A black and red logo

Description automatically generated

**ITD214 – Applied Data Science Project**

Chen ChongYuan Ken

8803640K

Date 12 August 2024

**1.Introduction and Background**

In a recent survey reported by Channel News Asia, it was found that the prevalence of mental health issues in Singapore is rising, with young adults having the highest proportion at 25.3%. In February 2024, Deputy Prime Minister Lawrence Wong passed a motion on advancing mental health, highlighting the growing importance of this issue both in Singapore and globally.

As part of a support group initiative, this project aims to develop a platform focused on providing emotional support, fostering a sense of community, and ensuring anonymity for those seeking help. By applying machine learning models such as Support Vector Machine (SVM), Random Forest and conducting correlation analysis, this study seeks to predict the likelihood of depression and identify the most influential features.

Leveraging these data-driven insights allows for the identification of key demographic and behavioural factors contributing to depression risk, enabling tailored interventions and content optimization. The goal is to create a positive and engaging environment that empowers users to address their mental health challenges effectively.

**1.1 Project Plan**

A screenshot of a computer

Description automatically generated

**1.2 Business Goals**

|  |  |  |
| --- | --- | --- |
| 1 | Support Initiative | As a member of a support group, we aim to set up an online support system to improve emotional support and provide a sense of community and anonymity to those taking part and seeking help. |
| 2 | Community Engagement | The online community allows individuals who feels lonely or socially isolated to connect with others experiencing similar challenges and share coping strategies. |
| 3 | Content Optimization | By understanding user needs, preferences, and behaviours of the R/foreveralone community, we can optimize the platform's content to better meet their expectations. |
| 4 | Positive Environment | To create a positive and supportive online environment that fosters a strong sense of community and empowers users to effectively address their challenges. |
| 5 | Goal Measurement | As a measurement of our goal, we hope to see an increase in positive sentiment expressed by users' post by 20%. |

**1.3 Business Objectives**

|  |  |
| --- | --- |
| **Depression Prediction** | Determine the likelihood of depression in users based on their demographic and behavioural attributes |

**1.4 Dataset**

|  |  |
| --- | --- |
| **Title** | The Demographic /r/ForeverAlone Dataset |
| **Description** | Last year a redditor created a survey to collect demographic data on the subreddit /r/ForeverAlone. Since then they have deleted their account but they left behind their data set. |
| **URL** | https://www.kaggle.com/datasets/kingburrito666/the-demographic-rforeveralone-dataset?resource=download |

**2.Work Accomplished**

**2.1 Data Preprocessing**

The below are the data preprocessing steps I completed prior to analysis:

|  |  |  |
| --- | --- | --- |
| No | Steps | Remarks |
| 1 | Checking for outliers | A graph with dots and lines  Description automatically generated  A white rectangular object with black text  Description automatically generated  There is a concentration between age 20-25.  Observed that there are outliers for number of friends in the 400+ and 600+ region |
| 2 | Handling Missing Values | There are two columns (‘job\_title’ and ‘improve\_yourself\_how’) with missing values.  Performed Fillna  Verified No missing values |
| 3 | Removing Special characters | Columns with special characters identified |
| 4 | Adding new columns | To categorize sensitive data such as age and income. |
| 5 | Dataset transformation | Transform columns (virgin, prostitution\_legal,pay\_for\_sex,social\_fear,depressed,attempt\_sucide) to Boolean for modelling purposes |
| 6 | Exploratory Data Analysis | In accordance with the results of the below we understand that    Histogram & Bar chart – Majority of the users are young (Gen Z & Millennial) with no income    Income and Age: There's a clear trend where higher income categories correspond to older median ages, possibly indicating more experience or advanced career stages.  Income and Education: Higher educational levels are associated with higher income categories. There's a clear clustering of higher income individuals in educational levels beyond a bachelor’s degree.  These visualizations provide insights into how age, income, and education level are interrelated, with higher income often correlating with older age and higher educational attainment.  A pie chart with different colors  Description automatically generated  Majority of the users are Caucasian at 72.1%, Asian 11.5% and Hispanic 6.8% |
| 7 | Further Data cleaning | Dropping more columns to align with AI ethics, PDPA guidelines and our business objectives. See below for steps taken.   1. Print Data frame column for checking 2. Check against business objective PDPA and Ethics guideline to drop the following columns a) attempt\_sucide b) Virgin c) race d)Social\_fear e) job\_title f) job\_title\_cleaned g) job\_tite\_category h) prostitution\_legal i) pay\_for\_sex j) sexuality k) gender l) friends 3. Edu\_level column was dropped as duplicated column was discovered |

**2.2 Training and discussion of models**

In this analysis, I will be evaluating and comparing the following three models then selecting the best.

1. Logistic Regression
2. Random Forest
3. Support Vector Machine

Following the model selection, I will perform a Correlation analysis to examine the relationship between the predicted probabilities and the features. This will enable us to identify Key Features Influencing the likelihood of depression.

We have conducted 2 attempts for each model

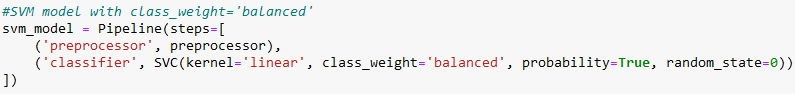
1. Without class weight adjustment
2. With balanced class weights.

The rationale for these two attempts is that the target variable might be imbalanced. By setting class weight='balanced', the model adjusts the weights assigned to each class during training based on their frequency. Specifically, the weights are inversely proportional to class frequencies, which helps to give more emphasis to the minority class (class 0 in this case) and improve the model's performance on imbalanced data.A close-up of a computer screen

Description automatically generated

A white background with black text

Description automatically generated



A computer screen shot of a program

Description automatically generated with medium confidence

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Metric | Class | Logistic Regression  (class balanced) | Logistic Regression | Random Forest (class balanced) | Random Forest | SVM (Class Balanced) | Support Vector Machine |
| Precision | Class 0 | 0.27 | 0.50 | 0.36 | 0.54 | 0.28 | 0.57 |
| Class 1 | 0.65 | 0.63 | 0.71 | 0.65 | 0.66 | 0.63 |
| Macro Avg | 0.46 | 0.56 | 0.54 | 0.60 | 0.47 | 0.60 |
| Weighted Avg | 0.53 | 0.58 | 0.60 | 0.61 | 0.54 | 0.61 |
| Recall | Class 0 | 0.45 | 0.13 | 0.31 | 0.28 | 0.45 | 0.07 |
| Class 1 | 0.46 | 0.92 | 0.75 | 0.85 | 0.48 | 0.97 |
| Macro Avg | 0.45 | 0.52 | 0.53 | 0.56 | 0.46 | 0.52 |
| Weighted Avg | 0.46 | 0.62 | 0.62 | 0.63 | 0.47 | 0.62 |
| F1-Score | Class 0 | 0.34 | 0.21 | 0.33 | 0.37 | 0.34 | 0.13 |
| Class 1 | 0.54 | 0.75 | 0.73 | 0.74 | 0.55 | 0.76 |
| Macro Avg | 0.44 | 0.48 | 0.53 | 0.55 | 0.45 | 0.45 |
| Weighted Avg | 0.48 | 0.54 | 0.61 | 0.60 | 0.49 | 0.52 |
| Accuracy | Overall | 46% | 61.7% | 62% | 63.1% | 47% | 62.4% |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | With Class Balanced | Initial Models |
| Support | Class 0 | 29 | 54 |
| Class 1 | 65 | 87 |

**Logistic Regression**

1. 61.7% indicates a moderate performance
2. High recall for class 1 (depressed) suggests the model effectively identifies individual with depression but not for class 0 (not depressed)
3. Significant disparity between precision and recall highlights challenges in balancing both false positives and negatives
4. Class score for F1 os higher than for class 0 which supports my 2nd observation

With balanced class weights

* Precision: Lower for Class 0 (0.27) post-balancing.
* Recall: Improved but still not high (0.45) for Class 0.
* F1-Score: Remains low for Class 0 (0.34).

**Random Forest**

1. 63.1% indicates a moderate performance
2. Model performs well in identifying Individuals with depression but not for non-depressed individuals
3. Precision for Class 1 is higher than for class 0 (supports point 2)
4. Class score for F1 also supports point 2

With balanced class weights

* Precision: Improved but still moderate for Class 0 (0.36).
* Recall: Improved significantly for Class 0 (0.31).
* F1-Score: Shows a balanced improvement (0.33).

**Support Vector Machine**

1. 62.4% indicates a moderate performance
2. There is a significant imbalance with the recall between the classes. It has exceptionally high recall for class 1 but very low for class 0
3. Precision for Class 1 is slightly better than class 0 but the high recall for class 1 indicates that it effectively identifies individual with depression but struggles to identify for non-depressed individuals accurately
4. Class 1 score for F1 is high  which supports the above points

With balanced class weights

* Precision: Improved but not as high (0.28) for Class 0.
* Recall: Shows significant improvement (0.45).
* F1-Score: Shows some improvement but still moderate (0.34).

**Confusion Matrix**

Initial

A graph showing different colored squares

Description automatically generated with medium confidence A graph of a forest confusion matrix

Description automatically generated with medium confidence A diagram of a model

Description automatically generated with medium confidence

With balance class weights

A blue squares with numbers

Description automatically generatedA graph of a graph with numbers and labels

Description automatically generated with medium confidenceA blue squares with white text

Description automatically generated

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metric | Random Forest (class balanced) | Random Forest | Logistic Regression (class Balanced) | Logistic Regression | SVM (Class balanced) | SVM |
| True Negative | 9 | 15 | 13 | 7 | 13 | 4 |
| False Negative | 16 | 13 | 35 | 7 | 34 | 3 |
| False Positive | 20 | 39 | 16 | 47 | 16 | 50 |
| True Positive | 49 | 74 | 30 | 80 | 31 | 84 |

**Observations:**

Initial

1. SVM has the highest TP which equates to best identifying depressed individuals
2. Random Forest provides a balance making it a reliable choice
3. Logistic regression excels at detecting depression but with a higher rate of false positives

With balanced class weights:

1. Random Forest: Best at identifying positives but struggles with negatives.
2. Logistic Regression: Balanced but has a high rate of missing positive cases.
3. SVM: Provides a good balance but also has a high rate of missing positive cases.
4. **Analysis**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Recall (class balance)** | **Recall** | **F1-Score (class balance)** | **F1-Score** |
| **Support Vector Machine** | **0.48** | **0.966** | **0.55** | **0.760** |
| **Random Forest** | **0.75** | **0.851** | **0.73** | **0.751** |
| **Logistic Regression** | **0.46** | **0.920** | **0.54** | **0.748** |

Without class weight adjustment

For predicting depression, recall is crucial as it identifies the most individuals at risk. The SVM model is chosen for its highest recall and balanced F1-Score, ensuring effective and comprehensive detection of depression.

**Why Not Random Forest or Logistic Regression?**

* Random Forest: Lower recall compared to SVM, resulting in more missed cases.
* Logistic Regression: Lower recall than SVM, affecting the ability to detect all at-risk individuals.
* SVM best meets our goal of maximizing depression detection.

With balanced class weights

* **Random Forest** has the highest recall for detecting depression in both initial and class-balanced versions (0.75), making it the most effective model for identifying individuals at risk of depression.
* **SVM** and **Logistic Regression** have similar recall rates in both scenarios, but still lower than Random Forest.

Given that recall is the most critical metric for the objective, we will use **Random Forest model**. It has consistently demonstrated the highest recall, thus effectively identifying the most individuals at risk of depression.

**Correlation Analysis**

|  |  |  |
| --- | --- | --- |
|  | **Initial SVM** | **Random Forest (Class Balanced)** |
| 1 | Retired | High school / Diploma or equivalent |
| 2 | Unable to work | Gym goers |
| 3 | Gym goers | No Income |
| 4 | Master’s Degree | Millennial |
| 5 | High school / Diploma or equivalent | Gen X |

**Initial SVM**

After selecting SVM, further analysis was conducted to identify the top 5 features correlated with the predicted probabilities.

A graph with different colored squares

Description automatically generated with medium confidence

* **Retired & Unable to work** - The highest correlation with depression risk was found among individuals who are retired or unable to work. This is likely due to factors such as loss of purpose, social isolation, and financial stress, which can significantly impact mental health. A study by the National Library of Medicine (NIH) highlights that many workers retire earlier than anticipated, often due to health issues that complicate work duties. Early retirement can lead to challenges, as illustrated by one retiree who said, “Initially, I found it soul-destroying that I wasn’t going to work. But as time went on and I became involved in voluntary work, I derived a great deal of satisfaction from it. I don’t think I could go back to a full-time job.”
* **Gym Goers** - Surprisingly, individuals who engage in physical activities, such as joining a gym, also show a higher correlation with depression risk. A study by Harvard suggests that exercise is a natural treatment for depression. This implies that some gym members or gym goers might be using physical activity as a means to combat existing depression. Similarly, a Reddit thread titled "Just Go to the Gym" indicates that some individuals use exercise as a coping mechanism for depression.
* **Master’s Degree** - Research published in The Lancet Public Health found evidence of higher levels of depression and anxiety among higher education students compared to their peers. Those with advanced degrees often face increased stress, high expectations, and feelings of isolation, which contribute to higher depression risk. This highlights the need for targeted mental health support for individuals dealing with these pressures.
* **High school/Diploma or equivalent** - “I put a lot of stress on myself to study harder and take time off relaxation to practice more. Sometimes, the stress gets so high that I'll sleep at 4 a.m.,” said a secondary school student preparing for O-Levels. A recent study by Channel News Asia found that 1 in 10 teens suffer from mental health disorders, with school stress and a lack of parental understanding cited as major contributing factors.

**After Class balanced:**

Below are the Top 5 features for Random Forest and surprisingly there are 2 features similar with our initial SVM analysis which is High School Diploma or equivalent and Gym Goers.

A colorful bar chart with text

Description automatically generated with medium confidence

**No Income** – supporting evidence from a study by Annals (Official Journal Of The Academy Of Medicine, Singapore) mentioned that in Japan, it was reported that lower income and unemployment were associated with higher prevalence of depression and greater probability of receiving antidepressant treatment. The lack of financial resources can indeed contribute to higher levels of stress and mental health issues.

**Millennial** - It seems that the millennial generation are really troubled, this can be supported by the following 1) An article by fortune magazine stating that Millennials are so worried about their finances that they are falling into depression 2) News article recently by straight times stating that In Singapore and even globally, it been noted that millennials have increasing incidences of depression and anxiety 3) Children’s Health Council stating 12 issues plaguing the millennials into depression Financial worries, job insecurity, and other stressors contribute to higher depression rates among this generation.

**Gen X**- often labelled as sandwich generation an article by yahoo finance indicated that GenXers are burnt out and stress about money another study by Ohio State university highlights that Gen X showed poorer physical health, higher levels of unhealthy behaviours such as alcohol use and smoking, and more depression. The “sandwich generation” often faces significant stress from both aging parents and their own families, contributing to mental health issues.

**3.1 Recommendations**

With our findings, please see below for the recommendations

|  |  |  |
| --- | --- | --- |
| 1 | Enhance Targeted Support for High-Risk Groups | The analysis identified retired individuals, those unable to work, and higher education students as being highly correlated with depression risk. The platform should focus on developing specialized resources and interventions tailored to these groups, such as retirement planning support, mental health resources for students, and programs addressing social isolation and financial stress. |
| 2 | Promote Physical Activity as a Coping Mechanism | Given the correlation between gym memberships and depression, it is recommended to include content and community support around the benefits of exercise for mental health. Encouraging users to engage in physical activity and providing resources on how to integrate exercise into daily routines can be valuable for those using fitness as a coping strategy. |
| 3 | Implement Data-Driven Content Personalization | Utilize the insights gained from the correlation analysis to personalize the platform experience for users. For example, users who indicate stress from education or work can be provided with tailored articles, forums, or virtual support groups focused on managing academic or job-related pressures. |
| 4 | Continuously Monitor User Sentiment and Engagement | The platform should include mechanisms to track user sentiment and engagement over time. Regular analysis of posts, comments, and interactions can provide early indicators of worsening mental health conditions, allowing for timely intervention. Aim to achieve the goal of a 20% increase in positive sentiment by refining content strategies based on ongoing user feedback. |
| 5 | Expand Community Building Features | Foster a stronger sense of community by integrating features such as peer support groups, mentorship programs, and anonymous chat options. These features can help reduce social isolation, allowing users to connect with others who share similar challenges. |
| 6 | Focus on Ethical and Inclusive AI Practices | As the platform leverages machine learning to predict depression risk, it is essential to ensure that the models are developed and deployed in an ethical manner. Regularly assess the impact of the algorithms to prevent bias, ensure data privacy, and uphold user trust, especially when dealing with sensitive mental health information. |
| 7 | Collaborate with Mental Health Professionals: | Consider collaborating with mental health professionals to validate the recommendations and resources provided on the platform. This partnership can also support users who may require professional intervention, ensuring that the platform serves as both a support and a gateway to clinical help if needed |

**4. AI Ethics**

A diagram of a project

Description automatically generated with medium confidence

"The ultimate goal of AI is to create machines that can think and act ethically, but the challenge lies in how we design these machines to align with human values." – Stuart Russell

Our company has complied and is aligned with Singapore Ethics & Governance Body of Knowledge (BOK)

**4.1 Data Management**

Data Privacy - To align with AI ethics, PDPA guidelines, and our business objectives, the following data columns have been excluded:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **PDPA Obligations \ AI Ethics Principles** | **Fairness & Non-Discrimination** | **Privacy & Confidentiality** | **Transparency & Explainability** | **Accountability & Responsible Use** | **Avoiding Harm** | **Inclusivity & Equity** |
| **Consent Obligation** | Race, Sexuality | Attempt\_suicide | Social\_fear | Job\_title, Job\_title\_cleaned, job\_title\_category, | Prostitution\_legal, Pay\_for\_sex | Sexuality, gender Virgin |
| **Purpose Limitation Obligation** | Race, Bodyweight | What\_help\_from\_others | Friends | Job\_title\_cleaned,job\_title\_category,Job\_title,job\_title\_category | Attempt\_suicide | Sexuality, gender, Virgin |
| **Protection Obligation** | Sexuality, Virgin | Attempt\_suicide | - | Job\_title, Job\_title\_cleaned, job\_title\_category,  Income | Prostitution\_legal, Pay\_for\_sex | - |
| **Retention Limitation Obligation** | Sexuality, gender | Attempt\_suicide | - | - | - | - |
| **Accuracy Obligation** | Race | - | Social\_fear |  | - | - |

**Data Security – To ensure safety and privacy of user data our platform will employ a range of security measures such as:**

|  |  |  |
| --- | --- | --- |
| 1 | Data Encryption | All sensitive data will be encrypted both in transit and when rest ensuring that even in event of interception the data remained protected. |
| 2 | Identity and Access management | Strict controls are in place with multi-factor authentication required for administrative access and role-based permissions to limit access to sensitive information |
| 3 | Incidence Response and Security audits | Regular security audits and penetration testing to be conducted for identifying and addressing vulnerabilities.  An official incident response plan is to be in place to handle potential breaches effectively |
| 4 | Data Masking & Minimization | Data collection is minimized to only what is necessary for the platform's functionality. We can mask sensitive data if necessary, such as age, name or income. |

**4.2 AI Internal Governance Structure**

We understand the importance of AI Internal Governance structure. Please see below for our Core Values, Mission Statement, Organize Structure. We have also conducted the Risk Impact Matrix to decide on the degree of human oversight in the decision-making process.

|  |  |  |
| --- | --- | --- |
| Core Value | Empathy and Respect | We prioritize understanding the diverse experiences and emotions of our users. Every single interaction is guided by compassion and respect. |
| Integrity and Transparency | Honesty and openness is the way and how we operate ensuring that our actions and decisions are aligned with ethical standards. Transparent to the community, users and stakeholders. |
| Privacy and Confidentiality | Protecting user data and maintain confidentiality is our top priority. Upholding the highest standards to secure and respect user information. |
| Inclusivity and Equity | We make sure to create an inclusive environment where all individuals feel valued and supported regardless of their background and circumstances. |
| Continuous improvement | We dedicate and pride ourselves with ongoing learning and improvement (Kaizen). Using data and feedback collected to enhance our platform and improving our services thus serving our community better. |
| Responsibility and Accountability | We take full responsibility for our decisions and actions. Ensuring that we are accountable to our users and community we serve. |
| Mission Statement | Our mission is to provide a super safe, supportive, and inclusive online platform that empowers individuals to connect, share, and seek help.  By leveraging data-driven insights whilst upholding the highest ethical standards, we aim to foster a strong sense of community and support mental well-being for every user. | |
| Organization Structure | Hybrid structure  As we are a company focusing on emotional support and mental health. Hybrid structure is the most effective as it allows us to maintain centralized control over core values, ethics and strategy while giving teams or departments the flexibility to respond and adapt to specific user needs. The balance can help to ensure our ethical standards and mission while allowing for responsiveness and innovation in meeting user needs. | |

We have identified the following risks and based on their likelihood(probability) and potential impact(severity) to plot the matrix.

A table with text overlay

Description automatically generated

Understanding the above, we have come out with Mitigation strategies and also adopt **Human-In-The-Loop(HITL)** approach on areas such as Model recommendations, ethical oversights, user interaction and technical and operational oversight.

**5. Measuring the business goal (Group)**

A screenshot of a computer screen

Description automatically generated

After launching the online community, we initiated a thread to encourage members to share their daily experiences. Our analysis of these posts aims to measure overall sentiment, with a target of achieving at least a 20% positive sentiment rate.

A graph of a bar graph

Description automatically generated with medium confidence

Sentiment scores are calculated using a predefined sentiment lexicon (TextBlob), where each word is associated with a sentiment polarity score. The sentiment scores are categorized as follows:

* Pink shades: Negative sentiment
* Yellow shades: Neutral sentiment
* Green shades: Positive sentiment

Clusters of sentences exhibit similar sentiment scores, with some individual sentences reflecting particularly strong positive or negative sentiments.

A graph of multiple colored lines

Description automatically generated

The graph depicts the \*Sentiment Subjectivity\* of sentences, where the X-axis represents the \*Sentence Index\* and the Y-axis represents the \*Sentiment Subjectivity\*.

Subjectivity Scale: The subjectivity ranges from 0 to 1. A score close to 1 indicates highly subjective content, while a score closer to 0 indicates more objective content.

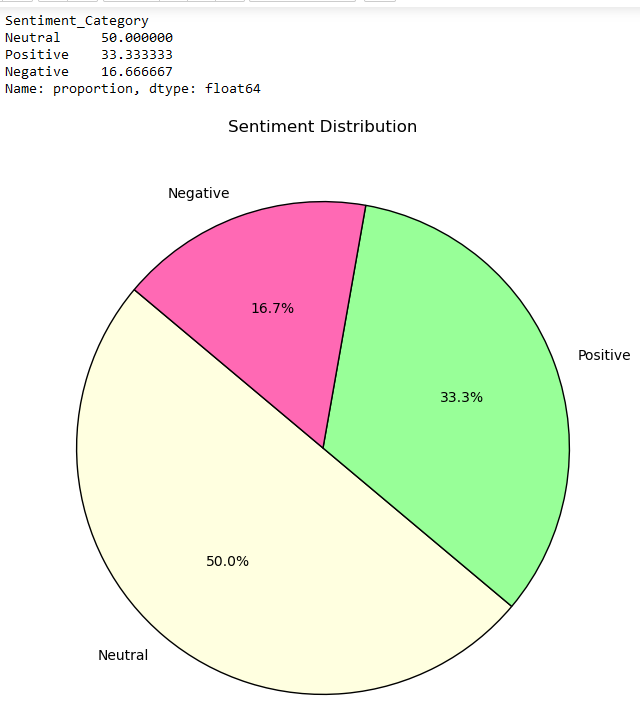
Color Gradient:

* The color gradient seems to move from darker shades of pink to green.
* Pink Shades (0-35): These indicate sentences that are more subjective (higher values).
* Green Shades (40-65): These indicate sentences that are more objective or contain varying levels of subjectivity.

Sentence Index Grouping:

* Early Sentences (0-10): High subjectivity, with many sentences scoring close to 1.
* Middle Sentences (15-35): These sentences are also subjective but vary more, with some sentences scoring lower on subjectivity.
* Later Sentences (45-65): More balanced with some sentences having lower subjectivity, suggesting a mix of subjective and objective content.

The content is highly subjective in the early and middle parts, gradually moving towards more balanced or objective statements in the latter part of the text. This could imply a transition in the tone or nature of the text, potentially starting with opinionated statements and moving towards more neutral or factual content.



We have successfully met our business goal, achieving a positive sentiment rate of 30%. The pie for neutral sentiment is the largest (50%), indicating that many of the data falls into this category. Overall, about 33% of the sentiments expressed by users are positive, reflecting a strong engagement and satisfaction with the community.

**6. Conclusion**

This assignment has successfully aligned the development of our mental health support platform with our core business objectives and goals while ensuring adherence to AI ethics and data privacy standards.

**Business Objectives and Goals:** Our primary objective was to create a platform that provides robust emotional support and fosters a sense of community for individuals facing mental health challenges. By leveraging advanced machine learning models and correlation analysis, we aimed to accurately predict depression risk and identify key factors influencing mental health. Through our evaluation, we determined that the Random Forest model outperforms other models, including Support Vector Machines (SVM), in this context. Random Forest's superior performance in terms of recall and balanced metrics is particularly effective for identifying individuals at risk, aligning with our goal of maximizing recall to proactively address mental health issues.

**AI Ethics and Data Privacy:** Our commitment to AI ethics is evident in our adherence to PDPA guidelines and AI Ethical guidelines. We have carefully considered fairness, transparency, and accountability throughout our development process. Data minimization and anonymization practices have been employed to protect user privacy and ensure that only necessary data is collected and used. Our approach integrates Human In the Loop (HITL) to balance automation with human oversight, thereby enhancing both ethical integrity and operational effectiveness.

In conclusion, this assignment not only meets our business objectives by delivering a platform that is both effective and ethical but also demonstrates our commitment to responsible AI practices. The integration of advanced machine learning techniques with a strong ethical framework ensures that our platform remains both a valuable tool for mental health support and a model of ethical AI deployment.

**7. Learning & observation**

Despite addressing class imbalance by balancing class weights, the results remain suboptimal. This may be due to factors such as:

* 1. Class imbalance alone is not the sole issue, other aspects of the data or model might also be affecting performance.
  2. Overfitting or underfitting- The models may not be well-tuned to the complexity of the data.

**Suggestions for improvements**

1. Hyperparameter tuning - Perform hyperparameter tuning to optimize model parameters for better performance. This can help improve the model’s fit and predictive power.
2. Try with different Algorithms - Experiment with alternative algorithms or ensemble methods, such as Gradient Boosting or XGBoost. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) could also be explored to generate synthetic samples for the minority class.
3. Analyse model residuals - Examine model residuals and errors to gain insights into where the models might be failing and guide further improvements in training and preprocessing.

By addressing these aspects, I can enhance the model's performance and ensure that it aligns better with the business objectives and problems.