**Capstone Project**  
Final Report

**Student Behaviour**

**Analysis**

**Group #8**Archit Navadiya - 0815897  
Dhrumil Patel - 0813662  
Felix Saji - 0803791  
Bhumit Navadiya - 0812548  
Arbaz Shaikh - 082312

Contents

[Abstract 3](#_Toc165063729)

[Introduction 4](#_Toc165063730)

[Literature Review 5](#_Toc165063731)

[Methodology 6](#_Toc165063732)

[Data Flow and Technology 6](#_Toc165063733)

[Survey 7](#_Toc165063734)

[Approach 8](#_Toc165063735)

[Statical Analysis 9](#_Toc165063736)

[Welch’s test: 9](#_Toc165063737)

[Machine Learning 10](#_Toc165063738)

[Natural Language Processing: 11](#_Toc165063739)

[Result: 11](#_Toc165063740)

[Conclusion 15](#_Toc165063741)

[Future Work 15](#_Toc165063742)

[Reference 16](#_Toc165063743)

# Abstract

For students, the college experience is a critical period that is greatly impacted by a variety of circumstances, including demographics, employment history, housing arrangements, College Services, and educational background. Yet, while these elements are recognized as crucial, a comprehensive understanding of their interconnections is still lacking.

By surveying students at St. Clair College's Zekelman School of IT, our study aims to close this disparity. We hope to learn more about the complex connections between demographics, employment history, housing situations, educational backgrounds, and the day-to-day experience of attending college through these surveys.

Our study aims to offer insightful information that can guide efforts to improve the academic experience and promote student achievement by examining these linkages. By carefully examining the collected data, we want to provide actionable suggestions for developing a setting that encourages student involvement, retention, and general well-being.   
  
Our main goal is to support continued initiatives to improve college experiences and provide students with the tools they need to succeed both academically and personally.

# Introduction

College experience plays an important role in the student's lives. This is the time of development, self-awareness, and building of the inner world. Getting a satisfactory college experience is always a priority for students, but the criteria for a perfect college experience may differ significantly among students. We wanted to take a deep dive into what makes students do well in college. We looked at lots of different things like where students come from, how they live if they have jobs, and if they use any services provided by the college.

To get started, we asked students to fill out surveys to tell us about themselves. We asked all kinds of questions, like about their demographics, educational background, prior experience, accommodation, and academic experience. We also asked about their experiences with classes, like if they can switch sections easily or if they feel like they have too much work. Plus, we asked if they know about and use any of the services the college offers. With all this info in hand, we set out to see if there's a link between these different things and how well students do in their classes.

The null hypothesis of our project is that there is no relationship between academic scores and all the factors which are demographics, education and work experience, accommodation type, workload expectations, and services offered by the college.

Our study hypothesizes that all the following factors matter in student life and overall academic performance, we found out that students who have past work experience are good at balancing their work-life balance. And they are more financially stable than those who don’t have work experience.

Moreover, we found out that students who have higher education are quite good in their academic performance. Students who have a part-time or full-time job and have private rooms have overall good experience and focused. This allows them to have a better engagement in their academic and social activities.

Our project was all about making things better for students in college. By understanding what factors matter most for success, we hoped to give colleges ideas on how to support their students better. So, come along with us as we dig into the details and uncover what makes a difference for students in higher education.

# Literature Review

* This research examines a novel intervention aimed at reducing the achievement gap among college students based on socioeconomic status. The intervention strengthened resource utilization and enhanced academic achievement for first-generation students, while also improving the psychosocial outcomes of all students, by highlighting the impact of various backgrounds on the college experience.
* The connection between minority students' tech use, social media, online courses, program of study, satisfaction, and academic performance. Findings showed tech use, online course load and program of study impacted performance, while satisfaction and social media did not. It highlights the importance of tailored approaches and multicultural representation in online education for minority students.
* This study investigated how minority students' technology use, online course enrollment, and program of study influence academic performance, underscoring the need for tailored approaches and multicultural representation in online education.
* It introduces DTLP, a novel deep learning-based approach, to automate student feedback analysis with improved accuracy. By leveraging convolutional neural networks, bidirectional LSTM, and attention mechanisms, DTLP addresses challenges like contextual polarity and word sense variations, outperforming existing systems and providing insights into student feedback.

# Methodology

## Data Flow and Technology

A screenshot of a computer

Description automatically generated

The Data Flow process comprises five key stages: Data Collection, Preprocessing, Transformation, Prediction, and Visualization.

**• Data Collection:** To gather our data, we utilized Microsoft Forms. Our survey was structured into three sections: demographic, accommodation, and academic.

**• Preprocessing:** After setting up the survey, we began collecting data from students, and storing it in an Azure SQL Database. Subsequently, we used Python to clean the data, ensuring its quality, and then re-stored it in the database.

• **Transformation:** In this phase, we delved into our data, employing various machine learning and statistical techniques to derive valuable insights.

• **Prediction:** Once our data transformed, we applied machine learning and Neural Network models to predict students' overall academic scores. Additionally, we conducted Natural Language Processing (NLP) on student feedback and suggestions.

• **Visualization:** Concurrently, we created visualizations to present our insights. Using Tableau, we connected our database and crafted visual representations of our data.

## Survey

We created our survey with the help of Microsoft Forms where in demographics we added several features such as country of origin, age, gender, educational background, and prior experience. In the accommodation section, we asked for accommodation type, rent, and various features Likert scale related to accommodation, and in the academic, we asked them for their current employment status, college services, and a few questions related to academics.

* Data Import: After creating the survey, then connected the survey with the Azure Database with the help of Power Automate, where we created three tables for Demographics, Accommodation, and Academics.
* Data Cleaning: Once the data is stored, the data cleaning process is required the improve the quality of our data. Where cleaning process includes several things such as:
* Province: Various columns require a standard format such as province where few of them are in lower case, upper case, and misspelled. Here we used the Fuzzy Wuzzy library to make it in a standard format and to correct the misspelled province.
* Postal Code Formatting: In this column also some of them are in lower case, and upper case, and no space between the letters. We standardize this by doing all the alphabets in upper case and applying space after three characters as a standard format of Canada’s postal code.
* Field of experience: In these columns, we have various fields where some of which are sub-fields of fields. So, we standardize it and assign all those sub-fields to those fields.
* Data Types: Some of the columns are assigned with incorrect data types. We make sure that they are assigned with correct data types.
* Once the data has been cleaned, we store it back in a new table so that we can use cleaned data for our visualization and ML process.
* We made a pipeline where it will check whether we got a new response or not. If we have a new response, then it will do pre-processing of data and after that, it will be stored in our new table.

## Approach

**Accommodation:** The accommodation table contains all the features of accommodation such as accommodation type, rent, and Likert scale variables such as affordability of accommodation, reliable commute, quality of accommodation, and daily needs. To obtain an overall accommodation score we assigned weights to the Likert scale variable, and we made it data driven as we asked questions where they needed to select which one is more important for them.

A screenshot of a computer

Description automatically generated

**Academic:** Here the academic table contains all information related to the academic and current employment. This table also contains Likert scale questions such as learning resources, section swapping, scheduling, and workload expectations. Similarly, to obtain overall academic scores we assigned weights to the features of the Likert scale based on the priority of students.

A screenshot of a computer code

Description automatically generated

**Demographics:** In the demographics table, we have several columns that are non-numeric. To convert it into numeric we did label encode using the label encoder library. Then removed all the non-numeric columns from the table.

**Data Frame:** After doing all the necessary transformations on all tables we merge them with the help of ID.

## Statical Analysis

We performed statistical analysis to check which variables of our data are impacting academic scores. Here our null hypothesis is there is no relationship between all these features and academic scores whereas our Hypothesis is there is a relationship between some features which is affecting our academic performance.

**Correlation:** This graph represents the relationship between the academic score and other features of the data.

A screenshot of a computer screen

Description automatically generated

## Welch’s test:

Welch's test for feature selection is a statistical method used to identify the most relevant features (or independent variables) for predicting a target variable in a machine learning or statistical modeling context.

In feature selection, the goal is to determine which subset of features will yield the best predictive performance while minimizing complexity and computational costs. Welch's test specifically assesses the importance of individual features by evaluating whether there are statistically significant differences in means between groups (or categories) of a categorical feature concerning the target variable.

While performing Welch’s test we analyzed the features such as rent, affordability of accommodation, duration of classes, workload expectations and college events are have p-values>0.05 which means these features are not necessary to predict overall academic scores. Whereas the rest of the features that have p-values <0.05 are important for predicting the overall academic score

Finally, after performing we can reject the null hypothesis as most of the data features have importance in obtaining the overall academic scores of students.

## Machine Learning

The main reason for using the machine learning regression model is to evaluate which model is best to predict the academic score based on the selected features.

* **KNN Regression:** We first tried the KNN regressor to predict academic scores where we got an MSE: of 0.29 and an R^2 score: of 0.73.
* **Decision tree Regressor:** Then we tried the Decision tree regressor in which we got MSE: 0.22 and R^2 score: 0.78 which is good compared to KNN.
* **Gradient Boosting Regressor:** In this as a best parameter we got max\_depth = 3, min\_sample\_keaf = 4, n\_estimators = 50. Where MSE: 0.102 and R^2 score = 0.90 which is quite accurate in predicting score compared to KNN and Decision Tree.
* **Simple Neural Network:**

We also used a neural network for the regression task where we have a total of 12 hidden layers with a swish activation function. Where we got MSE: 0.22 and R^2: 0.79

* **Convolutional Neural Network:**

Here we have a total of 4 layers and one output layer where for the first layer of Conv1D we set the filter size as 50, the padding is the same, and the kernel size to 3, and for the second layer, we have set the filter as 30. The third layer is for flattening and the last layer is the dense layer. In CNN we got MSE: 0.17 and R^2: 0.84

* **Conclusion:**

Comparing all the three machine learning models, Neural Network and CNN models were the best model among all is the Gradient Boosting Regressor as it has a low MSE score and a high R^2 score.

## Natural Language Processing:

We used NLP for sentiment analysis based on the feedback provided by students regarding their courses, professors, lectures, and college.

To do the sentimental analysis first, we need to do data preprocessing where we first normalize data by making all the lowercase, then we remove all special characters and emojis from the data.

After that we used a library known as stop word to remove all the common English words like the, they, then, is, and many more. Then we did a tokenization for words where every single word was assigned with token.

For Sentiment Analysis we used the library nltk. sentiment.vader which will assign a score to a word from -1 to +1. Based on this it will calculate the score of the feedback provided by the students.

# Result:

We got 191 positive feedback, 32 negative feedback, and 182 neutral feedback.**Demographics Dashboard:** This dashboard represents the demographic details about the students such as country of origin, Gender, Age distribution with work experience, sector worked, and number of students who have experience.

Insights:

* Overall, most of the students have work experience. Where almost half of the students have experience in the IT/Computer field.
* Next most of the students are from India and Canada.
* The experience by age represents that almost every student whose age is more than 26 has experience.

A screenshot of a computer

Description automatically generated

**Accommodation Dashboard:** This dashboard represents the charts of accommodation types, rent for accommodation, and a map of area-wise accommodation scores.

Insights:

* As the graph represents the east side of Windsor has the highest score among all the areas of Windsor.
* The pie chart shows that most of the students prefer to live in a shared room as the price of sharing a room is low compared to all other types of accommodation.

A screenshot of a map

Description automatically generated

**Academic Dashboard:** The last dashboard shows the various KPIs like total students, Academic score, accommodation score, not aware of services, and aware and using it. Then it represents the word cloud for feedback and suggestions. Bar graphs of employment status and type of service used.

Insights:

* This dashboard shows that the overall academic score and accommodation score are 3.1 and 3.3 which is quite good.
* Almost all students know about the services provided by the college and most of them are using it.
* The highest service used by the students is tutoring service, then health and wellness.
* Talking about the employment status then almost half of the students have part-time or full-time jobs.

A screenshot of a computer

Description automatically generated

# Conclusion

As students, we need a system that can help us understand how different things affect our college experience. This system should be able to show us how different factors are connected and predict how they might affect us in the future. Making this system involves several steps: collecting data, processing it, analyzing it, and then showing the results in a way that's easy to understand.

In the beginning, we gather data from via surveys. Then, we tidy up and clean the data to make sure it's good quality and ready for analysis. After that, we use this data to create models that can predict how different things might affect a student's college experience.

Next, we look closely at the data and try to find connections and patterns between different things and how students experience college. We use graphs and dashboards to show these results in a way that's easy to understand. These interactive tools give us a clear picture of how different factors are linked to a student's college experience.

In summary, this system offers a complete solution for unraveling the intricate connections among various factors and a student's college journey. With its ability to shed light on the impact of different factors, the system enables universities and colleges to develop strategies and initiatives that foster a more encouraging and beneficial academic atmosphere. Ultimately, this contributes to improved outcomes for students.

# Future Work

* **Our future steps** involve incorporating **all programs** offered by St. Clair College, **not just limited to Windsor**. We aim to expand our research to include campuses in additional cities, enabling a comprehensive **understanding of student's needs** and preferences **across various programs and locations.**
* Another step is to make who system automate if new data is coming it will automatically clean it and store it in a new table.

# GitHub Link:

<https://github.com/architnavadiya74/students_behaviour_analysis>

# Reference

1. Stephens, Hamedani, & Destin, Closing the Social-Class Achievement Gap: A Difference Education Intervention Improves First-Generation Students’ Academic Performance and All Students’ College Transition. Sage Journals, 2014. <https://doi.org/10.1177/0956797613518349>
2. Yeboah, Alex Kumi, and Patriann Smith. "Relationships between minority students online learning experiences and academic performance." Online Learning 20.4 (2016): n4. <https://eric.ed.gov/?id=EJ1124650>
3. Kanwar, A., Sanjeeva, M. Student satisfaction survey: a key for quality improvement in the higher education institution. *J Innov Entrep* **11**, 27 (2022). <https://doi.org/10.1186/s13731-022-00196-6>
4. Farzana Afrin, Mohammad Saiedur Rahaman, Margaret Hamilton Computer Science and Information Technology, School of Science, RMIT University, VIC, Australia<https://arxiv.org/pdf/2006.07860>
5. Abdi, A., Sedrakyan, G., Veldkamp, B. *et al.* Students feedback analysis model using deep learning-based method and linguistic knowledge for intelligent educational systems. *Soft Comput* **27**, 14073–14094 (2023). <https://doi.org/10.1007/s00500-023-07926->2
6. <https://github.com/JayrajRadadiya/Capstone-Project>
7. <https://scikit-learn.org/stable/index.html>
8. <https://keras.io/guides/>
9. https://stats.libretexts.org/Bookshelves/Applied\_Statistics/Book%3A\_Learning\_Statistics\_with\_R\_-\_A\_tutorial\_for\_Psychology\_Students\_and\_other\_Beginners\_(Navarro)/13%3A\_Comparing\_Two\_Means/13.04%3A\_The\_Independent\_Samples\_t-test\_(Welch\_Test)