Exploratory Analysis of Student Academic Performance

PREPARED BY:

Shahaf Wieder

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION

In this report, we go on an exploratory analysis of student academic performance, delving into various factors with the aim of determining whether they have a noticeable impact on students' achievements. Additionally, we seek to find out whether these factors could potentially serve as predictors for student success based on the same set of features.

CHAPTER 2. DATASET DESCRIPTION

We used the Students Exam Scores: Extended Dataset[1], comprising 30,640 instances and 14 attributes, for our analysis.
This dataset was chosen precisely due to its inclusion of relevant attributes, making it well-suited for investigating the factors that potentially impact student performance. As outlined in the introduction paragraph, our exploration aims to determine whether these factors hold significance and whether they can be used as potential predictors for student success.

CHAPTER 3. ANALYSIS & FINDINGS

In this analysis, we aimed to explore factors that might influence students' academic performance. Specifically, we were interested in understanding the relationships between various student attributes, such as gender, lunch type, ethnic group, parental education, test preparation, practice of sports, weekly study hours, and the score in Math, Writing and Reading.

The questions of interest were:

- Do certain attributes correlate with higher overall mean scores?
- Can we predict students' academic success based on these attributes?

3.1 Exploratory Data Analysis:

Our initial exploration involved understanding the data distribution, visualizing correlations, and identifying potential trends. The distribution of academic scores and their relationship between all other attributes were visualized using bar plots.

[1] Dataset Link - Students Exam Scores: Extended Dataset

3.2 Missing Values:

During the data exploration we notice that there are missing values in some attributes as depict in the following table:

Attribute	Missing Values (count)	Percentage (%)
Ethnic Group	1840	6
Parent Education	1845	6
Test Preparation	1830	5.97
Practice Sport	631	2
Transport Means	3134	10
Weekly Study Hours	955	3

In addressing the issue of missing data, the approach we took is random imputation with ratio preservation.

This involves replacing missing values with random selections from the existing distribution within each categorical variable.

Given the relatively small proportion of missing data (up to 10% per category) and the importance of maintaining the ratio of values within each category, this approach offers a practical solution.

However, it's important to acknowledge that while this method maintains distribution ratios, it may introduce variability and noise, potentially impacting subsequent analyses.

3.3 Correlation Analysis:

We found correlations between certain variables and academic scores:

- Ethnic Group *(Figure 1 - Students' Scores by Ethnic Group*)
- Parent Education (*Figure 2 - Students' Scores by Parental Education*)
- Lunch Type (*Figure 3 - Students' Scores by Lunch Type*)
- Test Preparation (*Figure 4 - Students' Scores by Test Preparation*)
- Practice Sport (*Figure 5 - Students' Scores by Practice Sport*)

The above variables showed notable correlations, suggesting that these factors may play a role in student performance .

3.4 Estimation and hypothesis testing:

We aimed to investigate the potential relationship between student achievements in Math, Writing, and Reading subjects and their tendency to practice sports. This question piqued our curiosity due to the ongoing debate and research regarding the impact of sports on academic performance.

Another reason to investigate this way is due to the observed weak trend between practicing sports and higher scores across all subjects, we aim to assess whether the trend holds statistical significance.

3.4.1 Hypothesis:

- Null Hypothesis (H0): The overall mean (Math, Writing, and Reading) score of students who practice sports is not significantly higher than the average overall mean score of all students.
- Alternative Hypothesis (H1): The overall mean score of students who practice sports is significantly higher than the average overall mean score of all students.

Our approach involved generating multiple random samples (5000 simulations) from the population, keeping the sample size identical to the number of students who practice sports. This method ensured that the variation properties were maintained. In each sample, we calculated the mean score by combining the scores in Math, Writing, and Reading for students who practice sports.

3.4.2 Results:

By comparing the distribution (*Figure 6 - Hypothesis Testing Simulation Distribution*) of the sample mean scores to the observed mean score for students who practice sports, we determined the test statistic to be 0.27. The resulting p-value was calculated to be 1.0.

3.4.3 Conclusion:

The simulation results, along with the calculated p-value of 1.0, led us to conclude that we cannot reject the null hypothesis (H0). This high p-value indicates that the observed test statistic falls well within the expected range under the null hypothesis, meaning there is no strong evidence to suggest a significant difference between the two groups in terms of their overall mean scores in Math, Writing, and Reading.

However, it is important to note that such a p-value implies that the reverse test (i.e., testing for the absence of a relationship) should have been performed. Given the high p-value, any observed differences in mean scores between students who practice sports and those who do not are likely due to random variability rather than a meaningful effect of practicing sports on academic performance. Therefore, the evidence does not support the alternative hypothesis, which would suggest a significant relationship between sports participation and academic achievement.

3.5 Model Implementation:

In the process of building our predictive model, we took several steps to prepare the dataset for training and testing. Firstly, we created a new attribute in the dataset named "OverallMean," which served as our target class for the classifier. We engineered this attribute to represent the classification of students' mean scores, where a value of 1 indicates a high mean score (80 - 100), and a value of 0 represents a lower mean score.

To proceed with the implementation, we encoded categorical variables using one-hot encoding, which created additional features for the model. The categorical attributes we utilized for prediction were:

- ❖ Note: Despite the lack of statistically significant difference in overall mean scores based on the hypothesis testing, the inclusion of the "PracticeSport" feature in the classifier could be justified by considering its potential practical relevance to the overall model's predictive performance.

- Gender
- EthnicGroup
- TestPrep
- LunchType
- ParentEduc
- WklyStudyHours
- PracticeSport

After the preprocessing steps, we split the data into training and testing sets, allocating 80% of the data for training and 20% for testing. The purpose of this division was to train the model on a portion of the data and assess its performance on unseen data.

In order to ensure consistent scaling and eliminate the influence of different scales among features, we standardized the dataset using z-scores. This transformation enabled each feature to have a mean of 0 and a standard deviation of 1. Such scaling helps the model to converge faster during training and makes the model less sensitive to the magnitudes of different features.

The final dataset was then ready to be fed into our chosen classifier, the K-Nearest Neighbors (KNN) algorithm. By utilizing the attributes mentioned above, we aimed to predict the target class "OverallMean" and classify students into groups based on their mean scores.

These steps laid the foundation for model implementation, allowing us to apply the KNN classifier to the preprocessed data and assess its performance in predicting students' mean scores. The following sections will delve into the results and insights gained from this implementation.

3.6 K-Nearest Neighbors Classifier:

We implemented a K-Nearest Neighbors (KNN) classifier with k=24 (*Figure 7 - Most accurate K for Knn Classifier*) to predict student academic success based on the selected variables. The accuracy of the model was approximately 0.77%.

3.7 Performance Evaluation:

The confusion matrix, accuracy, precision, and recall metrics were used to evaluate the model's performance. The model achieved an accuracy of 0.77%. However, the recall score was relatively low (0.16%), indicating challenges in identifying positive cases (*Figure 8 - Knn Results*).

CHAPTER 4. LIMITATIONS

Our analysis, while informative, is subject to certain limitations that must be considered when interpreting the results. The dataset upon which our study is based may introduce biases and restrictions that could impact the outcomes of our analysis.

4.1 Missing Data:

One of the potential limitations of our analysis is the presence of missing data within the dataset. While we took steps to handle missing values during preprocessing, the nature and extent of missing data could influence the accuracy and generalizability of our findings. The imputation methods used to address missing values might introduce unintended biases, potentially affecting the relationships between variables and the outcomes.

4.2 Sampling Bias:

The dataset may also exhibit sampling bias, which could branch from the source of data collection or the sample population itself. If the sample population does not adequately represent the larger student population, the generalizability of our findings to the broader context may be limited. Biases introduced by the selection process could affect the observed relationships between variables.

4.3 Assumptions during Preprocessing:

In our effort to prepare the dataset for analysis, we made certain assumptions during preprocessing. These assumptions could influence the distribution and relationships of the data. For example, our decision to encode the "OverallMean" attribute into binary classes might simplify the classification process but could overlook nuances in the data. The choice of encoding thresholds could also impact the balance between the two classes.

4.4 External Factors:

Our analysis does not consider potential external factors that might influence students' performance, such as socio-economic conditions, family background, or personal motivations. Neglecting these factors could lead to an incomplete understanding of the true determinants of academic achievement.

4.5 Impact on Findings:

The limitations mentioned above could affect the robustness and validity of our findings. Biases, missing data, and assumptions may lead to inaccurate conclusions or overgeneralization. It's crucial to approach the results with a cautious perspective and consider the potential influence of these limitations.

In conclusion, while our analysis provides valuable insights, it is essential to recognize the inherent limitations that could affect the accuracy and generalizability of our findings. These limitations highlight the need for further research and data collection to refine our understanding of the relationships between variables and student achievement.

CHAPTER 5. FUTURE DIRECTIONS

Our analysis has illuminated several intriguing insights into the relationships between student attributes and academic performance. However, as with any exploration, new questions have emerged that could provide valuable paths for further investigation. Additionally, certain questions may require data beyond what is currently available to address fully.

5.1 Unexplored Variables:

While we have examined various attributes such as gender, parental education, and practice of sports, there are other factors that could influence student achievement that remain unexplored. Variables like students' socio-economic background, access to educational resources, and additional activities may significantly impact academic performance. Future research could focus on including and analyzing these variables to gain a more comprehensive understanding of the factors influencing student success.

5.2 Larger and Diverse Data:

To address certain questions, it may be necessary to collect a more extensive and diverse dataset. For instance, understanding the long-term impact of socio-economic conditions on student achievement would require data spanning multiple years and across various regions. Gathering such data would provide a more exact view of the relationships under consideration.

5.3 Beyond the Academic Context:

Our current analysis focuses solely on academic attributes and their impact on student performance. Exploring the interplay between academic performance and students' personal lives, mental well-being, and future career aspirations could provide a holistic understanding of the factors influencing their overall development.

5.4 Ethical Considerations:

As we delve deeper into understanding students' attributes and academic achievements, ethical considerations must be considered. Ensuring data privacy, obtaining informed consent, and mitigating biases in data collection and analysis are essential to conducting responsible and unbiased research.

5.5 A Question Beyond Data:

A significant question that our current dataset cannot answer pertains to the "why" behind the observed correlations. While our analysis establishes relationships between variables, it does not provide causal explanations. Investigating the causal mechanisms underlying these relationships would require a combination of data-driven analysis and qualitative research methods, such as interviews or surveys.

In conclusion, our analysis offers a valuable starting point for understanding the dynamics between student attributes and academic performance. However, it also highlights the need for continued research that goes beyond the current dataset's limitations, explores new variables, employs advanced techniques, and considers the broader context of students' lives. By addressing these directions, we can deepen our understanding and contribute to more informed educational policies and practices.

CHAPTER 6. APPENDIX

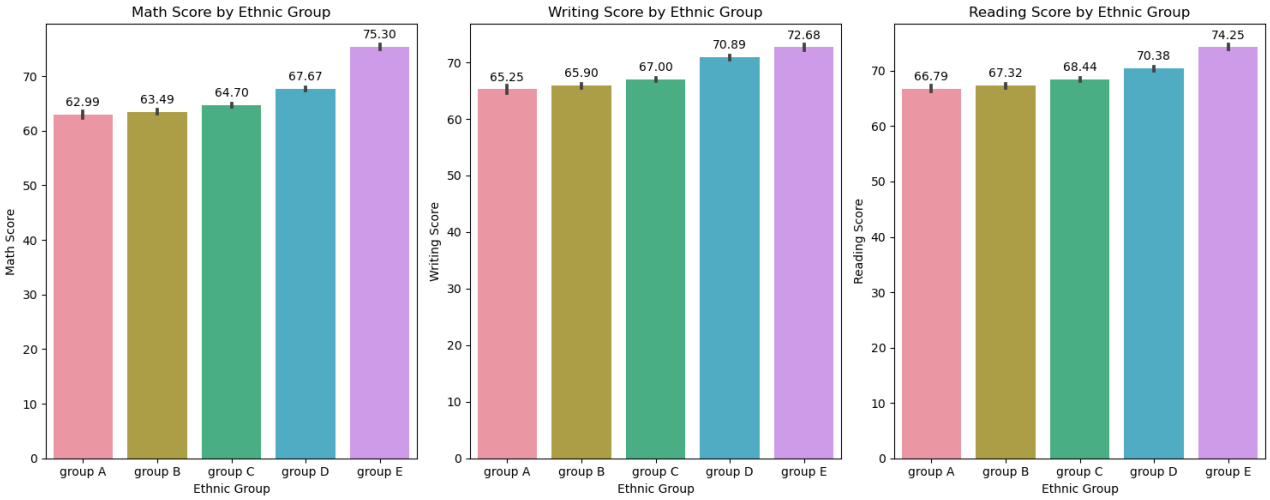

Figure 1 - Students' Scores by Ethnic Group

It seems that students from Ethnic Group 'A' tend to have lower scores on average, while students from Ethnic Group 'E' tend to have higher scores on average.
This observation could potentially suggest that there's a relationship between the 'Ethnic Group' and the academic performance of students.

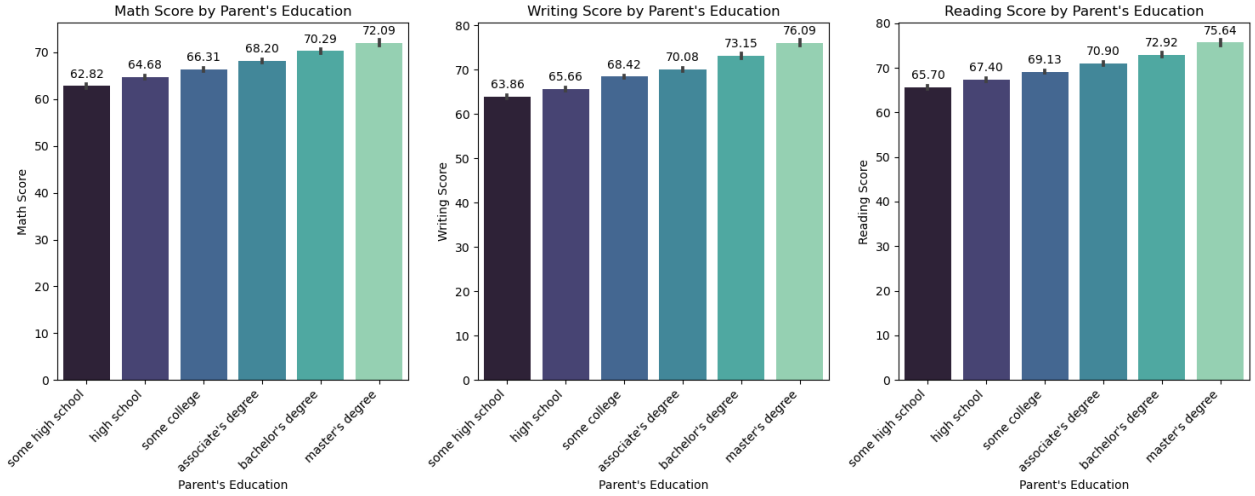

Figure 2 - Students' Scores by Parental Education

A notable trend emerges from the visualizations, revealing a positive correlation between parent education levels and student scores.
Specifically, as parent education levels increase from "some high school" to "master's degree," student scores in math, writing, and reading subjects also exhibit an upward trend.
This observation suggests a potential influence of parental educational background on their children's academic performance.

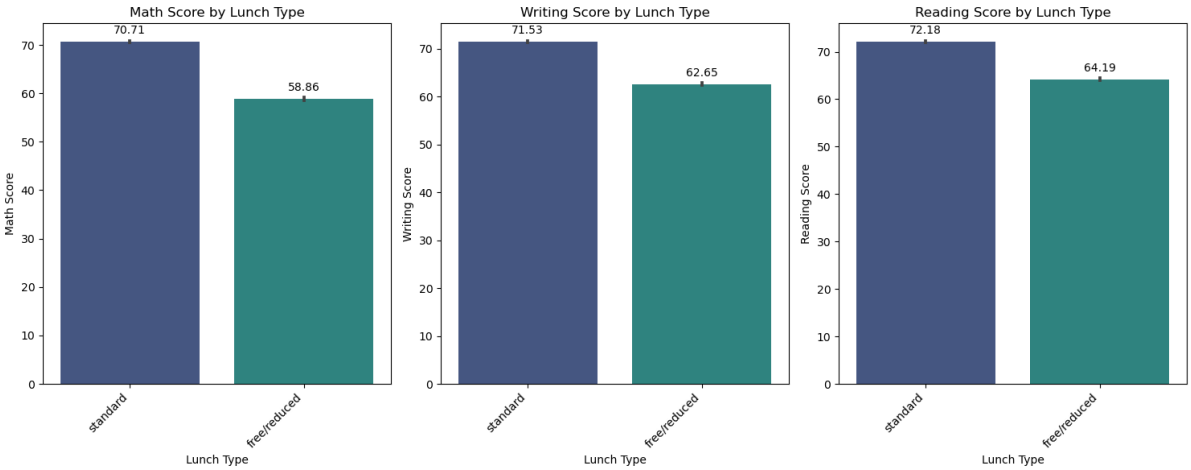

Figure 3 - Students' Scores by Lunch Type

We observed that students who receive reduced or free lunch tend to have lower scores compared to students who have standard lunch across all subjects. - - This observation could indicate a potential socioeconomic factor influencing students' academic performance.

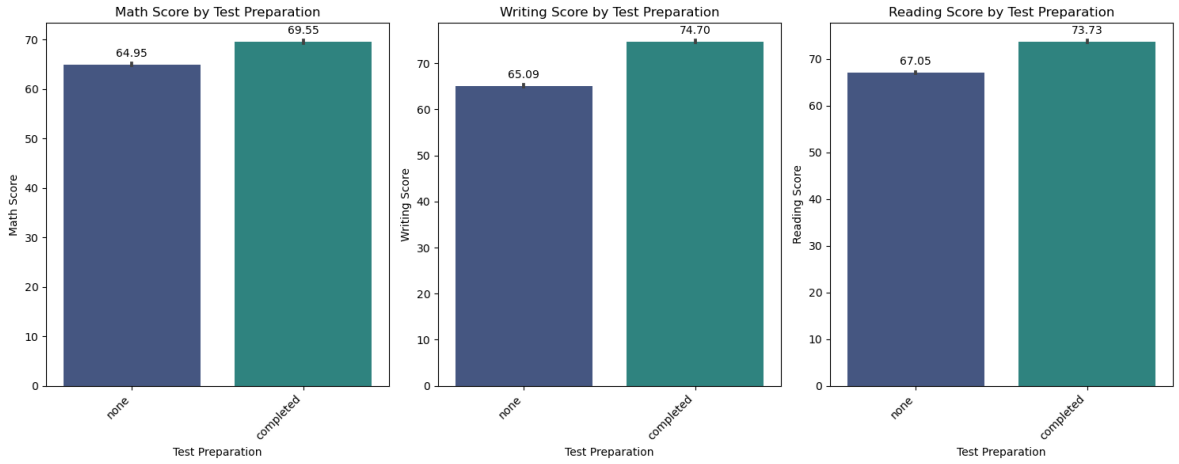

Figure 4 - Students' Scores by Test Preparation

We observed relationship between test preparation and subject scores reveals a notable trend.

Students who underwent test preparation exhibit consistently higher scores across all subjects Math, Writing, and Reading compared to students who did not engage in any test preparation.

This trend strongly suggests that test preparation plays a significant role in enhancing student performance, as reflected in their academic achievements.

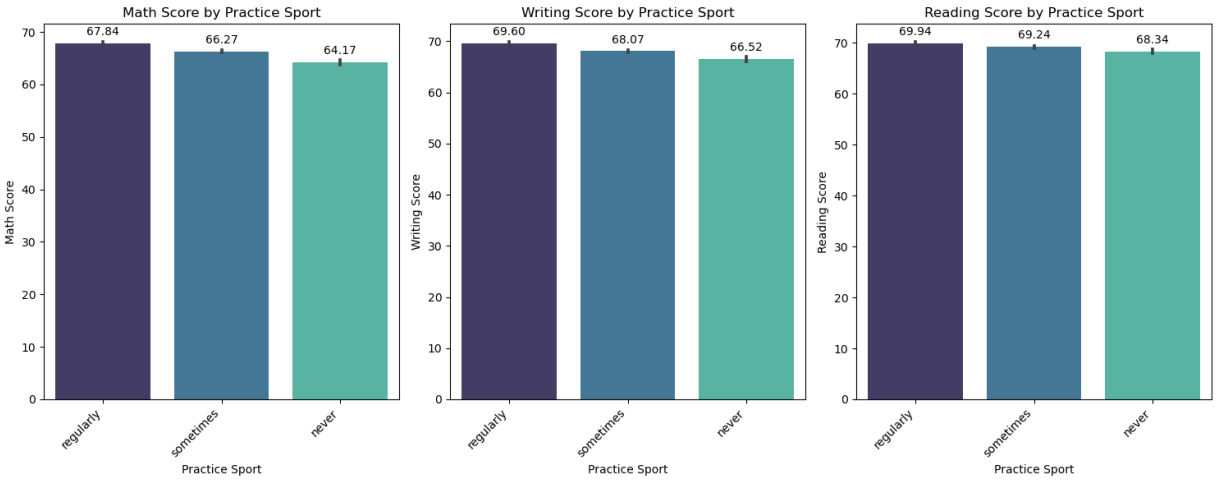

Figure 5 - Students' Scores by Practice Sport

Upon analyzing the data, a striking trend emerges indicating a weak positive association between regular practice of sports and higher scores in all subjects. Students who engage in sports regularly tend to achieve higher Math, Writing, and Reading scores compared to their peers who practice sports less frequently.

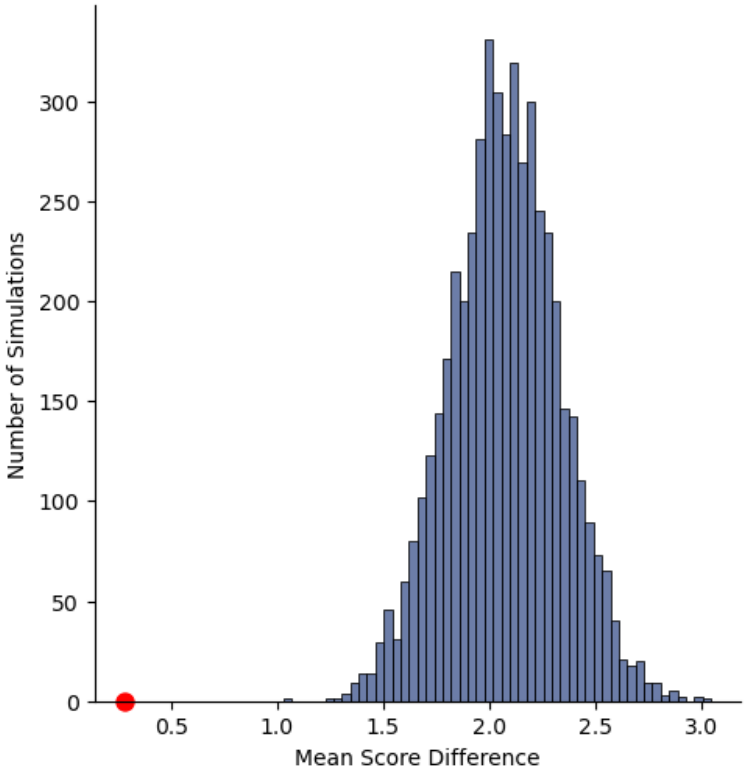

Figure 6 - Hypothesis Testing Simulation Distribution

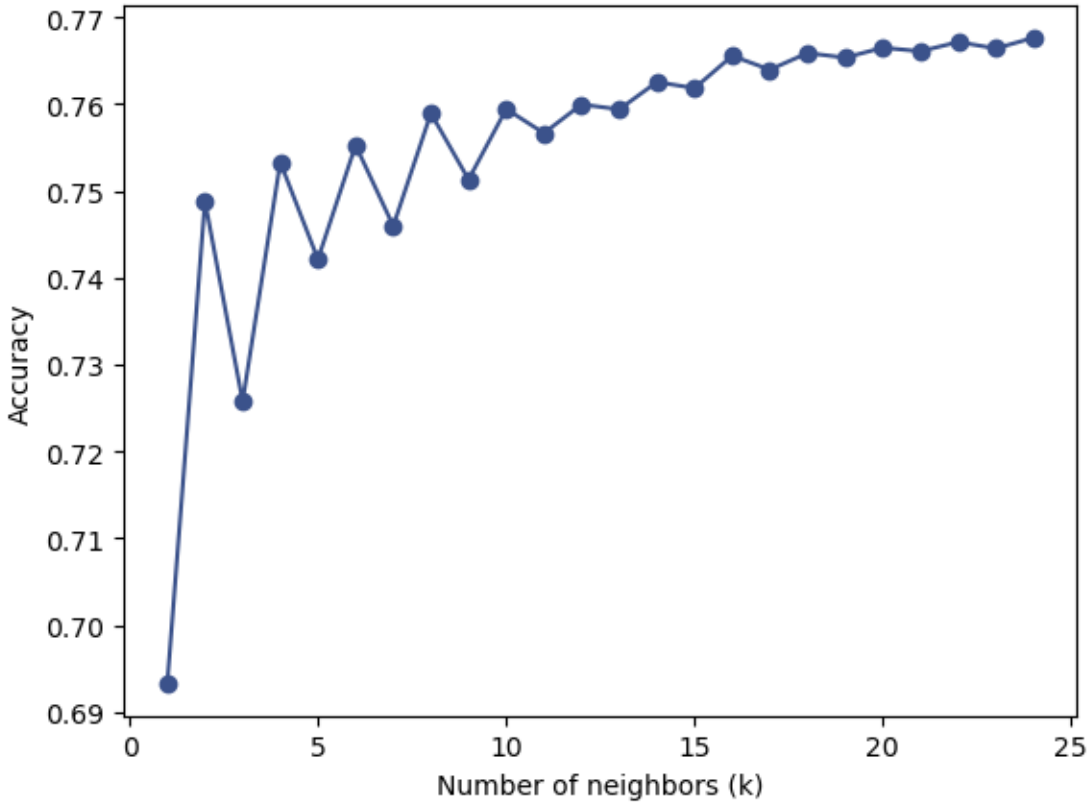

Figure 7 - Most accurate K for Knn Classifier

```
Accuracy of the classifier is 0.7732093326806984
Confusion matrix:
 [[4513  198]
 [1192  226]]
Precision: 0.5330188679245284
Recall: 0.15937940761636107
```

Figure 8 - Knn Results