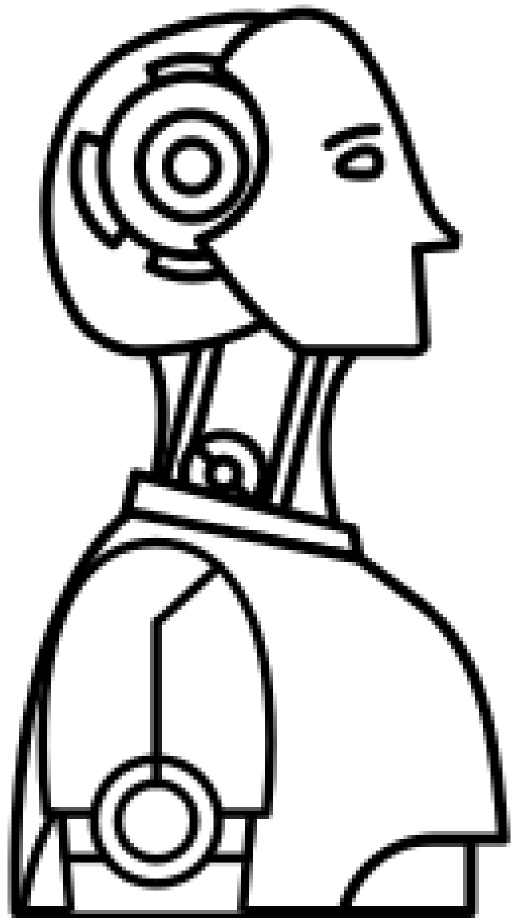
# VIRTUAL INTERNSHIPS

## PRESENTATION




GROUP VI2

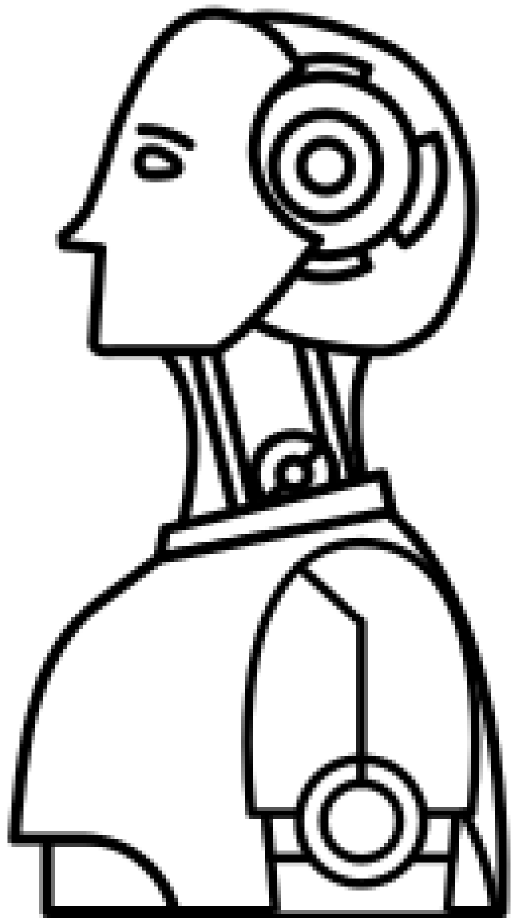
PRESENTED BY

Gue Zhen Xue (33521352)

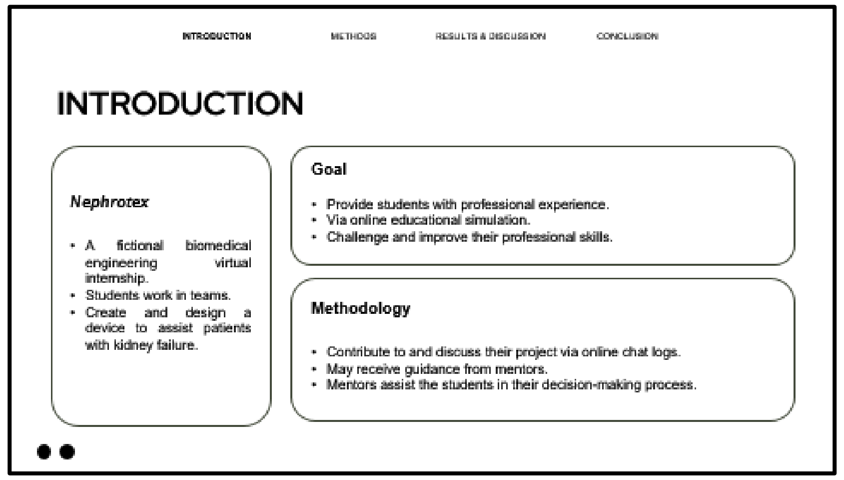
Andres Xue (34987274)

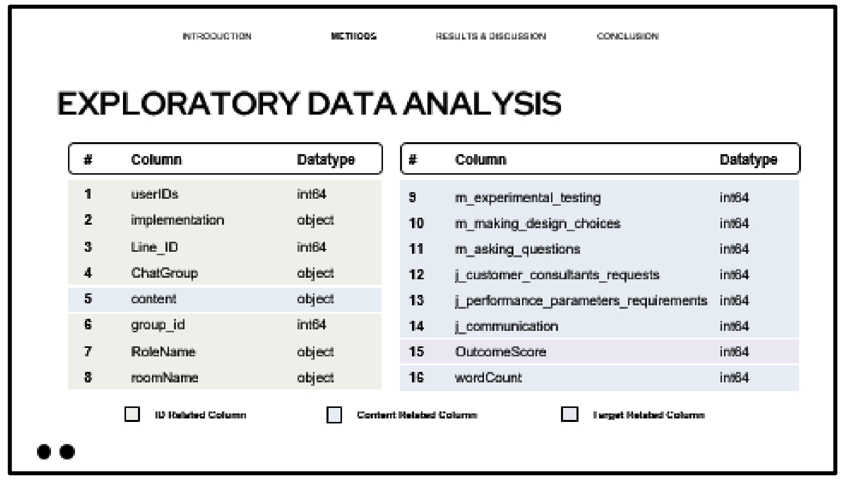
Zohaib Javed (34290826)

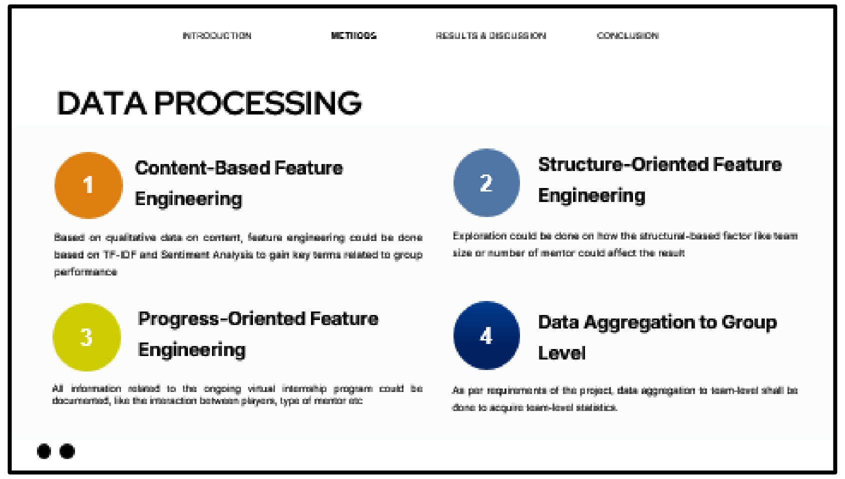
Denisha Fam Wen Hsiu (34091637)

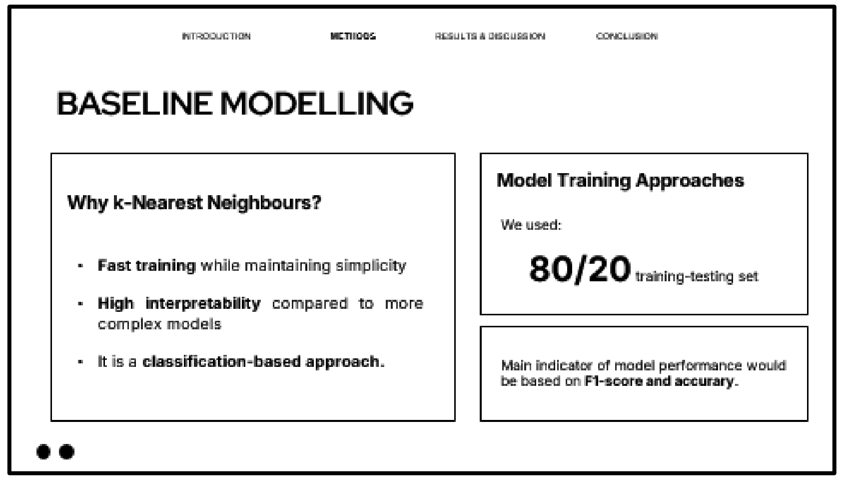



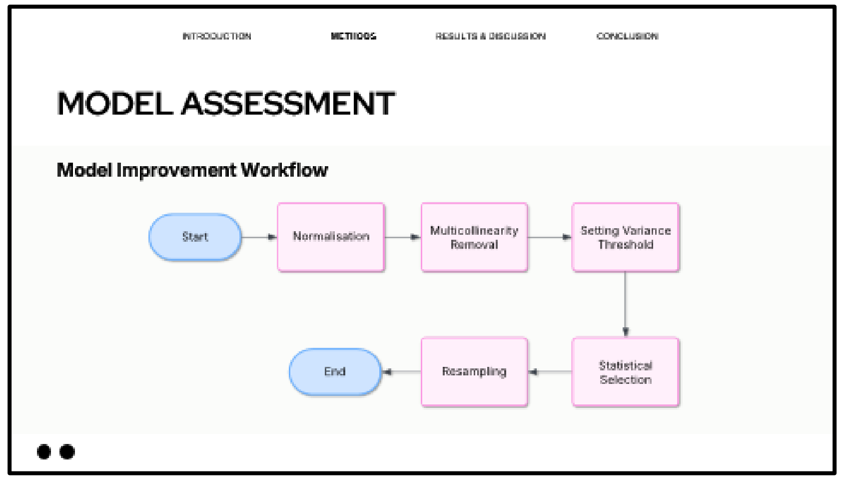
# TABLE OF CONTENTS

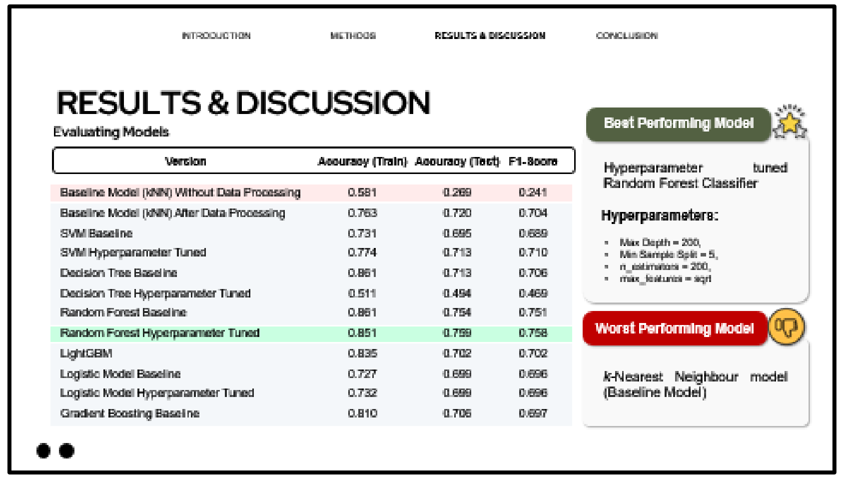


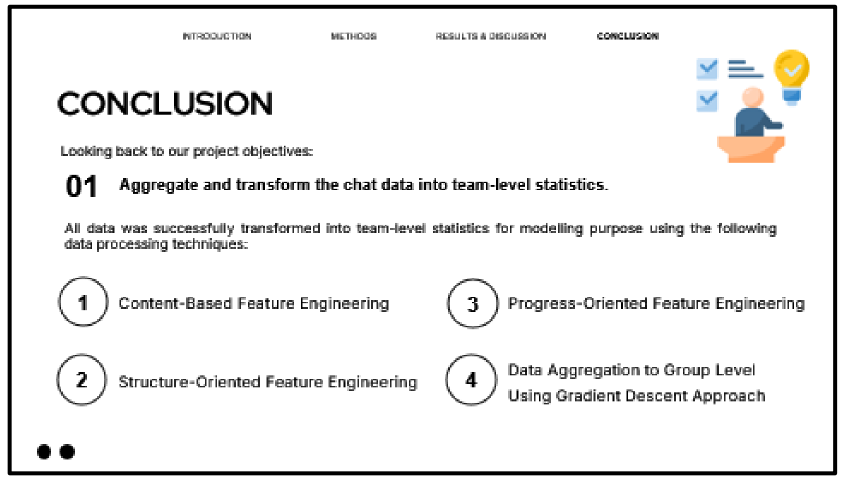



# INTRODUCTION

### *Nephrotex*

- A fictional biomedical engineering virtual internship.
- Students work in teams.
- Create and design a device to assist patients with kidney failure.

### Goal

- Provide students with professional experience.
- Via online educational simulation.
- Challenge and improve their professional skills.

### Methodology

- Contribute to and discuss their project via online chat logs.
- May receive guidance from mentors.
- Mentors assist the students in their decision-making process.

# BACKGROUND

## Description

- Students who partook in educational stimulation of a virtual internship had to come up with a biomedical engineered device that will serve to assist individuals with kidney failure.
- The challenge was to balance their engineering work whilst mastering the underlying biomedical aspects of the project.

## General Tasks

- Background research and understanding stakeholder needs
- Testing and evaluating their prototypes
- Justify their creative and technical decisions

# RAW DATA

## Dimensions:

19181 rows x 17 columns

## First Few Observations of Data:

| | userIDs | implementation | Line_ID | ChatGroup | content |
|---|---|---|---|---|---|
| 1 | 1 | a | 1 | PRNLT | Hello team. Welcome to Nephrotex! |
| 2 | 1 | a | 2 | PRNLT | I'm Maria Williams. I'll be your design advisor for your internship. |
| 3 | 1 | a | 3 | PRNLT | I'm here to help if you have any questions. |
| 4 | 1 | a | 4 | PRNLT | Please introduce yourselves with the name you prefer to be called. WorkPro records all the work we do, and we review it with an external consultant to improve the quality of our internship program. So we ask you to use you |
| 5 | 1 | a | 5 | PRNLT | I just want to make sure everyone has found the chat interface. Please send a chat to "check in" with the group. You can make your chat window bigger by clicking the + icon in the top right corner. |

| group_id | RoleName | roomName | m_experimental_testing | m_making_design_choices | m_asking_questions | j_customer_consultants_requests | j_performance_parameters_requirements | j_communication | OutcomeScore | wordCount |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | Mentor | Introduction and Workflow Tutorial with Entrance Interview | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 5 |
| 2 | Mentor | Introduction and Workflow Tutorial with Entrance Interview | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 11 |
| 2 | Mentor | Introduction and Workflow Tutorial with Entrance Interview | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 9 |
| 2 | Mentor | Introduction and Workflow Tutorial with Entrance Interview | 0 | 0 | 0 | 1 | 0 | 0 | 4 | 51 |
| 2 | Mentor | Introduction and Workflow Tutorial with Entrance Interview | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 39 |

# OBJECTIVES

01 **Aggregate and transform the chat data into team-level statistics.**

02 **Build predictive models to predict final report scores based on team communication behaviors.**

03 **Interpret the results of the models to understand how communication features relate to the team report performance.**

# EXPLORATORY DATA ANALYSIS

| # | Column | Datatype |
|---|---|---|
| 1 | userIDs | int64 |
| 2 | implementation | object |
| 3 | Line_ID | int64 |
| 4 | ChatGroup | object |
| 5 | content | object |
| 6 | group_id | int64 |
| 7 | RoleName | object |
| 8 | roomName | object |

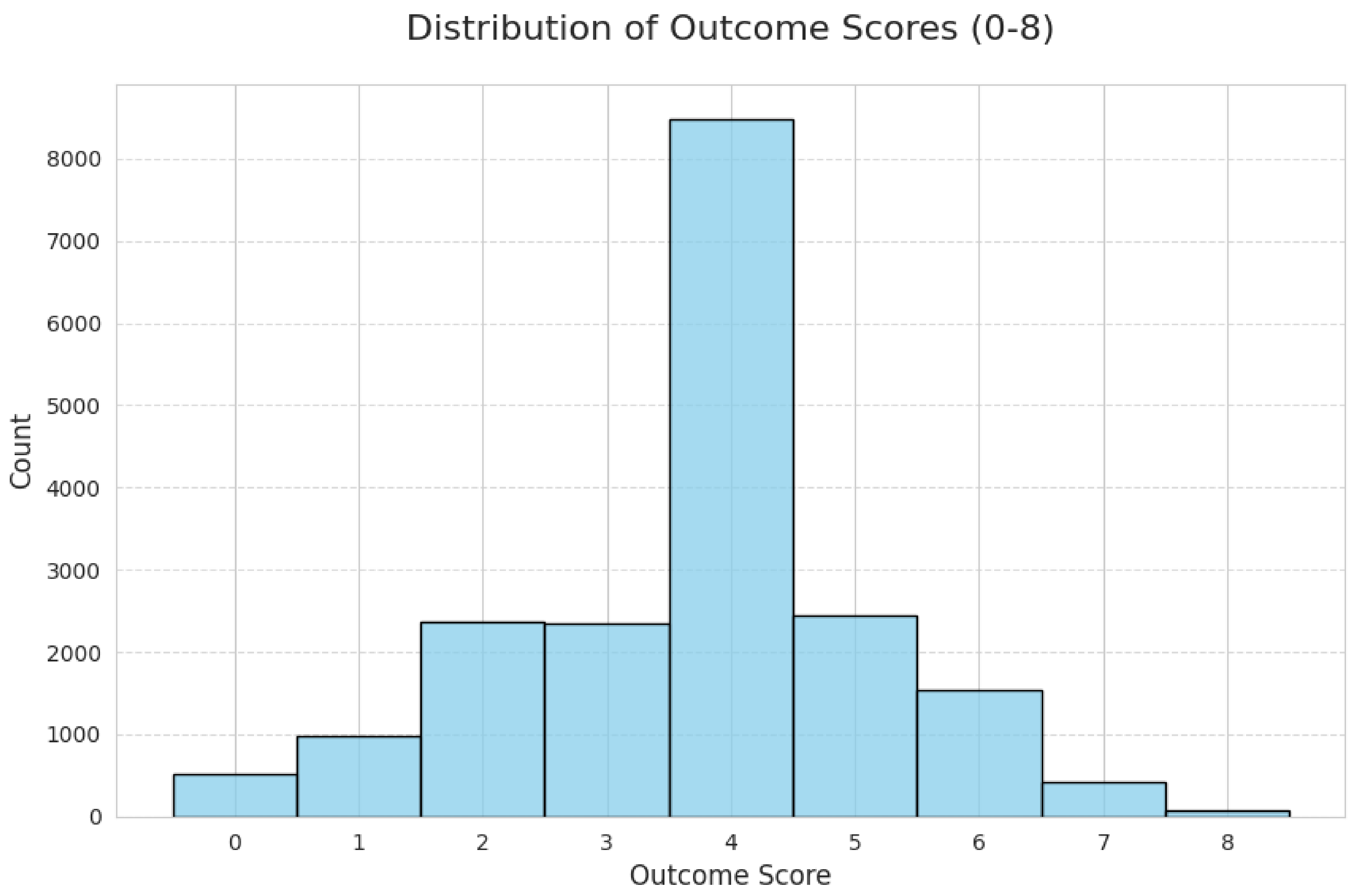
| # | Column | Datatype |
|---|---|---|
| 9 | m_experimental_testing | int64 |
| 10 | m_making_design_choices | int64 |
| 11 | m_asking_questions | int64 |
| 12 | j_customer_consultants_requests | int64 |
| 13 | j_performance_parameters_requirements | int64 |
| 14 | j_communication | int64 |
| 15 | OutcomeScore | int64 |
| 16 | wordCount | int64 |

ID Related Column

Content Related Column

Target Related Column

# EXPLORATORY DATA ANALYSIS



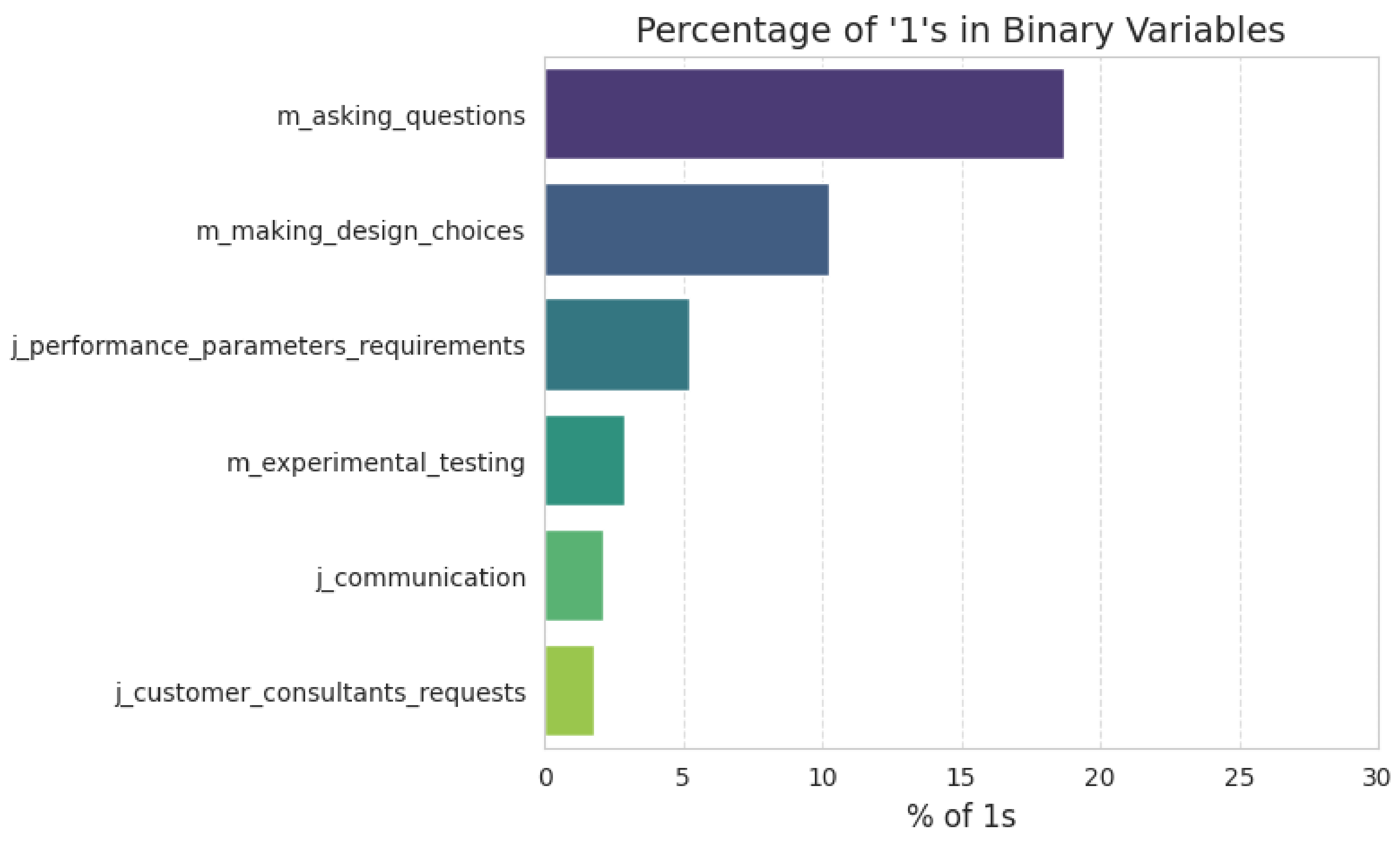

## Observations

- A very high peak appeared at 4
- Moderate decline at 0, and 8

## Risk

- Potential imbalance of the data
- Modelling might favour to class 4 instead of other classes

# EXPLORATORY DATA ANALYSIS




## Observations

- Overall **low occurrence**
- m_asking_questions with **highest** occurrence
- j_customer_consultants_requests with **lowest** occurrence

## Risk

- Low occurrence of these binary variables might cause severe **underfitting** on models

# DATA PROCESSING

## 1 Content-Based Feature Engineering

Based on qualitative data on content, feature engineering could be done based on TF-IDF and Sentiment Analysis to gain key terms related to group performance

## 2 Structure-Oriented Feature Engineering

Exploration could be done on how the structural-based factor like team size or number of mentor could affect the result

## 3 Progress-Oriented Feature Engineering

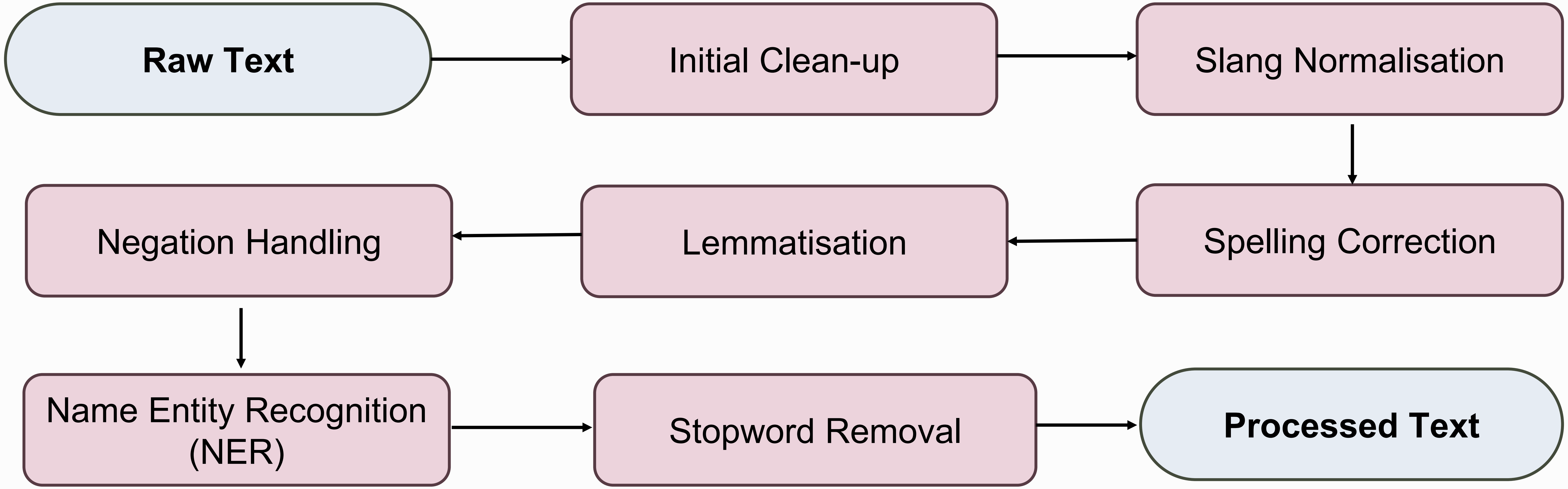
All information related to the ongoing virtual internship program could be documented, like the interaction between players, type of mentor etc

## 4 Data Aggregation to Group Level

As per requirements of the project, data aggregation to team-level shall be done to acquire team-level statistics.

# 1 Content-Based Feature Engineering

## Text Preprocessing Workflow



# 1 Content-Based Feature Engineering

| | Term Frequency – Inverse Term Frequency (TF-IDF) | Sentiment Analysis |
|---|---|---|
| | **Identify important terms** in chatlogs while down-weighting common words. | **Quantify emotional tone** of team communications, ranges between -1 to 1. |
| | **Generated 4400+ feature** in which each feature represent each term appeared in conversations | **Generated 1 useful feature** that allows quantify polarity scores (emotions) in group over time. |

2

# Structure-Based Feature Engineering

## Group Size

Group Size was gained by calculating number of players exist in each group. It is a constant throughout the internship.

## Mentor Count

Mentor Count was gained by calculating number of mentor exist in each group, which would be constant throughout the internship.

## Mentor-to-Player Ratio

Mentor-to-Player Ratio was gained by getting the proportion of mentor in each group. Similarly, it is a constant throughout whole internship

# 3 Progress-Based Feature Engineering

## Activeness & Engagement

Activeness metrics tracking message frequency for both players and mentors; while engagement metrics were computed by taking sum of topic-specific interaction flags.

## Mentors' Mentoring Style

Specific attention was paid on engagement matrices in which the frequency of questioning and asking for mentors were being paid attention. The frequency of mentor questioning and asking were totaled up respectively into 2 columns.

# 4 Data Aggregation to Group Level

We turned to the critical task of aggregating individual-level data to the team level required for our analysis. Different feature types demanded different aggregation strategies.

1

Based on the **nature of the features'** data, different aggregation way was conducted by either taking **sum, mean or first** value.

2

**Mean Squared Error** between the original and aggregated **OutcomeScore** was being prioritised for target column.

4

# Data Aggregation to Group Level

| Aggregation Approach | Mean-Squared Error |
|---|---|
| Raw data (for comparison) | 0.00 |
| Mode | 2.94 |
| Mean | 1.86 |
| Median | 1.87 |
| Weighted Average Mark (WAM) | 2.07 |
| Batch-Gradient Descent | 1.75** |

**Best Aggregation Approach:** Batch-Gradient Descent

# BASELINE MODELLING

## Why k-Nearest Neighbours?

- **Fast training** while maintaining simplicity
- **High interpretability** compared to more complex models
- It is a **classification-based approach.**

## Model Training Approaches

We used:

**80/20** training-testing set

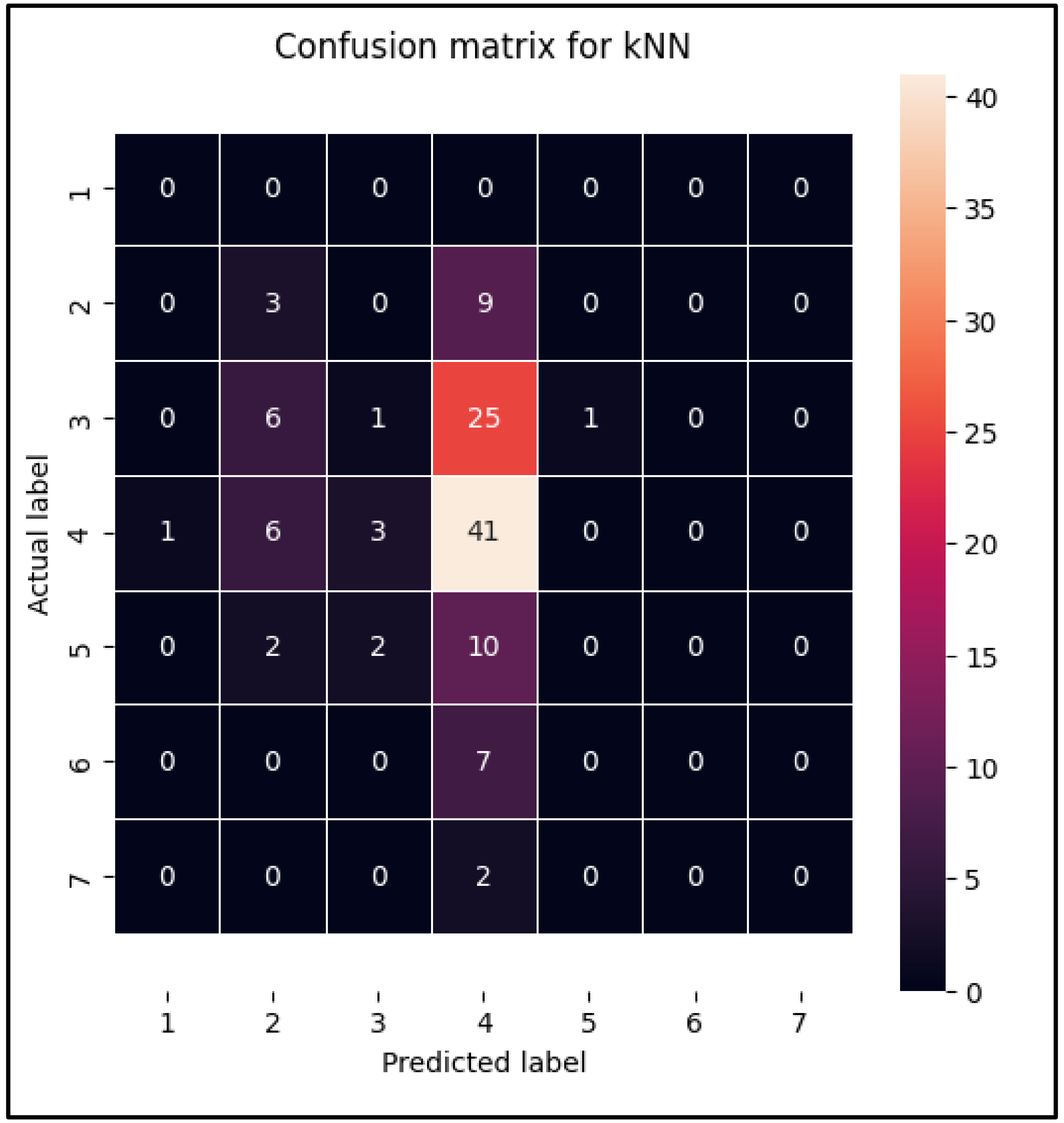
Main indicator of model performance would be based on **F1-score and accurary**.

# BASELINE MODELLING

| Number of Features | Training Samples | Accuracy (Train) | Accuracy (Test) | F1-Score |
|---|---|---|---|---|
| 4456 | 475 | 0.581 | 0.269 | 0.241 |

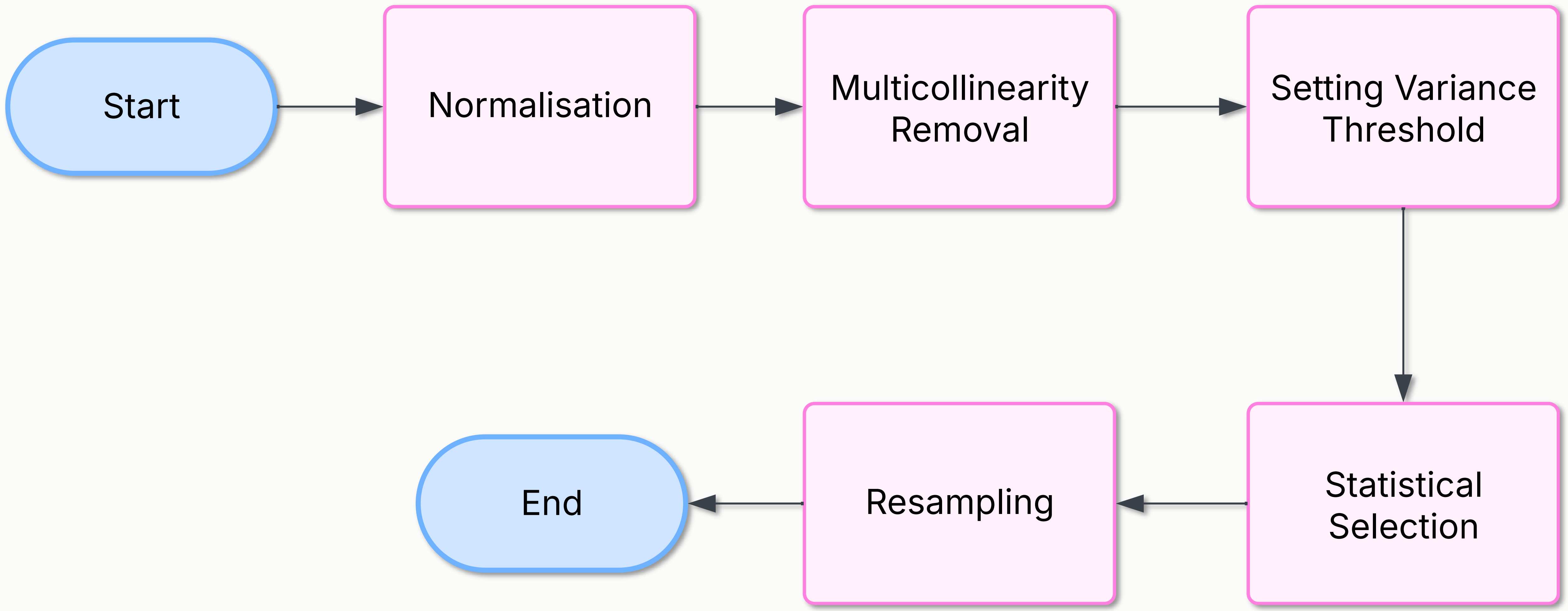
## Potential Issues

- High number of features
- Imbalance of samples
- Difference in scale
- Lack of training samples



# MODEL ASSESSMENT

## Model Improvement Workflow



# MODEL ASSESSMENT

## Evaluation of Model Improvement Approach

| Version | Number of Features | Training Samples | Accuracy (Train) | Accuracy (Test) | F1-Score |
|---|---|---|---|---|---|
| **Baseline Model** | 4456 | 475 | 0.581 | 0.269 | 0.241 |
| **Improved Model** | 136 | 1752 | 0.763 | 0.720 | 0.704 |

# MODEL SELECTION

## Our Aim on Models

- Supervised Learning Models
- Performing well on unseen data
- Avoids both overfitting and underfitting
- Stable on noisy data.

## WE AIMED FOR

- **Classification Methods**

## NOT OUR AIM

- **Regression Methods**
- **Clustering Methods**

# MODEL SELECTION

**The Baseline Model*

0 *k*-Nearest Neighbours

1 Support Vector Machine

2 Decision Tree

3 Random Forest

4 Gradient Boosting

5 LightGBM

6 Logistic Regression

# 1 Support Vector Machine

## Why?

- Finds the **optimal decision boundary** by maximising the margin between classes.

## Approach

- Ran **baseline SVM model** with default settings.
- Applied **grid-search hyperparameter tuning** on kernel type.
- **cross-validation** was used in each tuning loop to ensure result consistency.

| Kernal | C | Gamma | Decision Function | Accuracy |
|---|---|---|---|---|
| poly | 5.0 | 0.1 | ovr | 0.713 |
| linear | 10.0 | auto | ovo | 0.695 |
| linear | 1.0 | auto | ovr | 0.695 |
| linear | 0.1 | auto | ovr | 0.681 |
| rbf | 10.0 | 0.001 | ovo | 0.677 |
| rbf | 100. | 0.0001 | ovr | 0.674 |
| rbf | 1.0 | 0.01 | ovr | 0.672 |
| sigmoid | 1.0 | 0.01 | ovr | 0.606 |
| poly | 1.0 | scale | ovo | 0.449 |

# 2 Decision Tree

## How we applied it:

- Started with a **baseline Decision Tree** model.
- Performed **grid-search hyperparameter tuning**.
- **5-fold cross-validation** used to validate performance and avoid overfitting.

| Metrics | Accuracy (Train) | Accuracy (Test) | F1-Score | Recall |
|---|---|---|---|---|
| **Baseline Decision Tree** | 0.861 | 0.713 | 0.705 | 0.719 |
| **Tuned Decision Tree** | 0.861 | 0.7016 | 0.690 | 0.701 |

# 3 Random Forest

## Why?

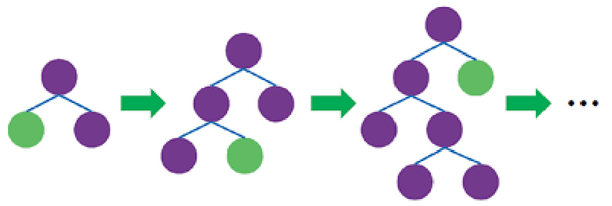
- Powerful ensemble method that reduces **overfitting** of individual trees.
- Works well with **high-dimensional and imbalanced data**.
- Helps improve **generalisation** and reduce variance.

## How we did it?

- Started with **Baseline Random Forest** model.
- Conducted grid-search hyperparameter tuning with 5-fold cross-validation.

| Version | Accuracy (Train) | Accuracy (Test) | F1-Score |
|---|---|---|---|
| **Baseline Random Forest** | 0.861 | 0.754 | 0.751 |
| **Tuned Random Forest** | 0.851 | 0.759 | 0.758 |

# 4 Gradient Boosting



## Why?

- Can capture nonlinear relationships
- Handle interactions between features

| Accuracy (Train) | Accuracy (Test) | F1-Score |
|---|---|---|
| 0.808 | 0.708 | 0.699 |

# 5 LightGBM

| Accuracy (Train) | Accuracy (Test) | F1-Score |
|---|---|---|
| 0.835 | 0.702 | 0.702 |

## Why?

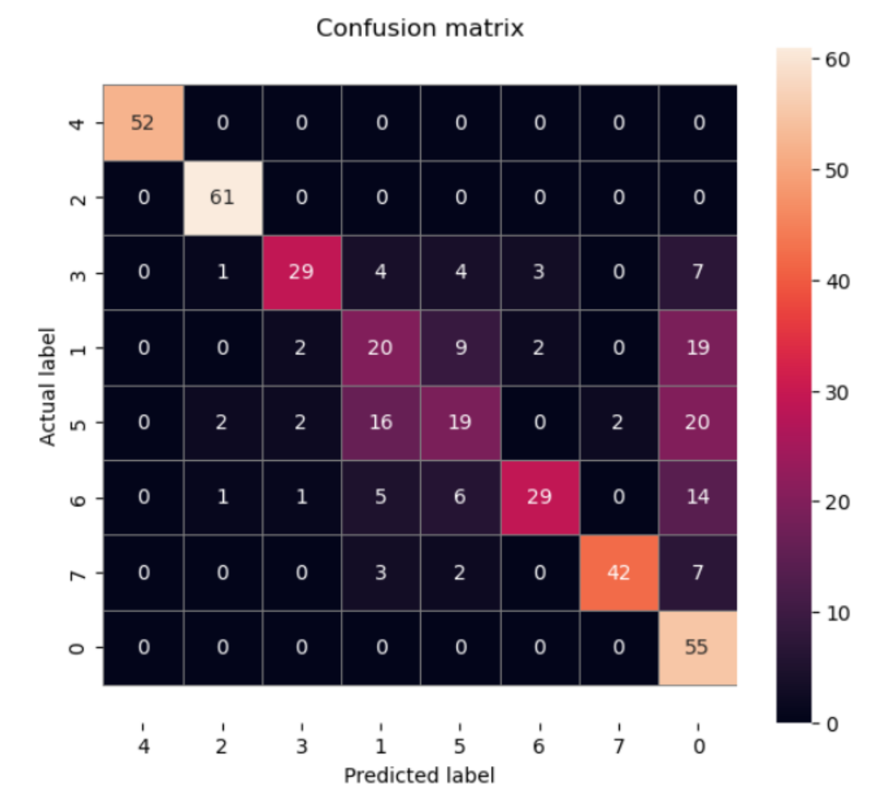
- Able to handle imbalance dataset.
- Focuses on important splits first that normally contribute to high accuracy

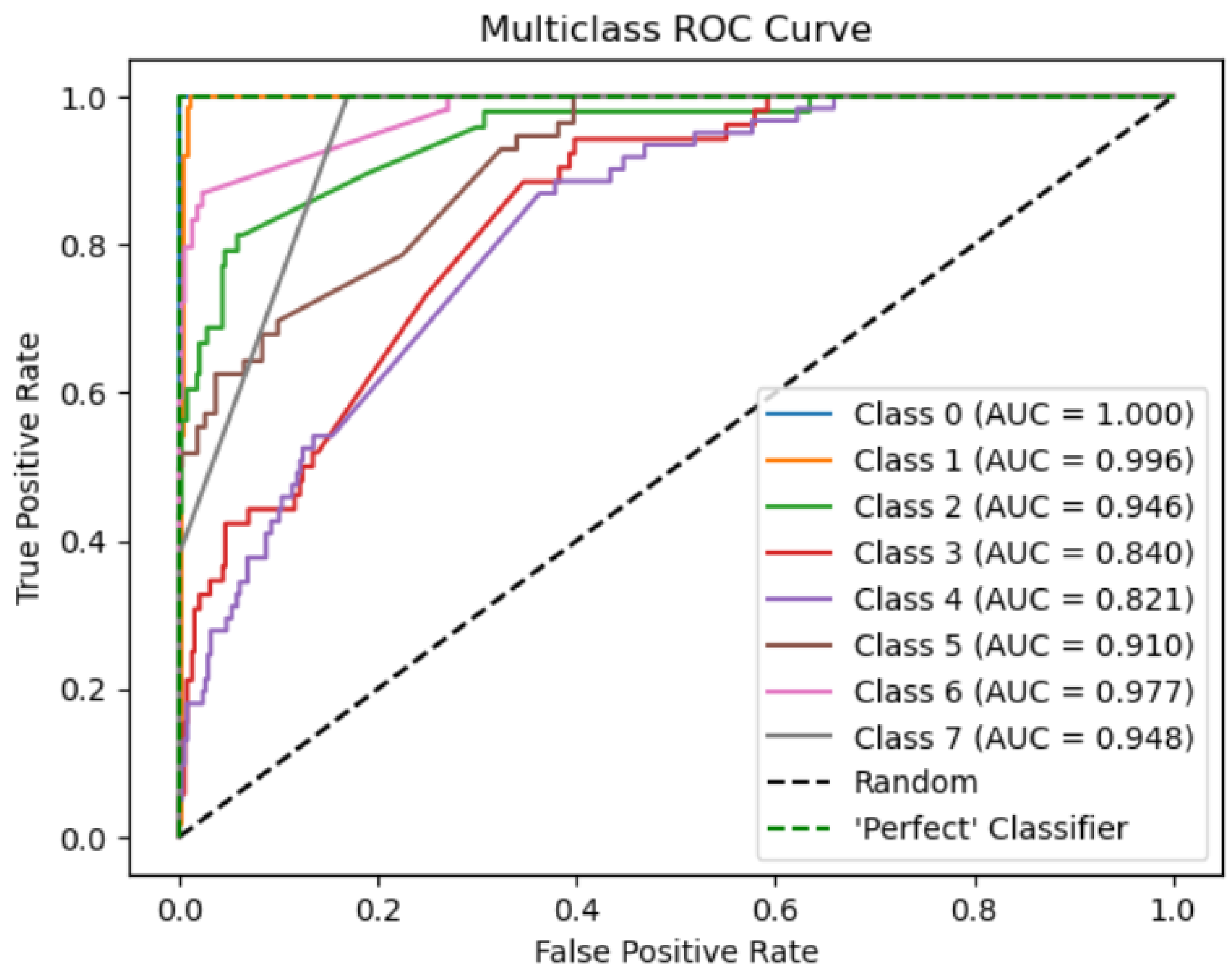
# 6 Logistic Regression

The Baseline Logistic Regression Model

## Why?

- It can be extended to **multilabel tasks** by training multiple classifiers.
- Fast and Efficient
- Built-in Regularisations




| Accuracy (Train) | Accuracy (Test) | F1-Score |
|---|---|---|
| 0.727 | 0.699 | 0.696 |

# 6 Logistic Regression

## How We Improve The Model?

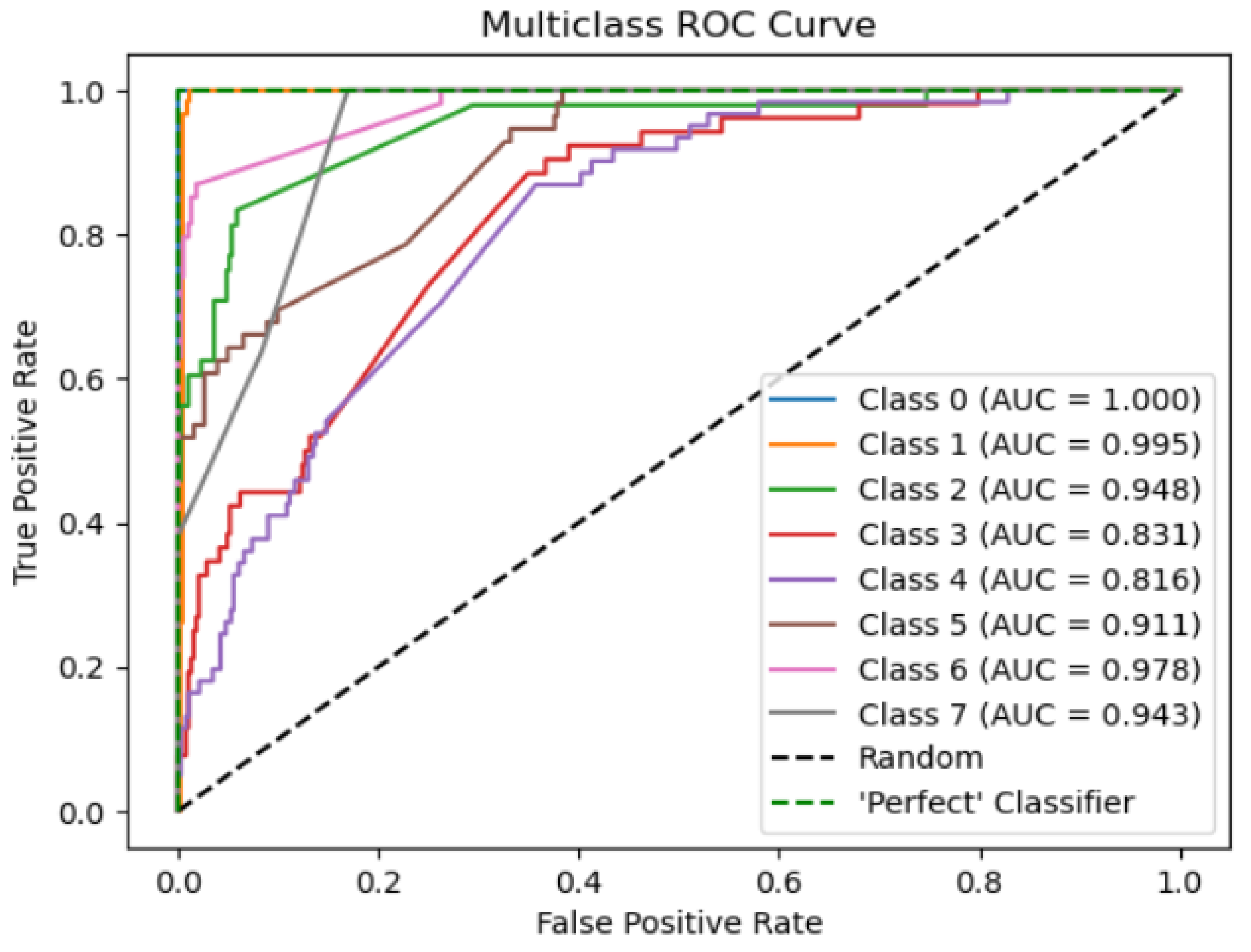
- **GridSearchCV** is employed to evaluate different regularisation and solver combinations.
- *5-fold cross-validation (cv=5)*

## Observations

- **The Best Set of Hyperparameters***: [L2, 'lbfgs',C=10]*
- **Minor improvements** according to ROC curve
- Only training accuracy is **improved slightly**.

The Hyperparameter Tuned Model

| Accuracy (Train) | Accuracy (Test) | F1-Score |
| --- | --- | --- |
| 0.732 | 0.699 | 0.696 |



# RESULTS & DISCUSSION

## Evaluating Models

| Version | Accuracy (Train) | Accuracy (Test) | F1-Score |
|---|---|---|---|
| Baseline Model (kNN) Without Data Processing | 0.581 | 0.269 | 0.241 |
| Baseline Model (kNN) After Data Processing | 0.763 | 0.720 | 0.704 |
| SVM Baseline | 0.731 | 0.695 | 0.689 |
| SVM Hyperparameter Tuned | 0.774 | 0.713 | 0.710 |
| Decision Tree Baseline | 0.861 | 0.713 | 0.706 |
| Decision Tree Hyperparameter Tuned | 0.511 | 0.494 | 0.469 |
| Random Forest Baseline | 0.861 | 0.754 | 0.751 |
| Random Forest Hyperparameter Tuned | 0.851 | 0.759 | 0.758 |
| LightGBM | 0.835 | 0.702 | 0.702 |
| Logistic Model Baseline | 0.727 | 0.699 | 0.696 |
| Logistic Model Hyperparameter Tuned | 0.732 | 0.699 | 0.696 |
| Gradient Boosting Baseline | 0.810 | 0.706 | 0.697 |

**Best Performing Model**

Hyperparameter tuned Random Forest Classifier

**Hyperparameters:**

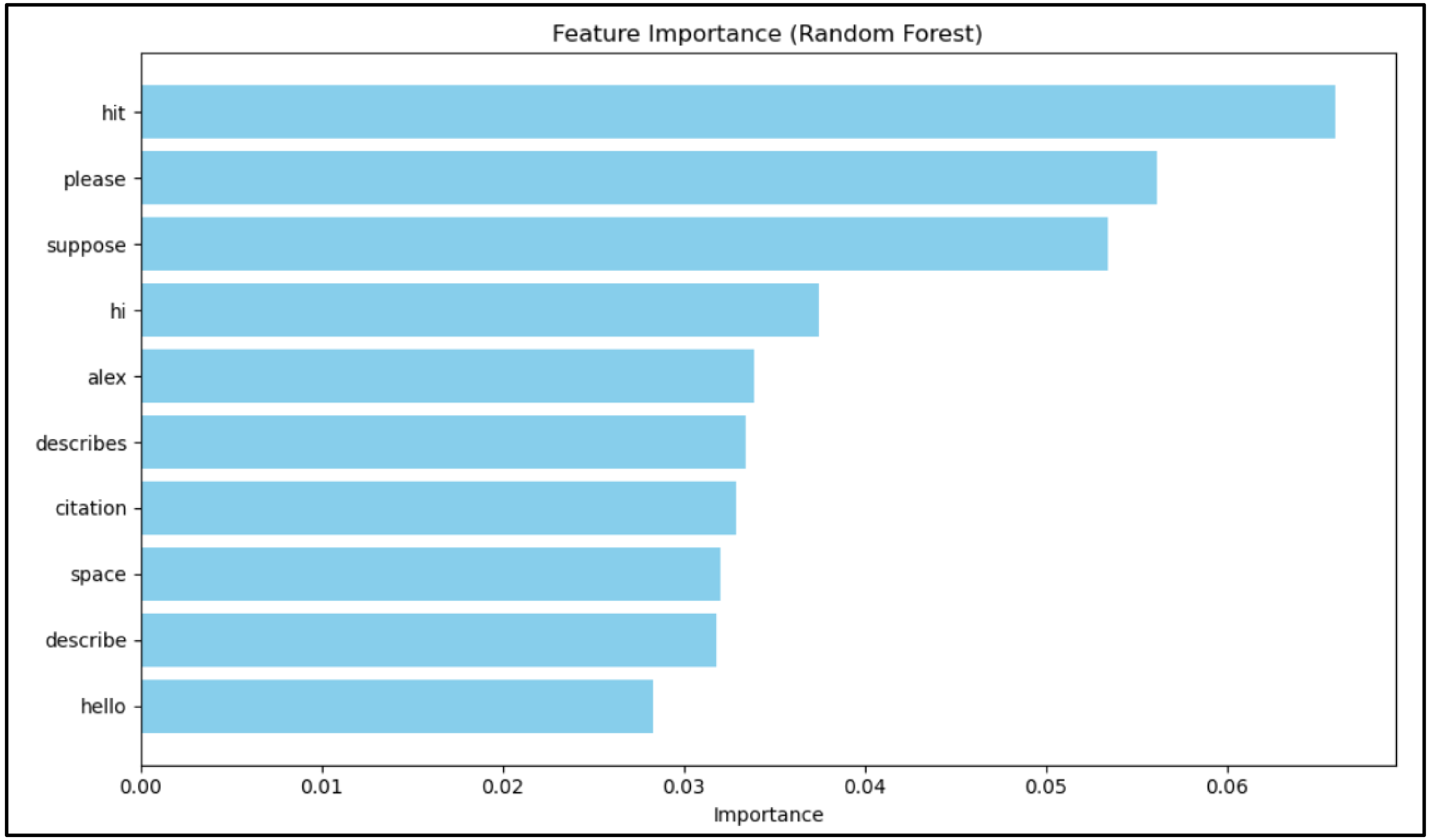
- Max Depth = 200,
- Min Sample Split = 5,
- n_estimators = 200,
- max_features = sqrt

**Worst Performing Model**

*k*-Nearest Neighbour model (Baseline Model)

# RESULTS & DISCUSSION

## How Feature Influences Models




## Observations

- We are extracting feature importances based on the best model --- **the fine-tuned Random Forest Model**
- Importance scores are overall relatively low.
- The keywords 'hit' and 'please' are most important in predicting target.
- There is no any mentor-related feature appeared in top 10.

# CONCLUSION

Looking back to our project objectives:

## 01 Aggregate and transform the chat data into team-level statistics.

All data was successfully transformed into team-level statistics for modelling purpose using the following data processing techniques:

1. Content-Based Feature Engineering
2. Structure-Oriented Feature Engineering
3. Progress-Oriented Feature Engineering
4. Data Aggregation to Group Level Using Gradient Descent Approach

# CONCLUSION

## 02 Build predictive models to predict final report scores based on team communication behaviours.

### Best Performing Model

- The highest performing model is **tuned Random Forest Classifier**.

### Performance of the Best Performing Model

- The **F1-Score of 0.758**, indicating a strong balance between precision and recall.
- The **testing accuracy score of 0.759**, suggesting the model has effectively captured underlying patterns within the dataset.

# CONCLUSION

## 03 Interpret the results of the models to understand how communication features relate to the team report performance.

- The keywords ‘hit’ and ‘please’ are most important in predicting target.
- However, overall feature importances are low
- Thus, all features contributes the same amount in predicting the ‘Outcome Score’

# THANK YOU!