Off-task behaviors in the classroom

The project analyzes and addresses off-task behavior in classrooms to improve student engagement and overall learning outcomes.

Abstract

This study investigates the influence of distracted behavior in early childhood classrooms through an "off_tasks.csv" dataset with 22 columns and 27,731 rows. The study uses pandas profiling, statistics, machine learning, and data exploration methods to assess and remedy the adverse effects of these types of behavior on student performance and the classroom environment. These findings can be of great help to teachers, administrators, and policymakers in improving student engagement and the overall learning experience. The project promotes the use of data-driven teaching methodologies to improve classroom dynamics and learning outcomes, highlighting a multidisciplinary educational approach that incorporates technology and data analysis. Lastly, recommendations for improving the education system are presented based on clear and verifiable evidence.

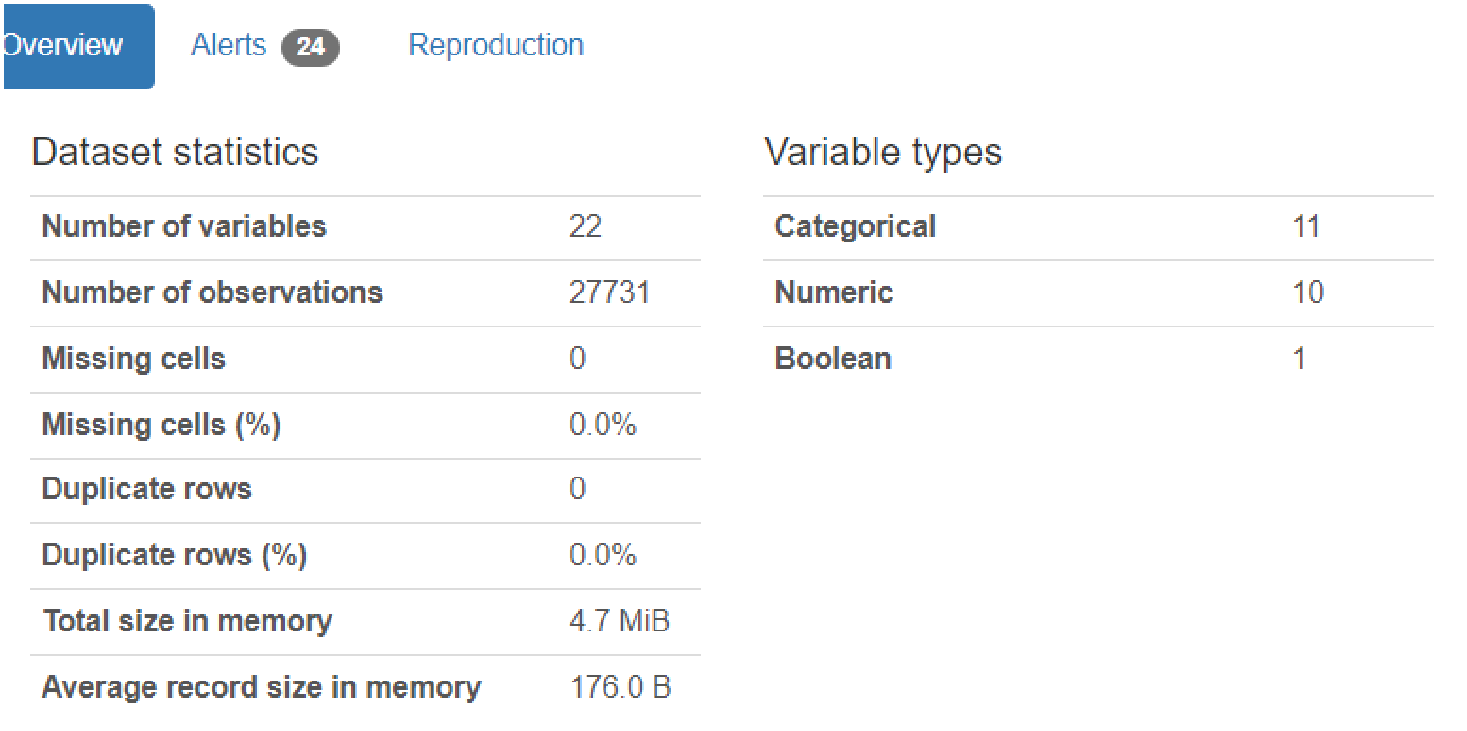
Motivation for the project

1. Recognizing that the issue of offhand conduct, which is a central concern affecting student performance and overall classroom standards, needs to be addressed by educators, administrators, and legislators.
2. To address the real-world dilemma faced by teachers and administrators, through insights and practical improvement measures designed to create positive changes in classroom interactions.
3. To provide evidence-based insights for more successful educational interventions, data technologies such as machine learning, statistical analysis, and exploratory data analysis are used.
4. Taking the lessons learned from this project and applying them across different educational settings, thus contributing to a broader field of education.
5. We want to strengthen the involvement of students, improve learning results, and benefit a wider student population by reducing off-task behavior.
6. By using an interdisciplinary approach that blends statistical, machine learning, and exploratory methodologies to reflect the changing role of technology and data in education.
7. To provide teachers, school management, and legislators with realistic data-based recommendations in support of continuous learning progress.

Exploratory Data Analysis

The EDA phase will most likely include analyzing the dataset to find patterns, anomalies, trends, and correlations. This can include statistical summaries, data visualization, and other tools for understanding the features of the data.

Overview of the Dataset



Dataset statistics

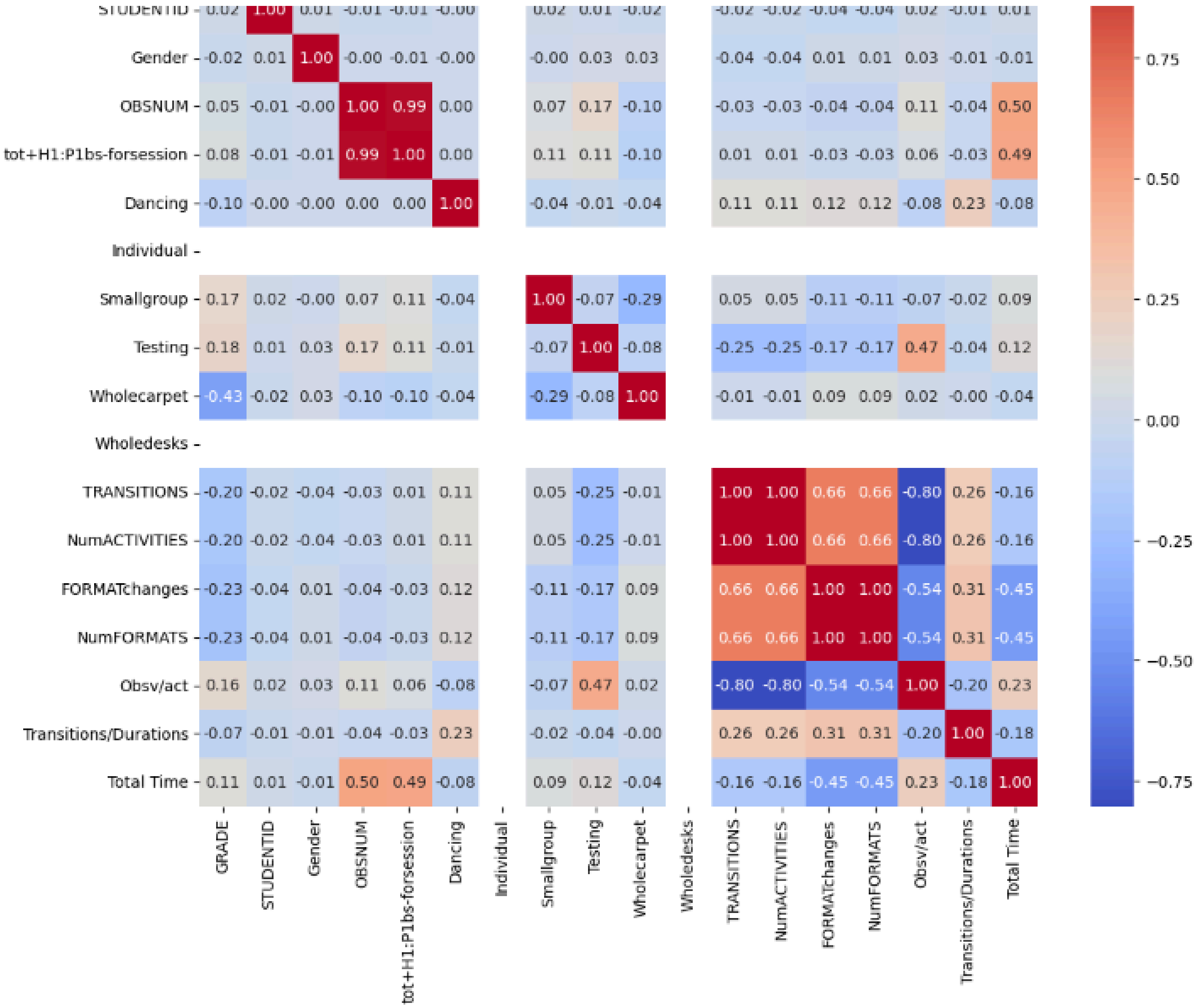
Number of variables	22
Number of observations	27731
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	4.7 MiB
Average record size in memory	176.0 B

Variable types

Categorical	11
Numeric	10
Boolean	1

We have used the Pandas Profiling Tool to get an overview of the Dataset. We can see that No of Variables, No of Observations, and Types of Variables like Categorical and numerical are present in the dataset.

Use of Correlation heatmap for the Data Cleaning Process



	GRADE	STUDENTID	Gender	OBSNUM	tot+H1:P1bs-forsession	Dancing	Individual	Smallgroup	Testing	Wholecarpet	Wholedesks	TRANSITIONS	NumACTIVITIES	FORMATchanges	NumFORMATS	Obsv/act	Transitions/Durations	Total Time
STUDENTID	0.02	1.00	0.01	-0.01	-0.01	-0.00		0.02	0.01	-0.02		-0.02	-0.02	-0.04	-0.04	0.02	-0.01	0.01
Gender	-0.02	0.01	1.00	-0.00	-0.01	-0.00		-0.00	0.03	0.03		-0.04	-0.04	0.01	0.01	0.03	-0.01	-0.01
OBSNUM	0.05	-0.01	-0.00	1.00	0.99	0.00		0.07	0.17	-0.10		-0.03	-0.03	-0.04	-0.04	0.11	-0.04	0.50
tot+H1:P1bs-forsession	0.08	-0.01	-0.01	0.99	1.00	0.00		0.11	0.11	-0.10		0.01	0.01	-0.03	-0.03	0.06	-0.03	0.49
Dancing	-0.10	-0.00	-0.00	0.00	0.00	1.00		-0.04	-0.01	-0.04		0.11	0.11	0.12	0.12	-0.08	0.23	-0.08
Individual																		
Smallgroup	0.17	0.02	-0.00	0.07	0.11	-0.04		1.00	-0.07	-0.29		0.05	0.05	-0.11	-0.11	-0.07	-0.02	0.09
Testing	0.18	0.01	0.03	0.17	0.11	-0.01		-0.07	1.00	-0.08		-0.25	-0.25	-0.17	-0.17	0.47	-0.04	0.12
Wholecarpet	-0.43	-0.02	0.03	-0.10	-0.10	-0.04		-0.29	-0.08	1.00		-0.01	-0.01	0.09	0.09	0.02	-0.00	-0.04
Wholedesks																		
TRANSITIONS	-0.20	-0.02	-0.04	-0.03	0.01	0.11		0.05	-0.25	-0.01		1.00	1.00	0.66	0.66	-0.80	0.26	-0.16
NumACTIVITIES	-0.20	-0.02	-0.04	-0.03	0.01	0.11		0.05	-0.25	-0.01		1.00	1.00	0.66	0.66	-0.80	0.26	-0.16
FORMATchanges	-0.23	-0.04	0.01	-0.04	-0.03	0.12		-0.11	-0.17	0.09		0.66	0.66	1.00	1.00	-0.54	0.31	-0.45
NumFORMATS	-0.23	-0.04	0.01	-0.04	-0.03	0.12		-0.11	-0.17	0.09		0.66	0.66	1.00	1.00	-0.54	0.31	-0.45
Obsv/act	0.16	0.02	0.03	0.11	0.06	-0.08		-0.07	0.47	0.02		-0.80	-0.80	-0.54	-0.54	1.00	-0.20	0.23
Transitions/Durations	-0.07	-0.01	-0.01	-0.04	-0.03	0.23		-0.02	-0.04	-0.00		0.26	0.26	0.31	0.31	-0.20	1.00	-0.18
Total Time	0.11	0.01	-0.01	0.50	0.49	-0.08		0.09	0.12	-0.04		-0.16	-0.16	-0.45	-0.45	0.23	-0.18	1.00

- Notably, the utilization of full group arrangements such as rugs and desks reduces as grade level increases. The frequency of transitions and the number of activities show a significant inverse relationship, indicating more focused participation.
- Student identity and gender were found to have little to no correlation with classroom activities, indicating that other characteristics are more important in off-task behavior patterns.
- The link between small group work and entire carpet activities suggests an instructional strategy trade-off. These findings can assist in informing attempts to improve learning environments by considering how different classroom dynamics influence student engagement.

Key Findings in the Project

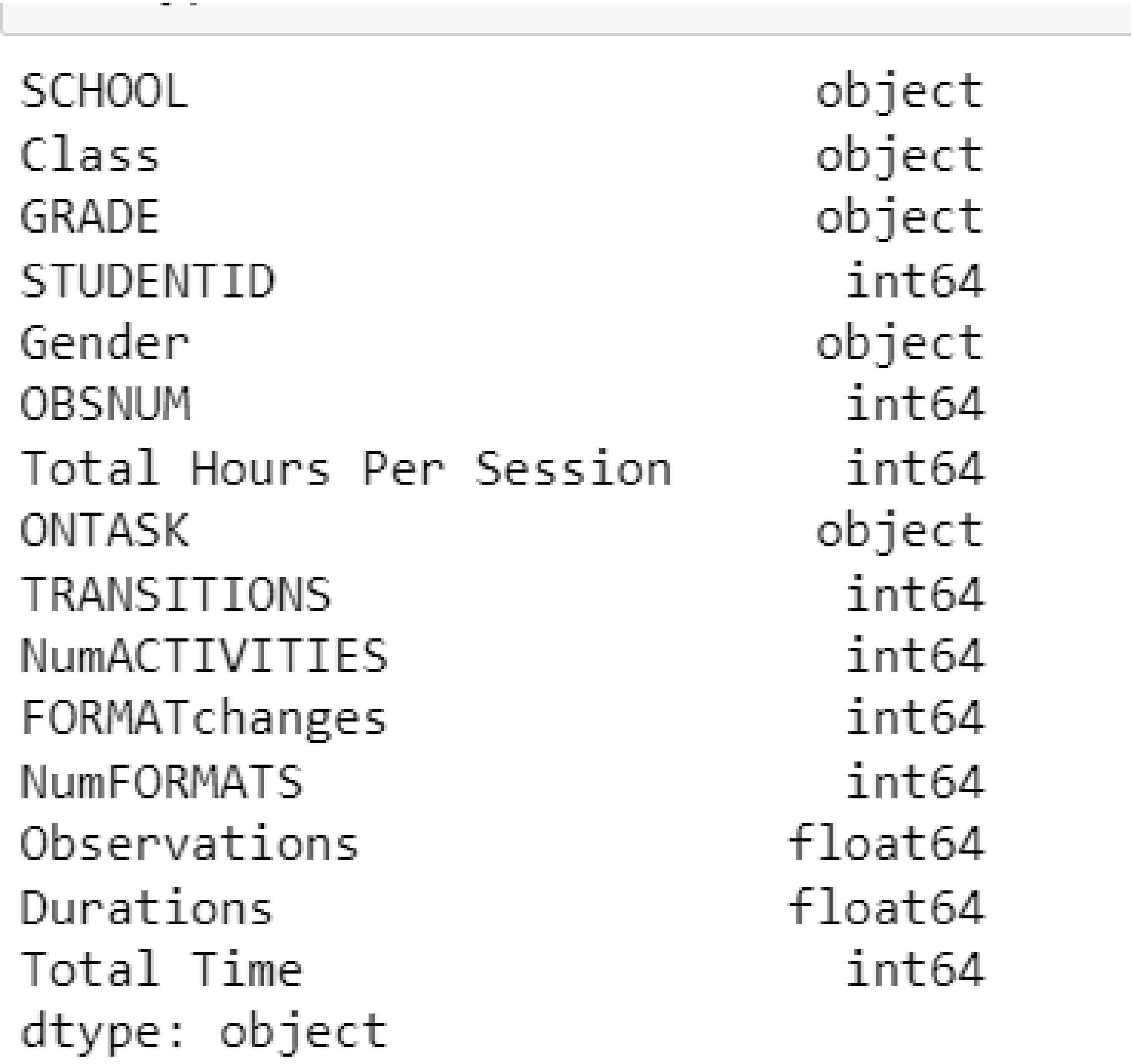
•We have an iterative data cleaning process for this dataset where one of the key findings is removing outliers present in the dataset.

•After finding out the outliers we were able to reduce the row count of the dataset. We have also apple shape condition to get an depth overview of the dataset.

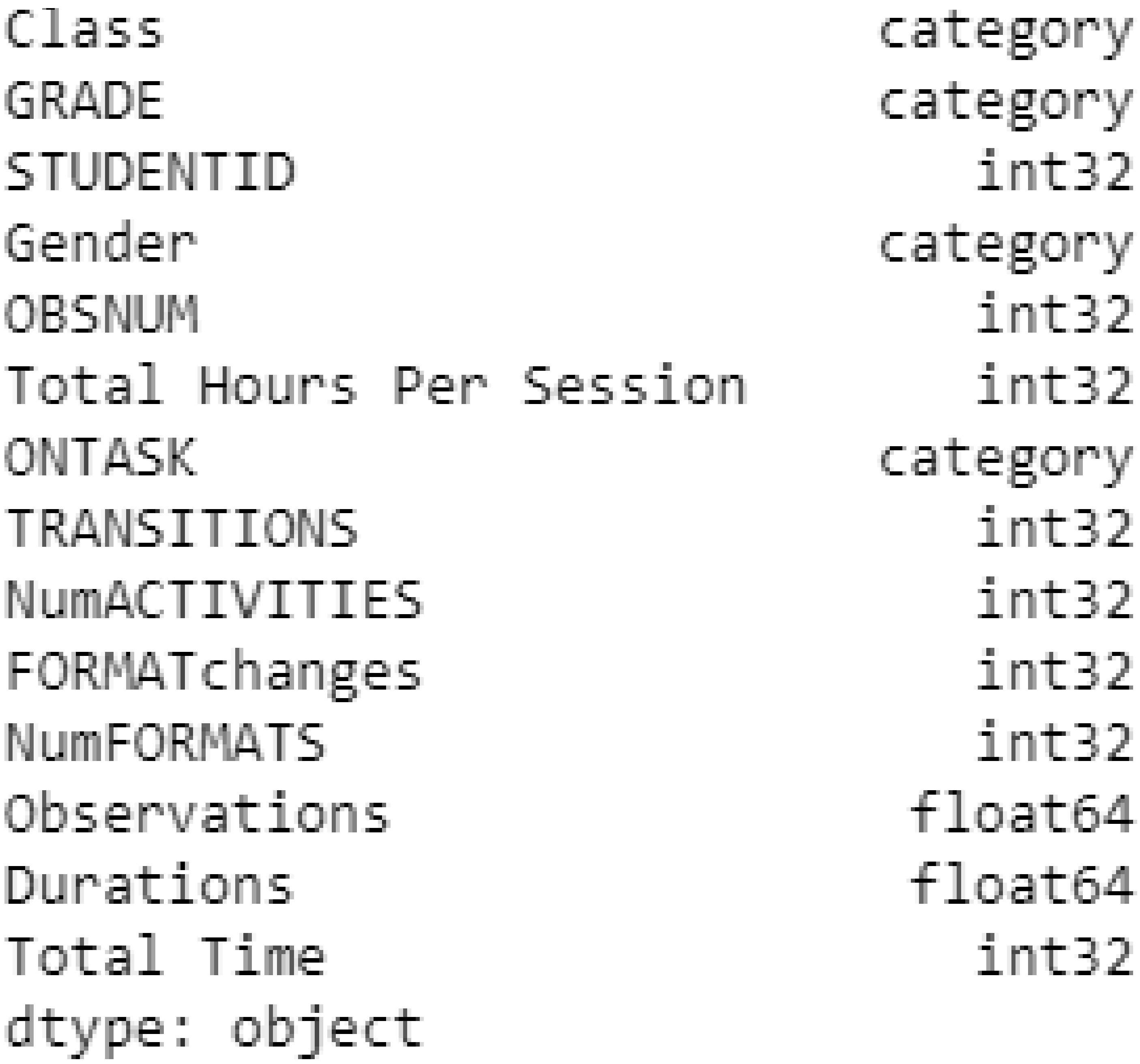
•We have reduced the shape of the dataset from where the shape of the dataset is reduced to 27271 rows x 15 columns

•We have also downcasted the datatypes of the columns

Down Casting of the Datatypes



```
SCHOOL                        object
Class                         object
GRADE                         object
STUDENTID                      int64
Gender                        object
OBSNUM                         int64
Total Hours Per Session        int64
ONTASK                        object
TRANSITIONS                    int64
NumACTIVITIES                  int64
FORMATchanges                  int64
NumFORMATS                     int64
Observations                 float64
Durations                    float64
Total Time                     int64
dtype: object
```

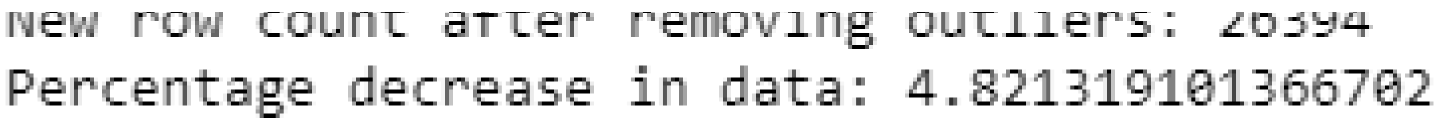


```
Class                         category
GRADE                         category
STUDENTID                        int32
Gender                        category
OBSNUM                           int32
Total Hours Per Session          int32
ONTASK                        category
TRANSITIONS                      int32
NumACTIVITIES                    int32
FORMATchanges                    int32
NumFORMATS                       int32
Observations                   float64
Durations                      float64
Total Time                       int32
dtype: object
```

Here we can see that we have down casted datatypes from int 64 to int32 for the columns that are present in the dataset.

We also dropped unnecessary columns present in the dataset and renamed the columns for the easier finding for the future purposes of the dataset.

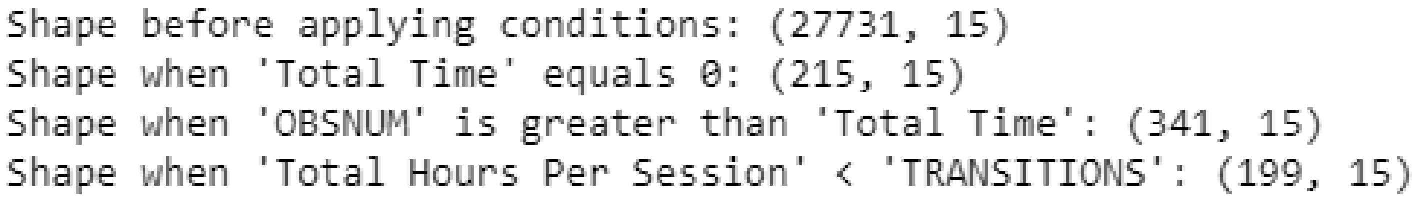
Finding out the Percentage decrease in data



```
New row count after removing outliers: 26394
Percentage decrease in data: 4.821319101366702
```

The revised row count after cleaning the dataset by removing outliers is 26,394, down from the original 27,731. This reflects a 4.82% reduction in data volume, showing a little drop in volume to ensure more accurate analysis by removing aberrant numbers.

Using Shape Conditions to find the Outliers



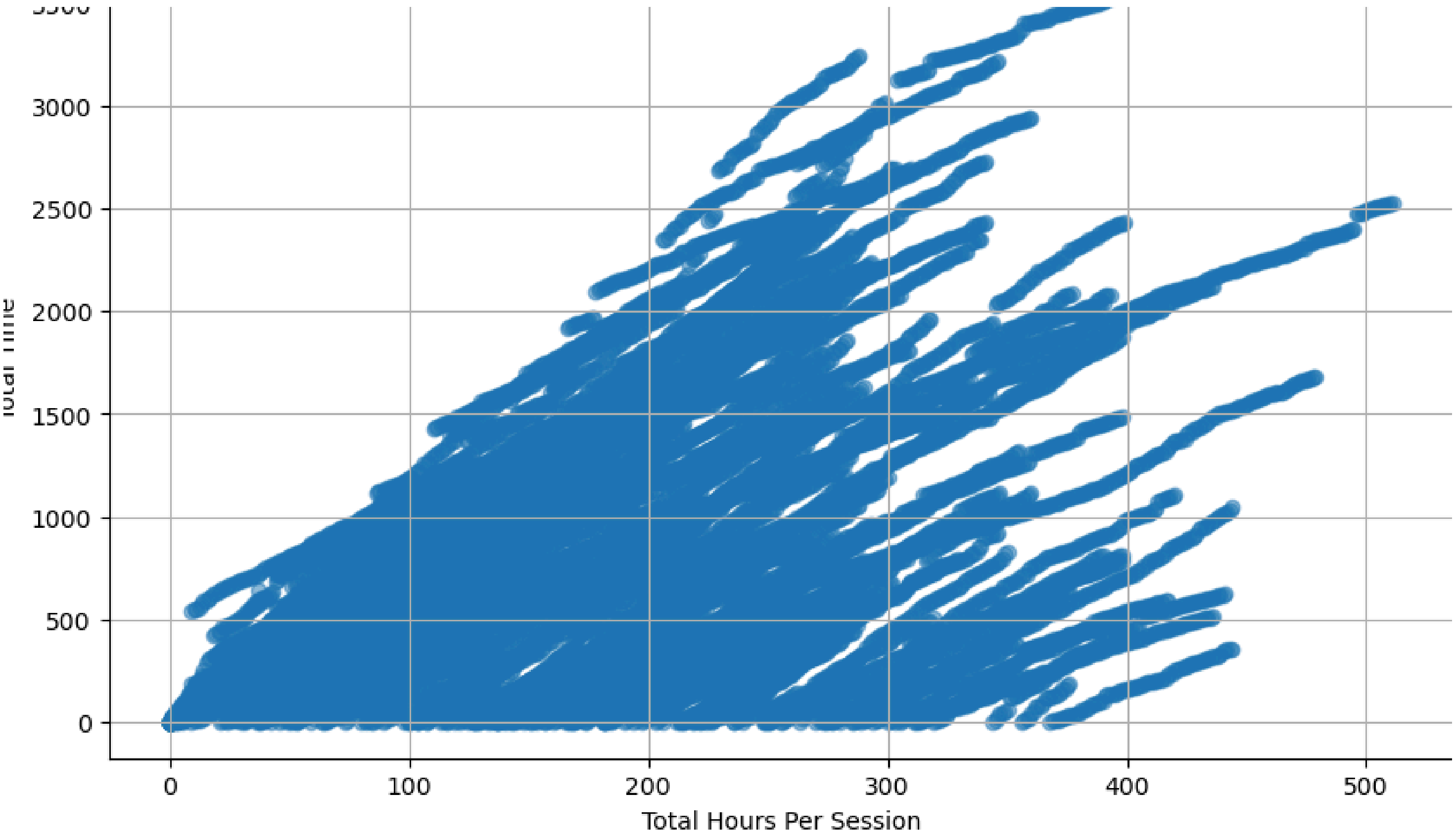
```
Shape before applying conditions: (27731, 15)
Shape when 'Total Time' equals 0: (215, 15)
Shape when 'OBSNUM' is greater than 'Total Time': (341, 15)
Shape when 'Total Hours Per Session' < 'TRANSITIONS': (199, 15)
```

The provided data illustrates how the shape of a dataset changes before and after specific criteria are applied:

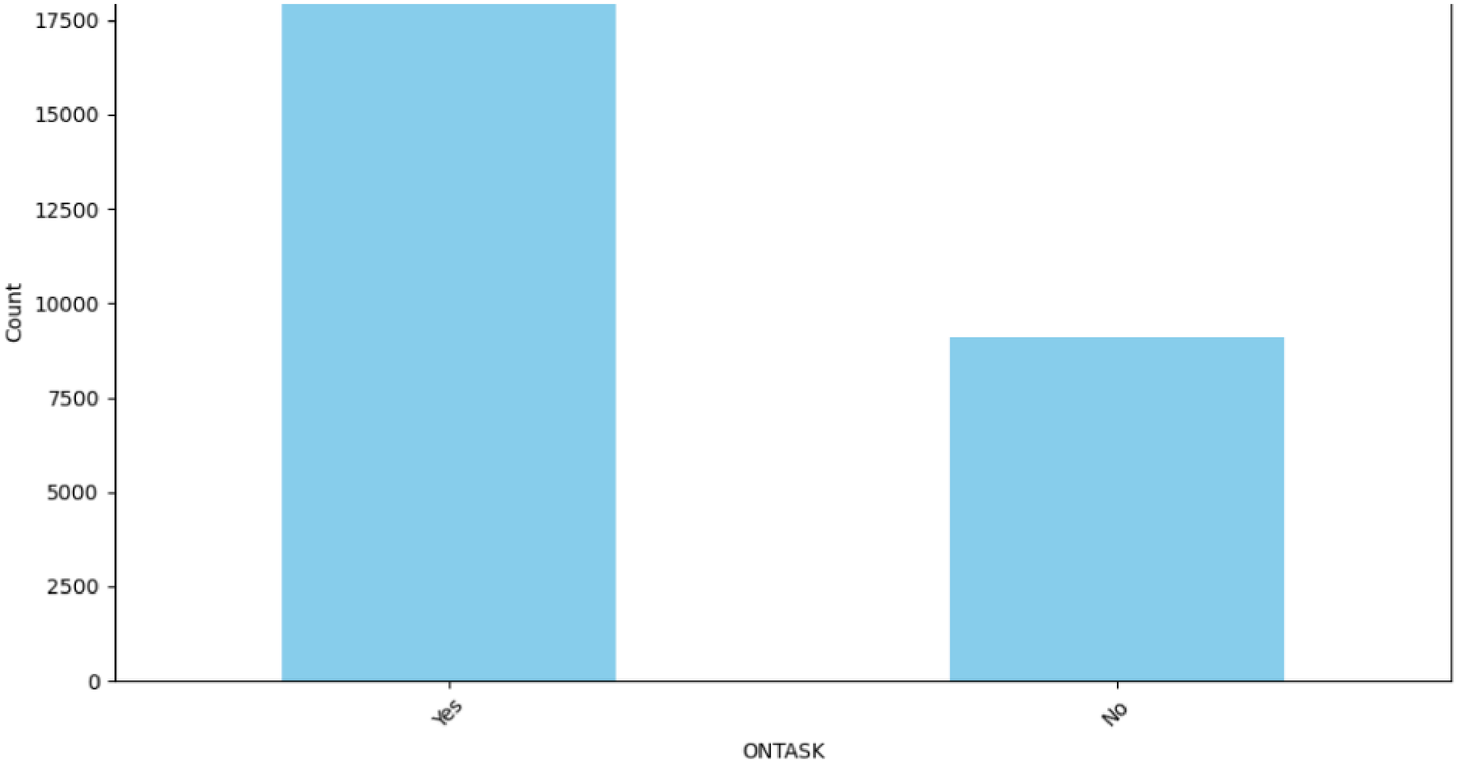
- The dataset began with 27,731 rows and 15 columns.
- The dataset was reduced to 215 rows after filtering for entries where 'Total Time' equals 0, suggesting that there were 215 occurrences where 'Total Time' was reported as zero.
- The dataset contains 341 rows when the 'OBSNUM' (observation number) was larger than the 'entire Time', indicating that the observation number was greater than the entire time recorded in 341 occurrences.
- Finally, by filtering for cases where 'Total Hours Per Session' was less than 'TRANSITIONS,' the dataset was reduced to 199 rows, indicating that there were 199 occurrences where the total hours per session were less than the total hours per session.

Visualization of the Data

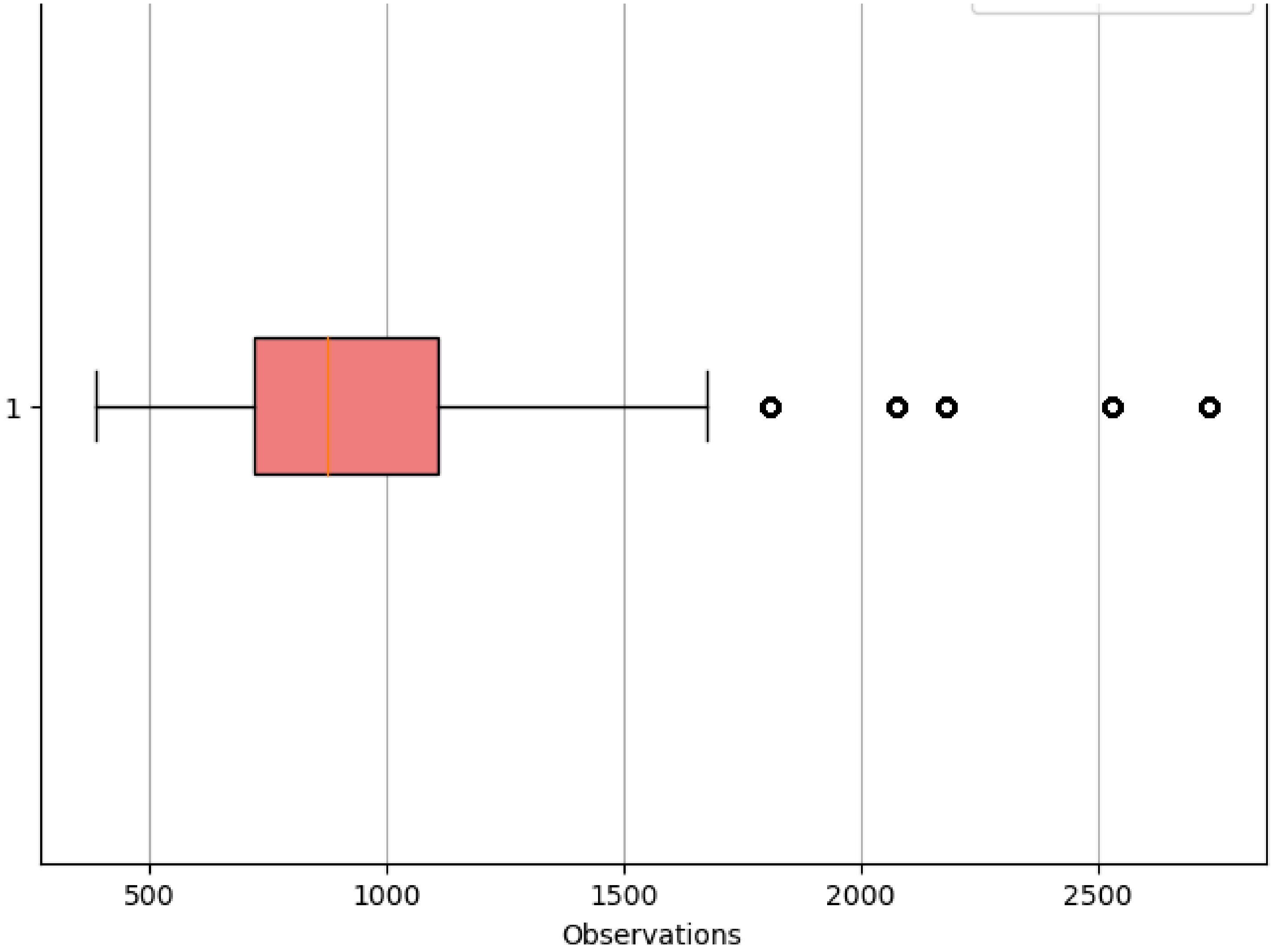
Several visualizations are used in the notebook to explain the dataset's properties and relationships. Line plots are used to track continuous data trends, and bar plots are used to compare categorical variables. Histograms describe data distributions, whereas box plots summarize this distribution with a focus on outliers. Scatter plots investigate variable correlations, pie charts depict categorical data proportions, and heatmaps depict data intensity, as in correlation matrices. Violin plots provide detailed data distributions, whereas pair plots investigate correlations across many variables, providing a multifaceted view of classroom dynamics related to off-task behavior.



The scatter plot depicts the association between 'Total Hours Per Session' and ' Total Time'. The data points are scattered in distinct bands, indicating that the data may contain discrete intervals or sessions. It implies a positive relationship between the two variables; as the number of hours each session grows, so does the total time. The banding pattern could indicate that session lengths are fixed or that data was collected in intervals. The dense clustering of points at lower hours per session shows a high frequency of shorter sessions, whereas the spread of points as sessions lengthen suggests unpredictability in the total duration associated with these longer sessions. The graphic emphasizes the structured form of the time-related data in the dataset, which may indicate the scheduling.

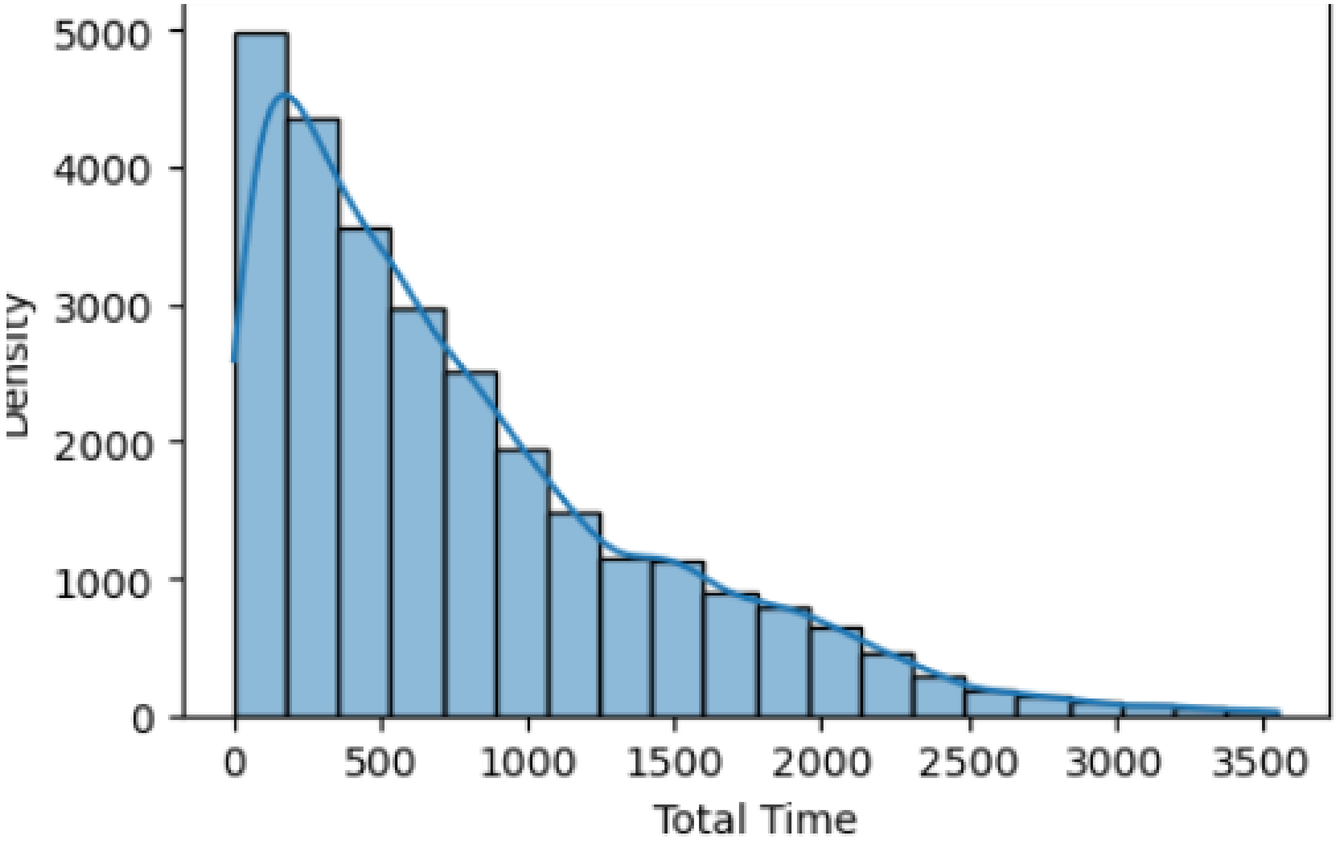


The bar chart depicts the distribution of on-task versus off-task conduct in a classroom setting, demonstrating a strong preference for on-task behavior. The graphic shows a good trend in classroom participation, with on-task instances greatly outnumbering off-task cases. However, the existence of significant off-task behavior indicates areas for potential improvement in sustaining student focus. This visual data can help educators identify tactics to boost students' on-task behavior even further.

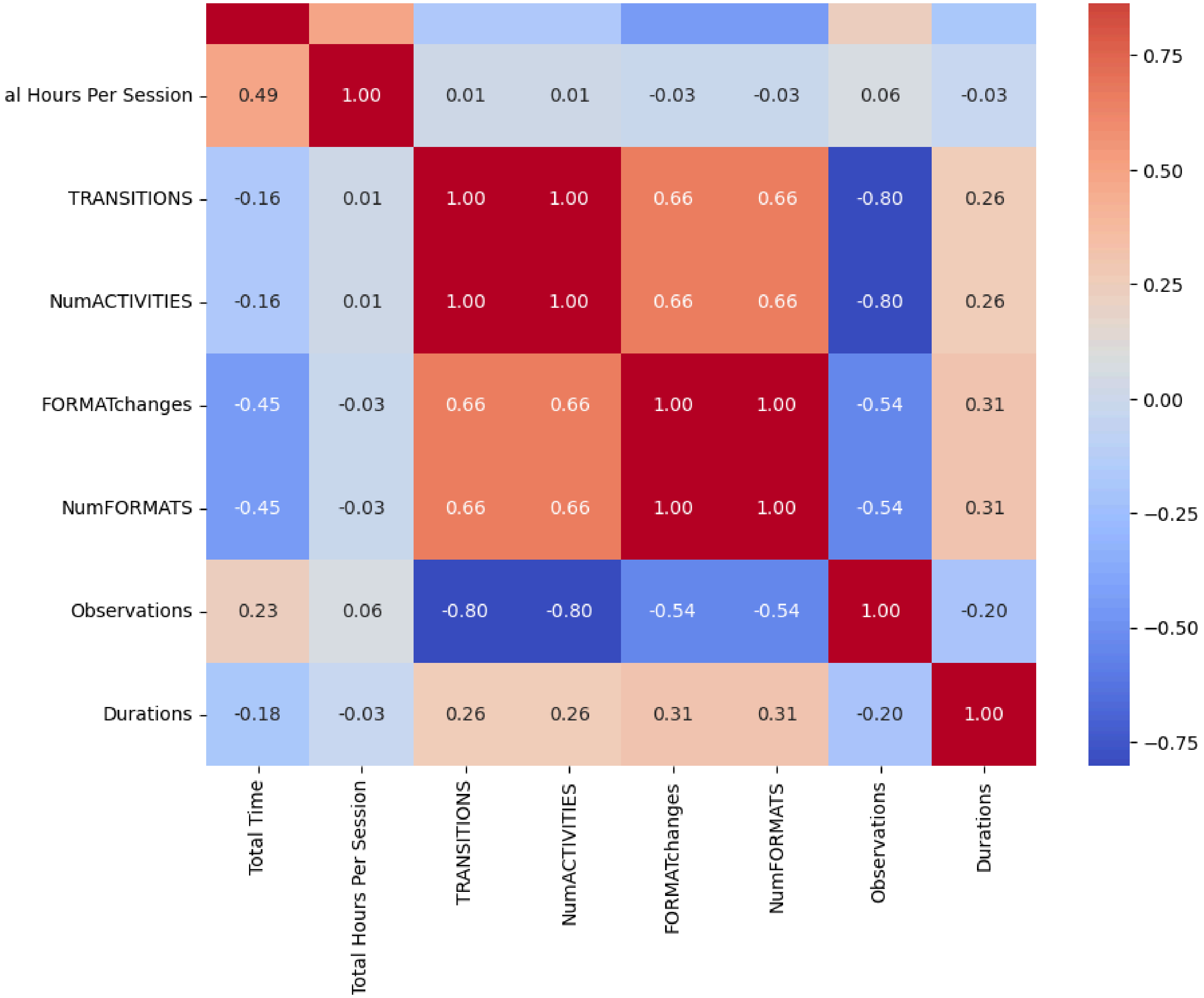


The box plot depicts the distribution of observations, with the central box indicating the interquartile range and the line within the box representing the median.

This indicates a roughly symmetrical distribution of observation counts, with a few outliers showing sessions with much higher observation counts than expected. The distribution of data points and outliers reveals information about the variability of the number of observations across different sessions or classes.

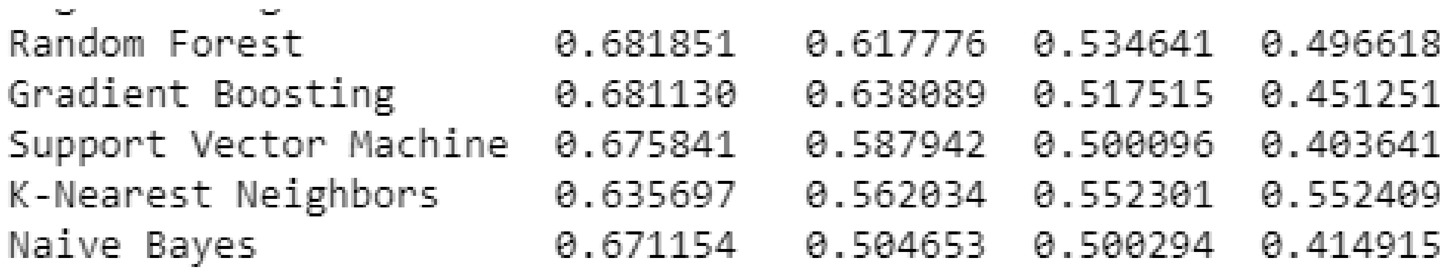


The histogram depicts the 'Total Time' distribution, with the majority of values concentrated towards the lower end of the scale, showing that shorter periods are more common in the dataset. The distribution's tail stretches to the right, indicating that there are fewer cases of extremely long total times. The histogram's form, combined with the superimposed curve, reveals that the data is right-skewed, implying that while most recorded times are brief, there is a large tail of less frequent extended times. This could be due to variations in session durations or observed activities.



The heatmap depicts correlations between classroom activity data. 'Total Time' marginally correlates with 'Total Hours Per Session' but adversely with activity categories, demonstrating a complex link between session length and activity diversification. Perfect correlations between 'TRANSITIONS' and 'NumACTIVITIES' indicate a direct relationship between the number of transitions and activities. Strong negative correlations between 'Observations' and transitions and activities suggest fewer observations in more dynamic contexts. These patterns may reflect underlying classroom dynamics, providing insights into how various facets of classroom management correlate with one another.

Model Development



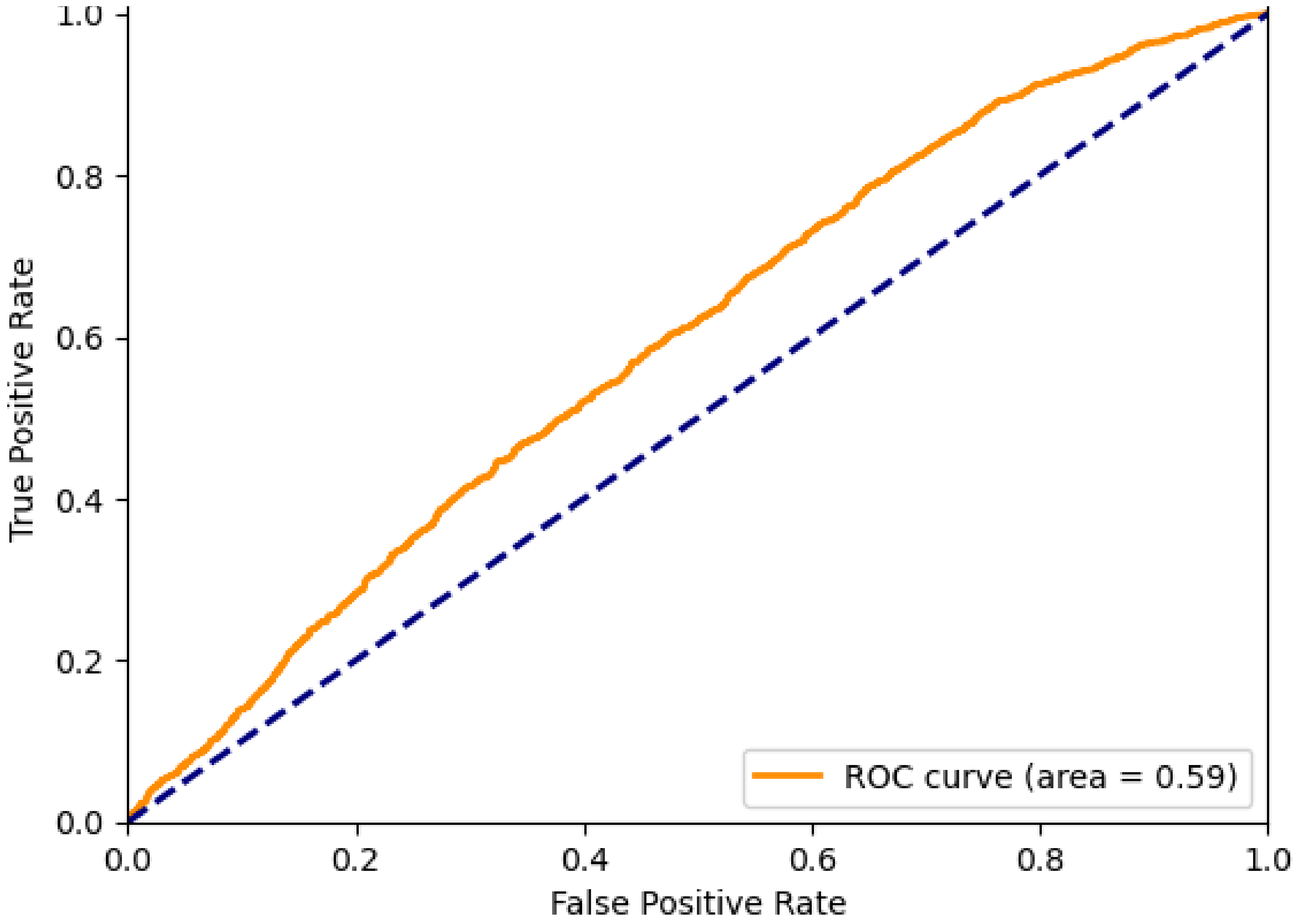
```
Random Forest           0.681851   0.617776  0.534641  0.496618
Gradient Boosting       0.681130   0.638089  0.517515  0.451251
Support Vector Machine  0.675841   0.587942  0.500096  0.403641
K-Nearest Neighbors     0.635697   0.562034  0.552301  0.552409
Naive Bayes             0.671154   0.504653  0.500294  0.414915
```

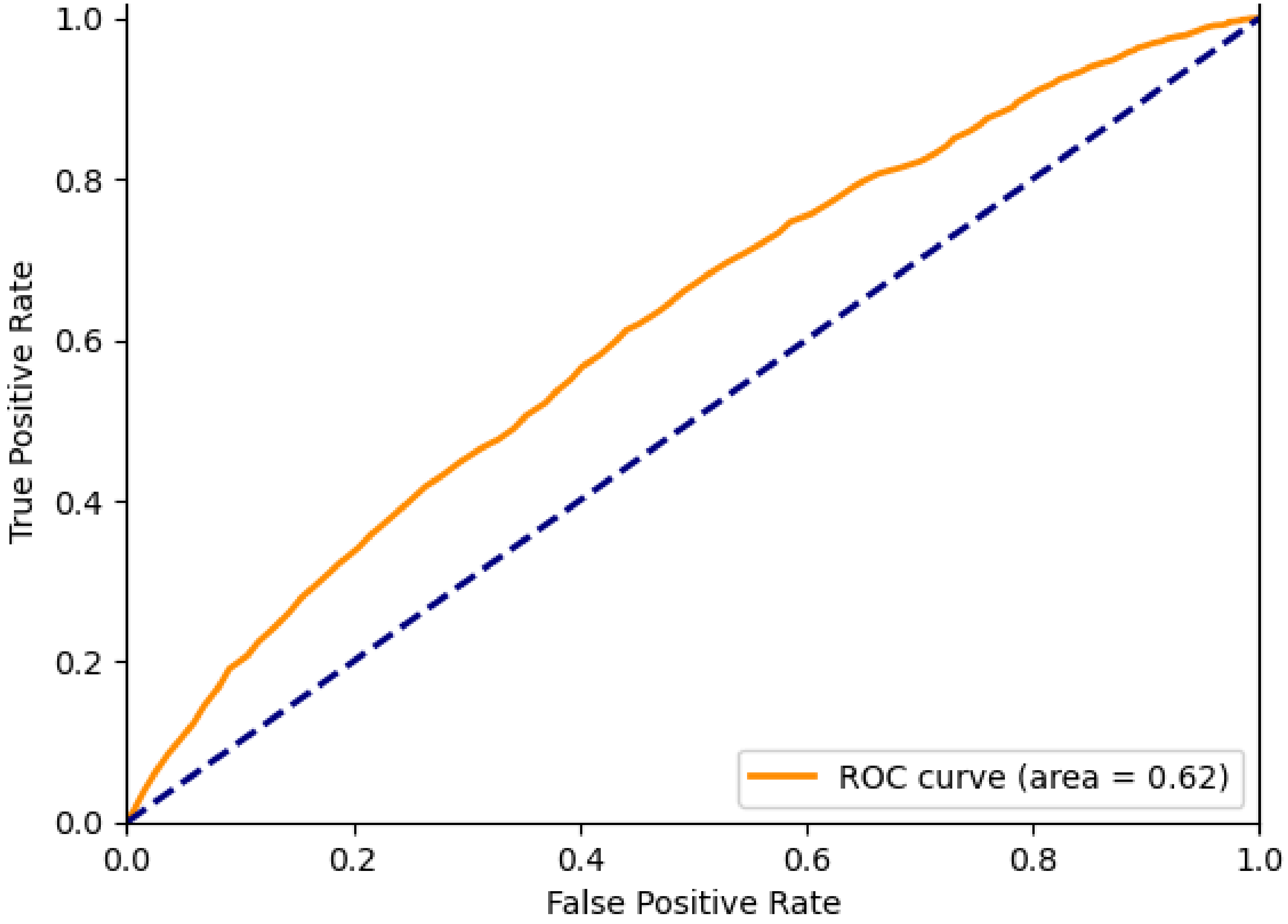
Six machine learning models for predicting 'ONTASK' behavior from classroom data are evaluated in the report. Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, K-nearest neighbors, and Naive Bayes are among the models. Preprocessing included standardizing numerical features and encoding category features in a single pass. Using a test set containing 30% of the data, the models were evaluated on accuracy, precision, recall, and F1 score measures. Random Forest had the highest accuracy (68.19%) and a solid balance of precision and recall, indicating an excellent overall prediction capability. Gradient Boosting had the highest precision (63.81%), however, K-Nearest Neighbors had the highest recall and F1 score, indicating that it was better at recognizing all relevant examples of the target variable. The accuracy of the Logistic Regression and Support Vector Machine was the same (67.58%). However, their precision and recall indicate that they may not be able to balance the two characteristics as effectively as Random Forest. While somewhat less accurate (67.12%), Naive Bayes displayed moderate precision and recall, showing its potential utility with additional tuning. According to the findings, ensemble approaches such as Random Forest and Gradient Boosting may be better suited for this prediction task.

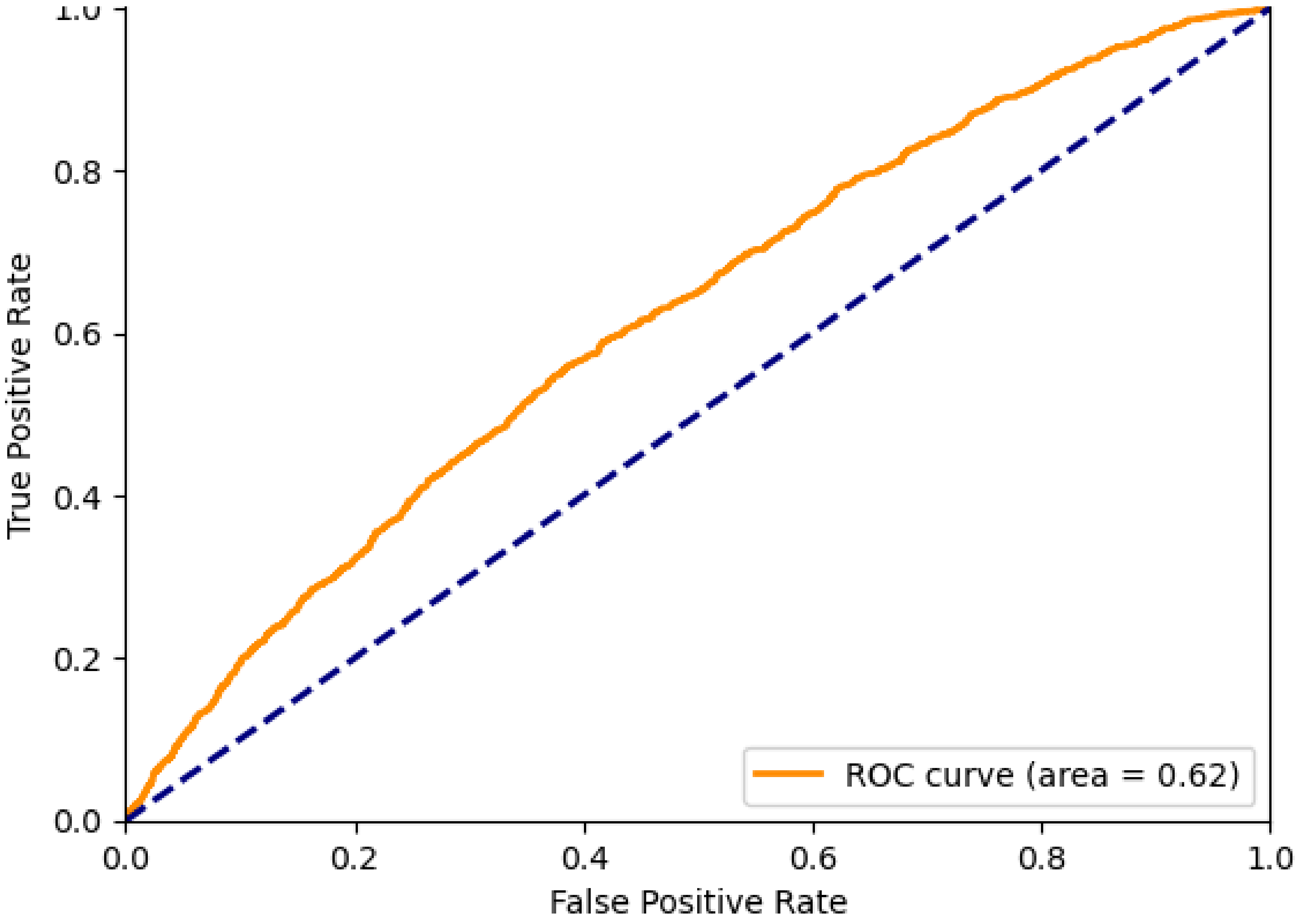
Interpreting the Models

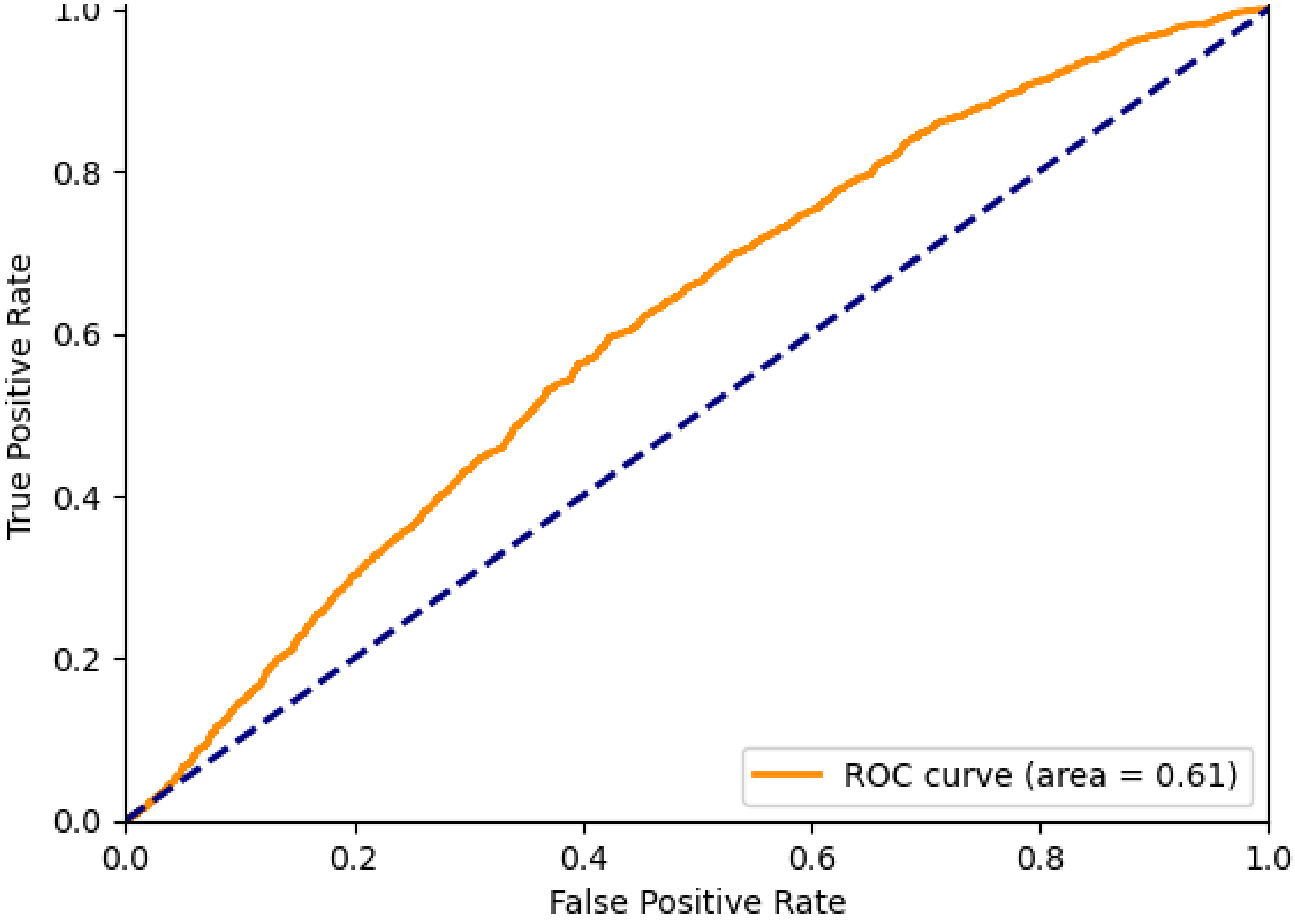
The image's ROC curves evaluate the diagnostic ability of multiple classifiers used to forecast a binary target. The AUCs for these models vary from 0.59 to 0.62, demonstrating modest predictive performance, with Gradient Boosting surpassing the others somewhat. The curves are above but close to the no-skill line, indicating that while the models can discriminate between classes better than random guessing, their effectiveness is limited, underlining the possible need for additional model tuning or more complicated modeling techniques.

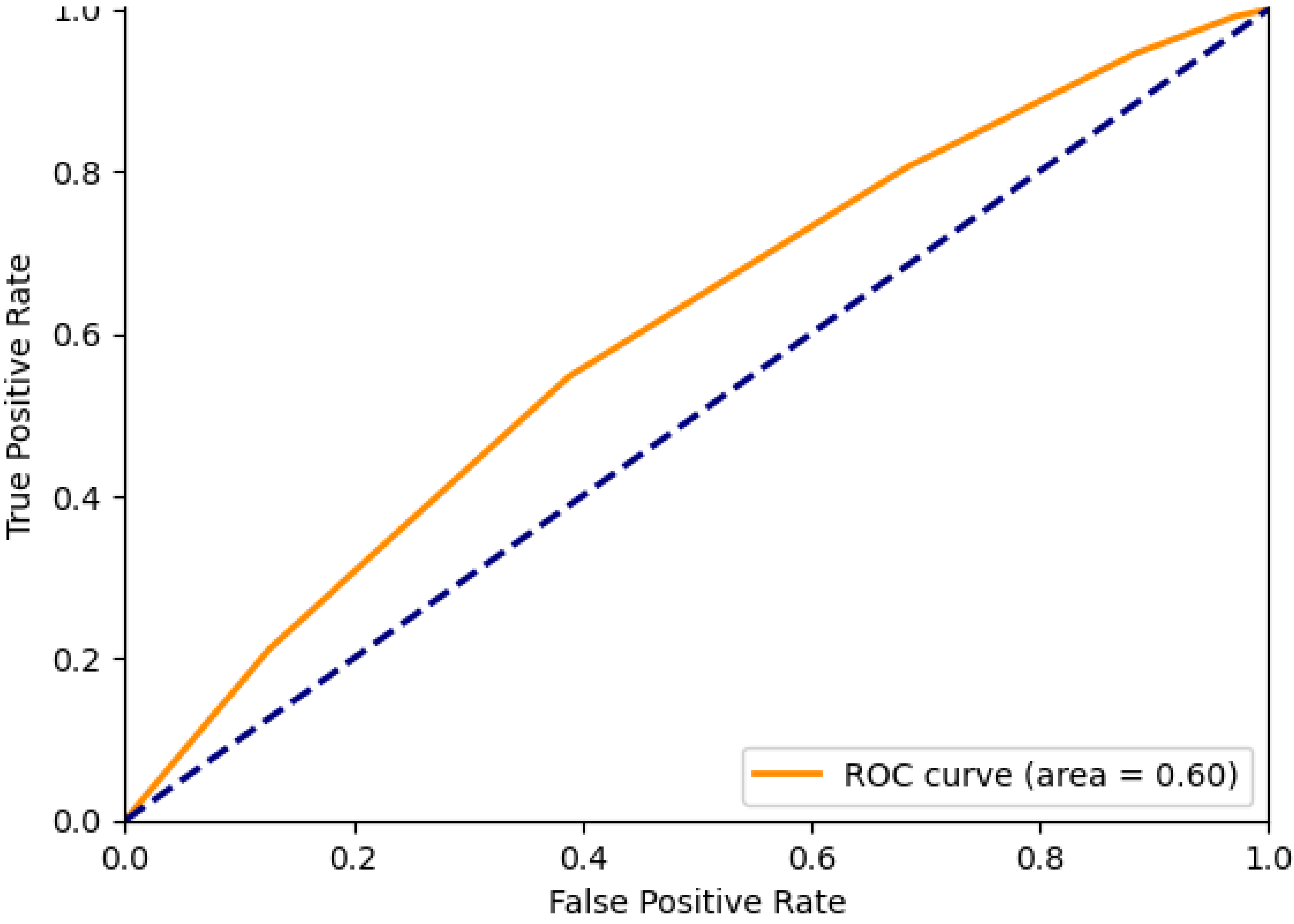
We have found the ROC Curve and AUC Score in the following Modeling Function to find out which model can be used for further analysis. Here are the attached snippets

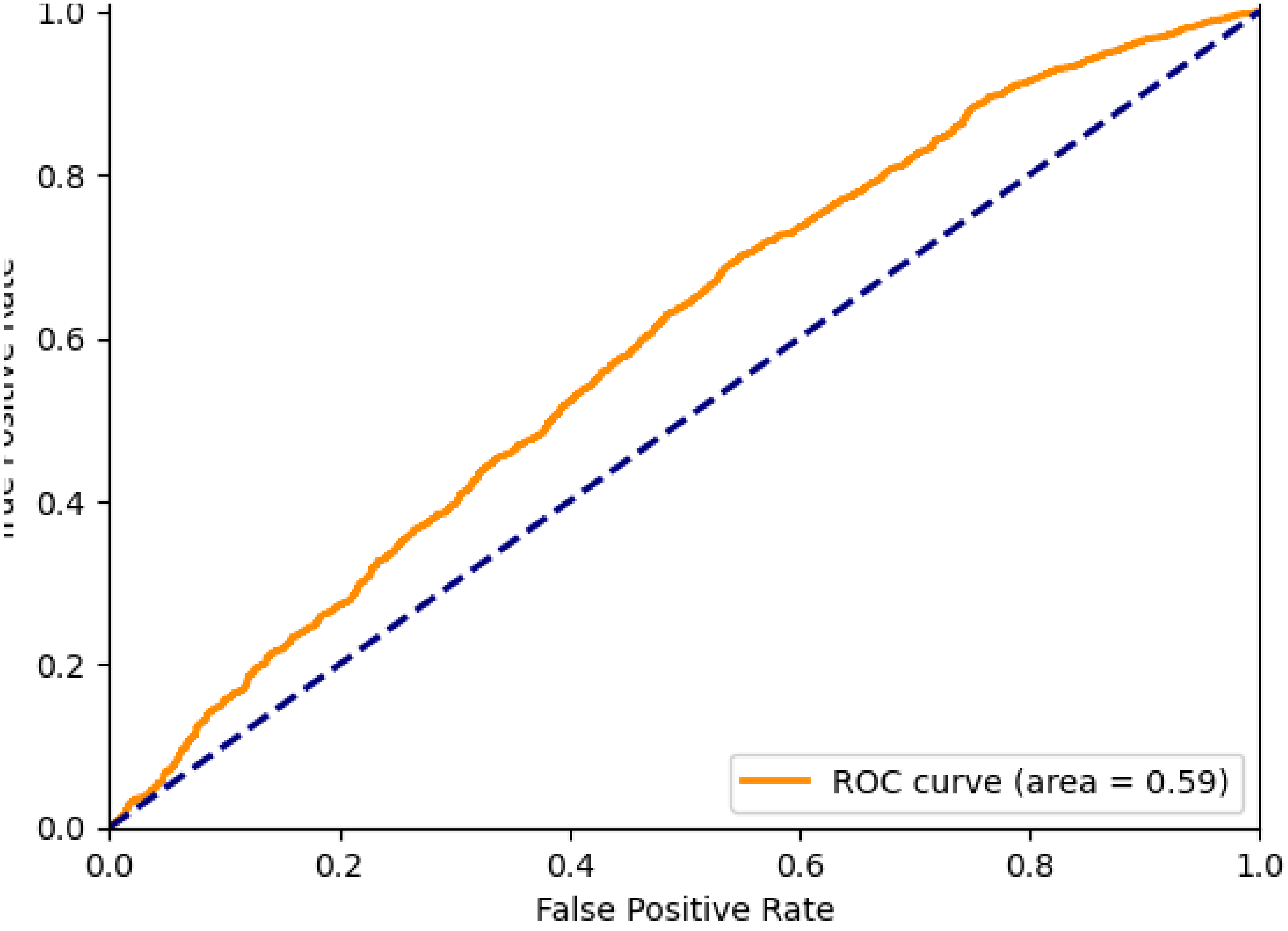

1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.59)


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.62)

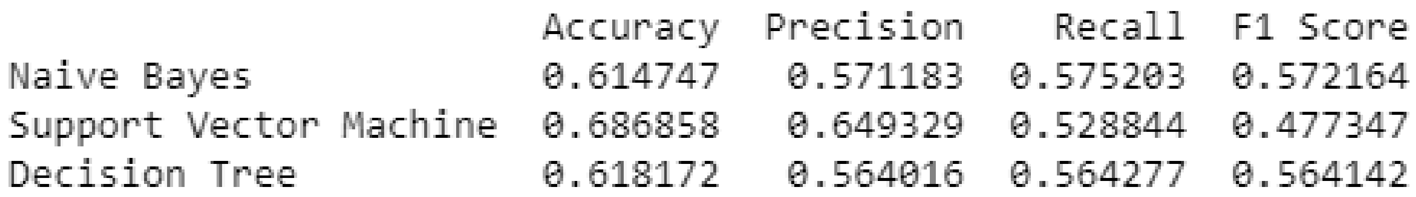

1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.62)


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.61)


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.60)



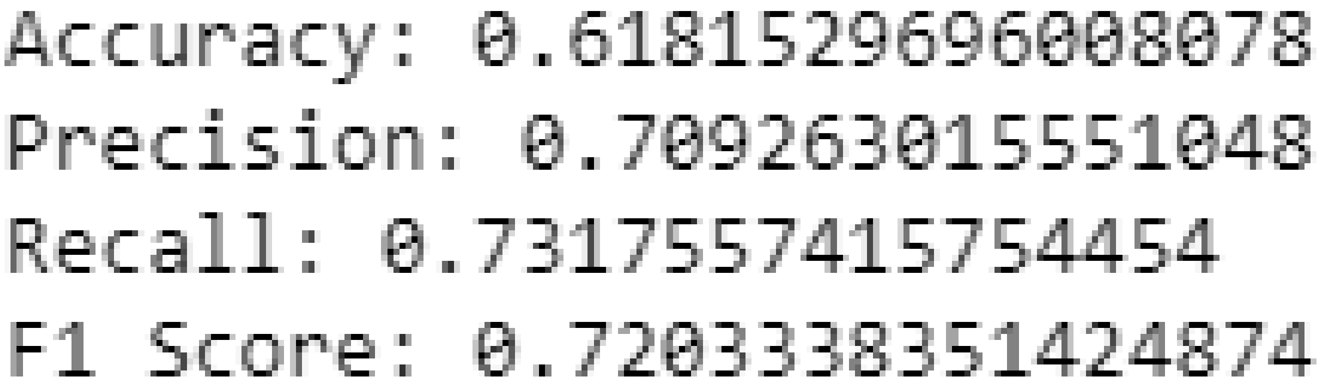
After finding the ROC and AUC score of the models we have narrowed the model function to Naive Bayes, Support Vector Machines and Decision Tree.



```
                        Accuracy  Precision    Recall  F1 Score
Naive Bayes             0.614747   0.571183  0.575203  0.572164
Support Vector Machine  0.686858   0.649329  0.528844  0.477347
Decision Tree           0.618172   0.564016  0.564277  0.564142
```

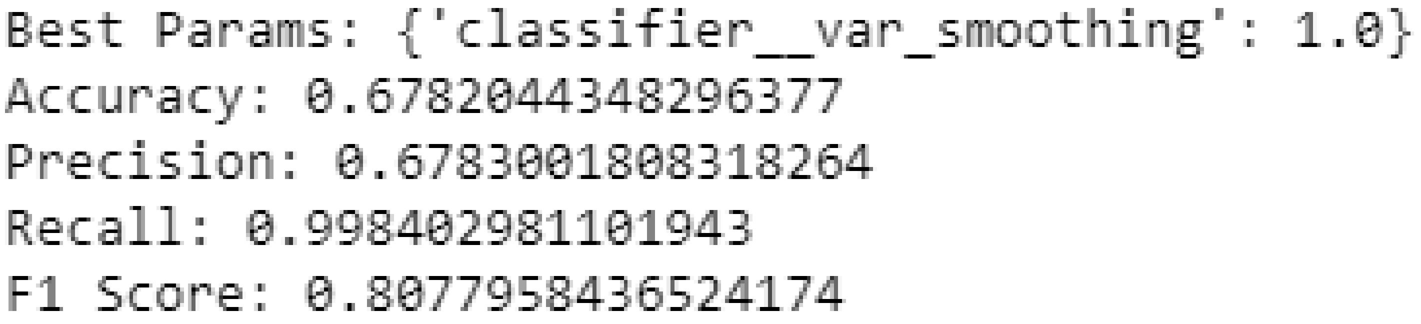
According to the performance metrics, the Support Vector Machine (SVM) has the highest accuracy (68.69%) of the three models, indicating that it is the most reliable overall for right predictions. However, its lower recall score in comparison to precision implies a proclivity for more false negatives.

The accuracy of Naive Bayes and Decision Tree models is comparable (about 61%), with Naive Bayes slightly exceeding in F1 score, indicating a superior balance of precision and recall. Although none of the models has an obvious advantage across all measures, each has a different balance of false positives and negatives, which may affect their selection depending on the application's requirements.



```
Accuracy: 0.6181529696008078
Precision: 0.7092630155551048
Recall: 0.7317557415754454
F1 Score: 0.7203338351424874
```

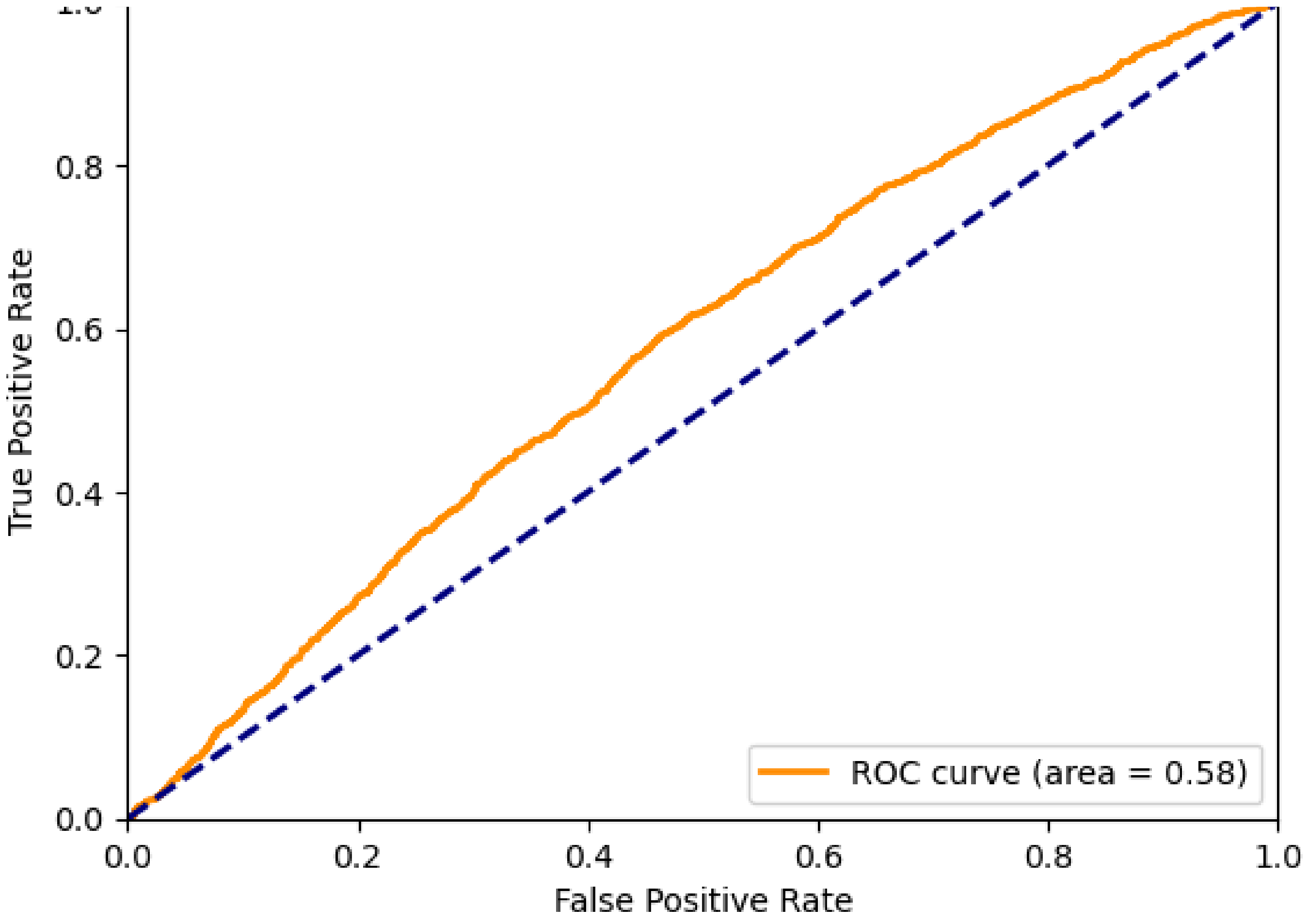
Before using the parameter tuning this are the results for the Naive bayes



```
Best Params: {'classifier__var_smoothing': 1.0}
Accuracy: 0.678204434829637
Precision: 0.6783001808318264
Recall: 0.998402981101943
F1 Score: 0.807795843652417
```

The output provides performance metrics for two different parameter configurations of a model, with the second configuration outperforming the first in all performance criteria. Specifically:

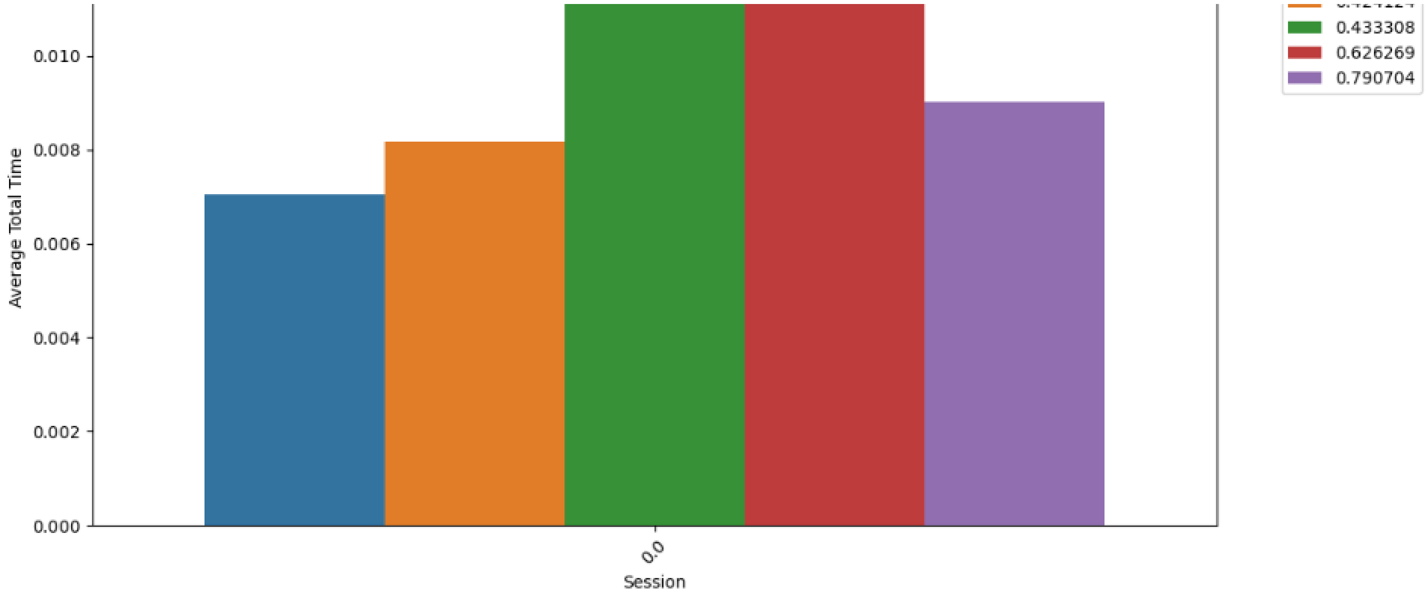
- Accuracy increased from 61.82% to 67.82%, suggesting a greater rate of right predictions overall.
- Precision increased somewhat from 70.93% to 67.83%, indicating that the model made more true positive predictions out of all positive predictions.
- Recall increased significantly from 73.18% to 99.84%, indicating that the model now captures a greater proportion of all actual positives.
- The F1 Score, which balances precision and recall, increased significantly from 72.03% to 80.78%, indicating improved overall model performance.
- The 'Best Params' column displays the best-performing hyperparameters discovered during model optimization. process, with 'classifier__var_smoothing' set to 1.0 being linked to better results. This indicates that the parameter adjustment was successful in improving the model's predicted accuracy.



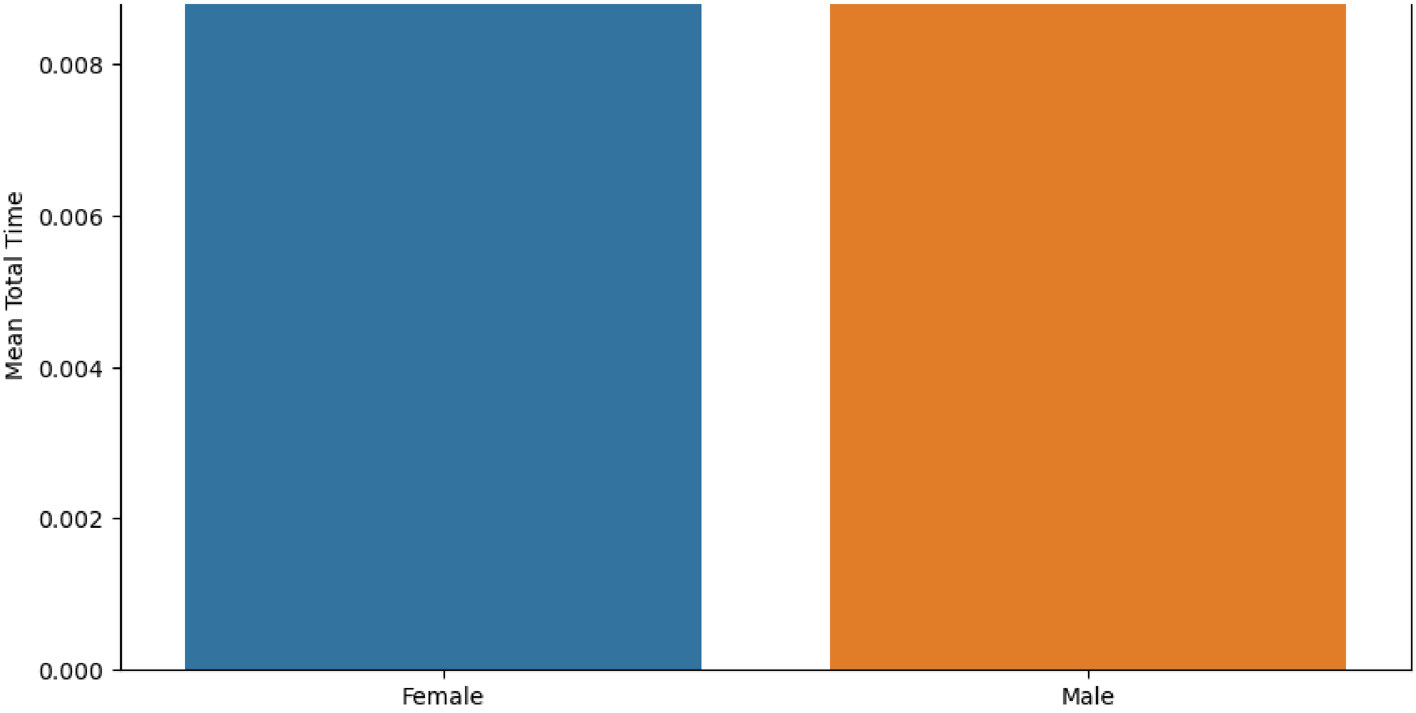
```
AUC Score: 0.5783581634580068
```

The AUC score of 0.578 on the ROC curve indicates that the classifier's ability to differentiate between classes is slightly better than random chance, demonstrating modest predictive performance.

Assigned Tasks



The bar chart depicts the average total time spent by students in various sessions. Each bar represents a different student, identified by their ID, and has a different average time. According to the graph, each student's involvement time varies, with some students having longer average total times than others. The varied colors correlate to different students, allowing for a comparison across individual session durations. This graphic comparison could be used to examine patterns in student involvement or the efficacy of various sessions for different students.



The bar graph compares total time spent by gender, with two bars indicating females and males. The height of each bar represents the average total time allocated to each gender, implying a comparison of female and male students' involvement or activity time. While the actual statistics are not shown, the graphic suggests that both genders have a considerable mean total time, with one gender likely having a marginally greater average. This information could be valuable in determining gender patterns in classroom participation or time allocation.

Conclusion

We used a variety of data processing techniques in our investigation of off-task behavior in classrooms, from initial data cleaning to exploratory data analysis, as well as the deployment of several machine learning models. Our findings shed light on the patterns and drivers of off-task activity, highlighting important aspects influencing student involvement. The predictive capability of the models examined varied, with some models performing better in terms of accuracy and others presenting a more balanced precision-recall trade-off. These findings provide educators and administrators with tangible solutions to improve student focus and learning outcomes. To further deepen our knowledge of classroom involvement, future research may investigate new modeling strategies or incorporate more detailed data.

Future Work

time. Personalized models for customized teaching approaches could also be created. Implementing interventional research to test the efficacy of these strategies, producing practical tools for educators, and widening the study to encompass a variety of educational and cultural settings would all aid in validating and applying the findings to improve educational outcomes.

Question?

Contact Us to get more information

Youtube Video Link