Brandon Mather

Predictive Modeling for High School Student Interest Trends

Summary:

This white paper will give an in-depth analysis of the online activities and interests of high school students based on data collected from a popular social network. I hope to uncover insights into the trend and behaviors of teenagers on social media from 2006 to 2009. The analysis includes data preprocessing, exploratory data analysis (EDA), predictive modeling, and ethical considerations. This will help give valuable insights for people in education, social media, and youth-focused industries.

Business Problem:

Marketing and educational institutions are very interested in understanding the interests of high school students. This information can help them with targeted marketing, curriculum developments, and engagement. However, gathering data on student interests can be challenging. Predictive modeling could be a solution by allowing us to predict interests by using demographic characteristics (such as age or gender) and the most used terms from social media.

History:

Over the last two decades, social media has become a major part of high school students' lives, providing important insights into their interests and behaviors. The large amount of data from these platforms has helped researchers analyze and understand interests, preferences, and demographics of many different age groups.

Data Explanation:

The dataset that is being used is a random sample of 15,000 high school students from a popular social media platform during 2006 to 2009. The data include demographic information such as graduation year, gender, and age as well as the number of friends each has. It also includes interest-related data, the counts of the 37 most common words found in the profiles. This data was collected by using text mining techniques.

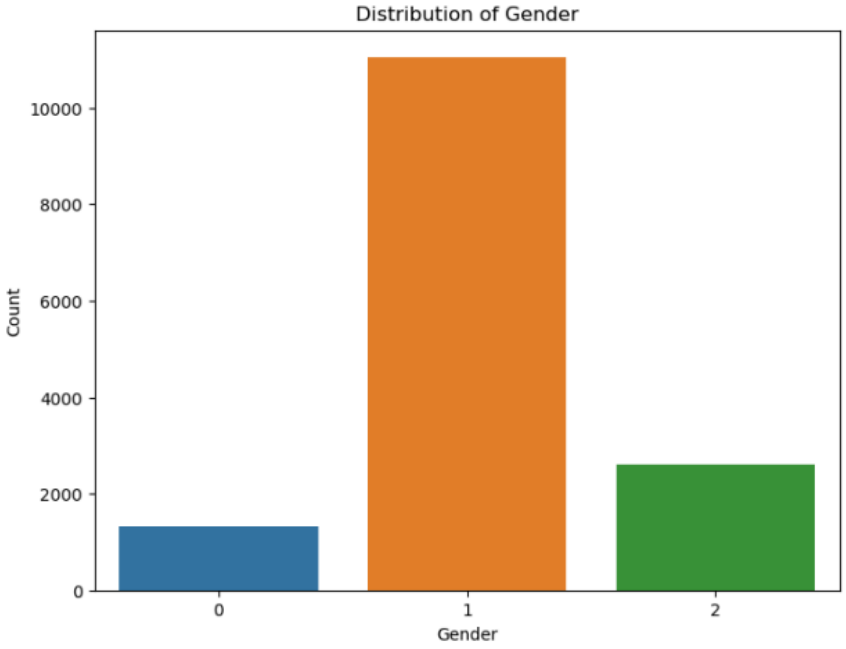
Methods:

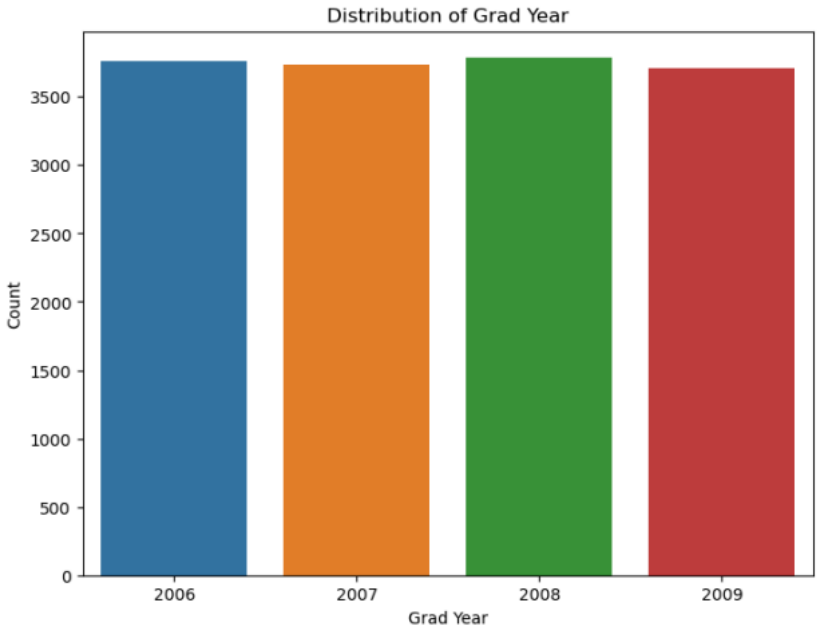
The methods used were a combination of data preprocessing, exploratory data analysis (EDA), and predictive modeling. Preprocessing was used to make sure the data was usable and clean, the EDA revealed insights into the data, and predictive modeling was used to forecast future trends in students.

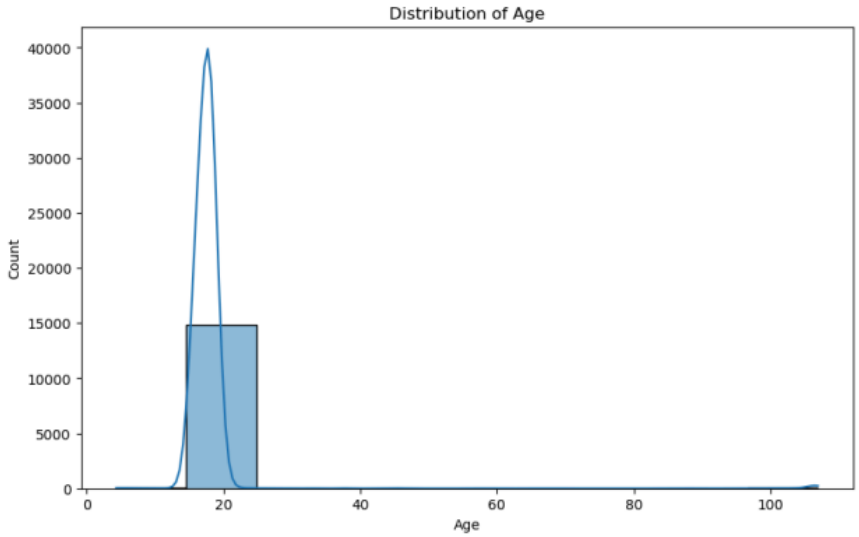
Analysis:

The first step in my analysis was to preprocess my data, and make sure it was usable. The larges problem was in the age category, there were over 1,000 blanks. What I decided to do was use the average age and use that to fill in the blanks. Since all the ages were within 4 years of each other this was going to be the most accurate assumption I could make. I decided I was going to focus more on graduation year than age as well, since there were no blanks in that category and would give me a better understanding of the trends each year that progressed. The other change I had to make was to the gender category. I needed to change them to numerical for modeling purposes. 0 = no gender indicated, 1 = female, and 2 = male.

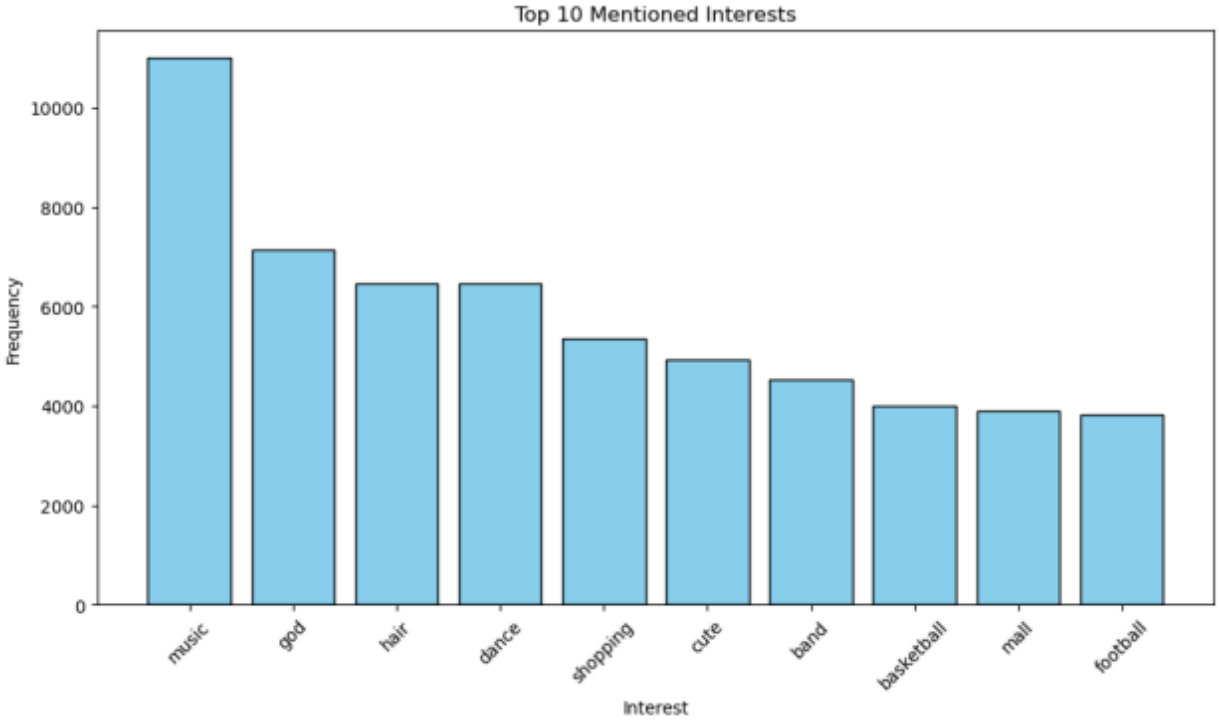
The second step was to complete my EDA, and to start, see what my data looked like from a demographic standpoint. Gender sways female to a high degree, graduation year was much more even, and age as expected was evenly split.

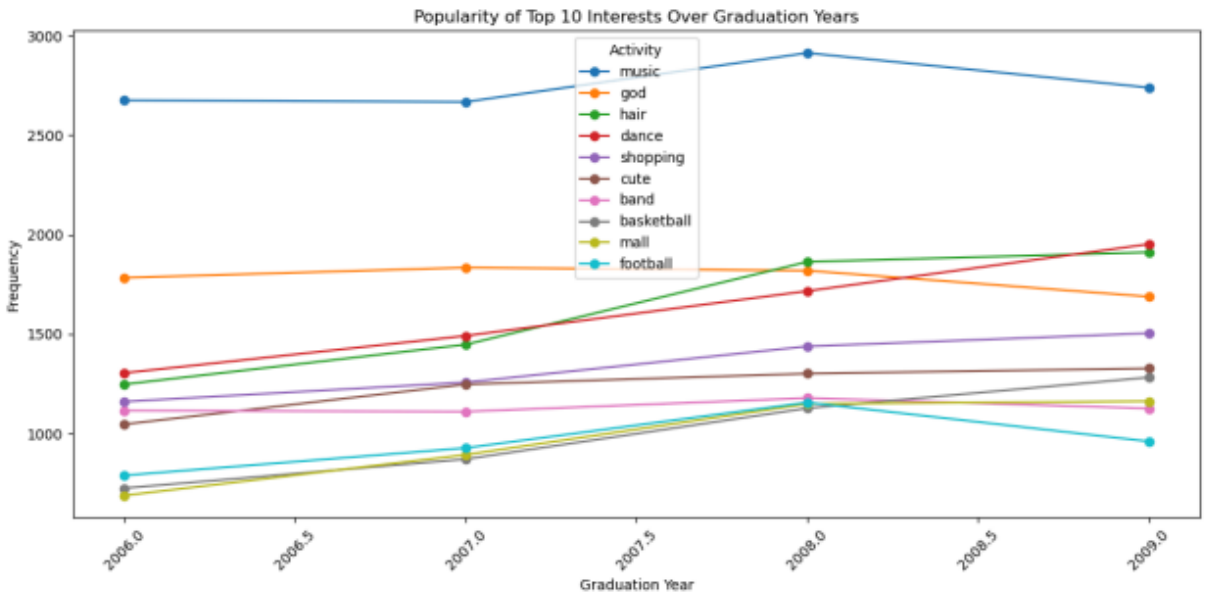


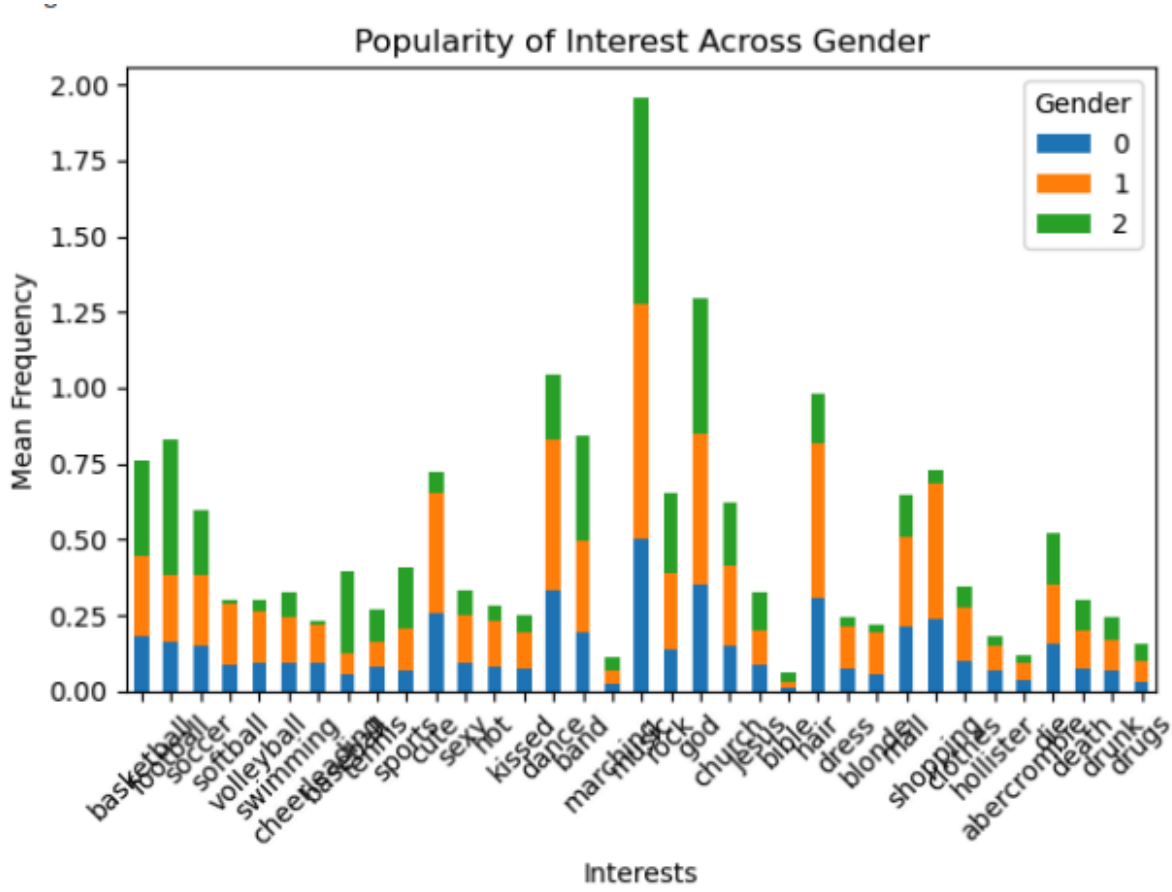




Now that I understood what the demographics looked like, I wanted to see what the interests looked like. I decided to look at the top 10 most mentioned interests, and then to compare that with graduation year and gender.

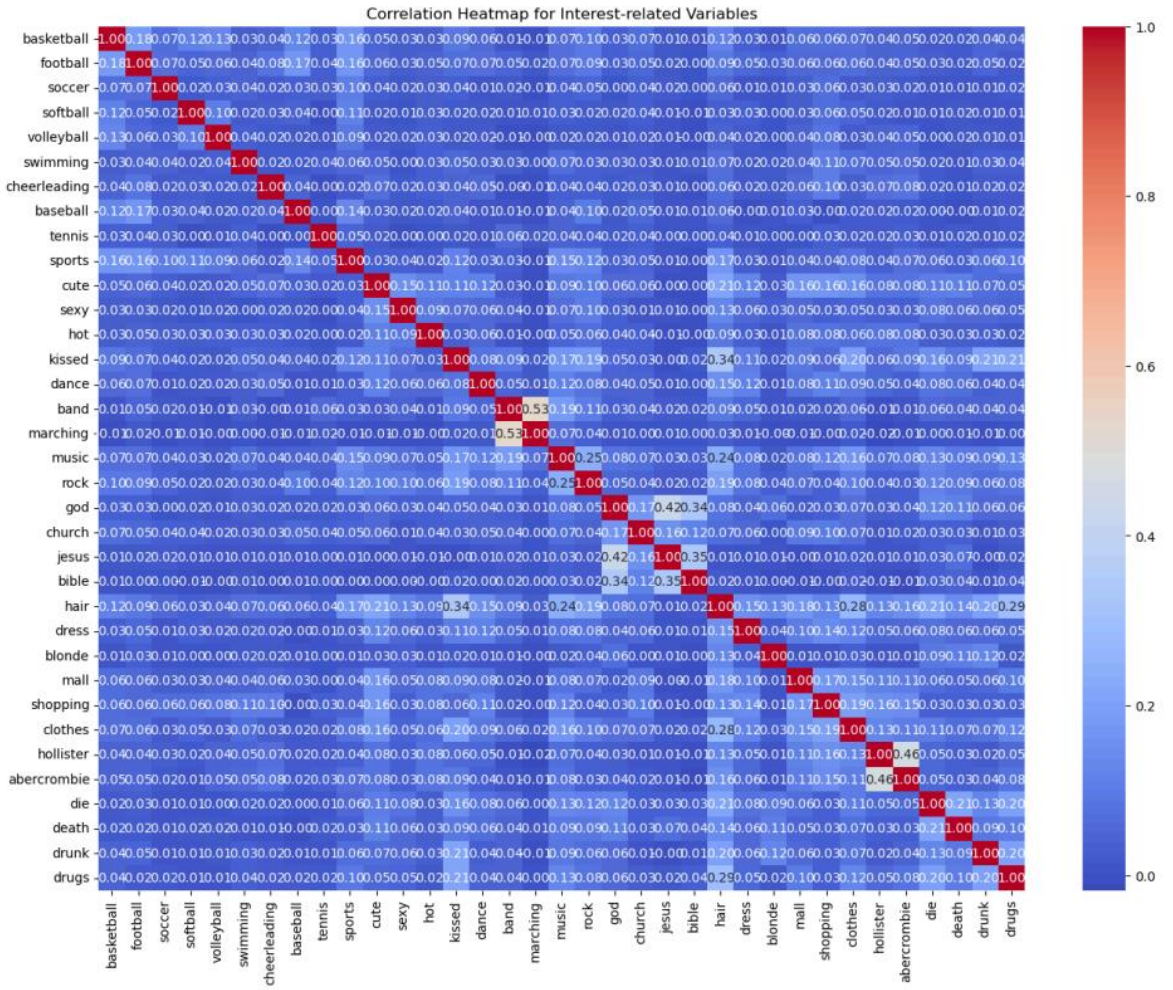



Popularity of Top 10 Interests Over Graduation Years
Activity
music
god
hair
dance
shopping
cute
band
basketball
mall
football
Frequency
Graduation Year
3000
2500
2000
1500
1000
2006.0
2006.5
2007.0
2007.5
2008.0
2008.5
2009.0

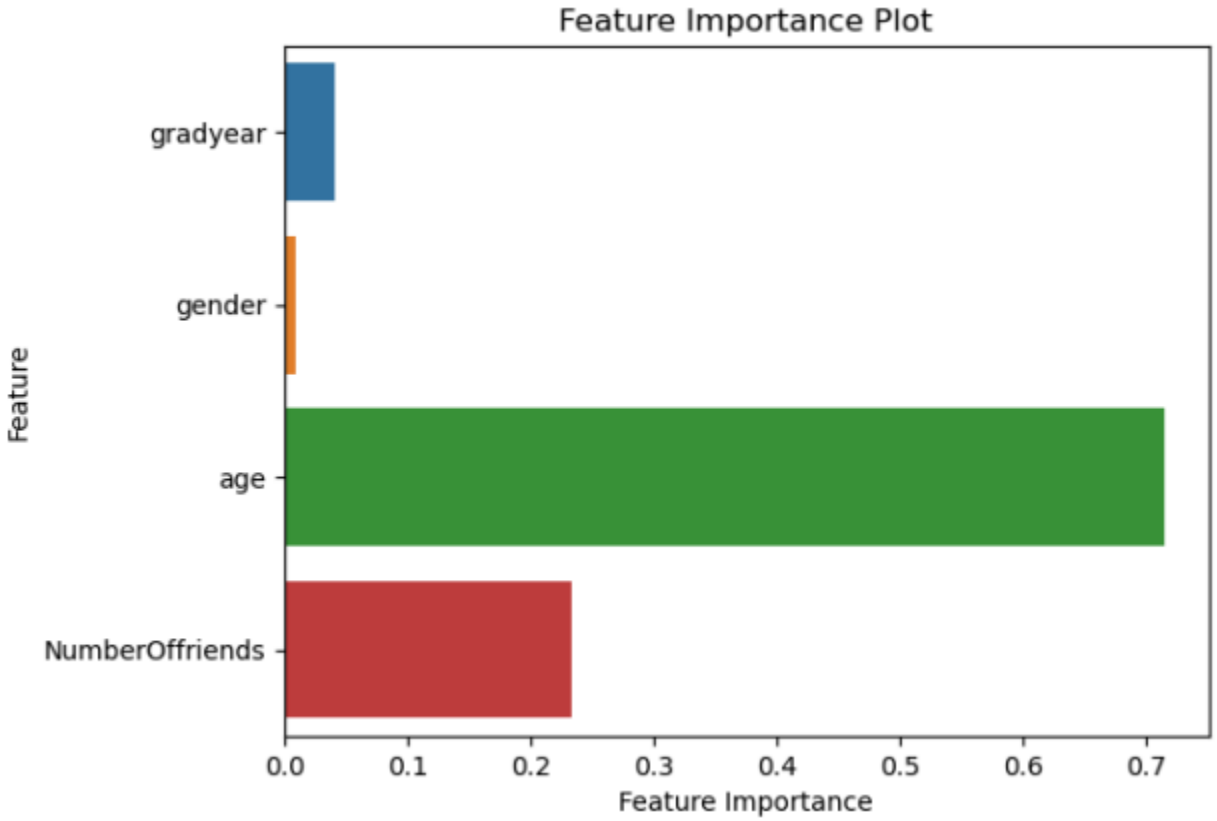

Popularity of Interest Across Gender
Gender
0
1
2
Mean Frequency
Interests
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
basketball
football
soccer
softball
volleyball
swimming
cheerleading
baseball
tennis
sports
cute
sexy
hot
kissed
dance
band
marching
music
rock
god
church
jesus
bible
hair
dress
blonde
mall
shopping
clothes
hollister
abercrombie
die
death
drunk
drugs

The popularity of the top 10 interests over graduation years I felt was the most insightful. Here you can tell if interests are starting to decrease or increase. For example, we can see music stays the most relevant out of them during this timeframe and never takes a dip until 2009, but that's only after it rose in 2008, it remained higher than 2006-2007. We can see dance as a category that gained popularity, rising every single year while god and band are going down in popularity.

The last analysis I wanted to see was the correlation between activities, which can be seen in the heatmap below. Here we can see if someone is interested in one topic, which other topics might interest them.



After my EDA I then created a prediction model. I decided to use a Random Forest Regression Model to predict the future trends based on the demographic of high school students. Here we can see which features were the most important in being able to make those predictions. I used Mean Squared Error, Cross-Validation RMSE scores and Mean CV RMSE to evaluate my model.



Conclusion:

The analysis sheds light on the online interests of high school students. This analysis and model will be able to make predictions of what interest's students based off their demographics. We will be able to see which interests relate to which gender, age, and other interests. By understanding these trends, we will be able to make high school students lives better in a the digital space.

Assumptions:

The biggest assumptions are that the mentions are related to actual interest in that subject, and that this dataset is representative of the broader high school student population. We are also assuming the data collection process adhered to ethical guidelines.

Limitations:

Limitations of the analysis include the potential for biases in the dataset, such as underrepresentation of certain demographic groups or inaccuracies in the self-reported information. The dataset also may not truly capture recent developments in social media since this was from a very specific 4 year timeframe.

Challenges:

Challenges in the analysis were handling the missing data and finding the best way to fill that in so I didn't lose the other data points for those.

Future Uses:

Future research could continue exploring trends in high school student interest on social media beyond this time period. The techniques could not only be used by social media platforms to improve their experience, but the techniques can also be used for marketing research, customer segmentation, and personalized recommendation systems.

Recommendations:

I would recommend further exploration of the relationship between student interests and academic performance and other demographic factors. Additionally, refinement of the predictive model to maintain accuracy and relevance.

Implementation Plan:

The findings from this analysis can help the development of targeted marketing campaigns, curriculum adjustments, and student engagement. Both educational institutions and marketers can use the model to tailor their approach based off the interests of each demographic.

Ethical Assessment:

It's important to consider ethical implications such as privacy and data security. Transparency in data collection and usage, informed consent, and data protection regulations are essential. The data should be protected for the sake of the well-being of the high school students.

References:

Zabihullah18. (2024b, February 29). *Students' social network profile clustering*. Kaggle. https://www.kaggle.com/datasets/zabihullah18/students-social-network-profile-clustering

10 Questions:

1. Can you provide more details about how privacy was respected in the collection of the data?
2. How did you ensure the reliability of your findings, given the nature of social media data?
3. What were the most significant insights uncovered in your analysis of high school students' interests?
4. How do you think these findings can be applied in real-world contexts, such as in education or marketing?
5. Were there any unexpected findings that emerged during your research?
6. How did you address potential biases in the dataset, such as demographics?
7. What are the implications of your findings for future research in this area?
8. How do you envision people using your research to inform their decision?
9. What ethical consideration did you consider regarding privacy and consent?
10. Do you have any unanswered questions that warrant further investigation?