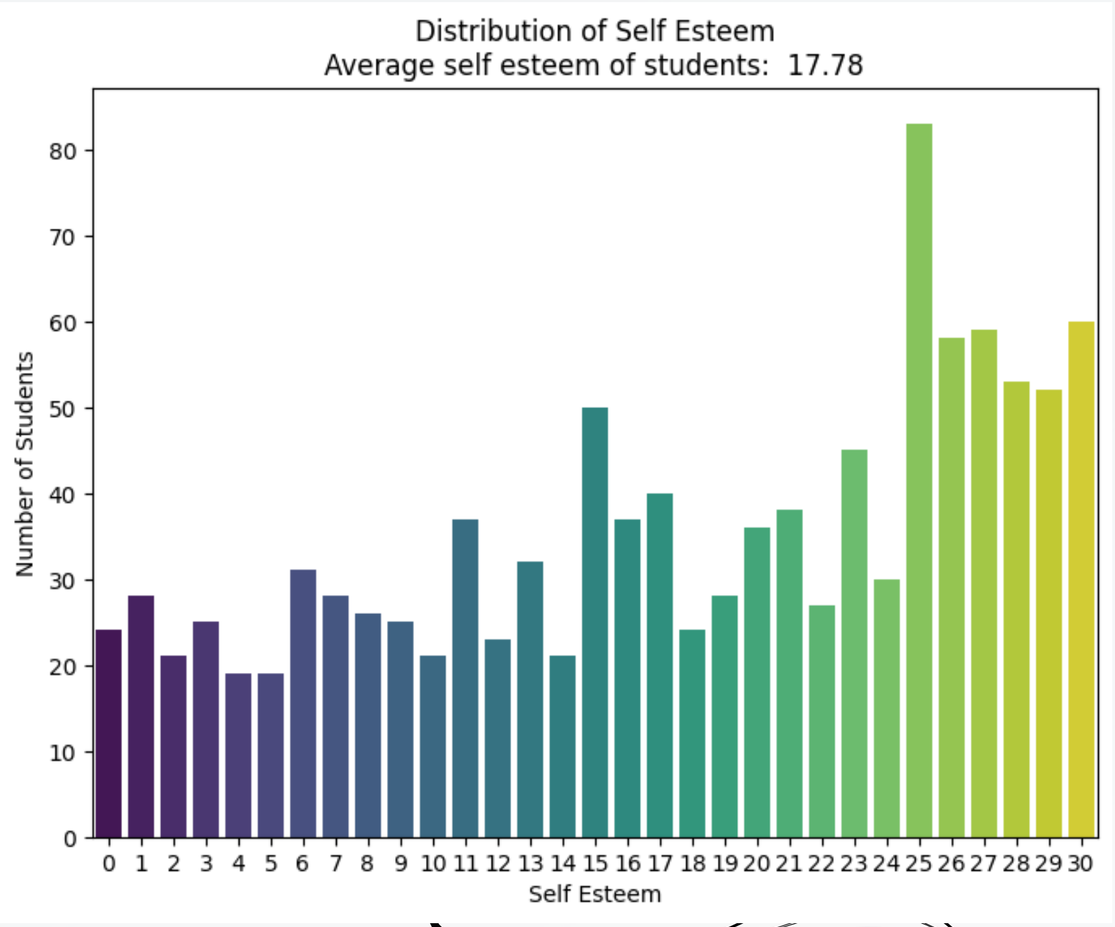
```
total students: 1100
```

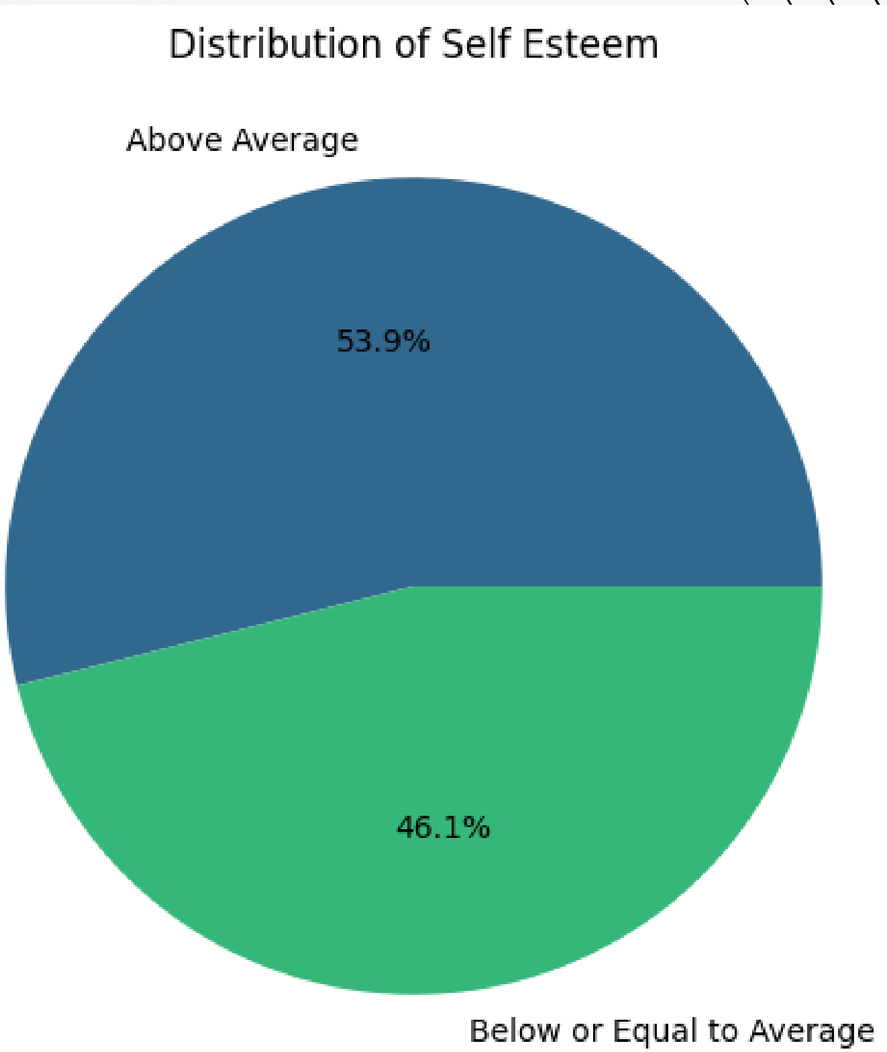




FEATURE DESCRIPTION

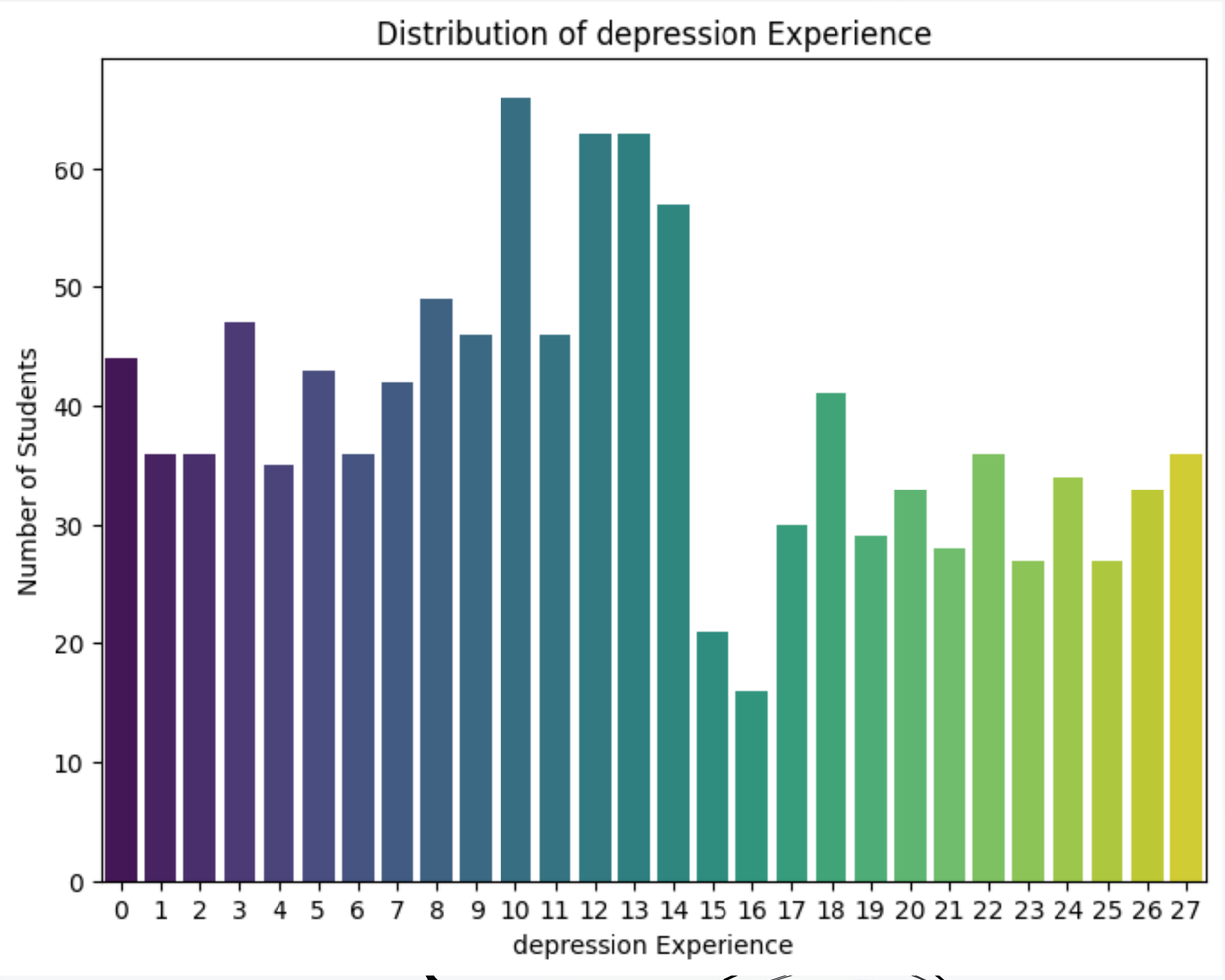
PSYCHOLOGICAL FACTORS

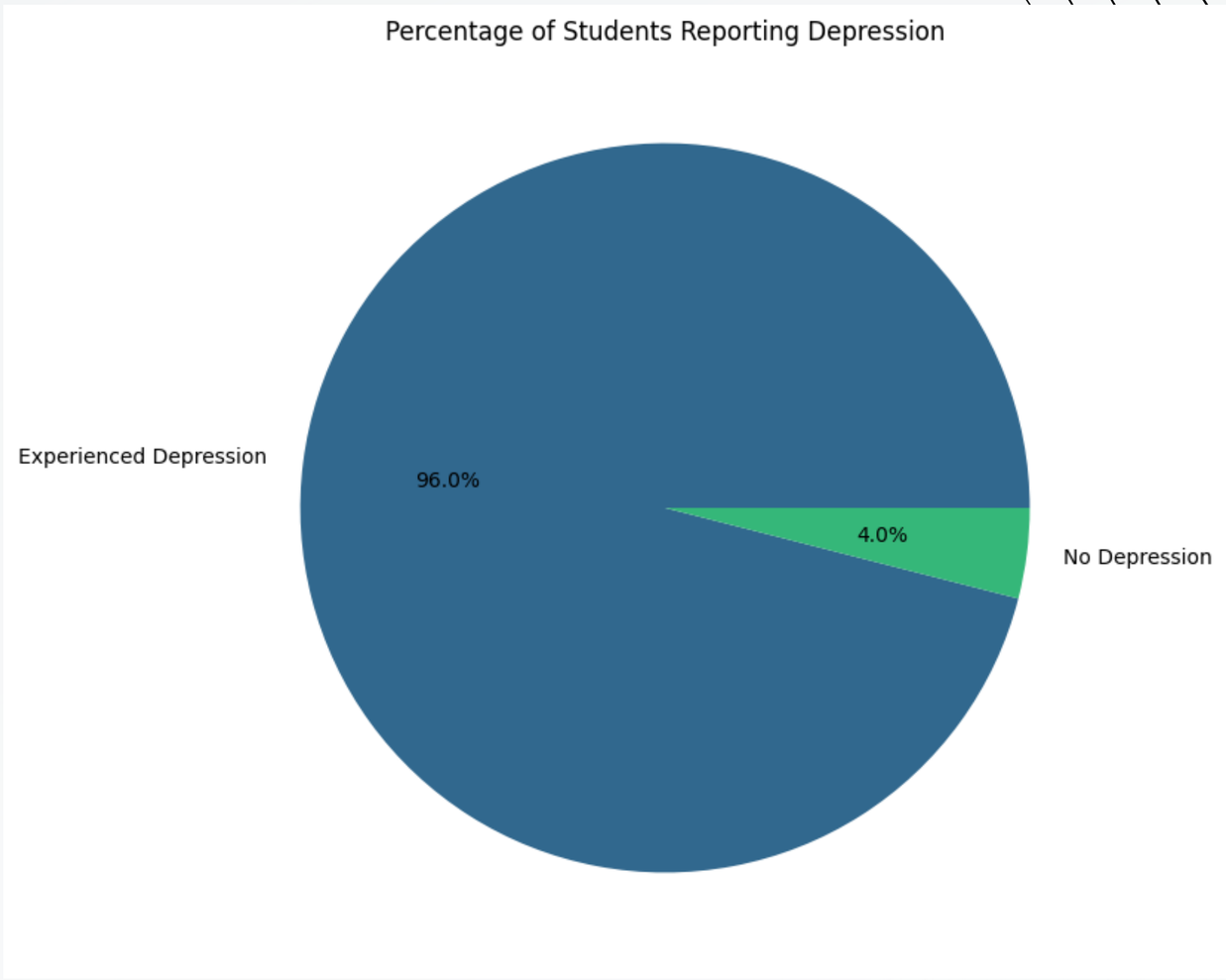




FEATURE DESCRIPTION

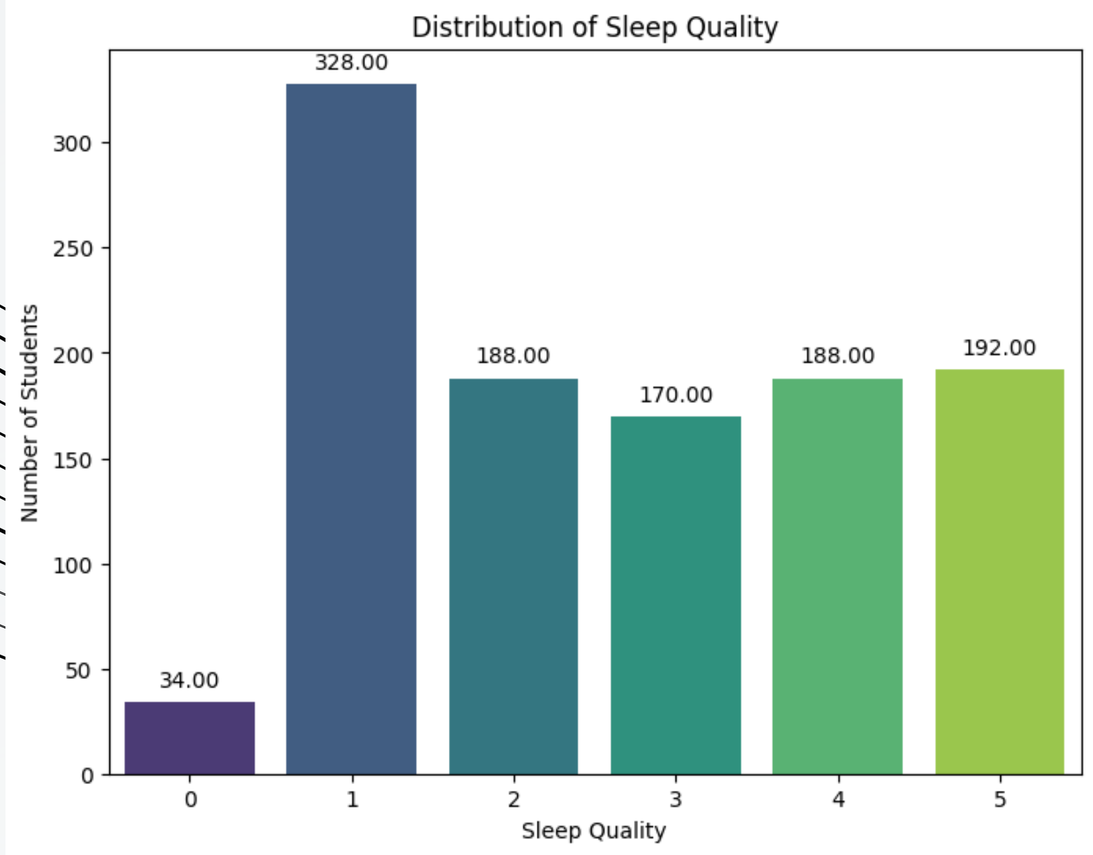
PSYCHOLOGICAL FACTORS

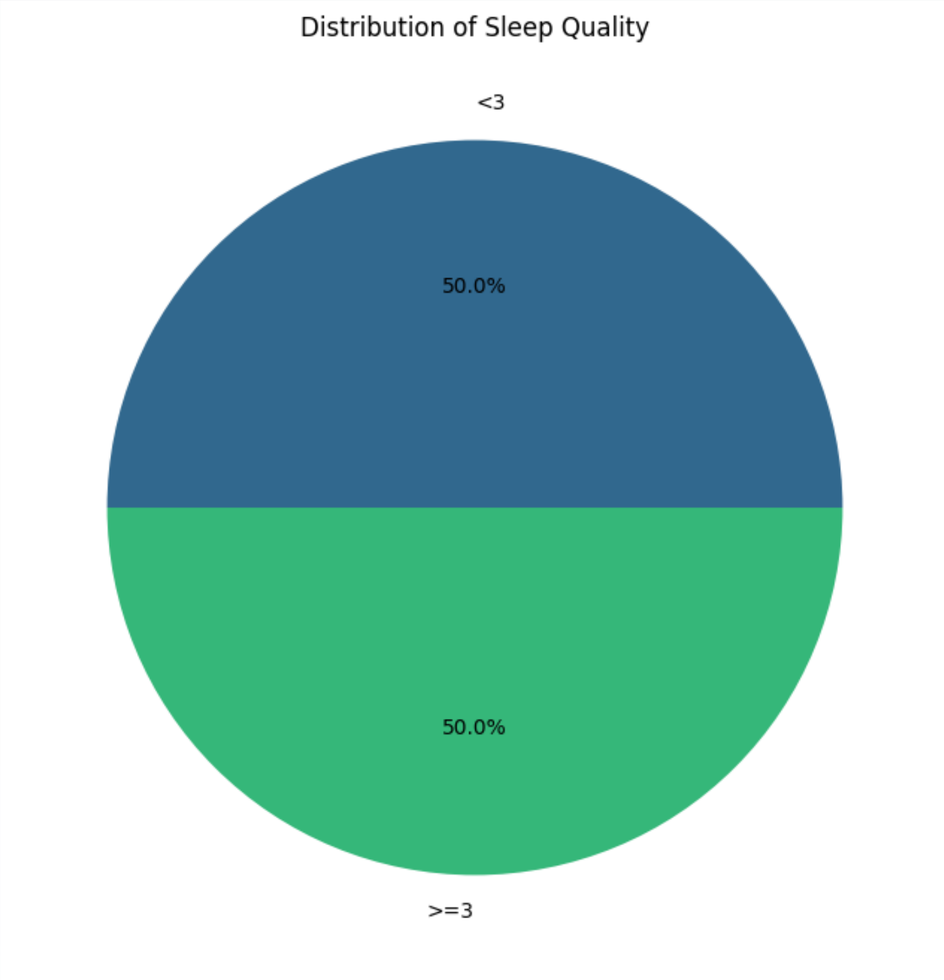




FEATURE DESCRIPTION

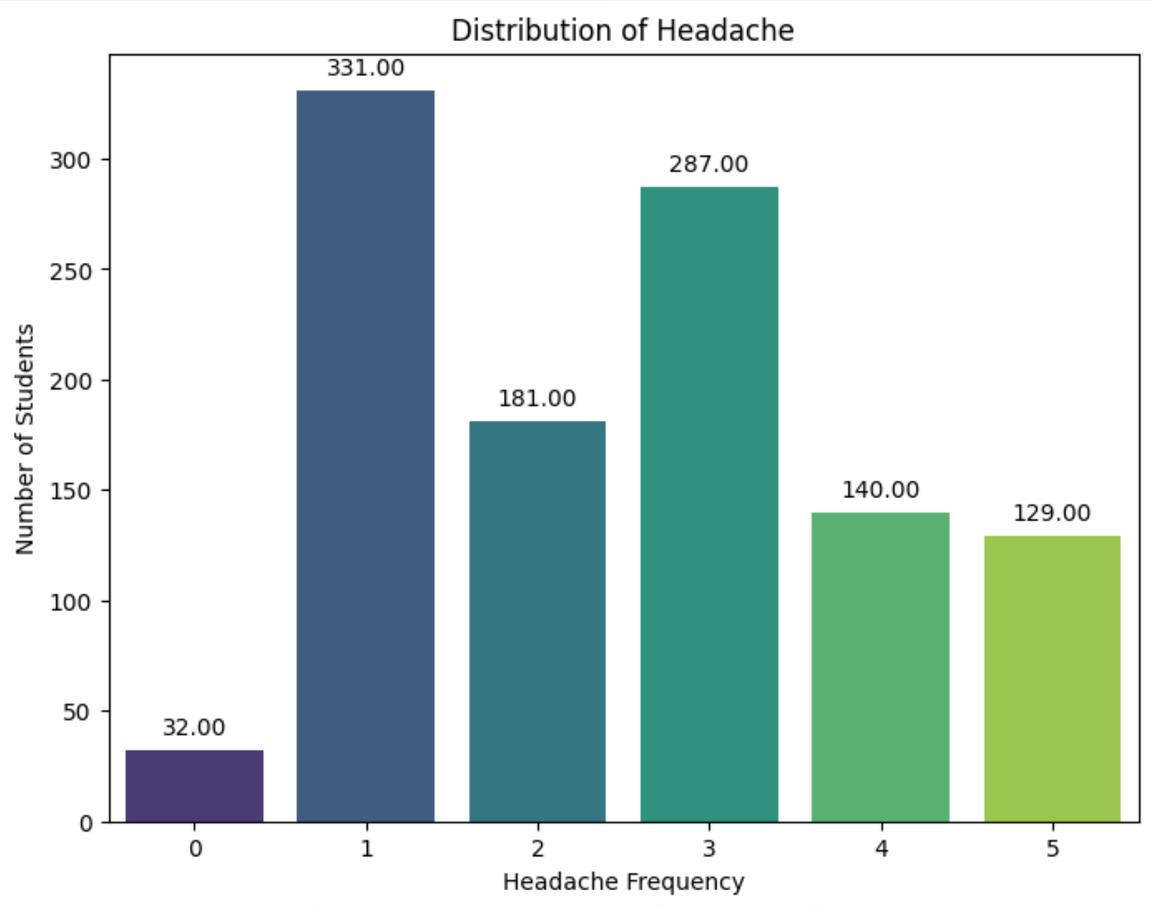
PHYSIOLOGICAL FACTORS

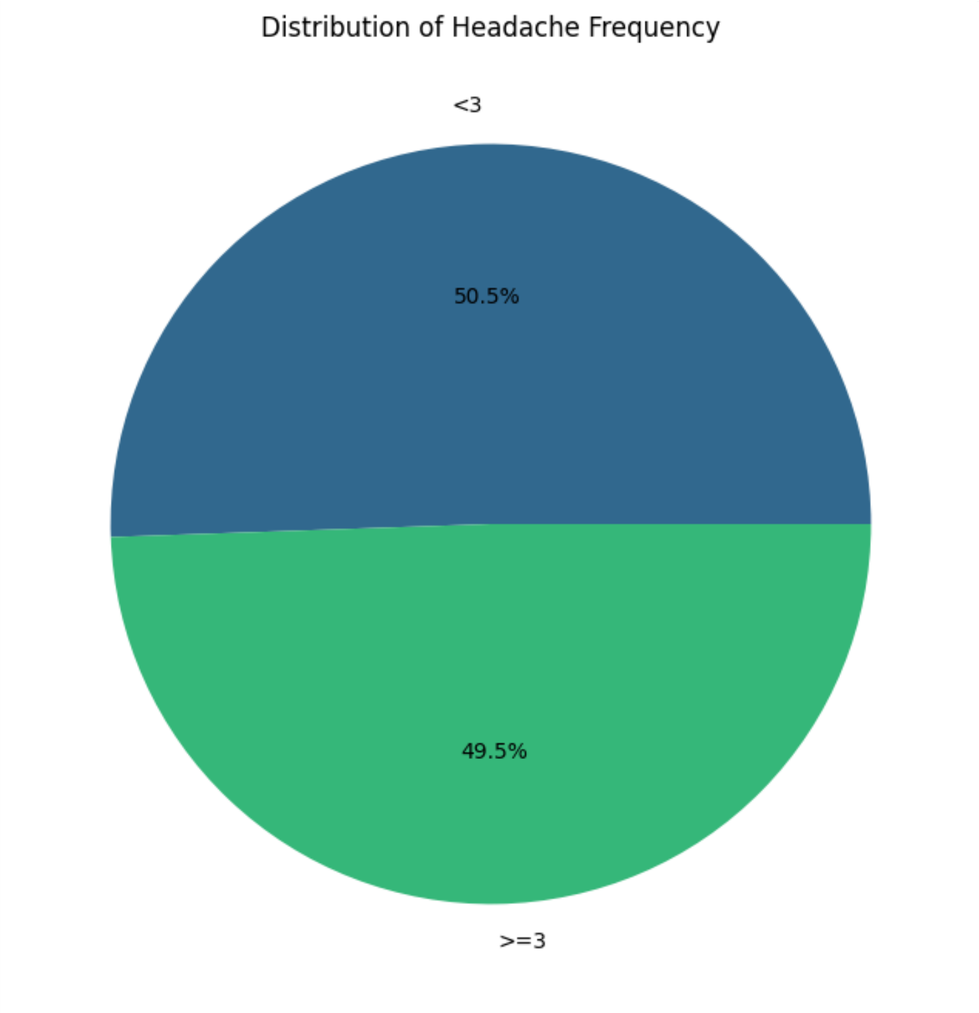




FEATURE DESCRIPTION

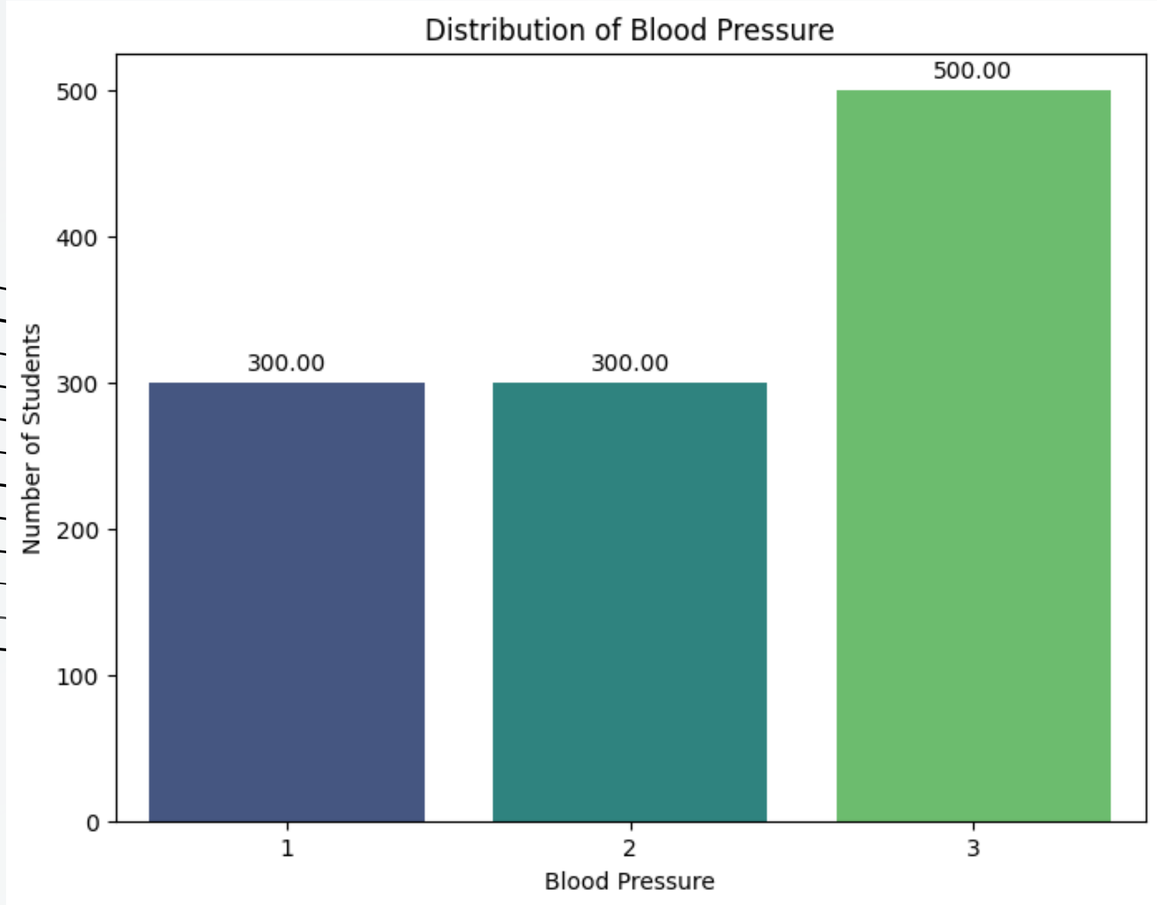
PHYSIOLOGICAL FACTORS

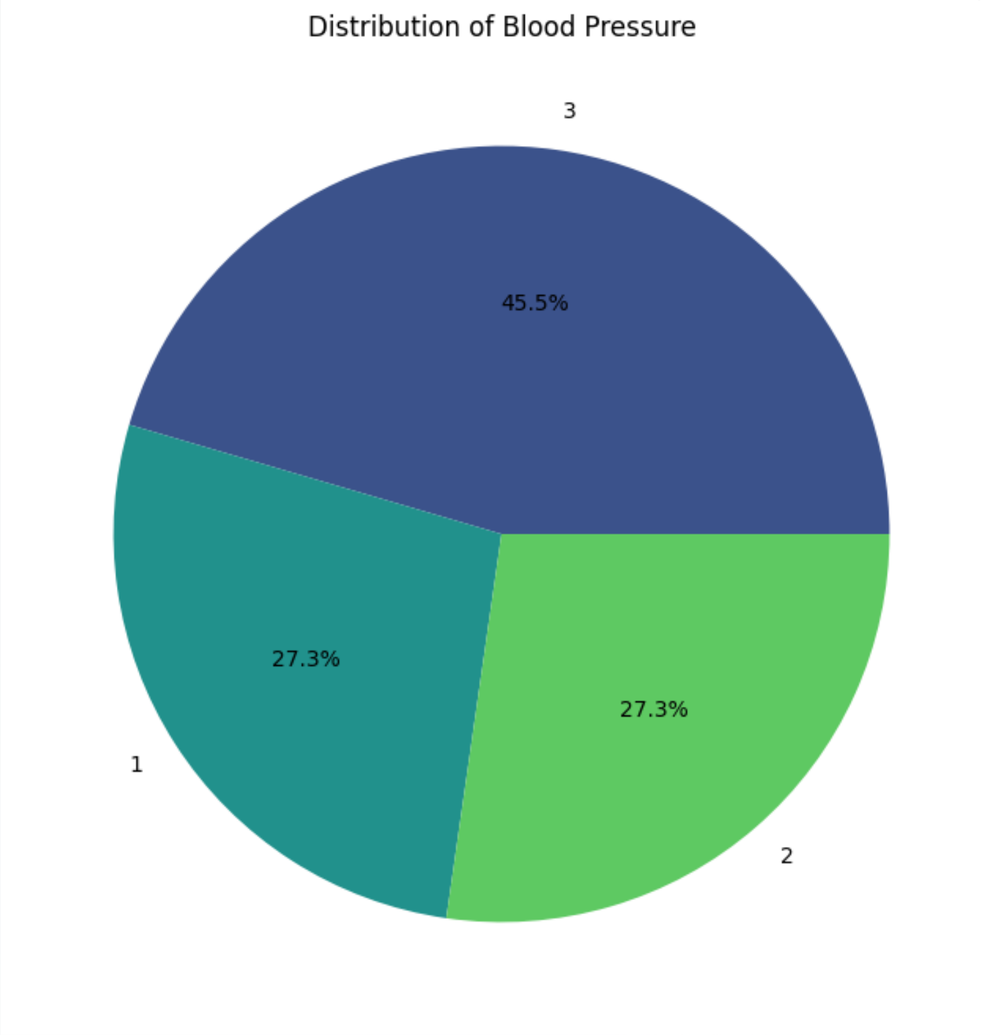




FEATURE DESCRIPTION

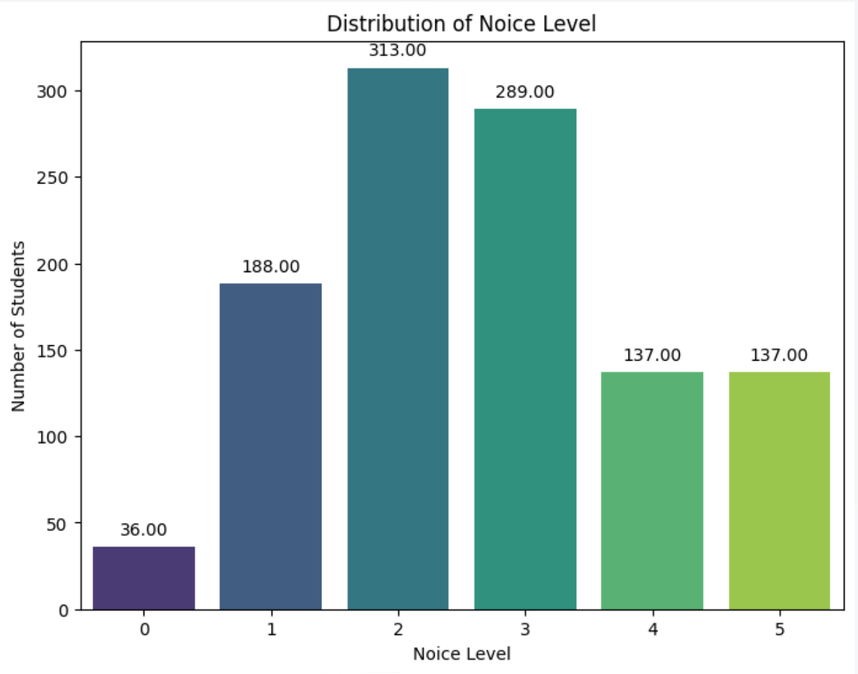
PHYSIOLOGICAL FACTORS

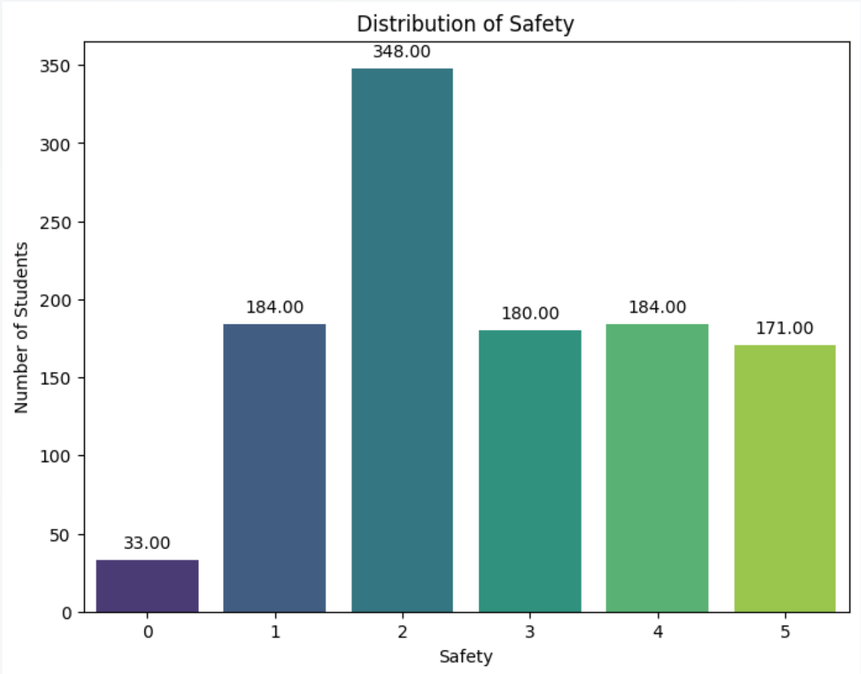


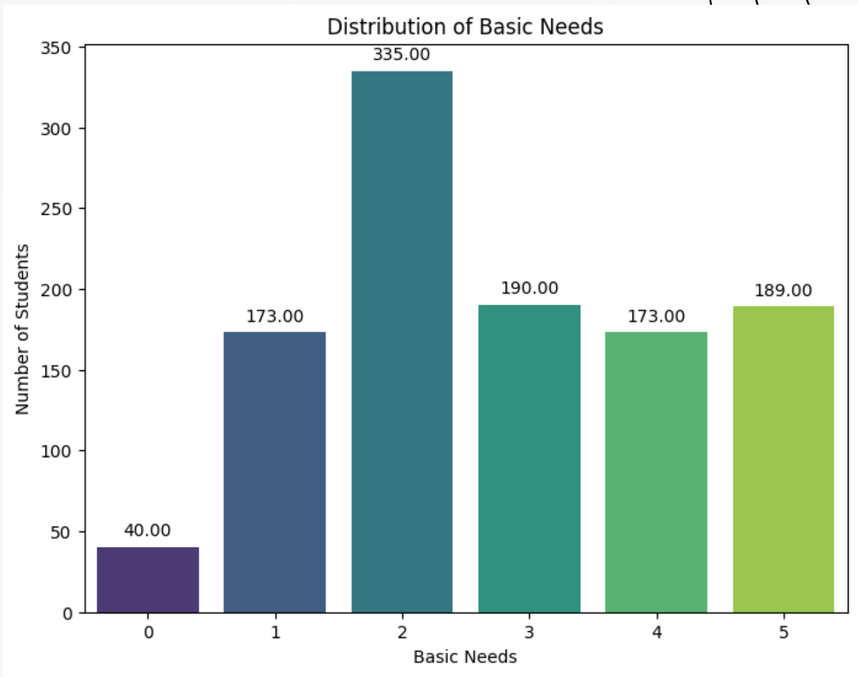


FEATURE DESCRIPTION

ENVIRONMENTAL FACTORS

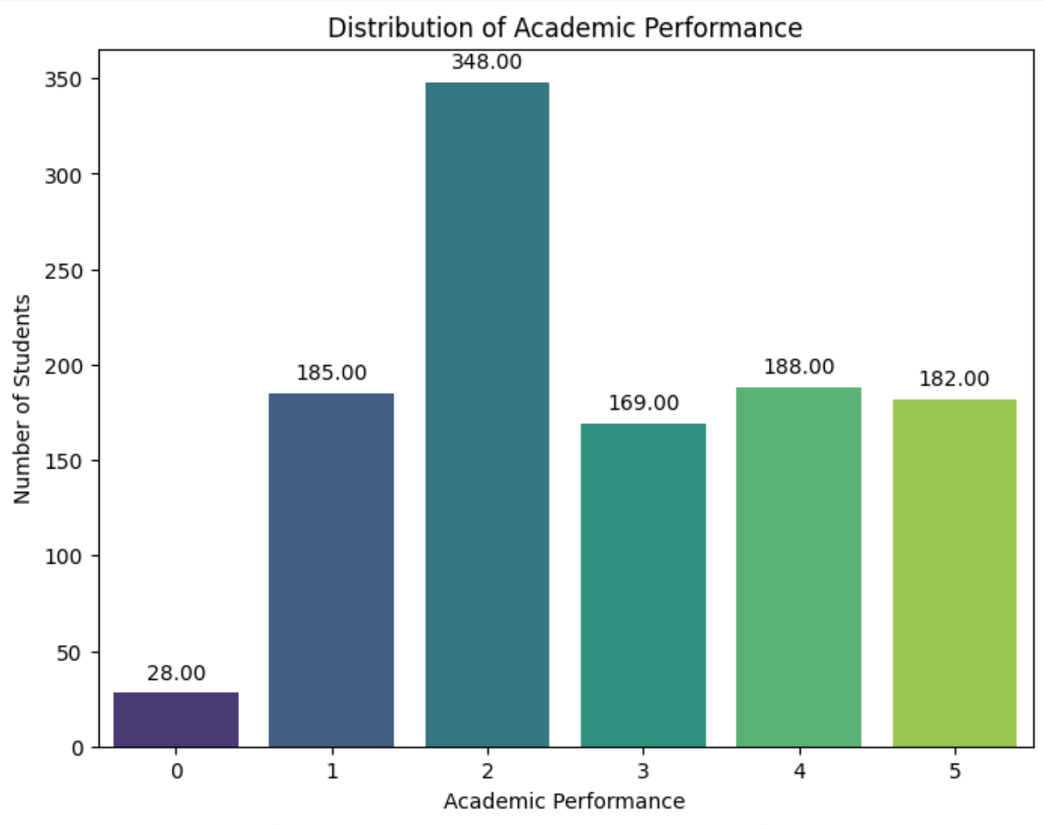


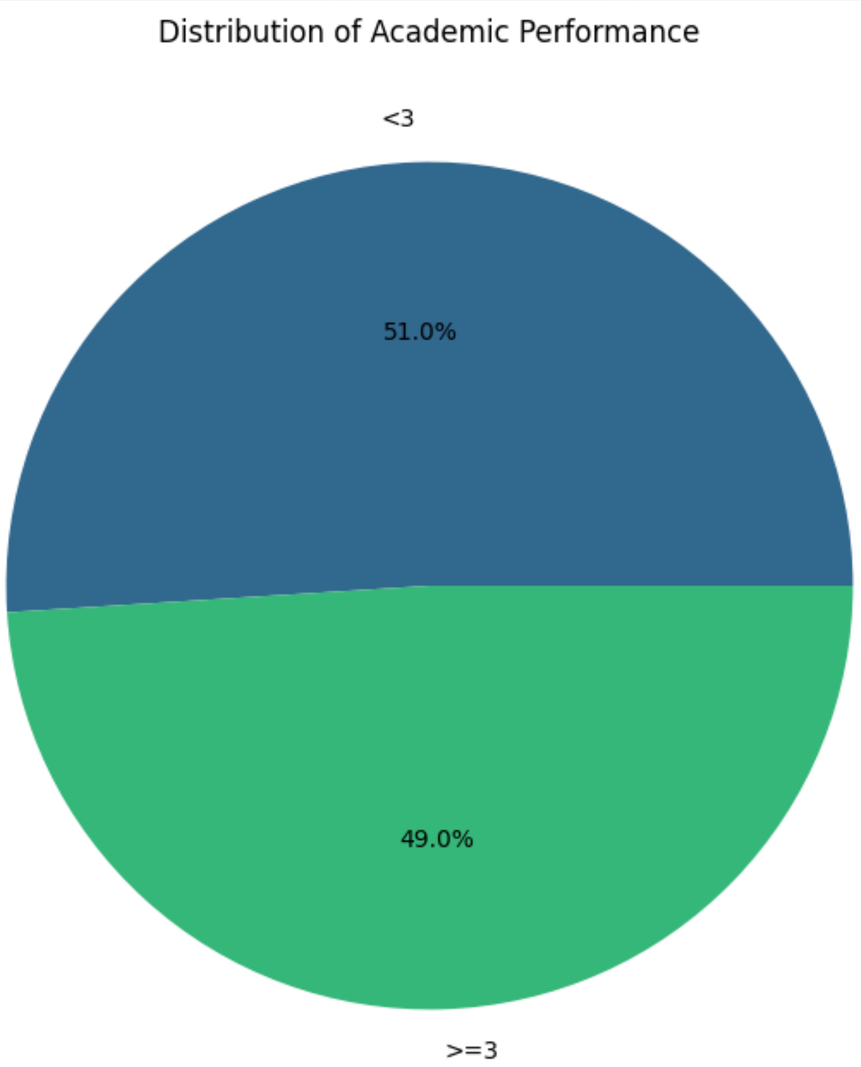




FEATURE DESCRIPTION

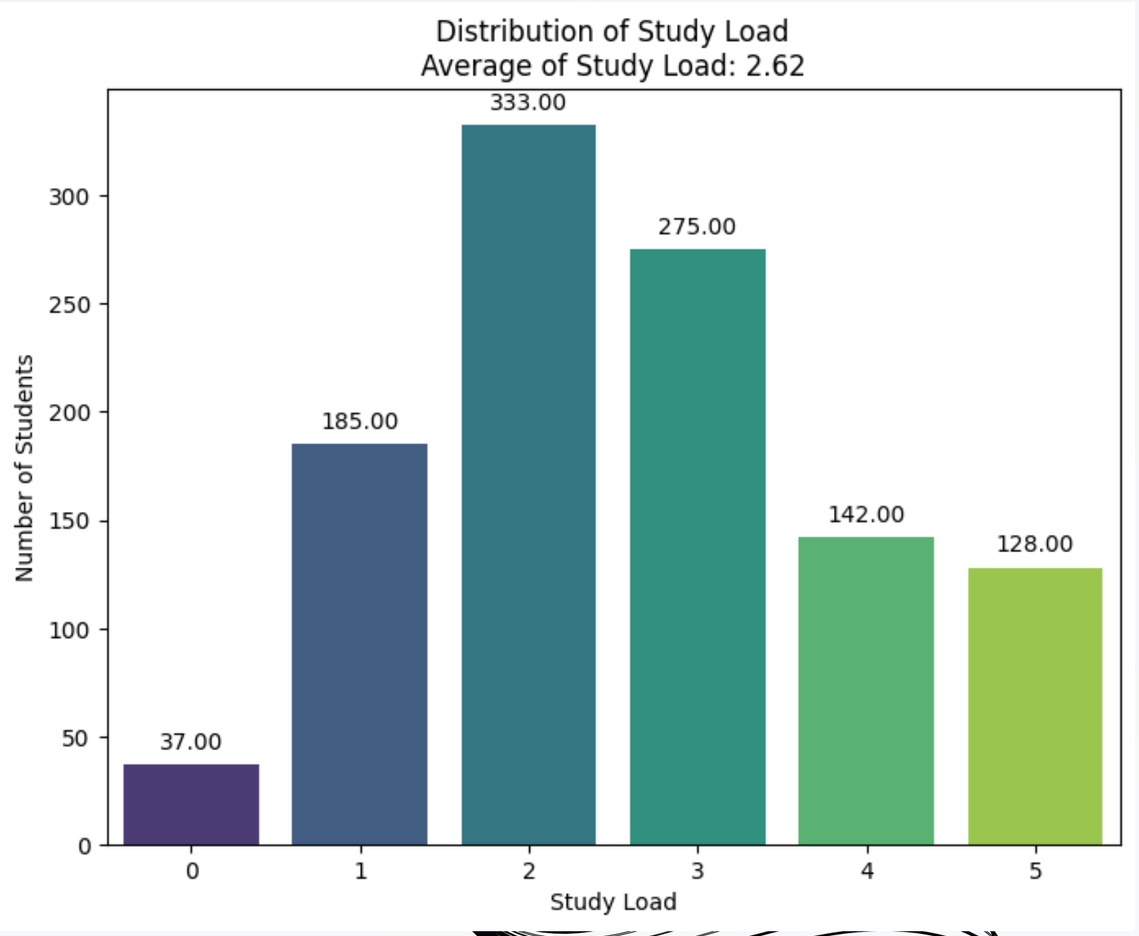
ACADEMIC FACTORS

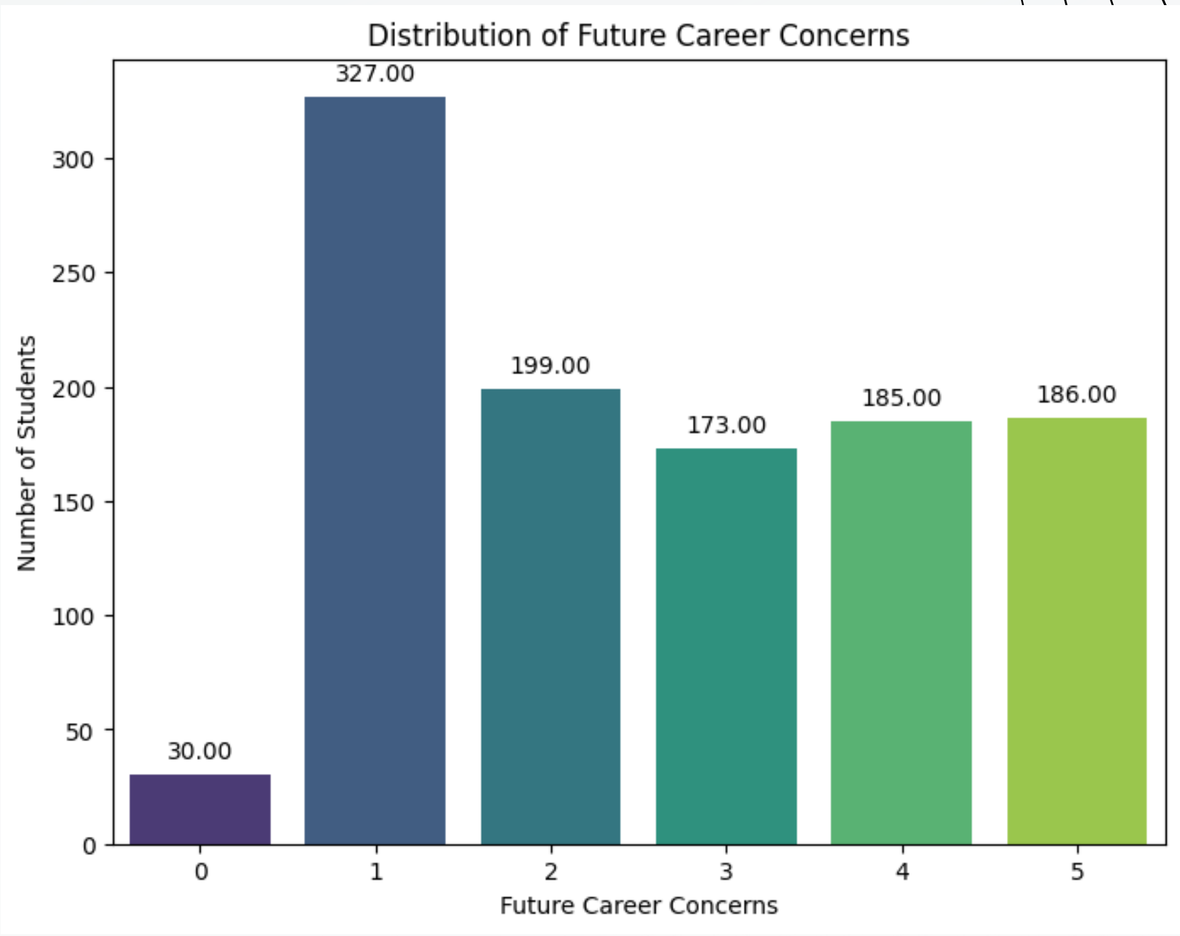




FEATURE DESCRIPTION

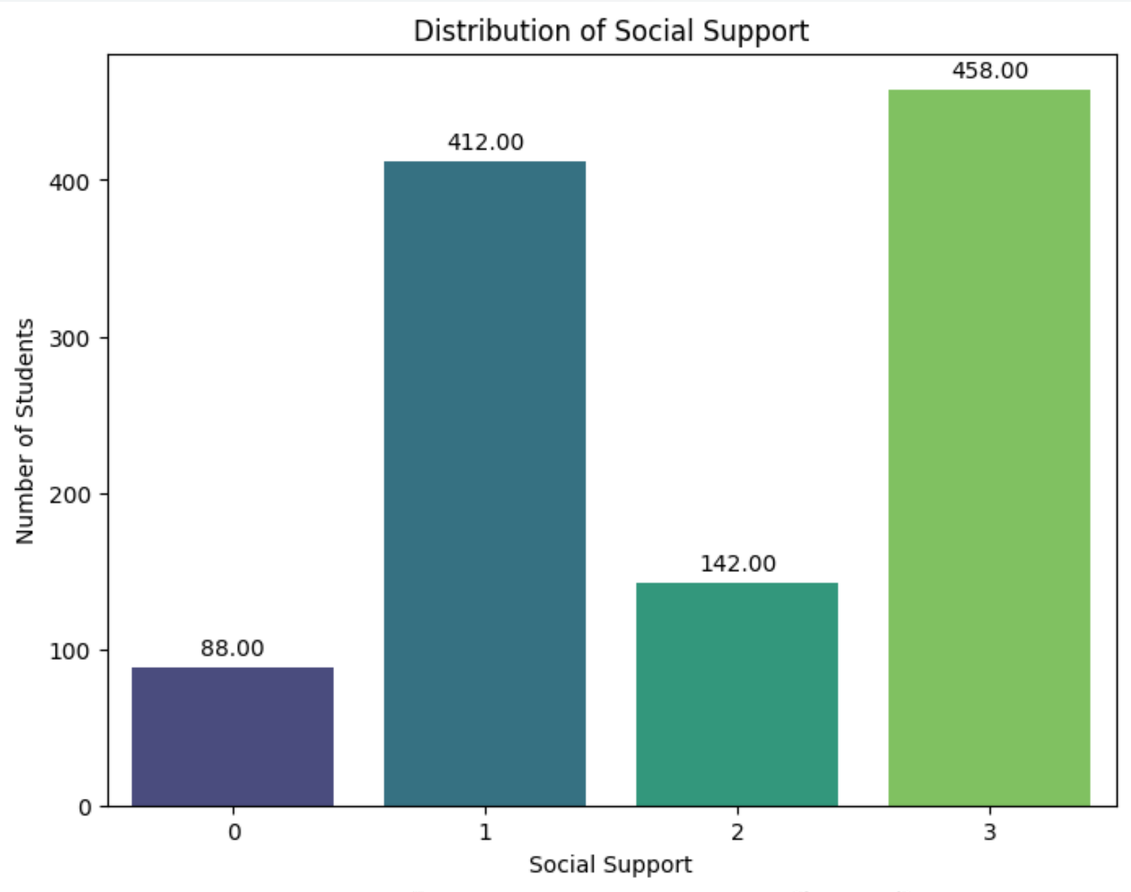
ACADEMIC FACTORS

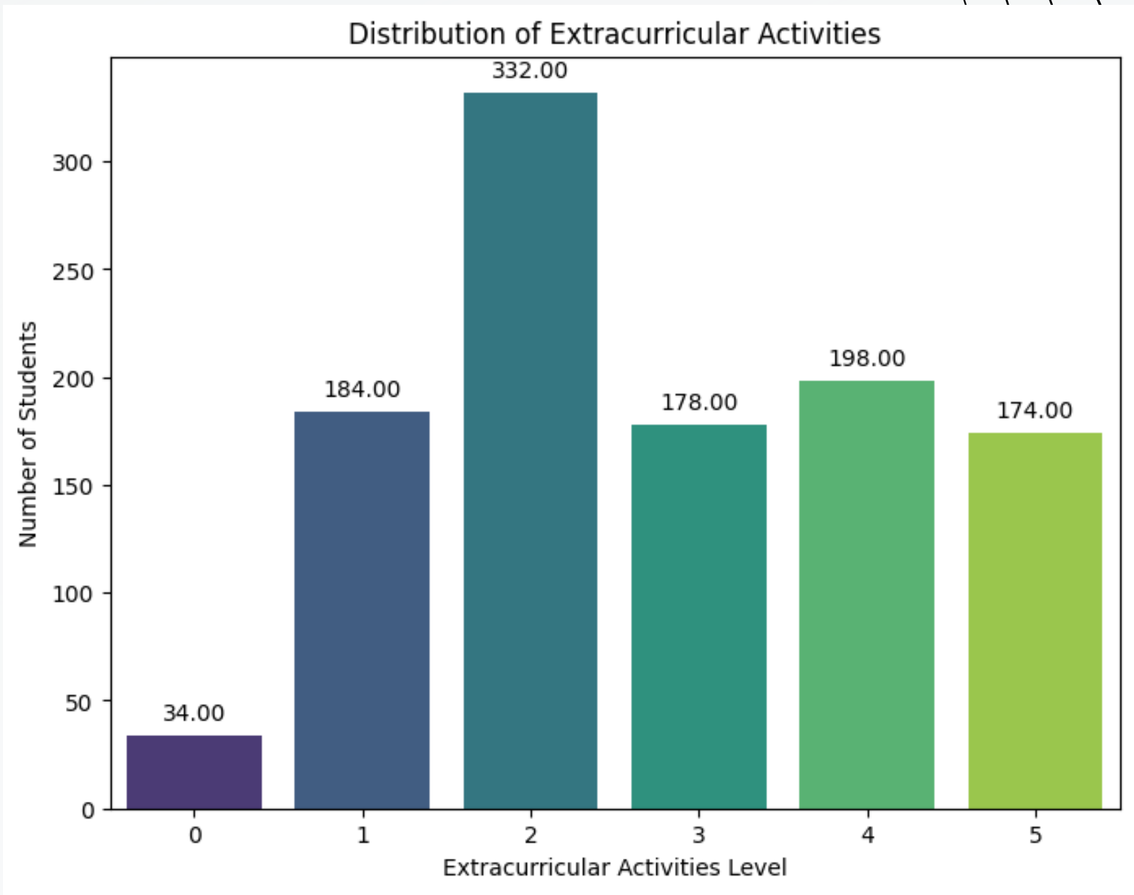




FEATURE DESCRIPTION

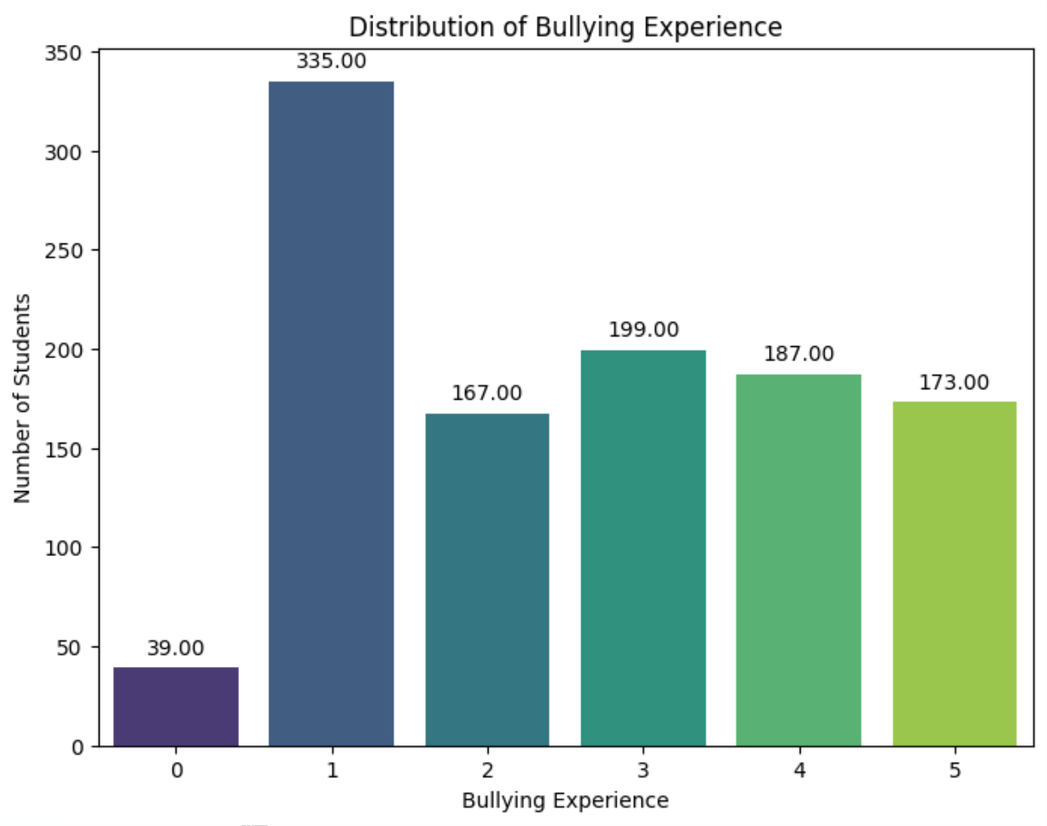
SOCIAL FACTORS

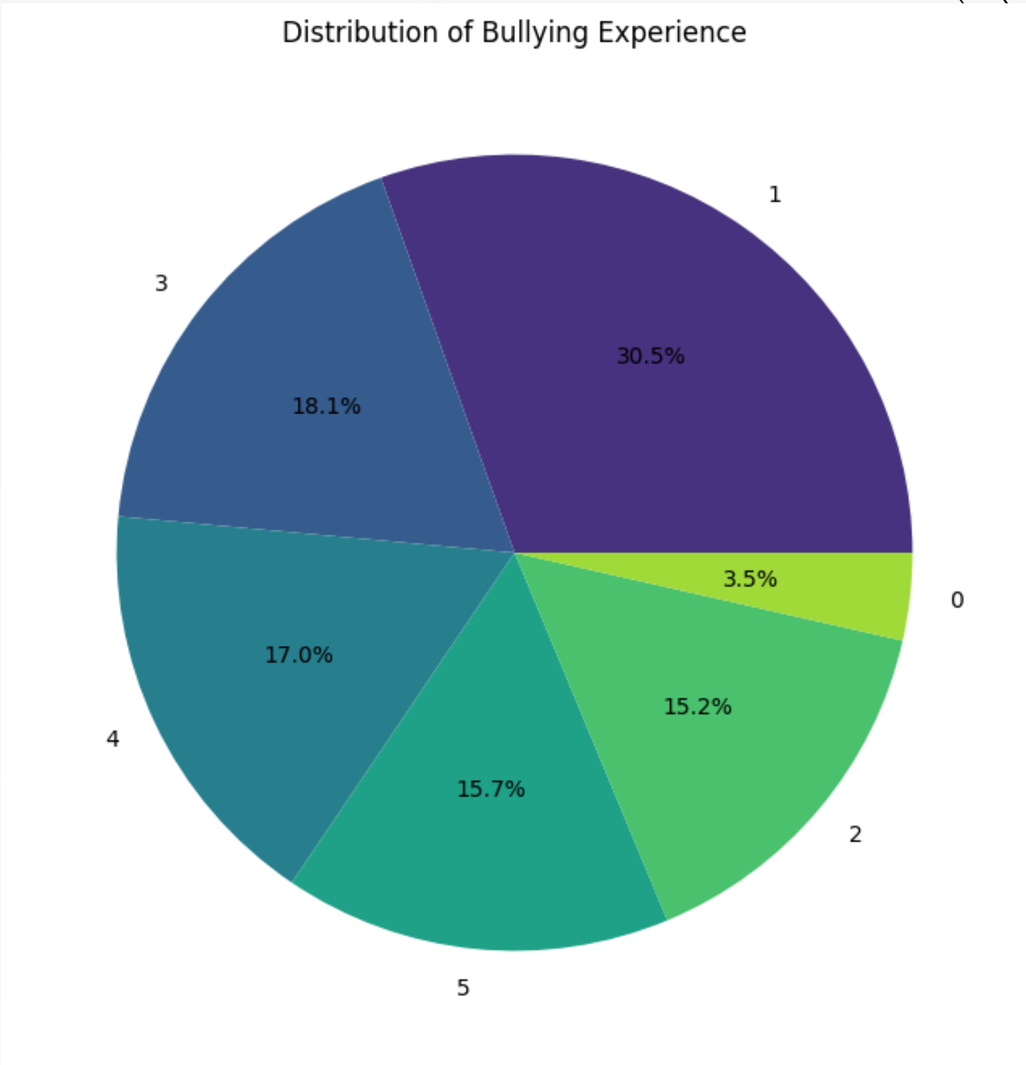




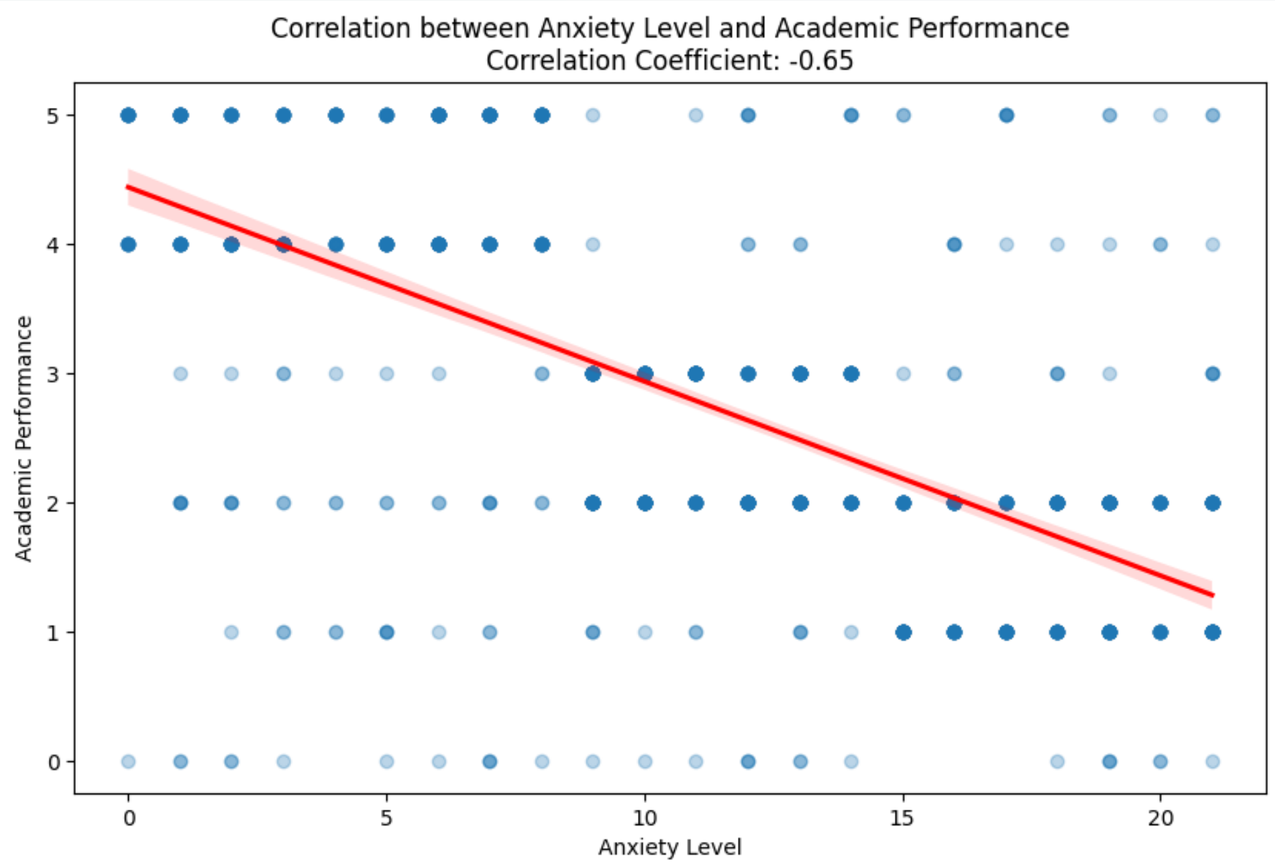
FEATURE DESCRIPTION

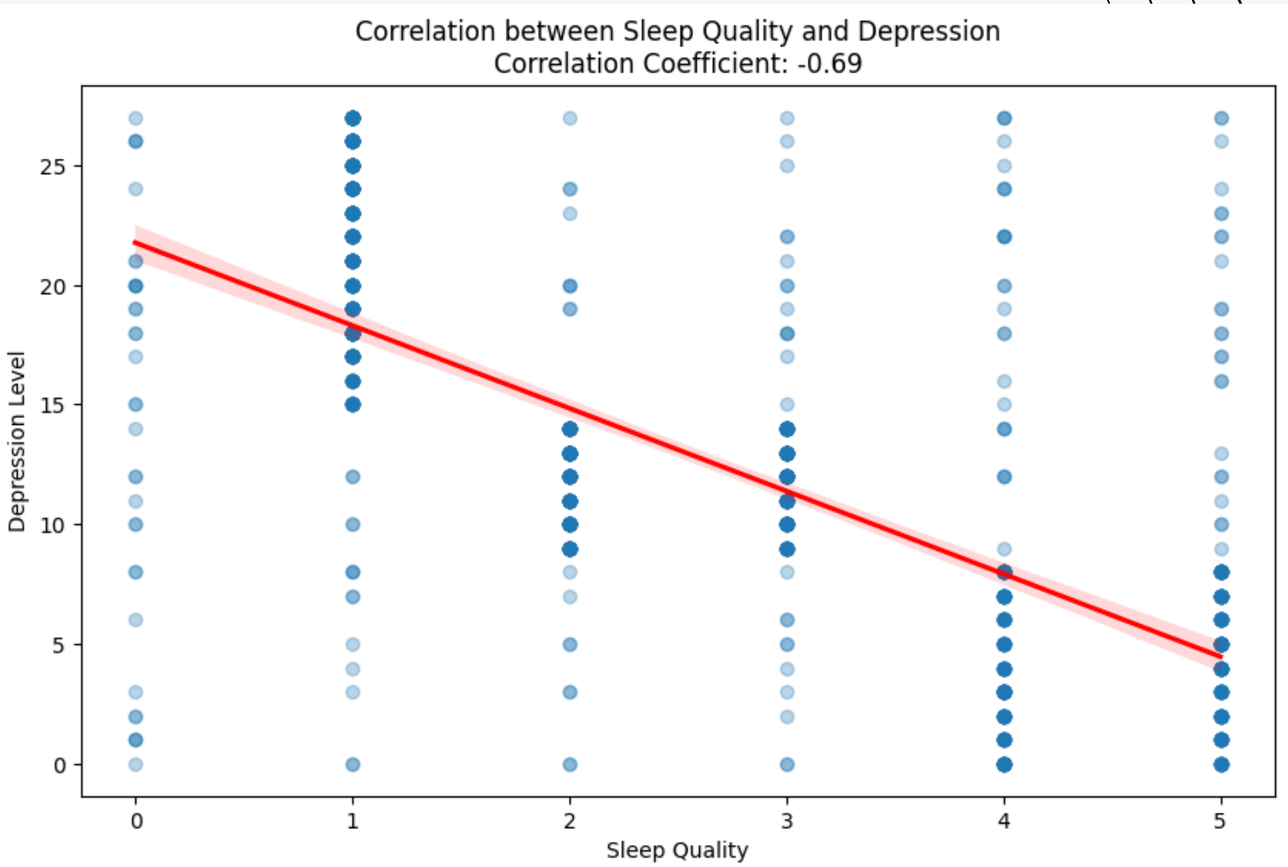
SOCIAL FACTORS





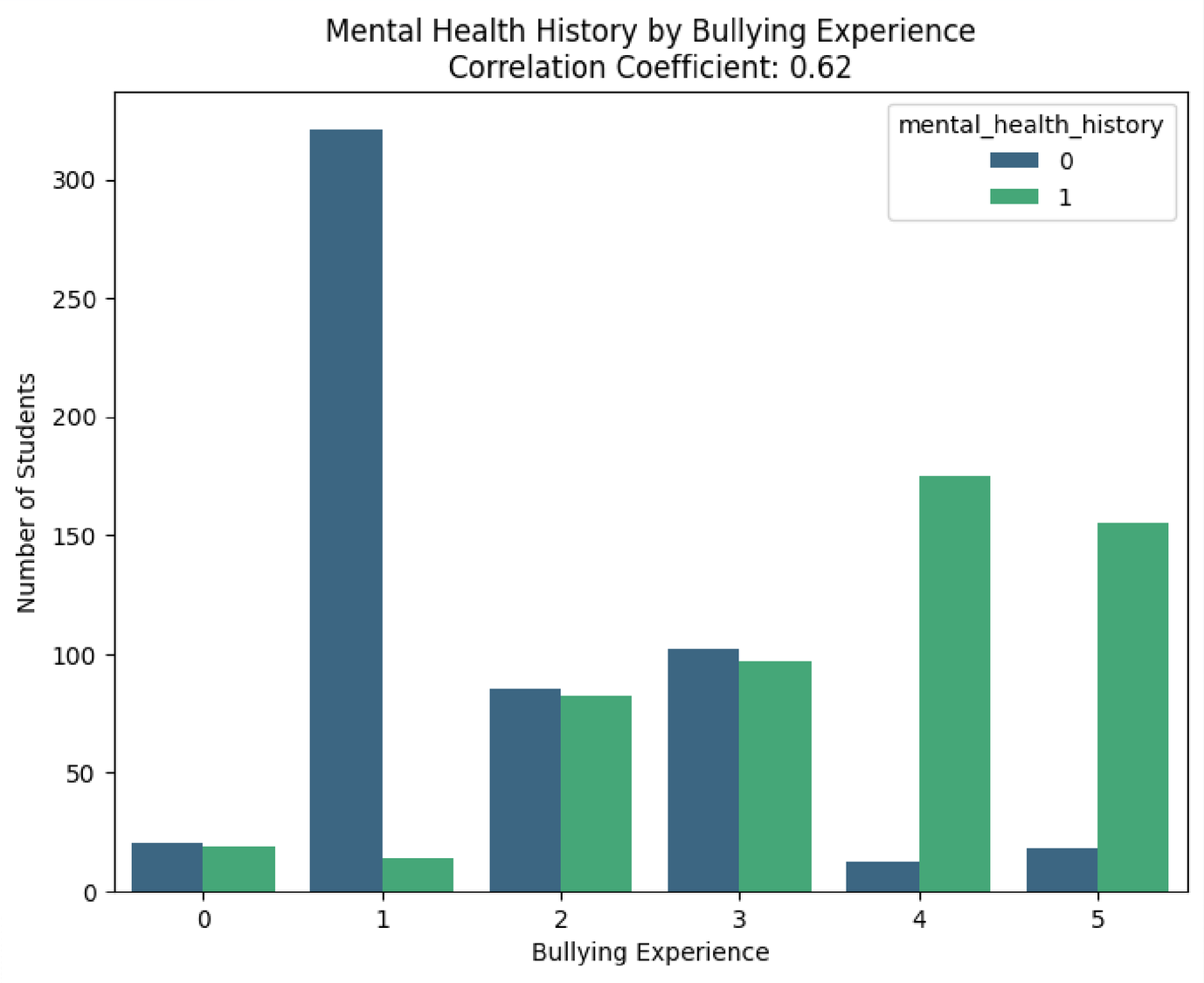
FEATURE DESCRIPTION
COMPARATIVE ANALYSIS



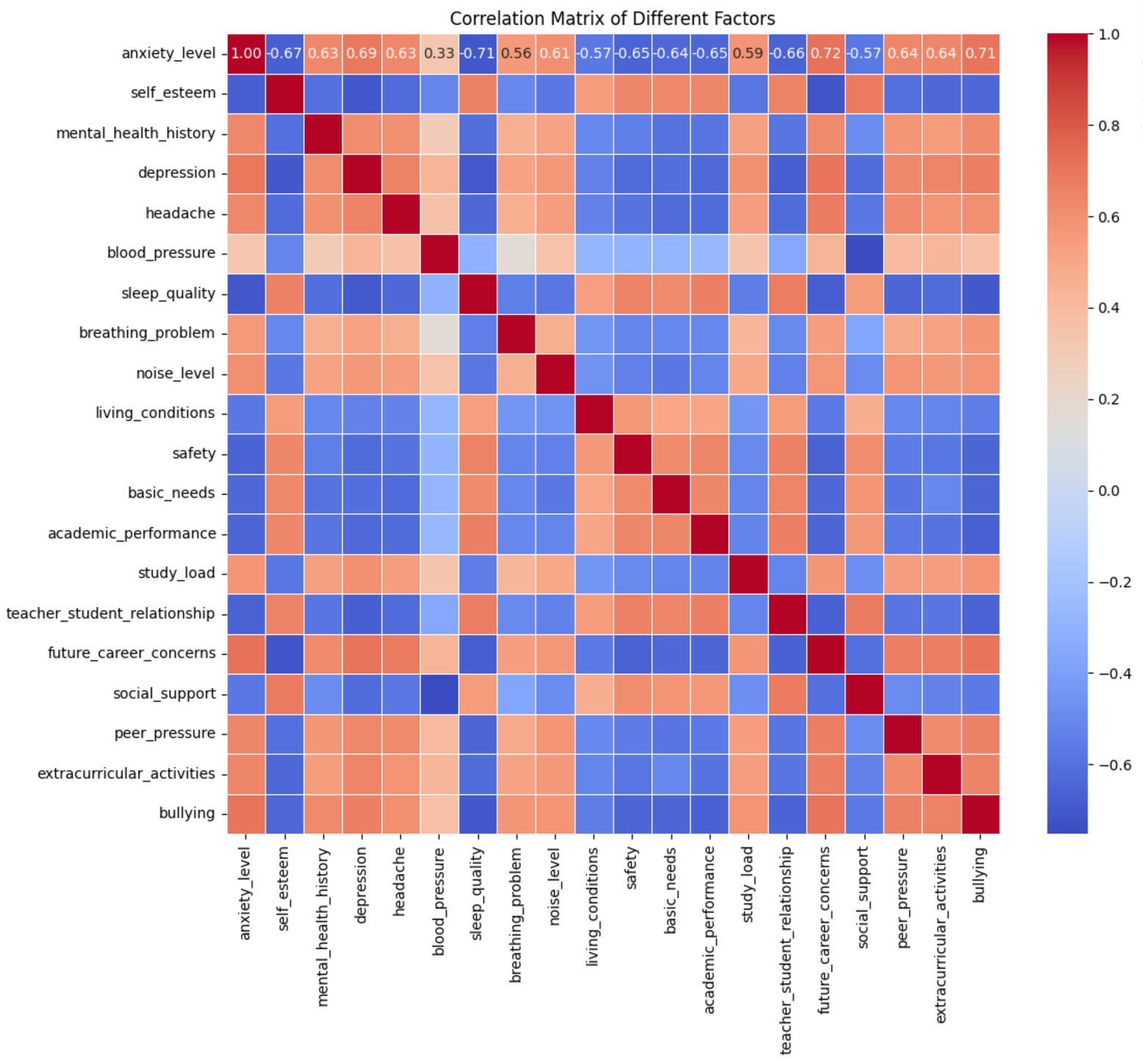


FEATURE DESCRIPTION

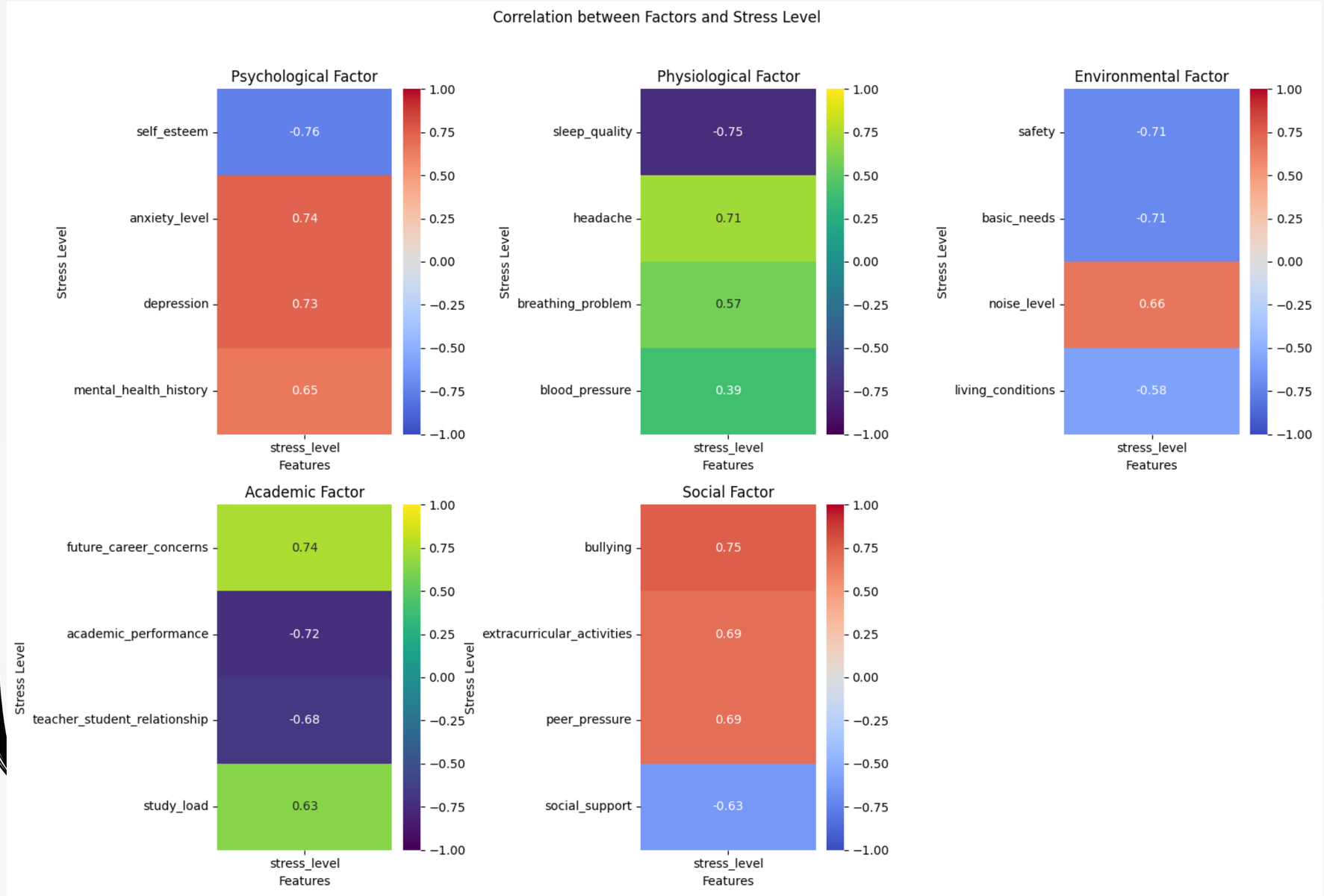
COMPARATIVE ANALYSIS



ANALYSIS

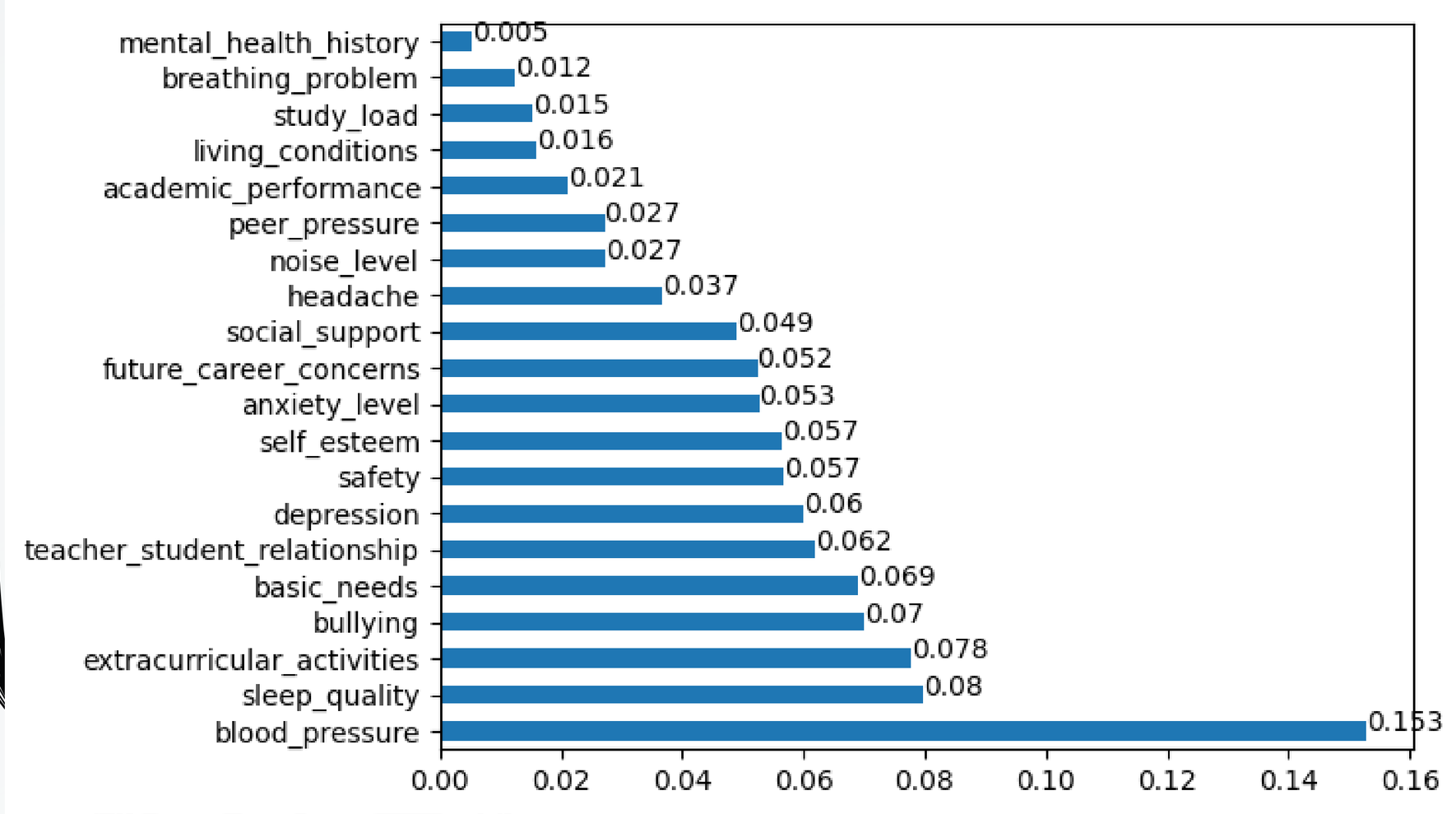

Correlation Matrix of Different Factors
anxiety_level
self_esteem
mental_health_history
depression
headache
blood_pressure
sleep_quality
breathing_problem
noise_level
living_conditions
safety
basic_needs
academic_performance
study_load
teacher_student_relationship
future_career_concerns
social_support
peer_pressure
extracurricular_activities
bullying
1.00 -0.67 0.63 0.69 0.63 0.33 -0.71 0.56 0.61 -0.57 -0.65 -0.64 -0.65 0.59 -0.66 0.72 -0.57 0.64 0.64 0.71
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

ANALYSIS


Correlation between Factors and Stress Level
Psychological Factor
self_esteem
anxiety_level
depression
mental_health_history
-0.76
0.74
0.73
0.65
Stress Level
stress_level
Features
Physiological Factor
sleep_quality
headache
breathing_problem
blood_pressure
-0.75
0.71
0.57
0.39
Environmental Factor
safety
basic_needs
noise_level
living_conditions
-0.71
-0.71
0.66
-0.58
Academic Factor
future_career_concerns
academic_performance
teacher_student_relationship
study_load
0.74
-0.72
-0.68
0.63
Social Factor
bullying
extracurricular_activities
peer_pressure
social_support
0.75
0.69
0.69
-0.63
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

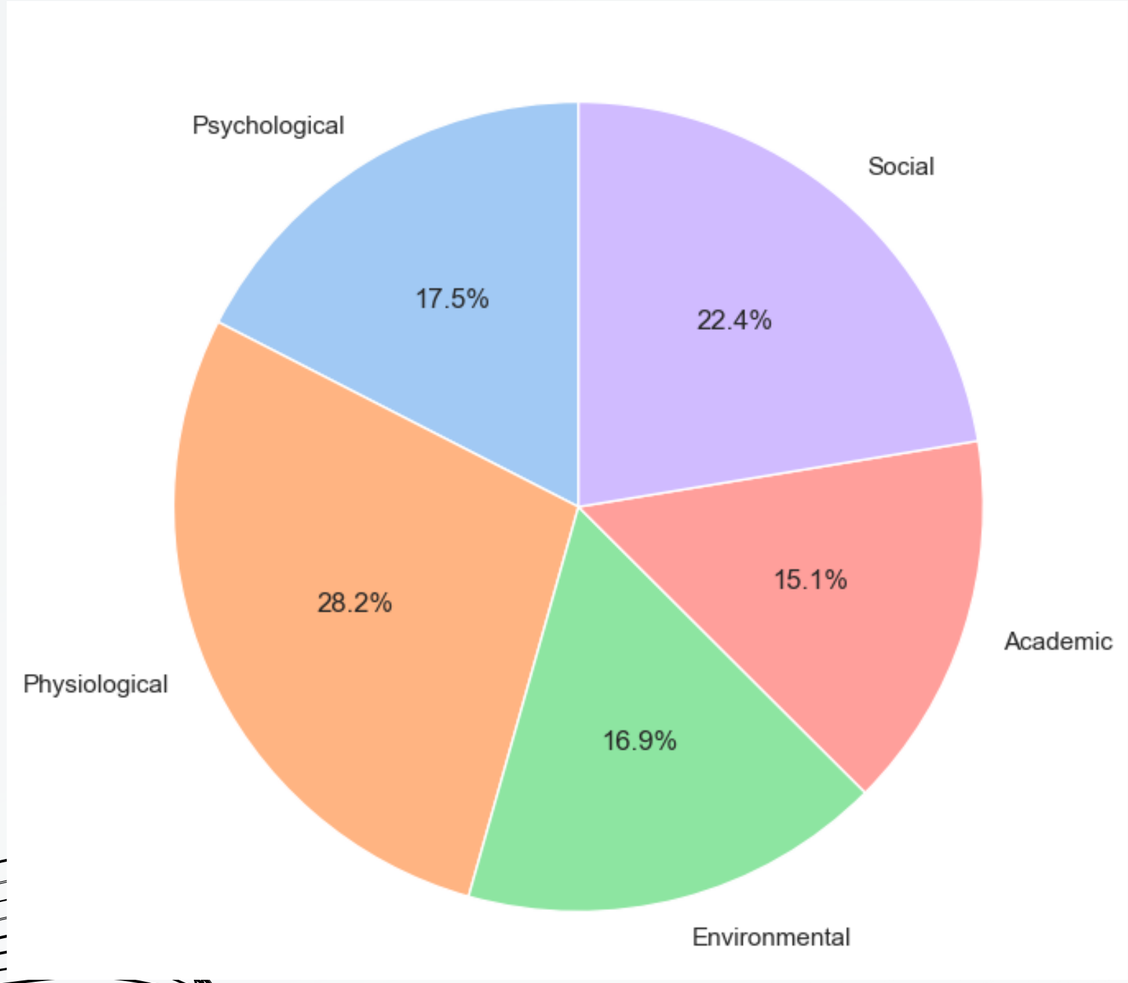
ANALYSIS

IMPORTANCE OF RANDOM FOREST



ANALYSIS

IMPORTANCE OF RANDOM FOREST



ANALYSIS

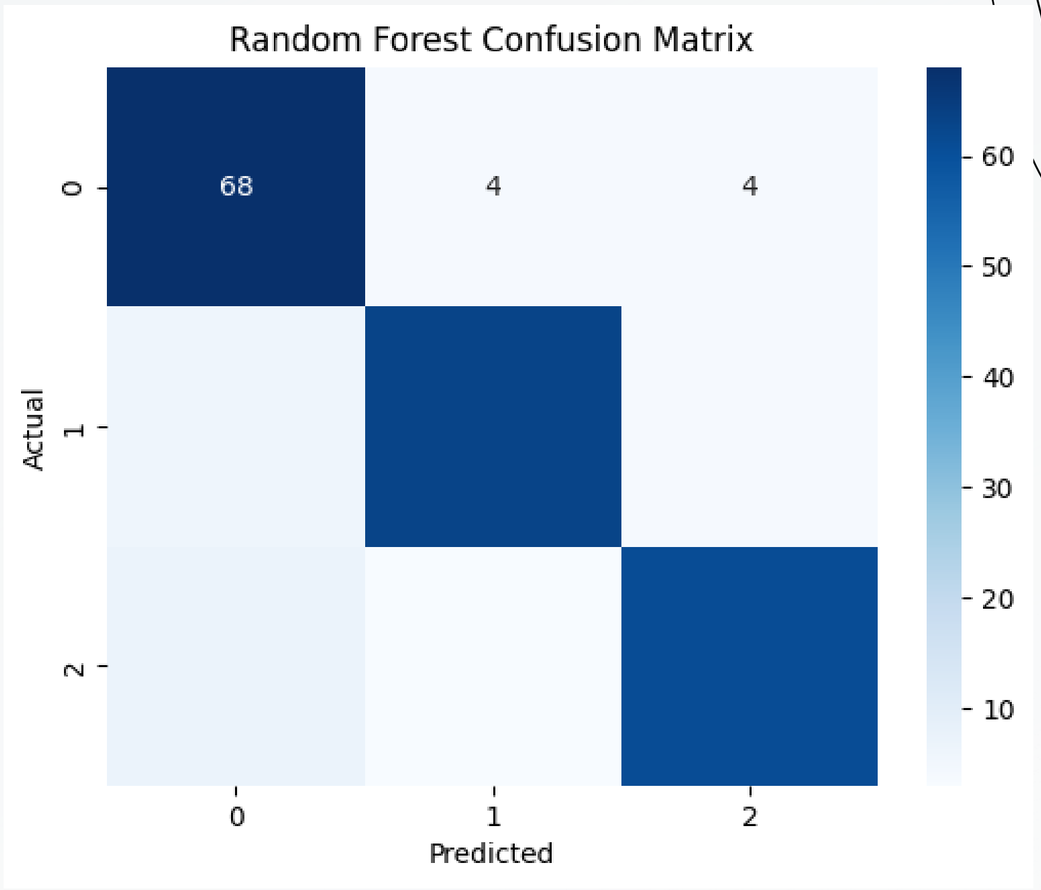
RANDOM FOREST RESULT

```
Model Accuracy: 0.8727272727272727
Confusion Matrix:
[[68  4  4]
 [ 6 63  4]
 [ 7  3 61]]
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.89      0.87        76
           1       0.90      0.86      0.88        73
           2       0.88      0.86      0.87        71

    accuracy                           0.87       220
   macro avg       0.87      0.87      0.87       220
weighted avg       0.87      0.87      0.87       220
```

ANALYSIS

RANDOM FOREST RESULT

0.87



ANALYSIS

REPO

DATASET

Dataset 1
https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
Dataset 2
https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis

THANKS