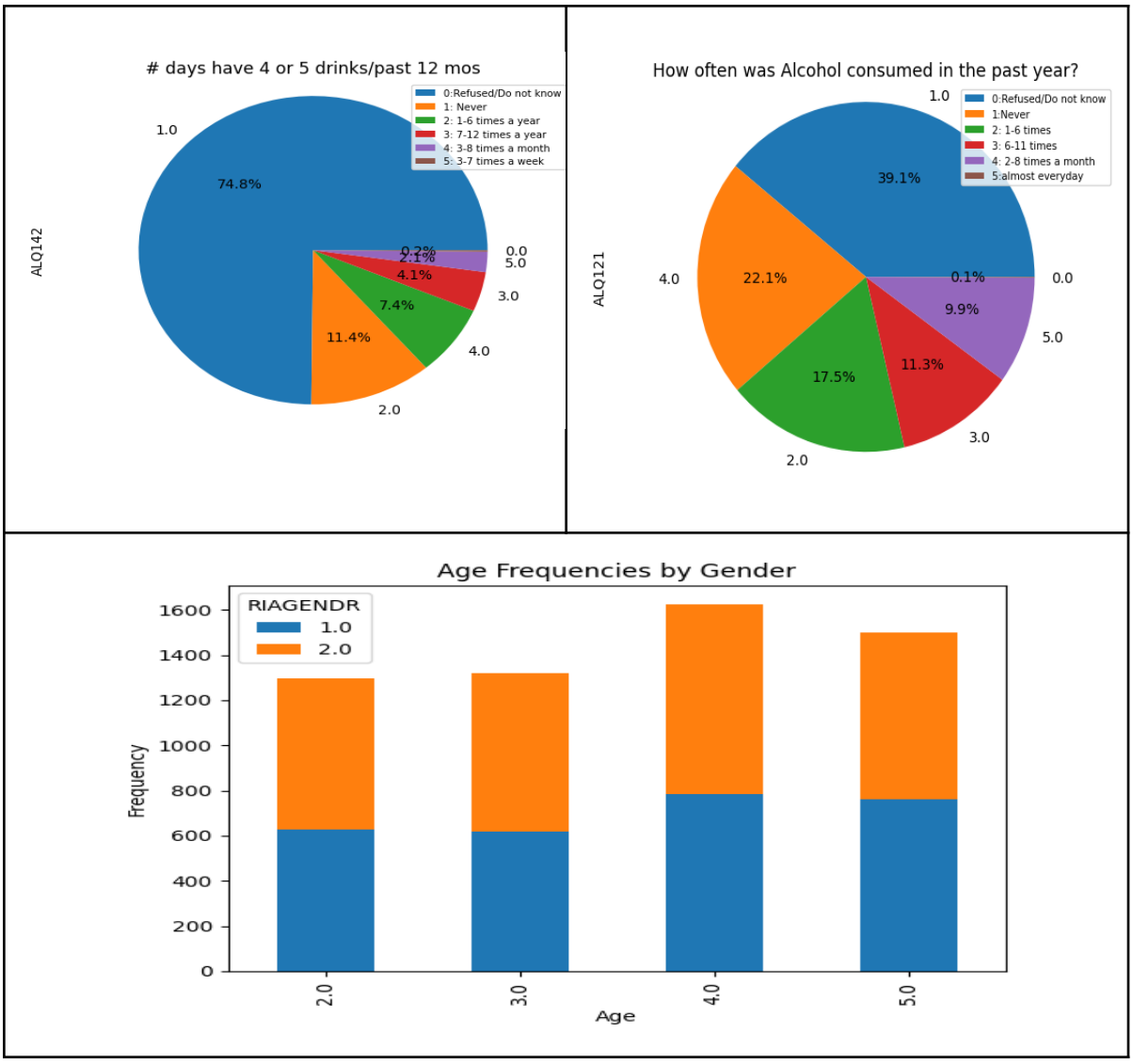
Key findings are divided into:

1. EDA Results
2. Correlation Results
3. Feature Engineering Results
4. Risk Score Generation Results
5. ML Model Implementation Results

1. EDA Results:

a) Pre-Risk score EDA results:



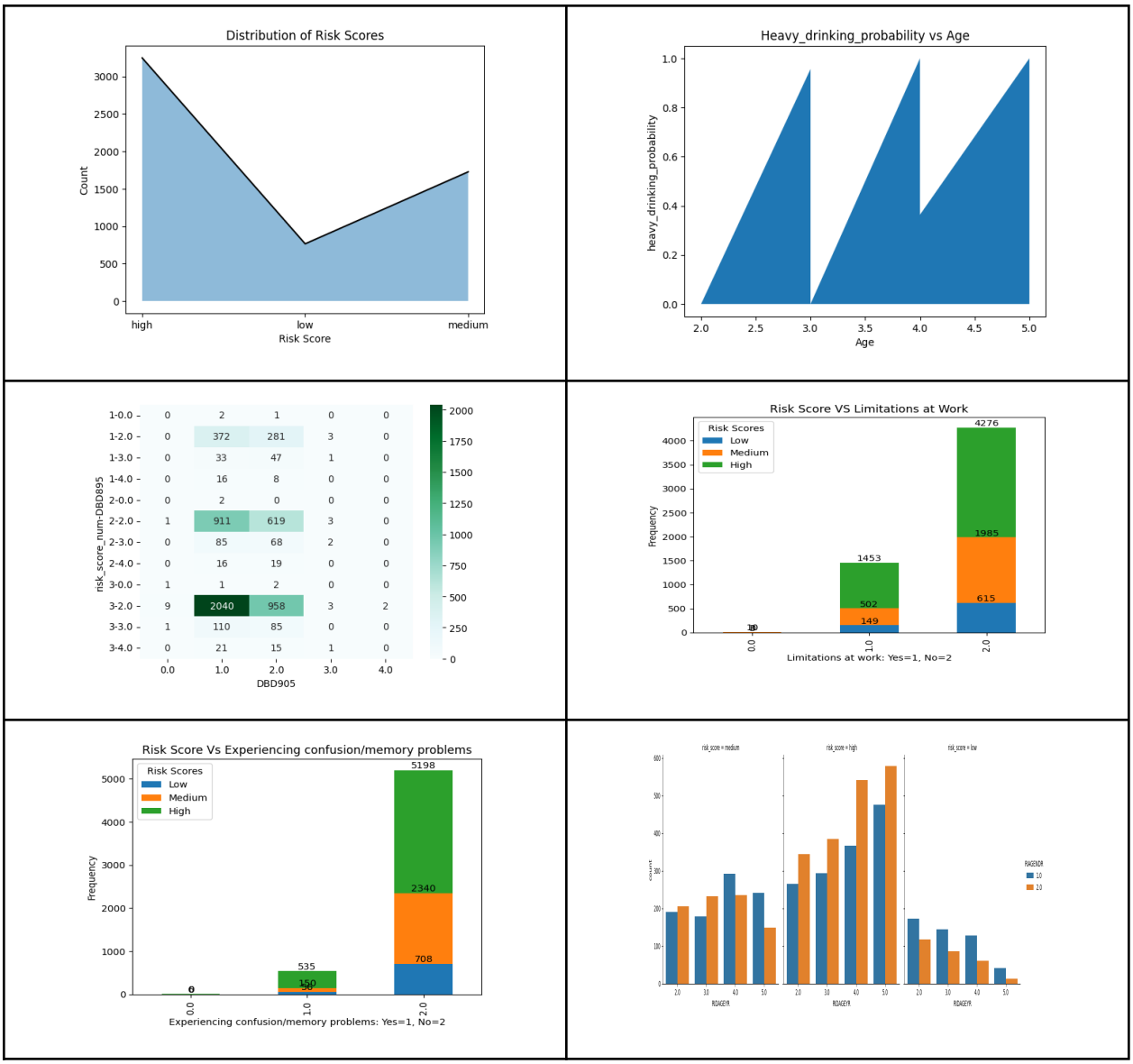
Key Findings:

- A significant proportion of participants declined to answer or lacked knowledge about drugs and alcohol, highlighting limitations in self-reported data.

- Moreover, the gender distribution was balanced across all age groups, avoiding gender bias. The low prevalence of depression made the analysis complex.

- Finally, a strong correlation was found among questions related to physical functionality, indicating the need for additional support in daily living.

b) Post-risk score EDA results:

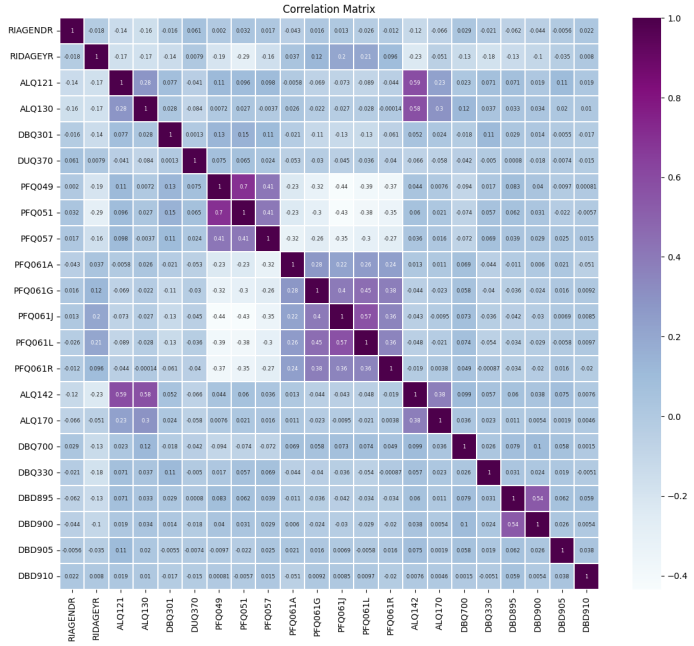


Key findings:

- The dataset is dominated by those at higher risk, with individuals between the ages of 40 and 70 having the highest probability of developing heavy drinking habits. However, there is a sharp drop in probability between ages 50 and 60.
- Women are present in higher numbers across all age groups in the high and medium risk categories for developing AUD.
- Individuals who consume homemade food that is mostly ready-to-eat are more prone to developing AUD than those who tend to eat outside, suggesting dietary habits may play a role in AUD development.
- Those who answered 'Yes' to experiencing confusion/memory problems and limitations at work had higher risk scores than those who answered 'No'.

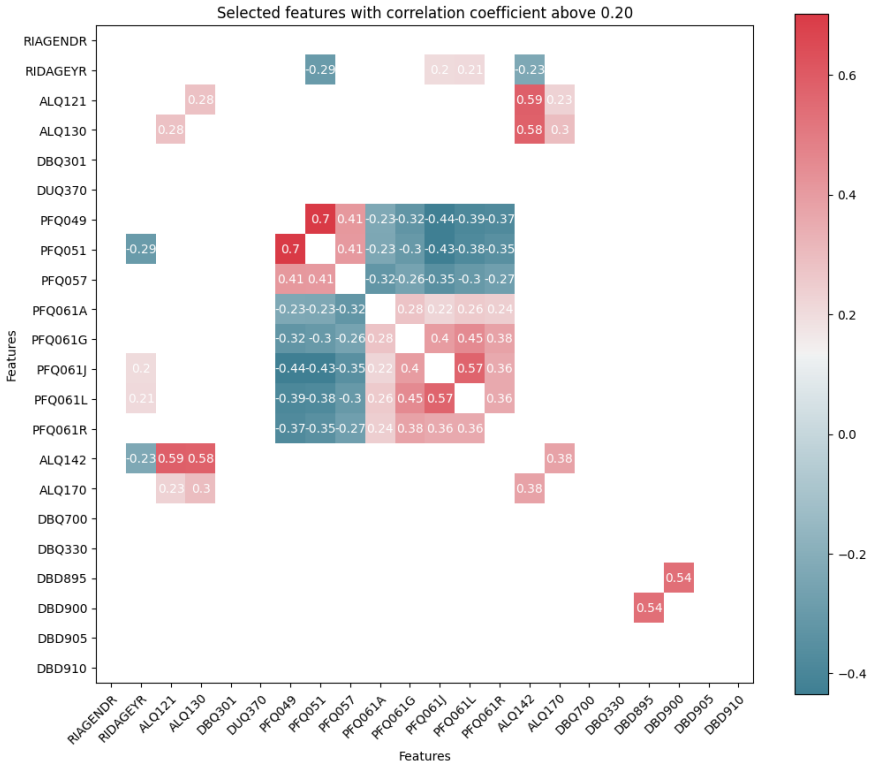
2. Correlation results:

The figure below shows an attempt at finding the correlation between the variables in the dataset. The dataset has variables that are poorly correlated to each other.



3. Feature Engineering Results:

Most of the attempts to perform feature engineering to select the attributes which have greater importance failed due to poor correlation. This includes our selection of variables with a correlation coefficient above 0.2. As a result, we had to give equal importance to all of our variables.



4. Risk Score Generation Results:

Key findings:

a) Risk scores were generated using Logistic Regression and KNN approaches

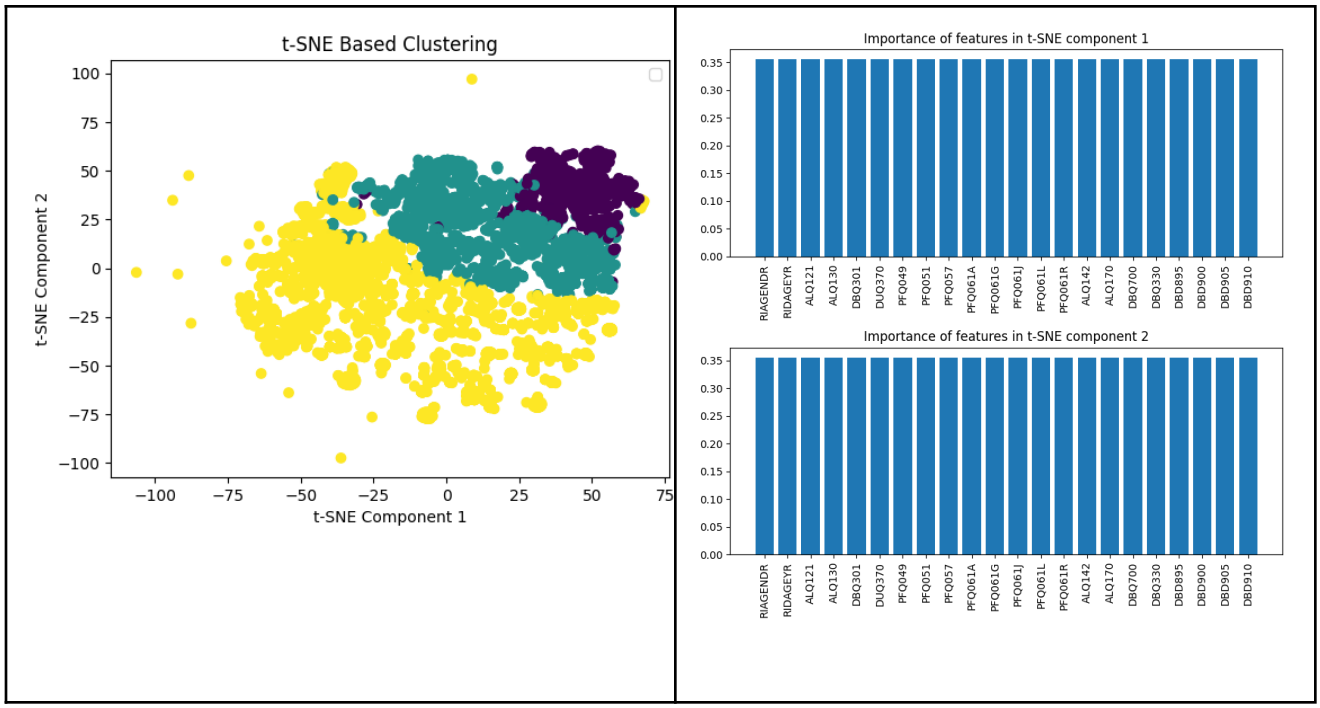
b) KNN approach was sub-classified into regular KNN and t-SNE-based KNN as other feature reduction methods failed

c) Logistic Regression model used selected features as predictors and AUD risk score as the target variable

d) KMeans clustering model was used for the basic KNN approach and assigned a score to each participant

e) t-SNE was used to reduce the dimensionality of the data and generate a risk score for each participant

f) t-SNE-based approach showed that features did not exhibit much variation in their importance as both t-SNE 1 and 2 had an importance of 0.35 for all features, indicating low correlation between the features.

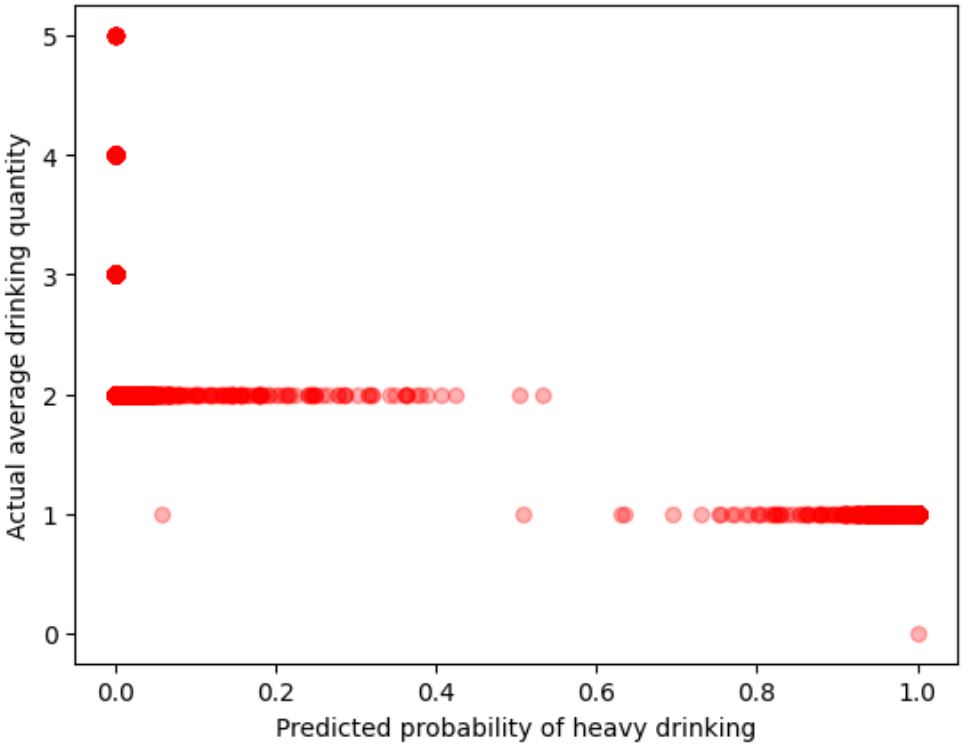


5. ML Model Implementation Results:

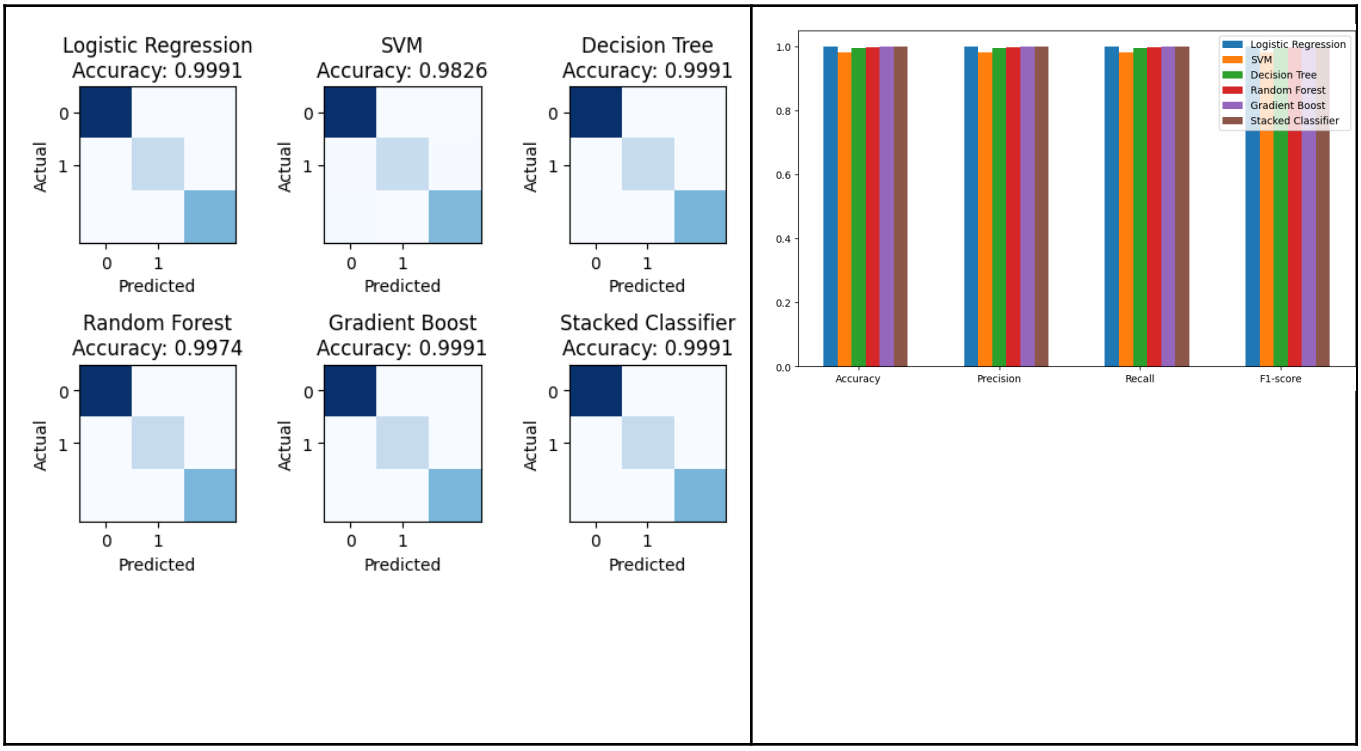
Key Findings:

- The logistic regression model showed a pattern of deviation from the regression line, indicating it may not accurately predict the target variable.
- KNN-based risk score results were considered more apt than logistic regression results. The confusion matrices showed perfect accuracy for all classifiers except SVM, which had an accuracy of 0.993. Evaluation metrics did not reveal significant differences between the models, suggesting they may have overfit the training data.
- The t-SNE-based approach showed lower accuracy than the original data, but the lower-dimensional representation of the data obtained by t-SNE can be more useful in visualizing and interpreting the data.

Logistic Regression Approach:



Basic KNN approach:



t-SNE-based approach: