A Novel Machine Learning Recommender System for Generating College Recommendations – College/University Admission Assistant

Disha Raghuvanshi

1. Abstract

This paper introduces Usher-AI, a Machine Learning recommender system utilizing the K-Nearest Neighbors (KNN) algorithm, designed to forecast university acceptance for prospective students and generate college recommendations. By analyzing historical student and university data of accepted and denied applicants, this method aims to assist individuals in estimating their chances of college acceptance, thereby enhancing their decision-making process. This paper details the steps in the method's development and implementation, including the data preprocessing, model training, and evaluation. The predictions generated by Usher-AI including, chances of acceptance using student fit and profile, demonstrate promising accuracy in predicting college acceptance, providing valuable insights for both applicants and university admissions. In addition, the Usher-AI-generated college recommendations inform student decisions in selecting colleges during the application cycle.

Keywords: college applications, college recommendations, college admissions, college guidance, usherai, knn machine learning model, admission likelihood, high school student, democratize college admissions guidance

2. Introduction

In today's highly competitive academic landscape, gaining admission to universities is a crucial milestone for individuals. However, many prospective students lack the tools and guidance to navigate the complex admissions process. While a wealth of independent guidance counselors are available, they are often accessible only to a privileged few. Moreover, available school guidance counselors may sometimes offer overly general guidance (due to the high workload), which is inadequate for students seeking individualized advice. To address these challenges, I have developed a Software as a Service (SaaS) solution called Usher-AI. Usher-AI aims to demystify the college admissions process by providing easily accessible and cost-effective resources to all students, thereby leveling the playing field. In the college admissions process, the first and arguably most crucial step is creating a well-balanced college list, which demands knowledge, guidance, and extensive research. Usher-AI implements a KNN-based predictive recommendation model, utilizing historical data on student admissions. By harnessing the power of AI, it aims to develop a solution that offers actionable insights into colleges with a higher likelihood of acceptance, empowering applicants to make informed decisions.

To identify a balanced list of colleges, schools can be classified into three categories. *Reach* Schools are colleges that commonly accept students with higher GPAs and standardized test scores, *Target* Schools are colleges that fit the student's GPAs and standardized test scores, and *Likely* schools are the colleges that consistently accept students below the student's GPAs and standardized test scores (1-4). To determine if a school is a reach, target, or likely school, a student needs to compare his/her academic performance to the average admitted student to see how the chances of being admitted stack up. The guidance for *Reach*, *Target,* and *Likely*

are summarized in Figure 1. If a school has an extremely low admissions rate (usually under 20%), that school should be considered a wildcard for everyone, meaning a majority of highly qualified applicants will not be admitted simply by virtue of the numbers. For instance, schools that fall under this category include Stanford, Harvard, University of Pennsylvania, MIT, and Pomona. These types of schools are called “highly rejective colleges” (1-4).

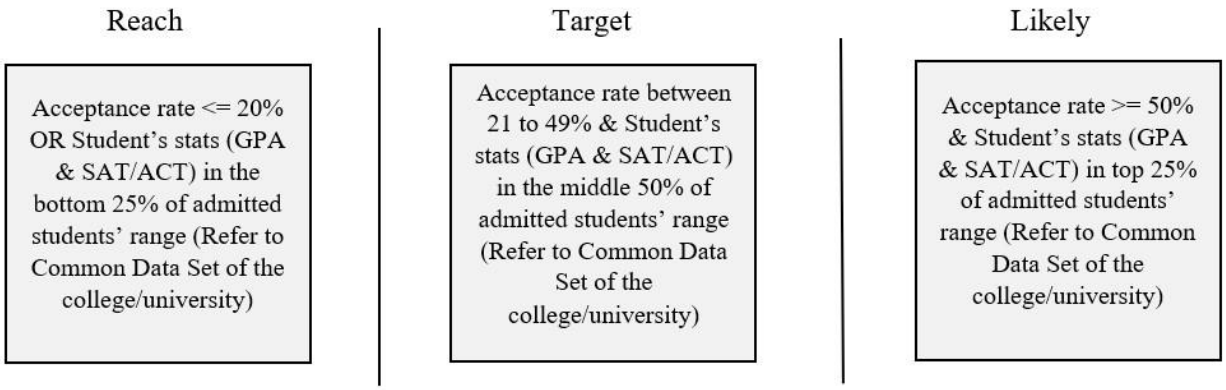

Figure 1: Guideline for Categorizing Colleges/Universities for College Applications

Figure 2 displays the SAT ranges for a selected group of colleges. To describe it further, let’s presume Stanford University shared the admitted students’ middle 50% score range for SAT as 1460 to 1560 in the common data set. This published score range provides reference to the score distribution which can be viewed as the Bottom 25%, Middle 50%, and Top 25%. In essence, it means 50% of the admitted students to Stanford University had SAT scores in the middle range of 1460 to 1560, 25% scored below 1460 and 25% scored higher than 1560. These ranges help students compare their scores with the admitted students to determine if a school should be categorized as *Reach*, *Target*, or *Likely*.

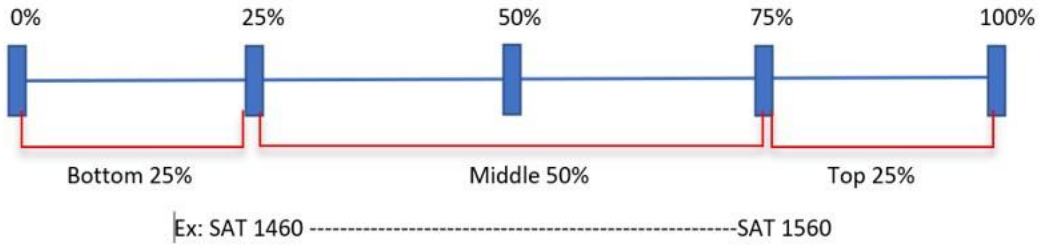

Figure 2: Example Depiction of SAT Score Ranges

A balanced list of colleges should include 30% reach, 40% target, and 30% likely schools as a general guideline (1-4). For instance, if the student wishes to apply to 15 schools, a balanced college list should include approximately 4/5 reach, 6 target, and 4/5 likely schools. The college admission process has become so competitive over the past decade that it is always important for students to create a balanced list to account for the randomness of college admissions.

However, many students do not have access, knowledge, or guidance with respect to creating a balanced college list. When constructing their lists, they are often influenced by a school's ranking or brand recognition, resulting in a list that is more heavily tilted toward highly rejective schools. This can result in disappointing outcomes not serving the student's expectations. The recommended approach is for students to research schools in each category that best match the student's interests and academic proficiencies.

In this research paper, I propose an approach called Usher-AI that incorporates a method for students to predict their likelihood of acceptance based on their profile and to craft a balanced college list using artificial intelligence (AI) models. Using AI algorithms such as K-Nearest Neighbors (KNN) and Python-Pandas libraries for data processing, this study aims to determine the optimal fit school for students based on their interests and profile. The model can take students' academic performance, extracurriculars, and achievements/honors into consideration to determine the best-fit school based on the previously admitted students' success.

3. Materials and Methods

Usher-AI functions as a tool for prospective high school students to assess their chances of admission to various universities and gain insights into a recommended list of colleges informing their decisions. Today, several tools on the web help students estimate their likelihood of admission based on quantitative data such as GPA and test scores. However, this prediction can be limiting and misleading to students since most colleges in the US practice a holistic review of applicants. A holistic review of a student's application involves reviewing the "whole" applicant. It takes into consideration applicants' experiences, attributes, and academic metrics as well as the value an applicant would contribute to learning, practice, and teaching. Therefore, predicting a student's success in college admissions by considering just the academic factors is far from accurate. Additionally, no tools available on the market today generate a college recommendation list tailored to the student's profile. To address this gap, Usher-AI considers both quantitative and qualitative factors regarding the applicant to predict the student's success in admission more accurately. By inputting relevant academic and extracurricular data, users receive machine learning model-generated college recommendations where the student has a higher likelihood of acceptance. This not only aids in managing expectations but also enables individuals to strategize their application approach effectively.

Figure 3 gives an example of the types of data that students can supply to Usher-AI which carries out an exhaustive and accurate assessment of their performance and achievements as well as making accurate predictions of outcomes.

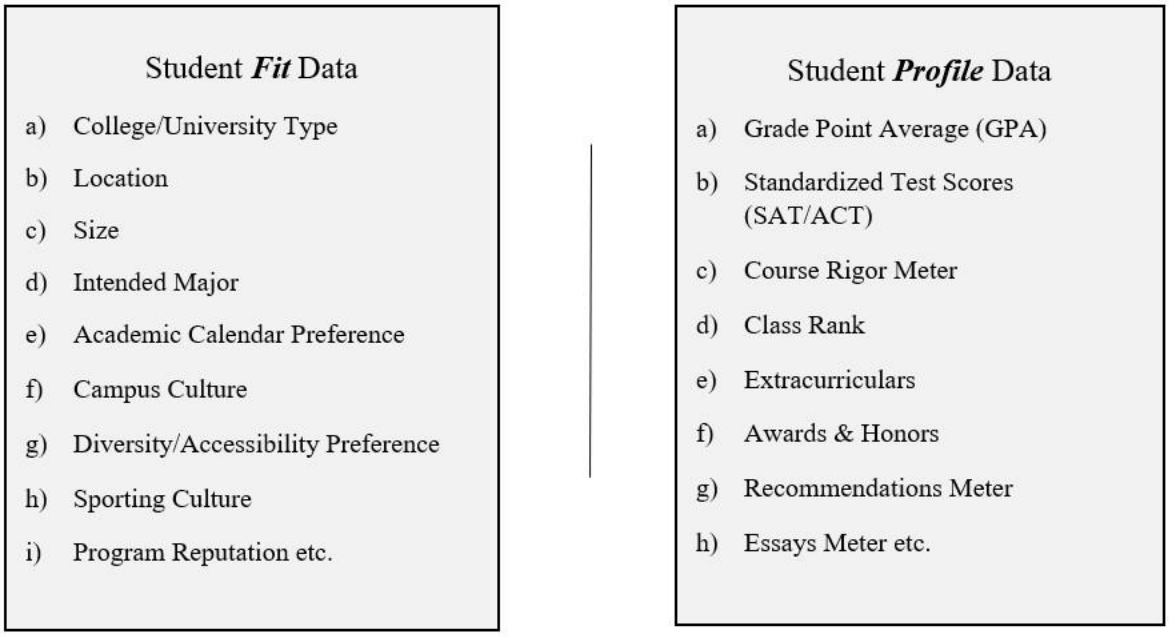

Figure 3: Types of data that can be supplied to Usher-AI

The AI model processes the input data and compares it against the already available outcome data of previously admitted/denied students to predict new student outcomes. It also generates a college list as a recommendation.

The logic underlying the proposed solution lies in the assumption that past admission decisions hold valuable insights into future outcomes. By analyzing characteristics of previously accepted and denied students by various universities, the model identifies patterns that contribute to college acceptance, thereby enabling accurate predictions for new applicants. This approach offers a data-driven and objective means of evaluating admission prospects.

3.1. Model

The predictive model employs the K-Nearest Neighbors (KNN) algorithm. Historical data comprising academic records, standardized test scores (SAT/ACT), extracurricular activities, achievements/honors, and previous admission outcomes are used to train the model. Data preprocessing involves normalization encoding i.e., Label Encoding and feature engineering to ensure uniformity across variables. The KNN algorithm calculates the distance between new applicants and reference students, assigning the K most similar ones as potential matches.

This research evaluated various machine learning algorithms such as K-Nearest Neighbors, K-Means, Support Vector Machine, and Random Forests, considering the algorithm's simplicity, versatility, and accuracy. After closely studying the pros and cons of various algorithms, the K-Nearest Neighbors (KNN) algorithm was selected for its versatility and applicability to this specific research problem.

The K-Nearest Neighbors (KNN) algorithm is a supervised machine learning method employed to tackle classification and regression problems. KNN algorithm is a versatile and widely used machine learning algorithm that is primarily used for its simplicity and ease of implementation. It does not require any assumptions about the underlying data distribution. It can also handle both numerical and categorical data, making it a flexible choice for various types of datasets in classification and regression tasks. It is a non-parametric method that makes predictions based on the similarity of data points in a given dataset (5-6).

The KNN algorithm works by finding the K nearest neighbors to a provided (*Target*) data point based on a distance metric, such as *Euclidean distance*. The class or value of the data point is then determined by the majority vote or average of the K neighbors. This approach allows the algorithm to adapt to different patterns and make predictions based on the local structure of the data. Please refer to Figure 4 for an illustration of the KNN algorithm (5-6).

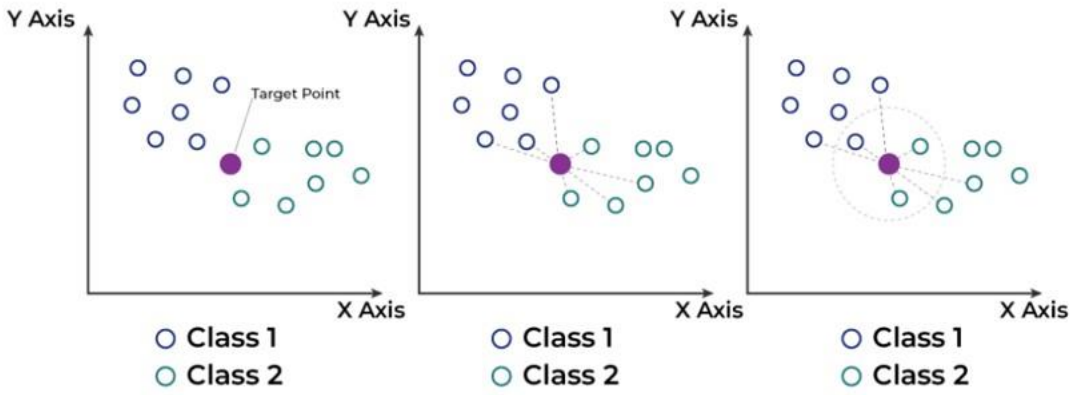

Figure 4: Illustration of KNN Algorithm

Euclidean Distance

The Euclidean distance is identical to the Cartesian distance between two points that are in a plane/hyperplane. Euclidean distance can also be visualized as the length of the straight line that joins the two points that are under consideration. This metric is also used to define the net displacement between the two states of an object (5-6).

$$distance(\mathrm{X},\mathrm{Y}) = \sqrt{\sum_{i=1}^{n}(\mathrm{x}_i - \mathrm{y}_i)^2} \tag{3.1}$$

X = Vector X where *i* is the index
Y = Vector Y where *i* is the index
n = Number of observations

3.2. Methods

a) ***Data Collection and Preparation*:** Raw data including college/university profiles, student profiles, and admission outcomes are collected and organized into structured datasets. Columns representing academic achievements, standardized test scores, and other relevant attributes are identified and categorized accordingly. All non-numeric values present in the raw data are converted into weighted numeric values using *Label Encoding.* This method is referred to as normalization encoding which is achieved to ensure uniformity across variables that is easily understood by the model.

b) ***Model Training*:** The KNN model is trained using historical admission data, where feature vectors of accepted and denied students serve as reference points. The algorithm calculates distances between these points and new applicants, determining the K nearest neighboring students. In the case of this research, after the hyperparameter tuning, the value of K=5 was chosen to yield optimal results.

c) ***Output Evaluation*:** To evaluate the model's performance, a portion of the dataset is reserved for testing. New applicant profiles are inputted and predicted acceptance outcomes are compared against actual outcomes. Metrics such as accuracy, precision, and recall are computed to assess the model's effectiveness. The flowchart of the Usher-AI KNN method is captured in Figure 5.

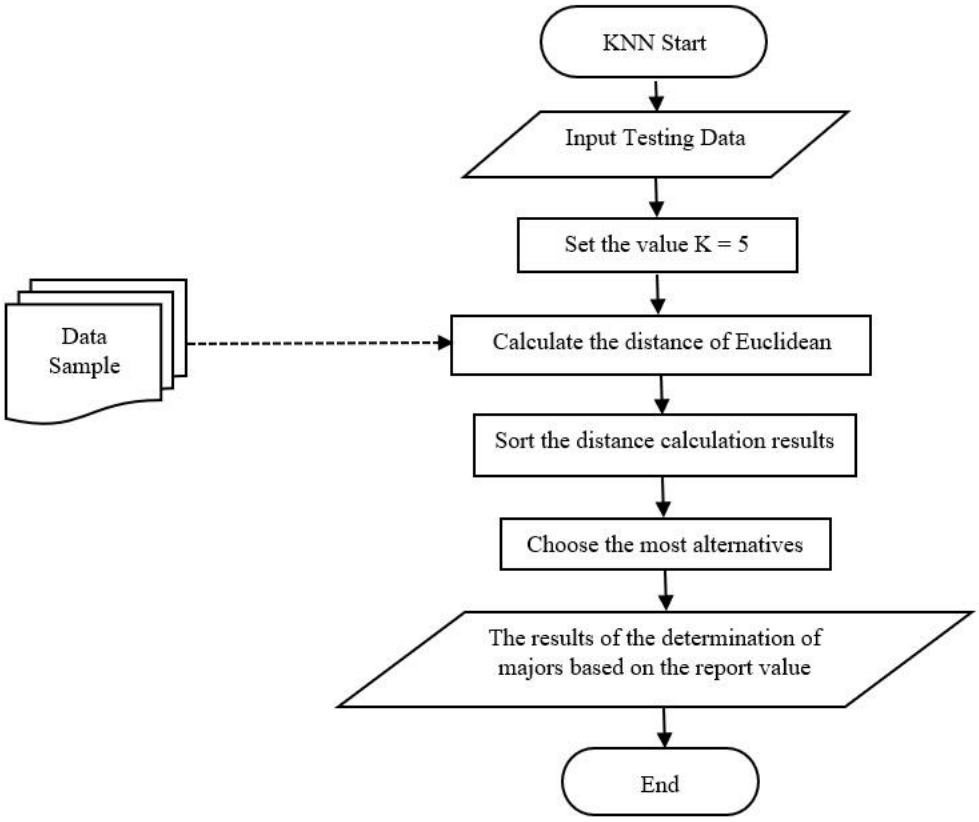

Figure 5: Flowchart of Usher-AI KNN Method

Figure 6 illustrates the high-level approach to training and finalizing the Usher-AI recommender model.

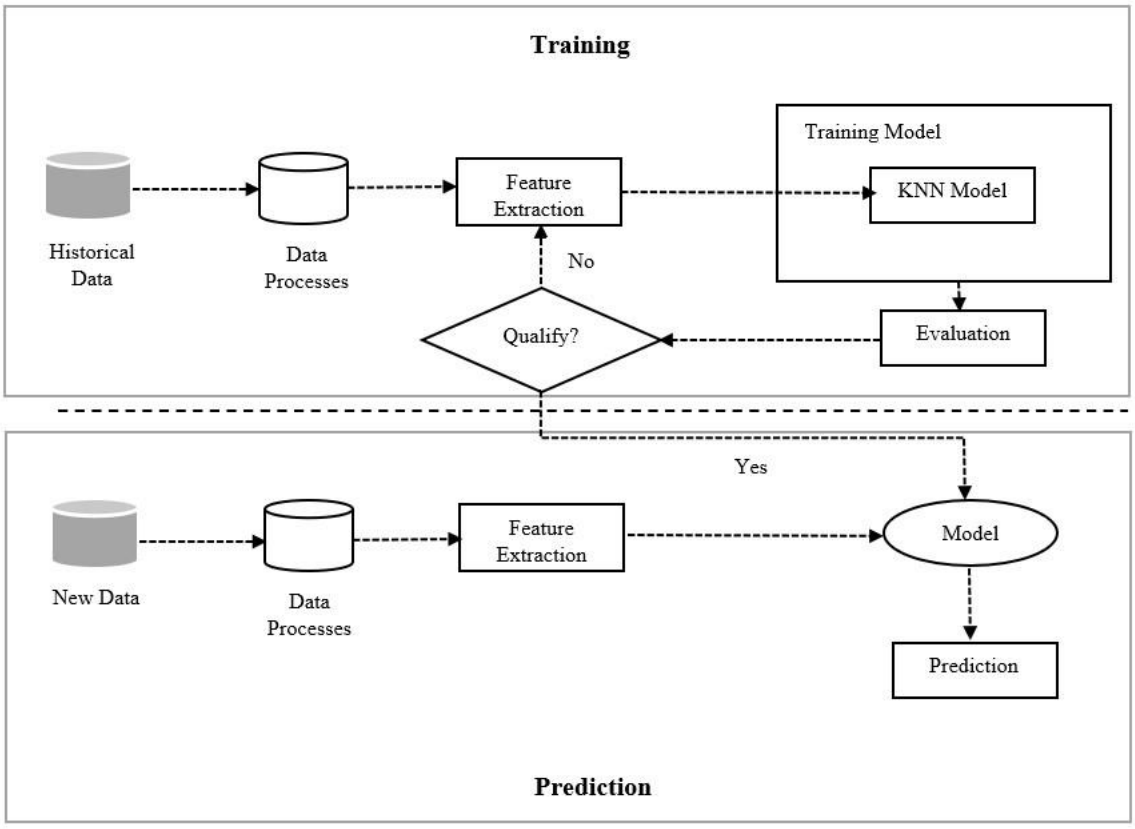

Figure 6: Usher-AI Model Training & Prediction Approach

4. Results

Following the training and testing of the model, the results demonstrate a high degree of accuracy in predicting admission outcomes. By implementing the methods described above, discernible patterns emerge, highlighting the significance of factors such as GPA, standardized test scores, and extracurricular involvement in influencing acceptance likelihoods. Specifically, applicants with higher academic achievements, diverse extracurricular backgrounds, and high honors tend to exhibit greater chances of acceptance to more selective institutions. The complete procedure for validating results is detailed below.

The model is trained and tested with three distinct datasets – (a) College/University Profile (b) Student Profile and (c) Student Admissions Outcome. The College/University Profile data is used to determine the student *fit (interests)* match. Student Profile data is used to determine the K-Nearest Neighbors students. Similarly, Student Admissions Outcome data is used to generate the list of colleges with higher chances of acceptance based on similarities in student profiles.

For research purposes, the Student Profile data is further divided into three sub-categories – (a) Students submitting SAT score (b) Students submitting ACT score, and (c) Students submitting no SAT/ACT score (i.e., Test-Optional). The model is trained on these three sub-categories of data. As the new student fit and profile data are entered into the solution as input, it first aims to determine the type of student based on the entered SAT/ACT score, or the Test-Optional preference. The model then processes the profile data of the new student and identifies the K-Nearest Neighbors students based on the similarities of the profiles. As an example, if the new student has submitted an SAT score, then the model identifies 5-Nearest Neighbors students who submitted SAT scores, considering the K value is set to 5. Similarly, if the new student entered an ACT score or plans to go test-optional then the model identifies 5-Nearest Neighbors students with the similarities in profiles. Once the nearest neighbors have been identified based on the student profile matching, the solution then loops through the Student Admissions Outcome dataset to find the colleges/universities where the 5-Nearest Neighbors were accepted or denied, thus generating two recommendation lists for the newly entered student – (a) Acceptance Recommendations and (b) Rejection Recommendations.

The solution further reduces the acceptance and rejection lists by matching the entered (test) student's fit or interest (e.g., type of college interested in) information with the attributes of colleges from the acceptance and rejection lists. The final results produced are a list of colleges/universities where the test student has a high likelihood of acceptance or rejection based on previous students' admission outcomes and those that also match the test student's college interests.

College/University, student, and admissions datasets include data or features as shown in Tables 1, 2, and 3 respectively.

Table 1: College-specific features considered by the Usher-AI recommender system

College ID	College/University Name
School Type	Setting
Region/Location	Size
Students of Color%	Gender Diversity
Academic Calendar	Campus Culture
Offer Merit Scholarship?	NCAA Division
Acceptance Rate (%)	Average GPA
SAT/ACT Requirement	SAT Middle 50%
ACT Middle 50%	

Table 2: Student-specific features considered by the Usher-AI recommender system

Student ID	GPA (Weighted)
SAT Score	ACT Score
Class Rank (decile)	#AP Offered
# AP Completed	# DE Classes
Average AP Score	Research Involvement?
Published Paper?	Community Service?

# Years of Service	School Clubs?
Highest Position Held	Founded Clubs?
Founded Non-Profit?	JV/Varsity Sports?
Captain?	# Years of Sports
Theater/Music?	# of Years Theater/Music
Highest Honor Level	Recommendations Meter
Essays Meter	Intended Major

Table 3: Admissions-specific features considered by the Usher-AI recommender system

Student ID	College/University Name
Outcome (Accepted/Denied)	

The following case study illustrates how the model can be applied in practice.

Step 1: Inputted a test student with a Weighted GPA = 4.62, SAT Score = 1540, Rank = Top 10%, high academic rigor, diverse extracurriculars and awards/honors, Computer Science as the intended major, and interested in colleges of medium to large size.

Step 2: The model identified the 5-Nearest Neighbors student IDs (with K = 5) that submitted SAT scores.

Nearest Neighbors for SAT Data:

[121, 8, 83, 175, 31]

Where the list contains the Student IDs from the Student Profile data (i.e., previous students).

Step 3: The model then identified all the colleges/universities the identified students were accepted to or denied from, generating two college lists.

Acceptance List:

[4, 1, 8, 33, 6, 29, 9, 47, 19, 26, 34, 13, 6, 33, 4, 12, 23, 18]

Rejection List:

[5, 3, 2, 10, 53, 50, 3, 2, 48, 22, 16, 11, 52, 24]

Where the list contains the College IDs from the College/University Profile data and is sorted in the order of acceptance or denial as determined by the Euclidean distance of the nearest neighbor students. In other words, the accepted colleges of the nearest neighbor student (Student ID: 121) are listed first followed by the accepted colleges of the next nearest neighbor (Student ID: 8). The colleges where the 5^{th} nearest neighbor (Student ID: 31) had been accepted are listed at the end.

Step 4: At the end, the program applied the test student's interests or fit (i.e., college size in this case study) to the above college list to generate the final recommendation of colleges where the specific test student has a high chance of acceptance.

Recommended College List:

[4, 1, 33, 47, 26, 13, 18]

Where College ID,

4 = Cornell University – Large Rural
1 = University of California, Berkeley – Very Large Urban
33 = University of California, San Diego – Very Large Suburban
47 = University of Massachusetts, Amherst – Very Large Suburban
26 = Rochester Institute of Technology – Large Suburban
13 = Georgia Institute of Technology – Very Large Urban
18 = Brown University – Medium Urban

The above recommendation was validated against the actual acceptance outcomes of the test student from the previous years' admission cycle. The model-generated recommendation was found to be 87% accurate. In other words, the test student actually had been accepted to 8 different colleges during the previous year's admissions cycle. 7 of the accepted colleges were recommended by the model. Multiple trials were run for the same user scenario changing the K (Number of Neighbors) value between 1 to 6 to determine the optimal value of K. The trial with K = 5 resulted in the most accurate result. Please refer to Figure 7 for the trial results.

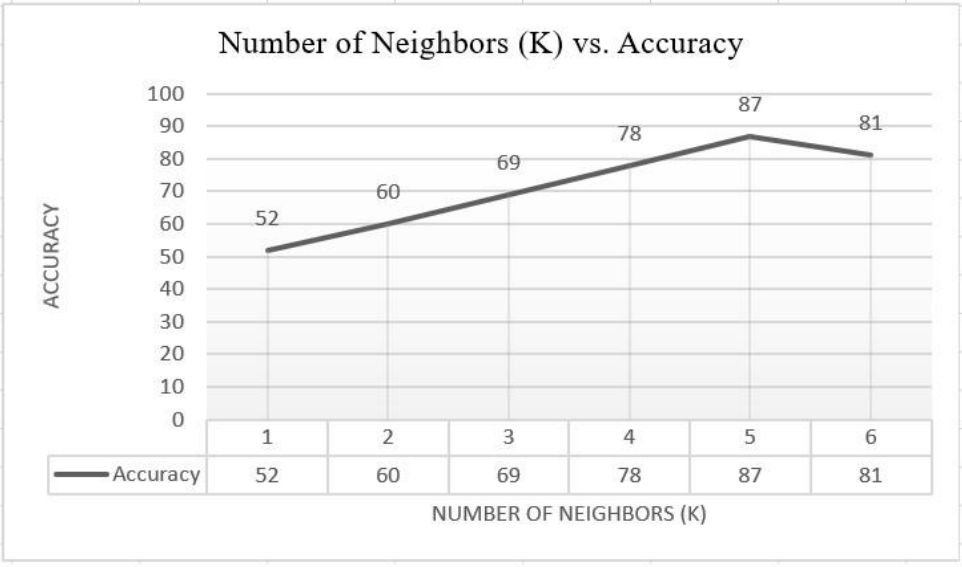

Figure 7: Number of neighbors versus Accuracy to determine the optimal value of K=5

These findings underscore the effectiveness of Usher-AI in providing actionable insights for prospective students. By accurately predicting admission outcomes based on historical student and university data, the solution empowers applicants to make informed decisions regarding their college choices.

5. Discussion

Usher-AI represents a promising solution for aiding prospective students in their application journey. By harnessing AI and leveraging admission data, the model offers valuable insights into colleges with a higher likelihood of admission, thereby facilitating informed decision-making. Moving forward, future iterations may incorporate additional features and datasets to enhance predictive accuracy and usability. Continued development and refinement of such models hold great potential in revolutionizing the college admissions process, making it more transparent and accessible to all aspiring students.

High school students today have access to relatively few tools like Usher-AI, which can be extremely useful in aiding them while making college choices. The goal of this research is to find a solution to this real-world problem using technology. In conclusion, the outcome of this research offers strong evidence that the challenges faced by high school students during the college admissions process can be alleviated using technology, especially by leveraging machine learning algorithm-based solutions. Moreover, the more student input factors the model will continue to include and the more historical admissions data it will be trained on the better its prediction accuracy is expected to be.

5.1. Future Work

To advance the research further, Usher-AI's capabilities can be expanded to encompass a broader range of universities and incorporate increased amounts of historical admissions data. The solution can include additional factors influencing college admissions, so the prediction accuracy is further enhanced. The solution can implement additional functionality to generate a "*balanced*" college list based on the college categorization guideline explained in the *Introduction* section. This can enable students to prioritize colleges during their selection process. Students can be allowed to personalize the number of colleges they wish to include in the recommendation depending on which balanced list can be generated. Additionally, introducing user-friendly interfaces and integrating user feedback mechanisms can further enhance the user experience. Bringing Usher-AI to mobile devices and the World Wide Web could make it more accessible to students worldwide. Ultimately, the goal is to empower applicants with the tools and knowledge needed to navigate the admissions process successfully.

5.2. Acknowledgments

I would like to give special thanks to my mentor, Michael Yang, for his guidance on this research project. He assisted me so much in this process of researching and writing by encouraging me to be rigorous and thorough in my work. He provided guidance far beyond my expertise. I am grateful for his mentoring.

6. References

1. Roberts, K. (2024, March 22). *Tips for Building a Balanced College List*. Retrieved from https://www.linkedin.com/pulse/tips-building-balanced-college-list-kerry-roberts-qrkhe/.
2. Sabky, B. M. (2021). *Valedictorians at the Gate: Standing Out, Getting In, and Staying Sane While Applying to College* (First). Henry Holt and Company.
3. Weinstein, B., & Tedards, S. (n.d.). *The College List – College Finder Series: Part II*. Retrieved from https://www.collegeadvisor.com/resources/college-list/.
4. Weintraub, L. (2023, July 12). *The Ultimate Guide to Making a Balanced College List*. Retrieved from https://blog.collegevine.com/how-to-make-a-college-list.
5. GeeksForGeeks. (2024, January 25). *K-Nearest Neighbor (KNN) Algorithm*. Retrieved from https://www.geeksforgeeks.org/k-nearest-neighbours/.
6. Harrison, O. (2018, September 10). *Machine Learning Basics with the K-Nearest Neighbors Algorithm*. Retrieved from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

7. Author

Disha Raghuvanshi, a junior at Moreau Catholic High School, in California, is enthusiastic about the intersection of business, technology, and data, all infused with an entrepreneurial spirit. She is passionate about simplifying the college admissions process for high school students and aspires to be a data-driven business leader in the future.