INTERVIEW STATUS PREDICTION

1. Problem Statement

1 Problem:

Conducting over 1 lakh interviews annually in our multinational corporation, maintaining the integrity and fairness of the hiring process poses a significant challenge. There is a need to identify potential biases and ensure a smooth interview experience for candidates.

2 Goal:

Introduce Interview-Intel, an advanced interview intelligence tool, to predict interview outcomes. The goal is to empower recruiters with insights that allow them to assess interview integrity, identify biases, and ultimately enhance the overall recruitment process. This aims to create a streamlined and unbiased interviewing experience for the more than 10,000 candidates hired each year.

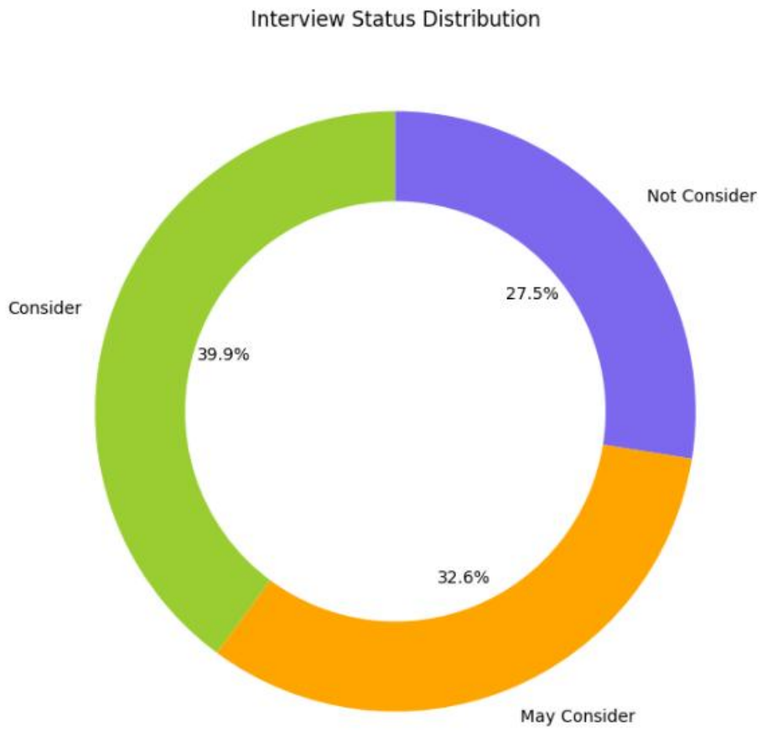
2. Exploratory Data Analysis

1. The dataset has “**5800”** rows and “**26”** features, including the target feature.

2. I've noticed there are categorical columns in the data which need to be converted in numeric format.

3. I found some null values in few categorical columns.

DOUGNUT DIAGRAM



Consider: 2311

Most candidates pass the interview are 39.9%

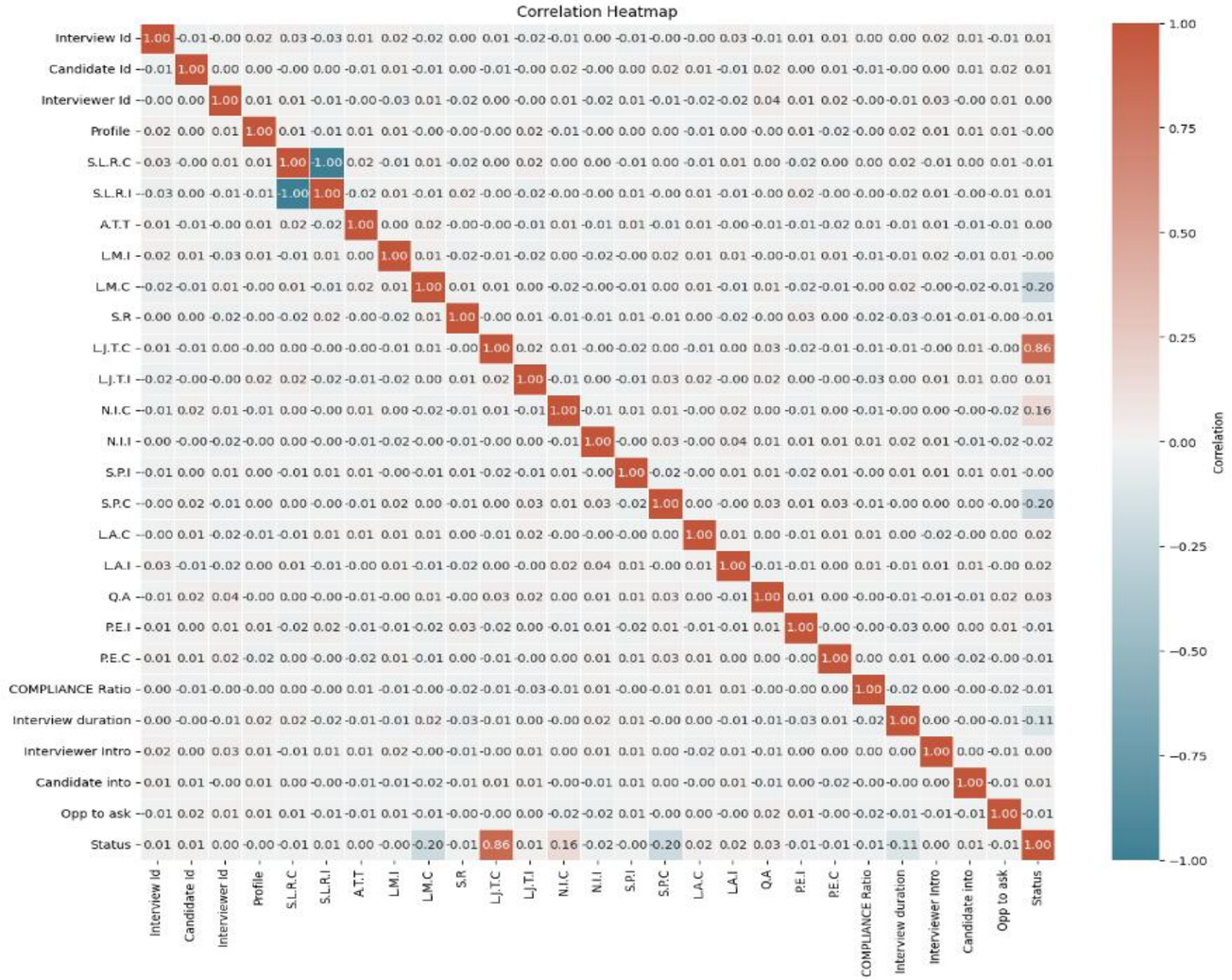
May Consider: 1890

Another group with potential are 32.6%

Not Consider: 1595

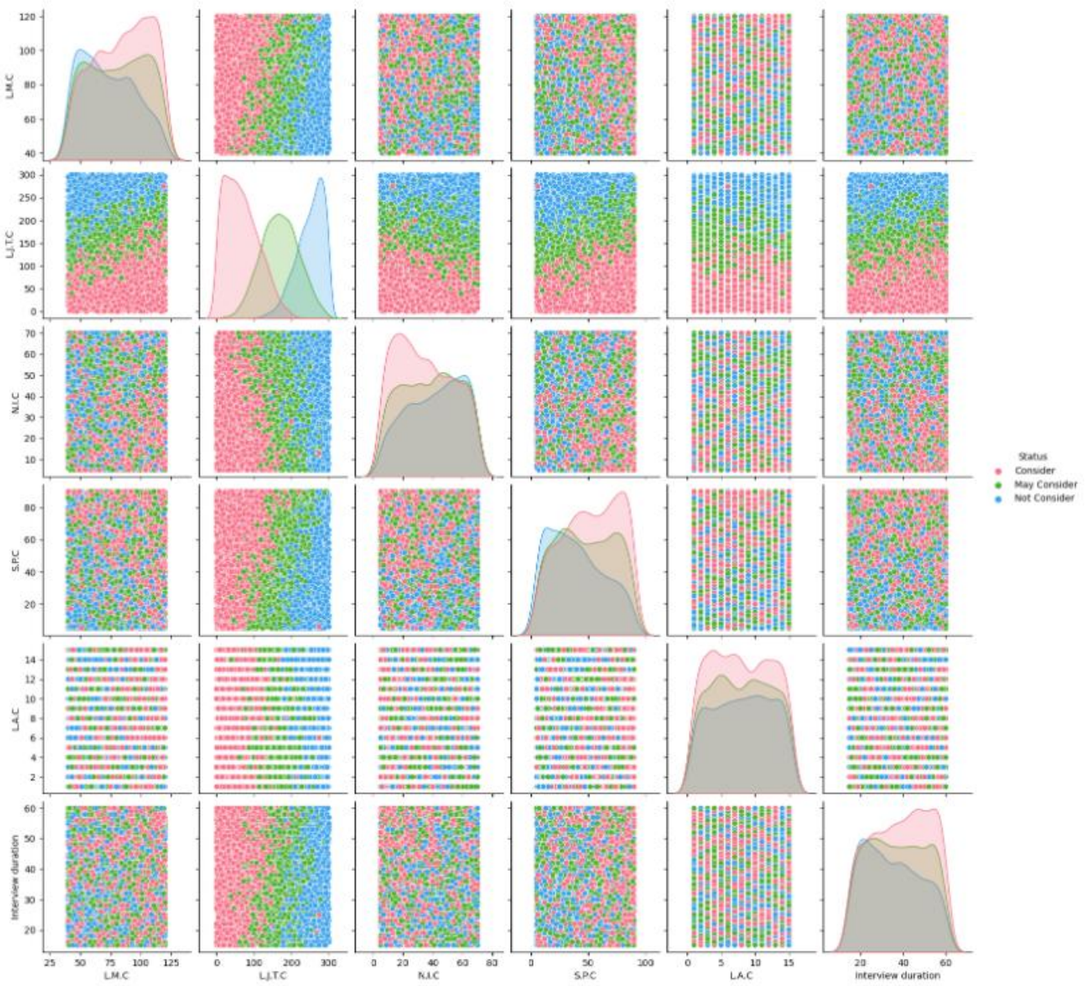
A smaller number didn't meet the criteria are 27.5%

• HEATMAP CORRELATION



We can visually identify the relationships between columns and ' Status ' on our heatmap. The colour variations help us see which columns have a notable impact on predicting the 'Status' outcome.

- **PAIR-PLOT OF CONTINUOUS FEATURE BY STATUS**



I noticed that only the 'Late Joining Time Candidate' (L.J.T.C) column provides a more visible insight on the pair plot. Additionally, it is highly correlated with the target column, as observed in the heatmap.

3. Feature Selection

1

I picked these columns after analyzing their correlation using a heatmap:

1) L.M.C - Longest Monologue Candidate.

2) L.J.T.C - Late Joining Time Candidate.

3)N.I.C - Noise Index Candidate.

4) S.P.C - Speaking Pace Candidate.

5)L.A.C - Live Absence candidate.

6) Interview Duration.

7) Status.

2

This selection ensures that these features contribute meaningfully to our machine learning model's ability to predict interview outcomes.

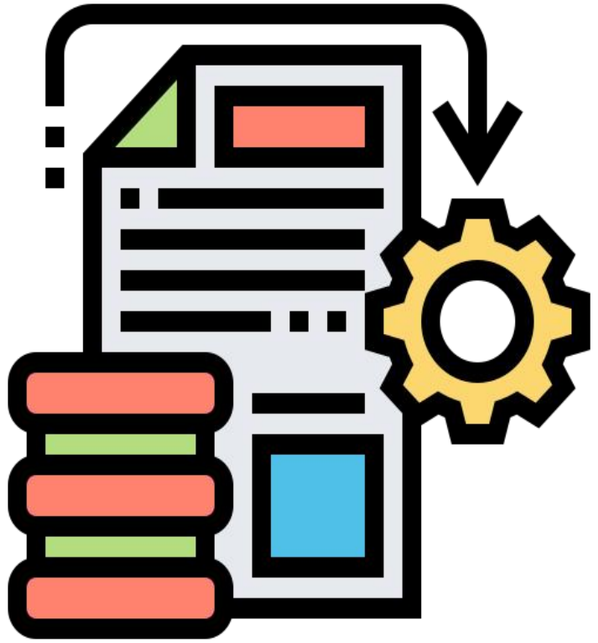
4. Data - Preprocessing

1. I realized I don't need the categorical columns and those with null values in our selected features.

2. Then I split the data into two parts ‘ **X** ’ and ‘ **Y** ’ .

   In which **X** holds important features we select after feature selection.

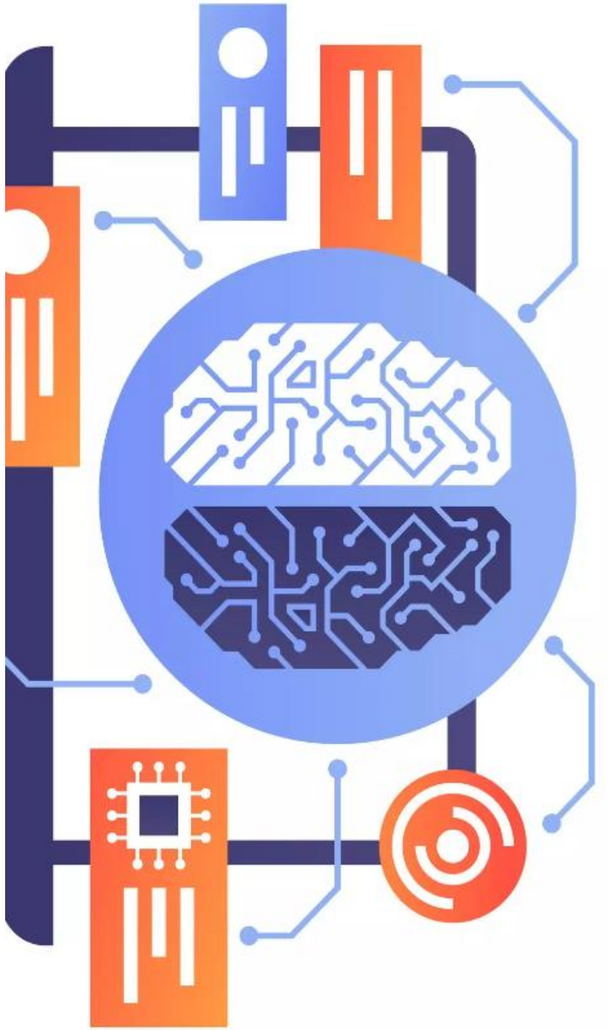
   Y is what we're trying to predict **("Status")**



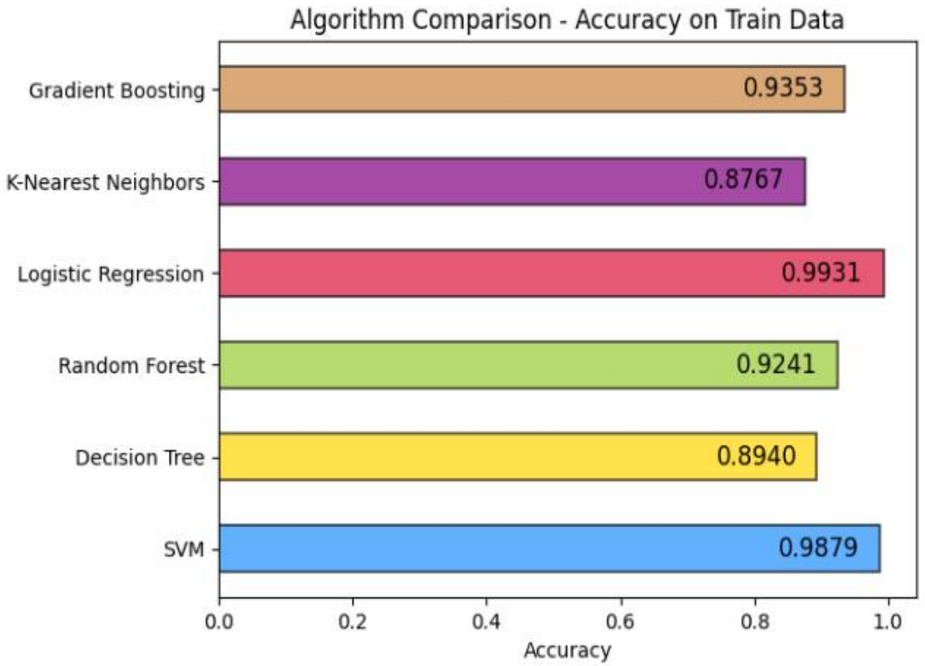
3. Then I applied **StandardScaler()** to standardize the units of each column in ‘ **X** ’

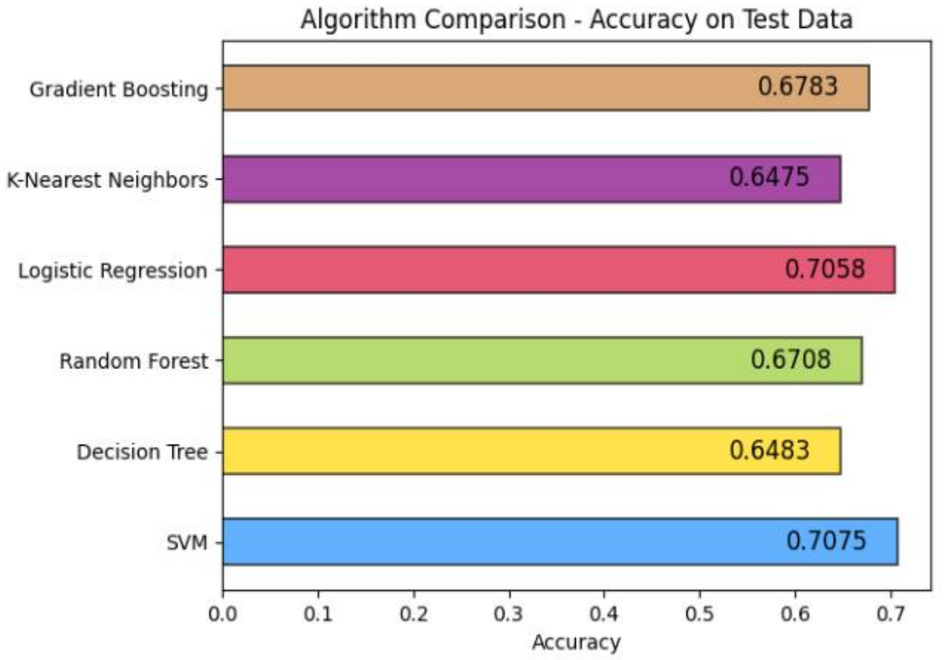
5. Model Building

- I split the data into training and testing sets with an **80:20** ratio.
- Compiled a list of chosen models for predictions: SVM, Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors, and Gradient Boosting with respect to their tuning technique.
- I observed the accuracy across the selected models are very high which might be over-fitted
- After that I check the models prediction on test dataset to check how well its predicting the unseen data. then I found that the model was really over-fitted.



- **Accuracy on Training and Test Data:**





So after trying many tuning parameters on all models only SVM and Logistic Regression has the better and same f1 score (" 75"," 72"," 63") for each class in target column and both achieve a good accuracy of approximately "71" percent.

CHALLENGE FACED:

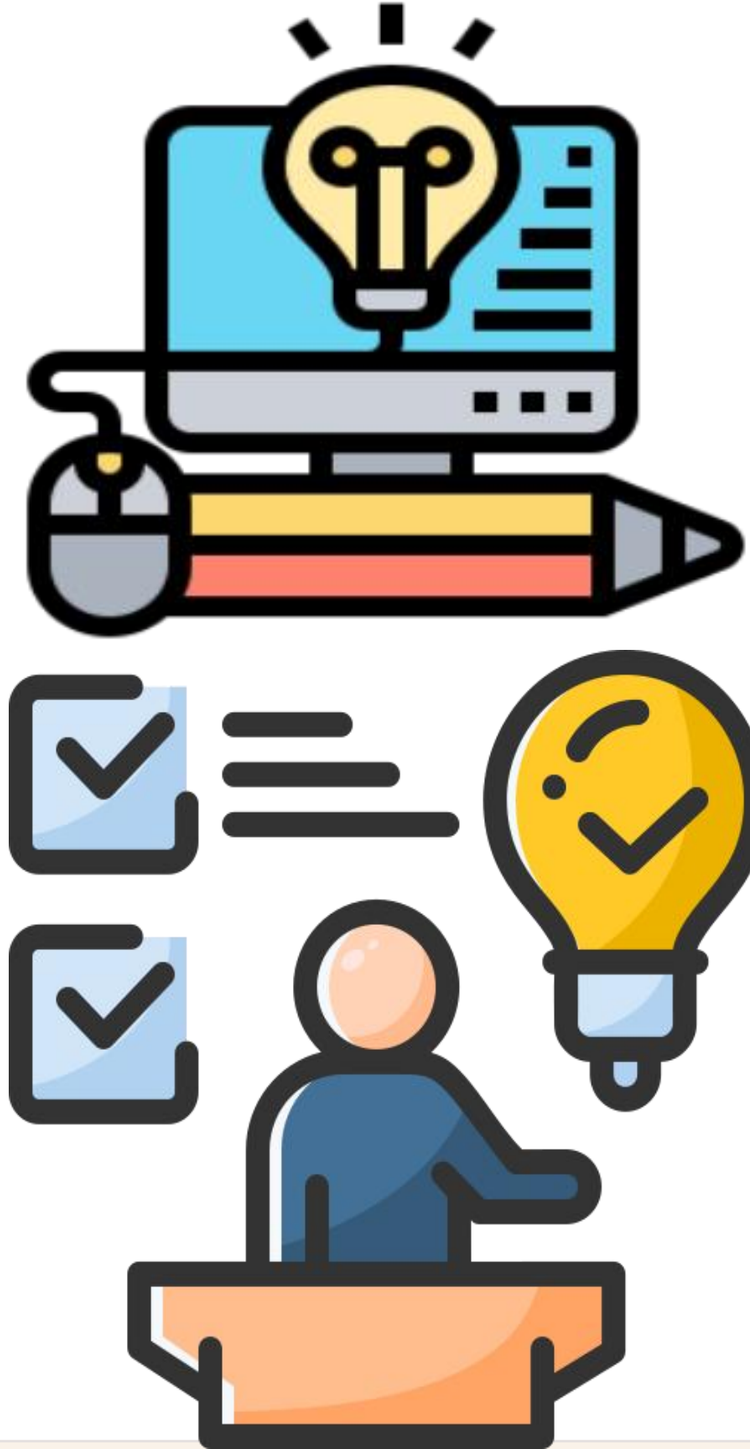
There were many problems I faced while working on this dataset:

1. Many features are not important for model building so need to select the features correctly.
2. All models were overfitted maybe beacuse of less data or status column highly related with L.J.T.C column .
3. Tuning taking alot of time and using tuning in every model was very tough so i mostly focus tuning logistic regression and svm
4. After using many algorithms and tuning them can only take the accuracy till close to "71" percent maybe can it can be increased if we use deep learning.

CONCLUSION:

So, I found that speaking for a longer time (L.M.C.) increases the chances of a positive consideration. However, being late (L.J.T.C.) or absent (L.A.C.) can negatively impact a candidate's status, depending on the circumstances. Surprisingly, interview duration doesn't play a significant role in hiring decisions.

Although machine learning achieved a decent 71 percent accuracy in predicting status.I might get more accurate prediction if I had more data. So, the complex nature of hiring suggests there may be room for improvement with deep learning.




THANK YOU