Bachelor of Applied Data Science | ADS2001 | Data Challenges 3

Virtual Internships

What performance indicators have the greatest influence on the outcome scores of students?

Report completed by:

Colin Chan (32996616)
Huda Tariq Ahmed Khraiss (32996659)
Simon Sun (30630460)
Fiona Lucy George (33154430)
Tan Nhut Nguyen (33051194)

TABLE OF **CONTENTS**

EXECUTIVE SUMMARY

Internships provide an invaluable setting for students to enhance their acting, critical thinking, learning, and discussion skills in a professional environment. In recent times, virtual internships have gained popularity due to the COVID-19 pandemic and the cost-saving benefits they offer. The "virtual internships'' platform, which is a type of MACROSIM (Massively Adaptive Complex Realistic Online Simulation with Interactive Mentoring), simulates not only the content students should learn, but also the ways some groups of people solve problems - their epistemologies. This platform offers a diverse range of experiences and internships that enable students to gain insights into the thinking processes of scientists, scholars, workers, or even artists. One particularly noteworthy program, which is the program of interest for today, is the virtual internship at Nephrotex, a fictitious biomedical engineering company. Designed to be a condensed 18-hour experience, students participating in this internship work collaboratively as a team, taking on the roles of actual biomedical engineers, with the goal of designing a prototype device to assist patients with kidney failure.

The primary objective of this project is to identify the key factors that significantly influence the outcome scores of the students. Understanding these factors is crucial not only for enhancing students' skills but also for enabling program providers to improve their internship offerings.

Among the main factors that might impact the outcome scores are the implementation approach, mentor participation, and the frequency of student engagement. Different implementation approaches can yield varied effectiveness and, consequently, better outcomes. The mentor's role in guiding and supporting students throughout the internship is of paramount importance as it facilitates reflective discussions, clarifies complex concepts, and provides supportive learning environments. The connection between mentor involvement and the outcome scores, can give us valuable insights into the role of mentorship in the overall success of the internship program. Furthermore, the frequency of students' active engagement during the internship might also be an indicator of their level of involvement. Each of these factors contributes to the final outcome and it is important to understand how each impacts it.

To ensure a comprehensive and successful analysis, several key steps need to be followed. These steps include data wrangling, exploratory analysis, model development, and interpreting the results. It is crucial to adhere to best practices, employ appropriate techniques and algorithms, and rigorously validate the models at each stage of the analysis. By doing so, we can ensure that

the findings are reliable and can be effectively utilised for program improvement and enhancing students' learning experiences.

The data-wrangling process for this project was relatively straightforward, primarily involving the removal of some missing values. During the exploratory analysis phase, it became evident that certain approaches consistently yielded higher overall scores compared to others. This finding has valuable implications for future internships, as it suggests that following the implementation strategies associated with higher scores could lead to improved outcomes.

When examining the relationship between mentor participation and student performance, it was observed that the frequency of mentor involvement did not directly correlate with the students' performance. However, the quality of guidance and support provided by mentors played a significant role in influencing the outcome scores. This highlights the importance of effective mentorship in facilitating a supportive learning environment, fostering reflective discussions, and clarifying complex concepts.

Similarly, the frequency of students' participation and involvement in chat interactions did not necessarily correlate with their overall performance. The emphasis should be placed on the quality of their engagement rather than the quantity. Simply being active in the chat does not guarantee greater productivity or better work. What matters is the depth and meaningfulness of their involvement.

Moreover, when it came to predicting the outcome scores, classification methods appeared to be more effective than regression algorithms. This finding suggests that treating the outcome score as a categorical variable within the range of 0-8 , or as categories low, medium, and high, yielded more accurate predictions. By employing classification techniques, the analysis achieved higher efficacy in determining the success or failure of the prototype device design.

INTRODUCTION

The virtual internship offers an ideal setting for students to enhance their skills in acting, thinking, learning, and engaging in discussions as professionals, especially in the biomedical engineering industry. By collaborating as a team, students will have the opportunity to engage in a wide range of tasks, simulating the role of actual engineers, to design a prototype device to assist patients with kidney failure.

The dataset used in this study was collected through an online platform during a virtual internship program called Nephrotex. The dataset includes chat records from 15 implementations of the program. The chat records were meticulously annotated to determine whether or not they included vital engineering concepts. Specifically, the annotations focused on design moves, design justifications, and their corresponding word counts. The room names were also annotated to classify the discussed topics. Additionally, the dataset includes the outcome score for each student's final design report during the internship.

The objective of this project is to develop a performance model that predicts students' performance on their design reports by leveraging the data extracted from their discussions throughout the internship.

Determining the most suitable modelling approach for predicting the outcome score was one of the main challenges that we faced, as different models may produce different levels of accuracy and other interpretability.

We will begin by assessing the data quality by checking the missing values and duplicates if there are any. Exploratory data analysis is also conducted to gain insight into the relationships between the different features. Subsequently, we will be employing a diverse range of models, in order to maximise our chances of identifying the most accurate and reliable approach with the lowest error rate. Then, we will use text analysis to extract valuable insights, and see if it helps improve our models. In the final step, the results will be interpreted.

DATA QUALITY

When working with any dataset, one initial step involves assessing the quality of the data. This involves examining factors such as missing values, duplicates, ensuring correct data types, and identifying and handling outliers that could potentially be misleading.

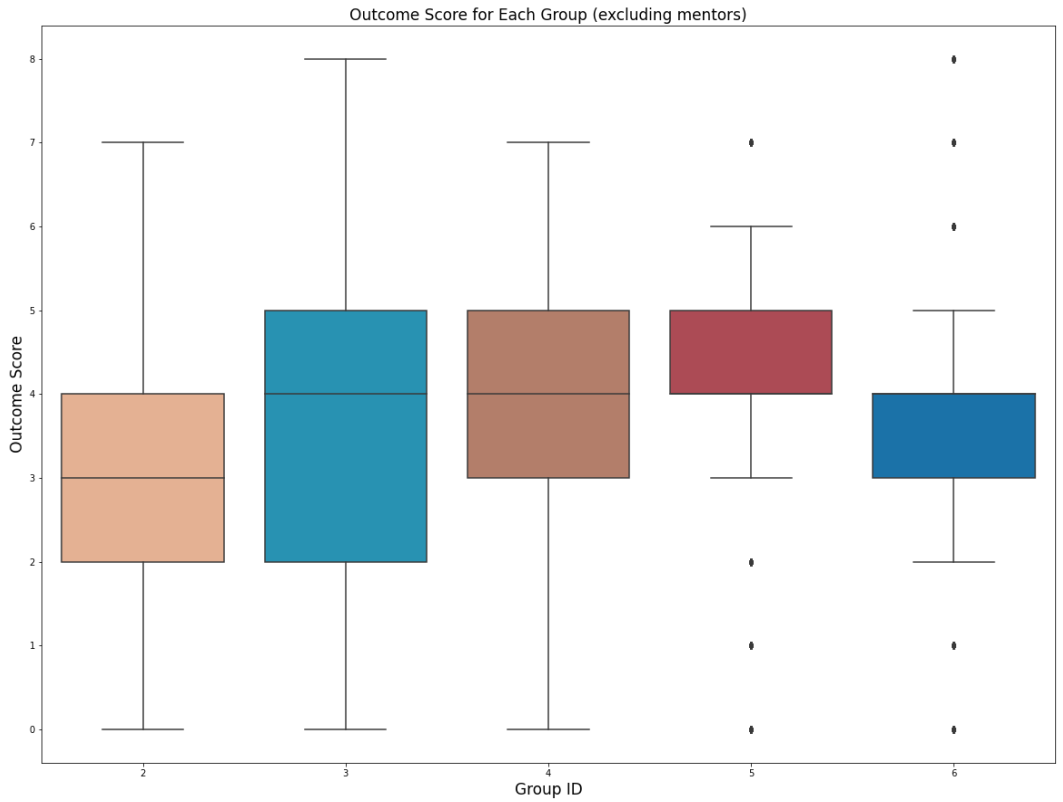
Luckily, the virtual internship dataset proved to be of high quality. All variables were in the appropriate data types, there were duplicates, and we did not come across any outliers. Additionally, we faced a small number of missing values -only four- which we effectively addressed by dropping those observations, so there was no need for additional data imputation techniques.

However, the data set needed to be aggregated, meaning that each student appears only once instead of having multiple observations for each student. This would be helpful for both data modelling and exploratory data analysis.

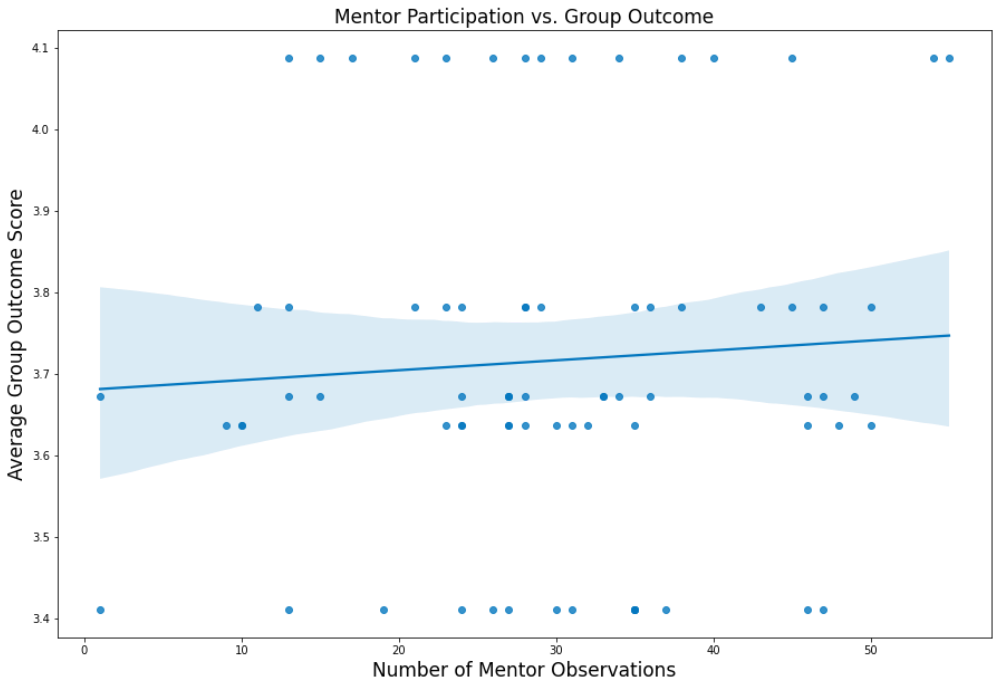
EXPLORATORY DATA ANALYSIS

Before the application of any prediction models, an exploratory analysis should be run to have a clear understanding of the data set and to check the relationship between different variables and the outcome score. Data visualisation will be mainly used in this step in order to bring out the most informative aspects of the data to easily understand what is going on and identify any patterns or trends in the data.

Since students were working in groups, the first thing to look at is the overall achievement of each group. The virtual internship consisted of five groups, each consisting of 71 to 78 students. Additionally, there were 15 mentors assigned to each group of students to provide guidance and support throughout the internship. To gain insights into the performance of the groups, a box plot visualisation was used. From the boxplot, it is clear that most of the groups achieved an average score between 3 and 5, indicating a moderate level of achievement among the students. Interestingly, group number 3 stood out with the highest maximum score of 8, indicating exceptional performance by students within the group. However, both groups 5 and 6 showed relatively minimal variation among their students' scores. This could be attributed to effective collaboration, cohesive teamwork, or consistent mentorship within these groups.

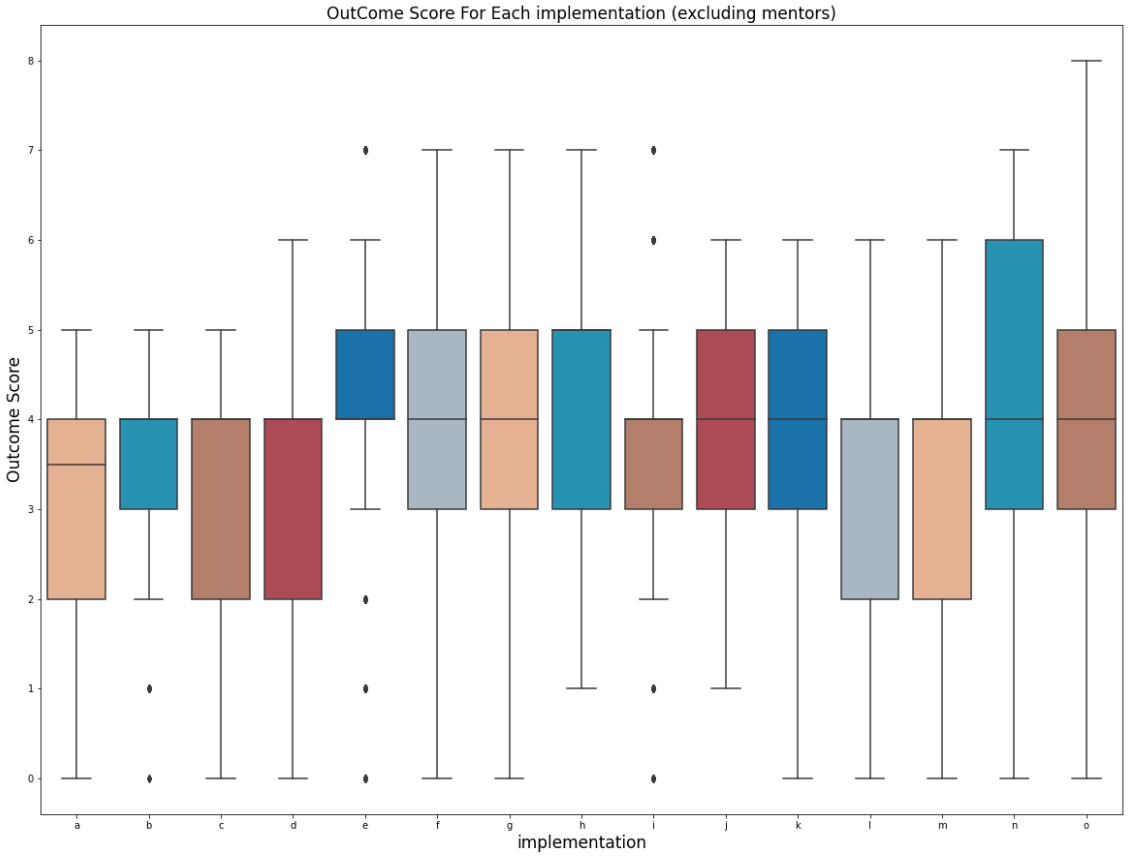


With the considerable number of mentors assigned to each group, it was important to understand how their participation influenced the overall performance of the students. Thus, a line plot is used to visualise the variations in mentor participation levels and how they impacted the average outcome scores achieved by each group. It is clear from the plot that some mentors participated less frequently, yet their groups still managed to achieve good average outcome scores. This suggests that these mentors are able to make a significant positive impact despite their limited participation. Conversely, other mentors actively participated in the chat but did not witness a corresponding increase in their group's outcome score. This indicates that the quantity of mentor participation alone does not guarantee higher group performance. Other factors, such as the quality of guidance and support provided, may have influenced the outcome scores. However, when considering the overall trend, mentor participation generally correlates positively with the group outcome scores.



While looking at the overall scores for each group can provide an initial understanding of their performance, delving deeper into the analysis by considering different implementations offers more meaningful insights. That is because implementations represent the way or the approach the student has used to design the prototype device that should assist patients with kidney failure. While the group ID is a numerical identifier lacking detailed information about the student's individual achievements. Within the virtual internship, there are 15 implementations,

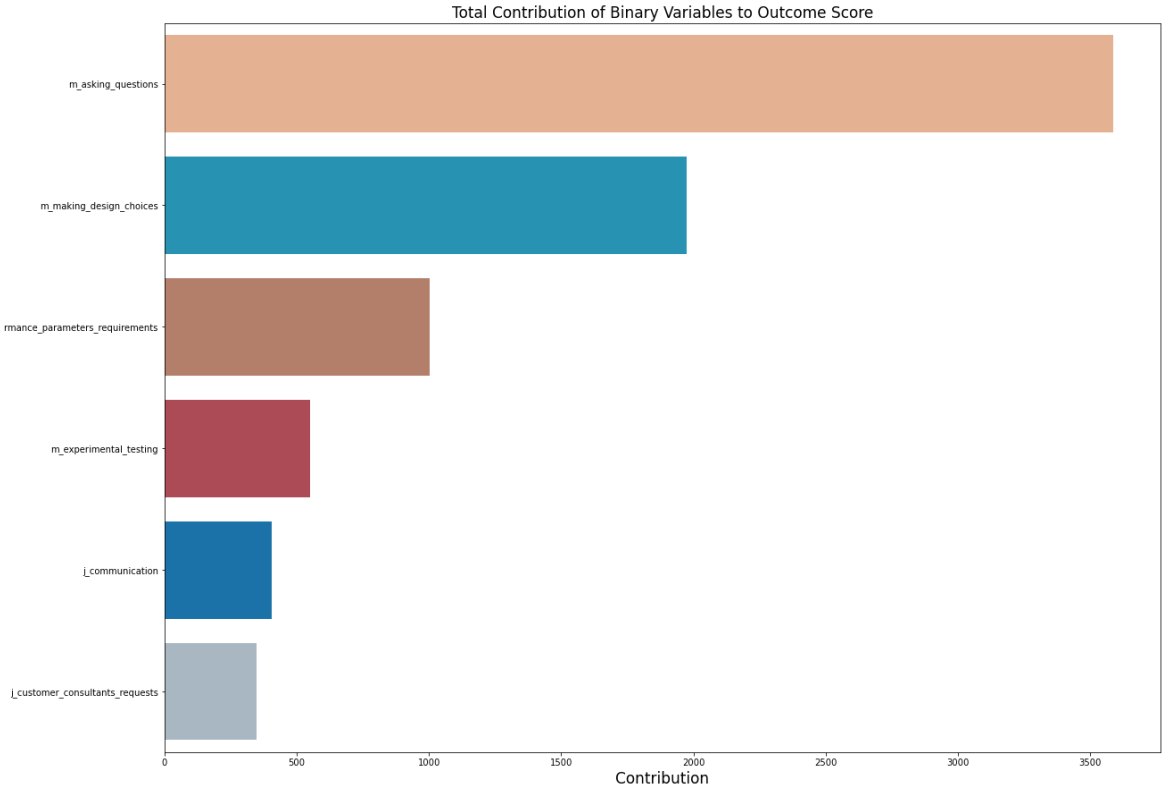
each implementation with a varying number of students ranging from 18 to 32 with 1-3 mentors. Most implementations achieved an average outcome score of 4. Whereas implementation O has the highest maximum of 8, suggesting that the approach employed in this implementation was successful in meeting the objectives of designing the prototype device.



Moving on to some basic text analysis, the dataset presents binary columns that offer valuable information regarding the presence or absence of various topics discussed in the chat. These binary columns are a fantastic way to get a general sense of the most common topics that were discussed. It was interesting to note that the most common topic in the chat was related to asking questions, which suggests that the participants were actively engaged in the learning process and were curious about the material.

Following closely behind, comes the topic of making design choices for the prototype device on which they were working. This is a crucial step in the learning process as it involves making important decisions that affect the final design of that prototype device. And seeing that this topic was not too far behind asking questions indicates that the participants were actively involved in both the theoretical and practical aspects of the project.

Analysing these binary columns provides valuable insights into the topics that were most important to the participants and can guide the next step which would be the predictions and modelling. It is important to consider the context of the data and the goals of the project in order to draw meaningful conclusions from the analysis.



MODEL DEVELOPMENT

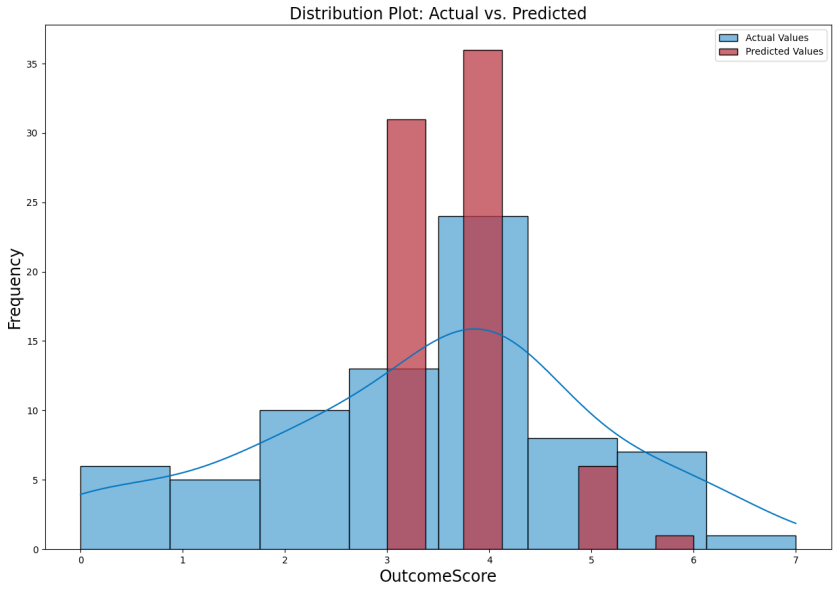
For the modelling aspect, we tried regression and classification methods, so we pursued five different types of algorithms: **Linear Regression, Logistic regression, Decision Tree, Random forests** and **Support Vector Machines (SVM)**.

INITIAL MODELS

LINEAR REGRESSION MODEL

The first model that we attempted was a linear regression model, for the estimation of students' outcome score based on these performance indicators: 'wordCount', "m_experimental_testing", "m_making_design_choices", "m_asking_questions", "j_customer_consultants_requests", "j_performance_parameters_requirements", "j_communication", as well as the chat room names that the students participated in.

We initially split the dataset that had been aggregated by 'userID' and 'implementation' into training and testing sets, where the training set was used to fit the linear regression model, while the testing set was used to evaluate the model's performance.



A training score of 0.127 and an R2 value of 0.108 was obtained, indicating that the model is not fitting that well into the training data, and only 10.8% of the variability in the outcome score

could be explained by the features provided. Furthermore, the model yielded a testing score of -0.047, suggesting that the model's performance on the testing data was not as accurate. It had an R2 value of -0.128, which implied that the linear regression model did not perform well in explaining the variance in the unseen data.

Based on the poor performance of this regression model, we concluded that employing classifier models would be more useful in understanding the extent of which factors contribute to the classification decision as they assign inputs to specific classes, which in this case is each possible outcome score. Hence, we decided to explore the following four classifier methods in an attempt to produce a final model with higher accuracy scores.

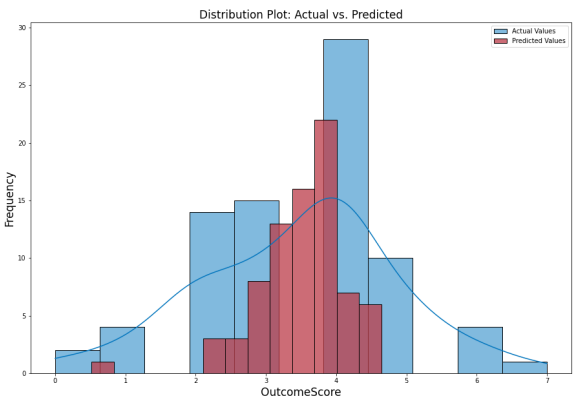
CLASSIFICATION MODELS

For the following models, we based our outcome score predictions based on the these performance indicators: 'wordCount', "m_experimental_testing", "m_making_design_choices", "m_asking_questions", "j_customer_consultants_requests", "j_performance_parameters_requirements", "j_communication". We chose to avoid the variables for chat room names as we found that there was little to no correlation between the two. Our findings are listed below.

Random Forest

Training score: 0.857 | R^2: - 0.023
Testing score: - 0.070 | R^2: - 0.070

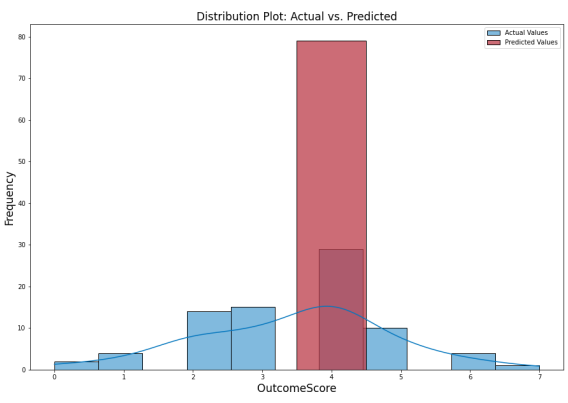
Accuracy: N/A



Decision Tree

Training score: 0.399 | R^2: 0.036
Testing score: 0.367 | R^2: - 0.147

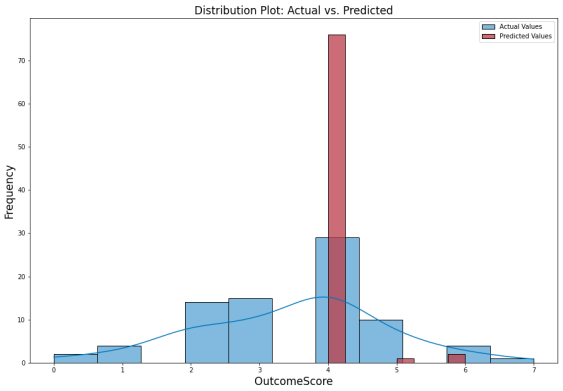
Accuracy: 0.367



Support Vector Machines (SVM)

Training score: 0.406 | R^2: - 0.022
Testing score: 0.342 | R^2: - 0.127

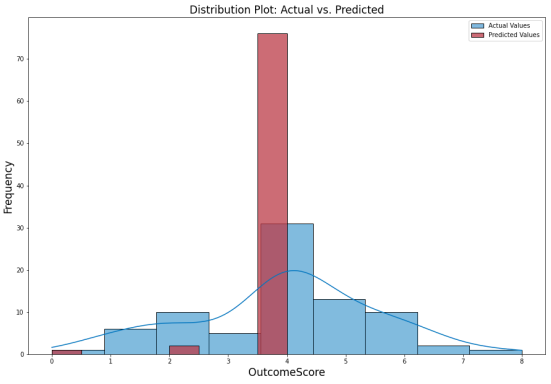
Accuracy: 0.342



Logistic Regression

Training score: 0.387 | R^2: - 0.085
Testing score: 0.392 | R^2: - 0.065

Accuracy: 0.392



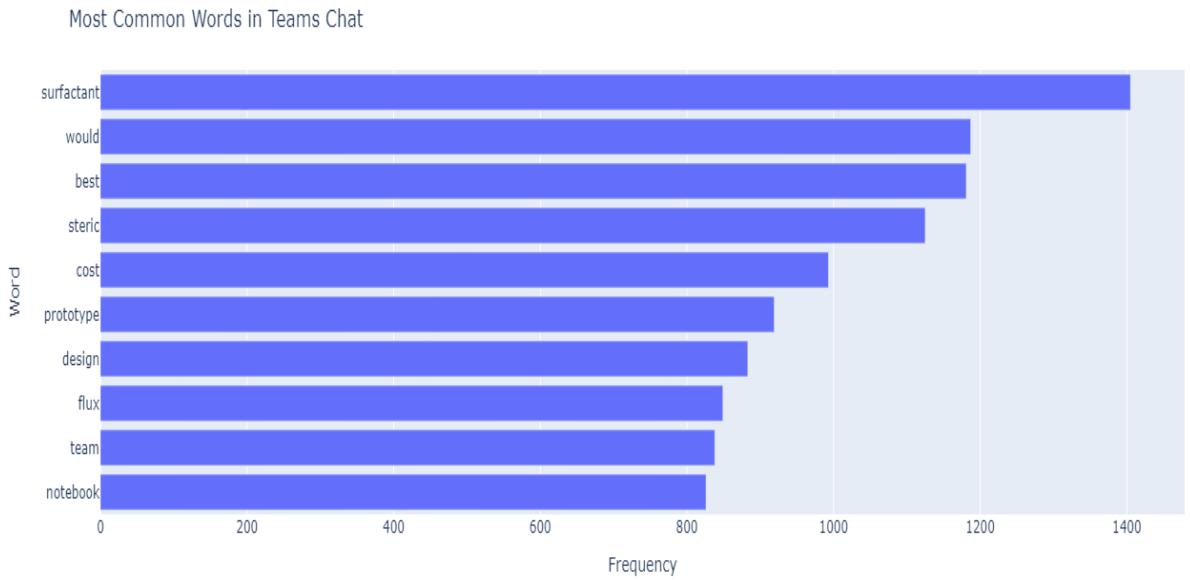
Unfortunately, none of the models produced scores that were deemed satisfactory for accurately predicting the students' outcome scores. As a result, our objective of identifying the performance indicators with significant impact on the outcome score, remains unfulfilled. Recognising this limitation, we proceeded to explore alternative approaches in an effort to enhance the performance of the models and obtain more reliable insights.

EVALUATION METHODS FOR MODEL PERFORMANCE

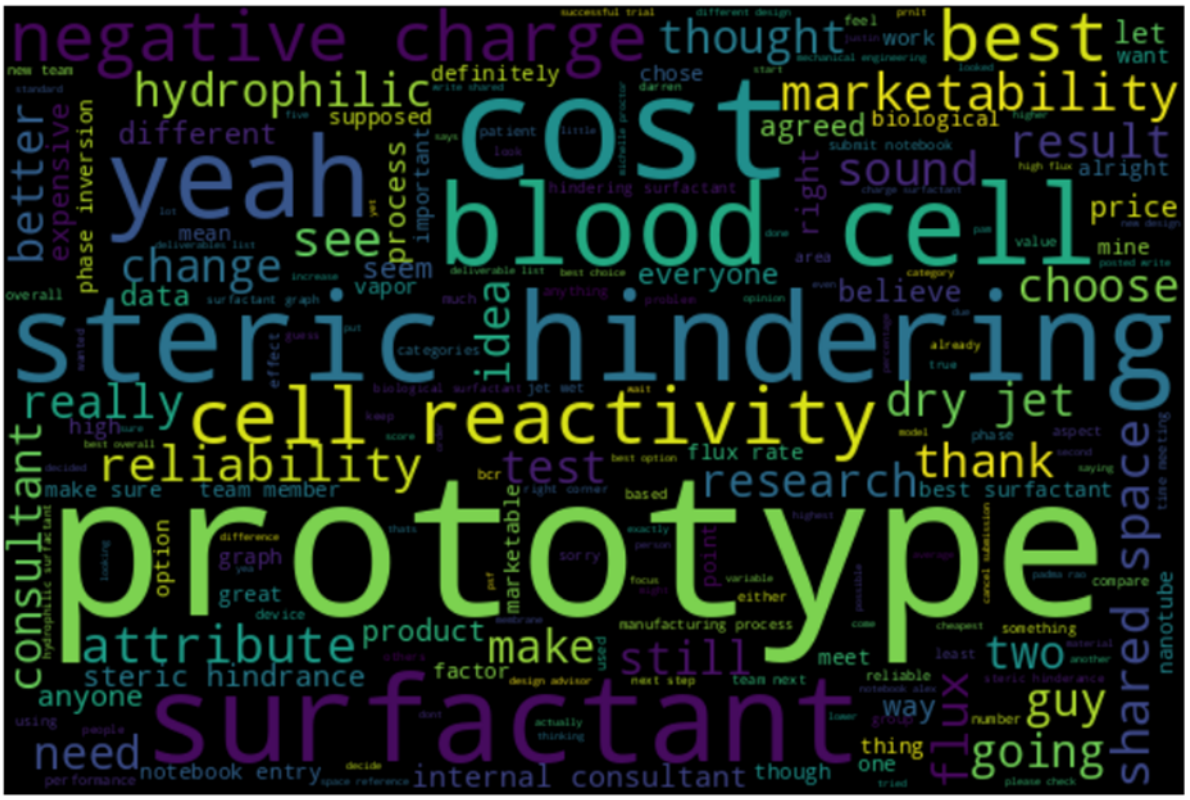
TEXT ANALYSIS

The data in the content column, which is basically the students' chats with their mentors and with each other, have the most potential to provide valuable insights into how their reports will perform. However, it was also the most challenging task. To be able to extract the relevant information from the text and to use it in predictions required high-level techniques. Initially, basic techniques were used from the NLTK module to establish a foundation of what was possible with the data.

The frequency distribution of notable words was created by initialising a variable that contained all the words in the messages. Each character was changed to lowercase and a list of stop words from the NLTK module was used to only include meaningful words in the distribution. A stop word can be considered words used in sentences that do not provide any insightful information, for example, “you”, “then”, “give” etc... The top ten most frequently used were then plotted into a bar chart.



A simple Word Cloud was also generated to visually map what the most frequently used words were in each chat.



In the end, the Word Cloud and the frequency distribution of words were difficult to incorporate into the prediction of the entire report as the high frequency of specific words that can be considered “productive” to the report was not indicative of high performance.

To further investigate the engagement of the participants, sentiment analysis was conducted on the content variable. Sentiment analysis is a technique in natural language processing (NLP) that categorises text as positive, negative, or neutral. Our aim was to analyse the speech patterns and utterances of each participant and mentor to determine if they had any impact on the outcome score for each student.

To perform sentiment analysis, we utilised the TextBlob library in Python, which provides built-in sentiment analysis tools. TextBlob evaluates each sentence and assigns a polarity integer: a positive value indicates positivity, a negative value indicates negativity, and a zero denotes neutrality. Afterwards, the sentiment strings are hot encoded to be represented as 1, -1 or 0.

Using the aggregated data set that we have, each participant had a total content integer representing the overall "positivity" or "negativity" of their language. However, we found that there was no significant relationship between the tone and emotion expressed through a participant's words and their outcome score. This lack of correlation can be attributed to the fact

that sentiment analysis is primarily used in opinionated texts, such as opinion articles and customer feedback, rather than in the context of a workplace environment. Therefore, the use of sentiment analysis may not have yielded meaningful insights. Consequently, we decided to proceed with our models without incorporating sentiment analysis.

An improved approach to text analysis would be to conduct productivity analysis which would evaluate the efficiency and effectiveness of each participant. However, the issue becomes apparent as there does not exist a library that has predetermined language as productive or not; even if such a library exists, it would need to be within the realm of biomedical engineers and kidney failure as the definition of productivity changes in every context. We would need to create our own library and model to determine productivity, which would be beyond the scope of this project.

OUTCOME SCORE CATEGORISATION - GRADE VARIABLE

Originally the models were trained to predict the outcome score in the range 0-8, however, this proved to have poor testing results. In order to improve those results, we created a new column entitled "Grades", which categorises 0-8 scores into "low", "medium", and "high", thereby we will be predicting the overall performance rather than the actual outcome score. As we considered scores less than or equal to 2 as low, scores greater than 2 and less than or equal to 5 as medium, and scores greater than 5 as high.

By converting the target variable from a continuous numeric variable to a categorical value, the test results were significantly higher, this could be due to a couple of reasons:

1. **Target Variable:** By creating the “Grade” variable to have distinct ranges and boundaries it meant that the variable was more appropriate to fit a classification model. Testing scores from models such as SVM, decision trees and logistic regression greatly benefited from this as it was better suited to accurately predict each category.
2. **Non-Linear Relationships:** It is difficult to assume a linear relationship between the features and the target variable. By assuming a non-linear relationship and utilising a categorical target variable, the models, in particular SVM can form more accurate boundaries for predictions.

3. **Data Distribution:** Each model had varying complexities that were better suited for different data distributions. The data distribution of the outcome scores as numerical scores may not have been appropriate compared to the data distribution of the “Grades” variable. The conversion of the outcome score variable may have slightly altered the distribution to better align with the scope of each of the models.

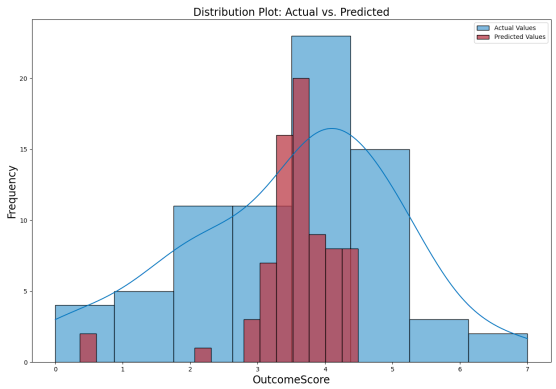
RESULTS

After introducing the Grade variable which categorises the outcome score into high, medium, and low categories, we pursued three different classification models: **Logistic regression, Decision Tree,** and **Support Vector Machines (SVM)**. From these three classification models, the model that had the highest accuracy rate, and lowest error rate was considered the best model.

Support Vector Machines (SVM)

Training score: 0.655
Testing score: 0.620

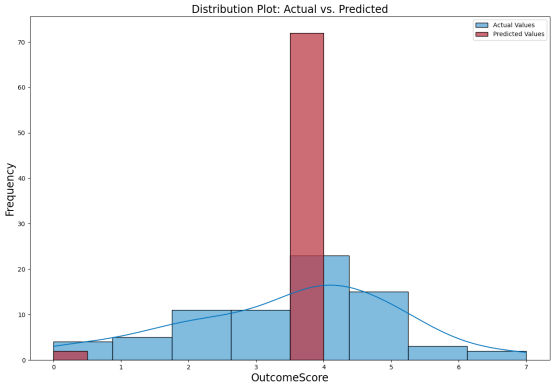
Accuracy: 0.620



Logistic Regression

Training score: 0.658
Testing score: 0.696

Accuracy: 0.696



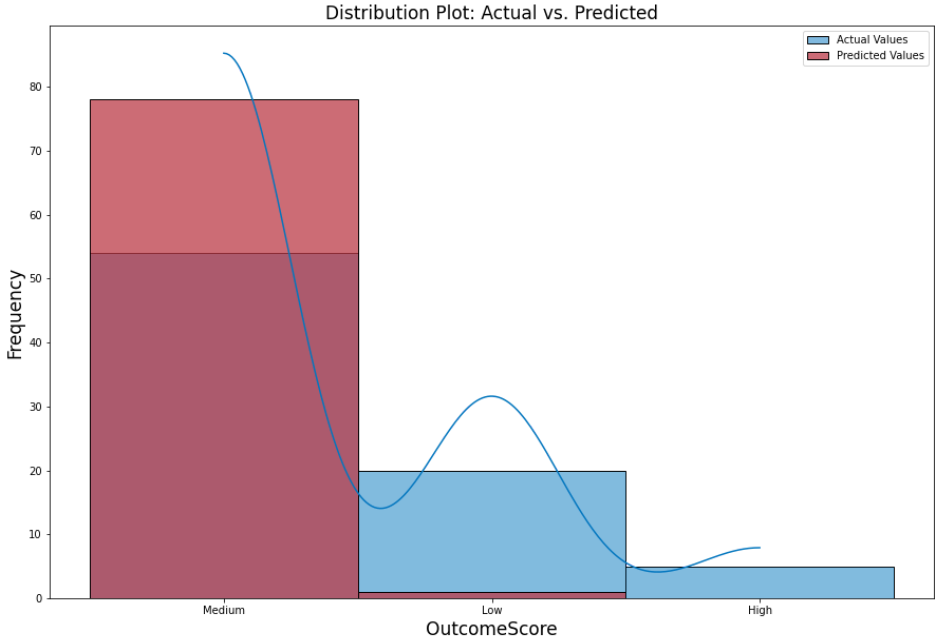
THE MODEL WITH OPTIMAL PERFORMANCE

Decision Tree Model

Being based off the same aggregated data, the decision trees model estimates the students' overall performance based on these indicators: 'wordCount', "m_experimental_testing", "m_making_design_choices", "m_asking_questions", "j_customer_consultants_requests", "j_performance_parameters_requirements" and "j_communication".

The decision tree model yielded a training score of 0.658, indicating a reasonable fit to the training data and capturing certain patterns within the input features and their relationship with

the overall performance. When evaluated on unseen data, the model achieved a testing score of 0.696, suggesting that it performs relatively well on new instances, demonstrating good generalisation capabilities.



Furthermore, the decision tree model achieved an accuracy rate of 0.69, highlighting its effectiveness in predicting the outcome score based on the selected features. This accuracy suggests that the model is able to classify the outcome scores into categories of "low", "medium", and "high" with a considerable degree of success.

Although logistic regression, and support vector machine models both exhibit satisfactory performance, the decision tree model showed the best accuracy. While this is the best approach we tried, we believe that there is still room for further improvement and refinement in the predictive modelling process.

CONCLUSION

The aim of the project was to develop a model to predict the student's report performance on the virtual internship program by extracting data from various features and discussions. By exploring the data and running different techniques on models, we gained valuable insights into what factors influenced outcome scores and ultimately the success of the virtual internship program.

The exploratory data analysis revealed multiple key findings. Firstly, the implementation approach had a significant impact on the student's outcome scores. Certain implementations had consistently higher and lower outcome scores. The implementation approach has a significant relationship with outcome scores, indicating that students that had a more efficient implementation strategy obtained better results.

Secondly, the impact of mentors and their participation played an important role in a student's performance. However, it was found that the frequency of mentor participation did not influence the student's score, but rather it was the quality of the mentorship that resulted in better performance.

With model development, multiple algorithms were explored with logistic regression producing the highest accuracy for predictions at 35.24%. Other models such as decision trees and support vector machines produced similar results at 33.78% and 31.08% respectively. The accuracy was poor due to the nature of the models and the type of target variable. Thus, we tried to come up with ways to evaluate the model performance.

We have seen that the text analysis could contribute greatly to the models in terms of accurately predicting the students' report performance, however, it was one of the most challenging and complex tasks for us. Through basic NLTK techniques, the results were similar to the mentorship factor; high frequency did not correlate with higher outcome scores, but the productivity and quality of engagement from students. After that we introduced the "Grade" variable. With the introduction of "Grades", the models produced more accurate results, with decision trees giving a 69.6% accuracy rate.

Overall, the findings highlight that it is crucial to have an effective implementation approach and "quality over quantity" mentorship and student engagement. These features significantly influenced and increased the outcome scores and the success of the virtual internship program.

While these features were analysed to have the most importance in determining the success of this internship program, it is difficult to conclude that these factors can be applied as a generalisation for all virtual internship programs. However, maximising the above factors can lead to valuable opportunities and potential for a successful internship program.

REFERENCES

Nephrotex. (2013). www.virtualinterns.org. https://www.virtualinterns.org/nephrotex/

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.

Waskom, M., Botvinnik, Olga, O'Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, … Qalieh, Adel. (2017). mwaskom/seaborn: v0.8.1 (September 2017). Zenodo. https://doi.org/10.5281/zenodo.883859

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."

Oesper, L., Merico, D., Isserlin, R., & Bader, G. D. (2011). WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. Source Code for Biology and Medicine, 6(1), 7.

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Inc., P. T. (2015). Collaborative data science. Montreal, QC: Plotly Technologies Inc. Retrieved from https://plot.ly