Interview Status Prediction by Machine Learning

By: Eng. Anwar Hashem (PhD Student).

Introduction:

The hiring entity or a specific service provider relies heavily on the personal interview in making employment or providing services decisions, as the interviewer tests the candidate's analytical skills, and thus can select candidates suitable for the organization's needs. But with the expansion of services and its conditions and the large number of applicants, the process becomes very cumbersome, and requires the use of machine learning to predict the results of the interview.

Machine learning algorithms have shown significant promise in accurately predicting interview-related outcomes. Various studies have explored the application of machine learning in predicting interview attendance.

This project focuses on predicting the status of job interviews using machine learning techniques. The goal is to create models that effectively determine whether a candidate would be considered, might be considered, or wouldn't be considered post an interview.

The Project Objective:

The "Interview Status Prediction" project focuses on building a predictive model to determine the status of job interviews, i.e. predict whether a candidate will pass or fail an interview. The goal is to create models that effectively determine whether a candidate would be considered, might be considered, or wouldn't be considered post an interview.

Data Description:

The dataset for the interview status prediction project consists of three CSV files as the following:

	CSV name	**The purpose**	**Samples**	**Features**	**Source**	**Notices**
1	train	to train the machine learning model (learn patterns and relationships relevant to predicting interview outcomes).	5800	27	https://fastupload.io/6c39710a81864f69	
2	test	to evaluate the performance of the trained model (predicts outcomes on this unseen data)	1200	26	https://fastupload.io/d1408114e1a38d33	
3	result	to store the predictions made by the model on the test set.	1200	2	https://fastupload.io/b6d9f28a70ae2722	

These files provide the necessary data for training, validating, and testing a predictive model for interview outcomes.

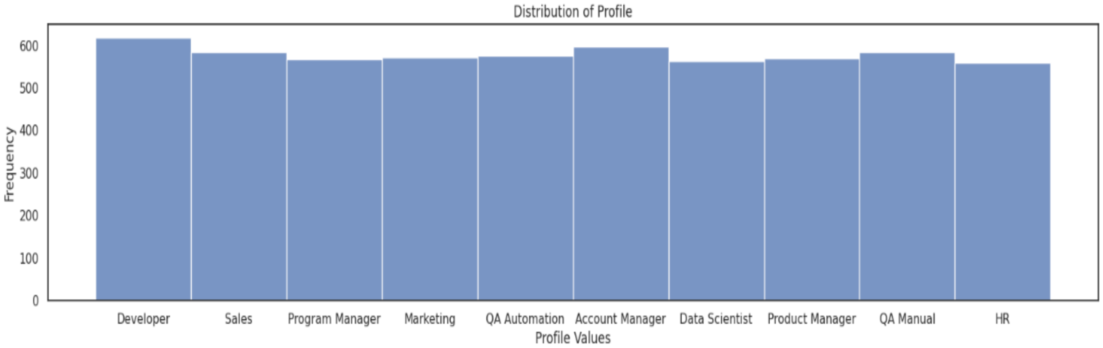
Methodology:

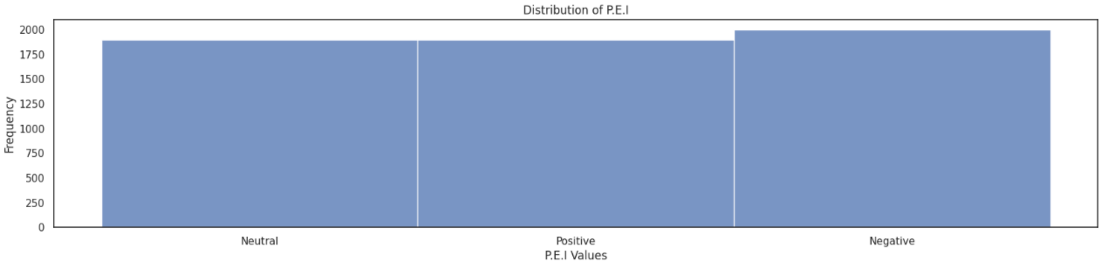
The steps taken during data preprocessing, EDA, and feature engineering. Include justifications for the methods used.

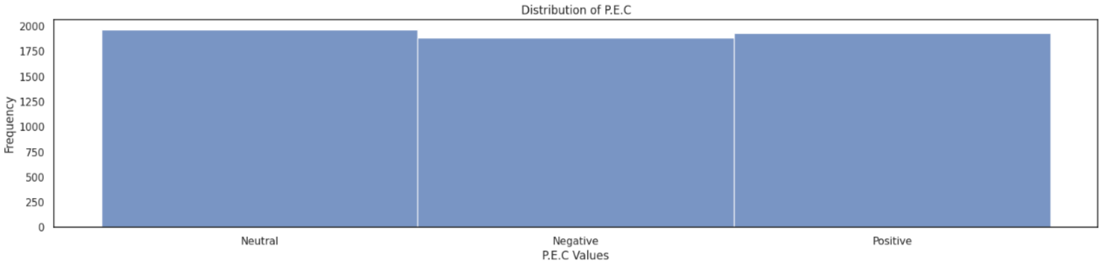
		Step	Details	Result of Implementation	Notices
Data Preprocessing	Data Cleaning	Removed missing values or filled nulls using mode/mean.	☒ **In Train Dataset**: - The rows (349, 3807, 3816, 5795) have null values in output (Status). - The rows (360) have null values in the column (Interviewer Intro. - The rows (16,50,1992) have null values in the column (P.E.C). - The rows (3814) have null values in the column (P.E.I). ☒ **In Test Dataset**: - the rows (8, 55, 56, 61, 152, 157, 215, 247, 268, 305, 318, 460, 524) have null values in Features columns.	✓ Remove the rows that have null values.	Since the missing records are very small and will not affect the integrity of the data, we can delete them.
	Data Cleaning	dropped redundant columns.	There are no duplicates in the training and test data.	No duplicates columns.	
	Exploratory Data Analysis (EDA)	Explored data distribution.	Explored data distribution for Categorical columns	It is as shown in the figures below(a).	
		Identified outliers	Identified outliers for Numerical columns by: Boxplot. - Calculating the outlier. -	It is as shown in the figures below(b). **We found that there are no outliers.**	
		Checked correlations between features	draw heatmap	It is as shown in the figure below (C). we found that **the highly correlated Features is**: S.L.R.C, S.L.R.I	
Feature Engineering	Label Encoding	Converted categorical variables into numerical representations for model compatibility.	The Categorical Variables are: `'Profile', 'P.E.I', 'P.E.C', 'Interviewer Intro', 'Candidate`	All Categorical columns were converted to numerical, and their encoding was verified.	

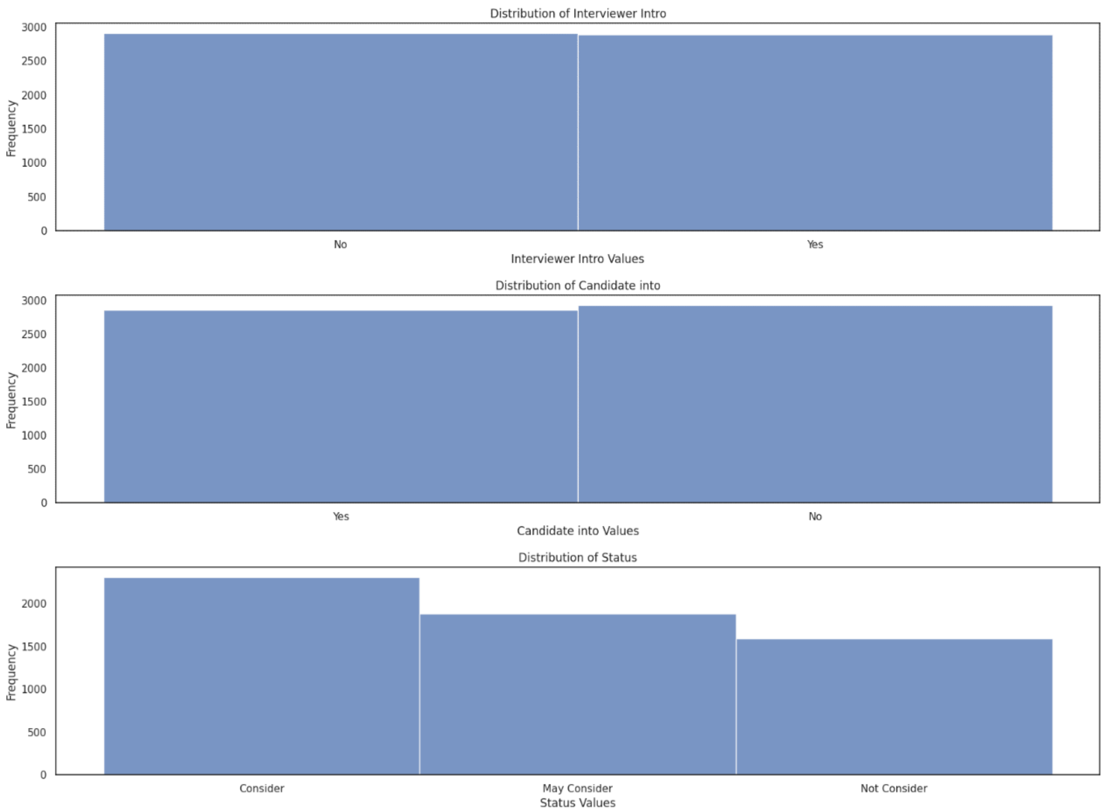
		Step	Details	Result of Implementation	Notices
			into', 'Opp to ask', 'Status'.		
	Feature Selection	Selected relevant features for training the models based on their impact on interview status prediction.	20% of the total features were selected to obtain the most important features.	We found that the importance order of features as the following: 1- L.J.T.C (-0.856376) 2- S.P.C (0.203198) 3- L.M.C (0.194775) 4- N.I.C (-0.155696) 5- Interview duration (0.114197)	

A- Explored data distribution for Categorical columns:

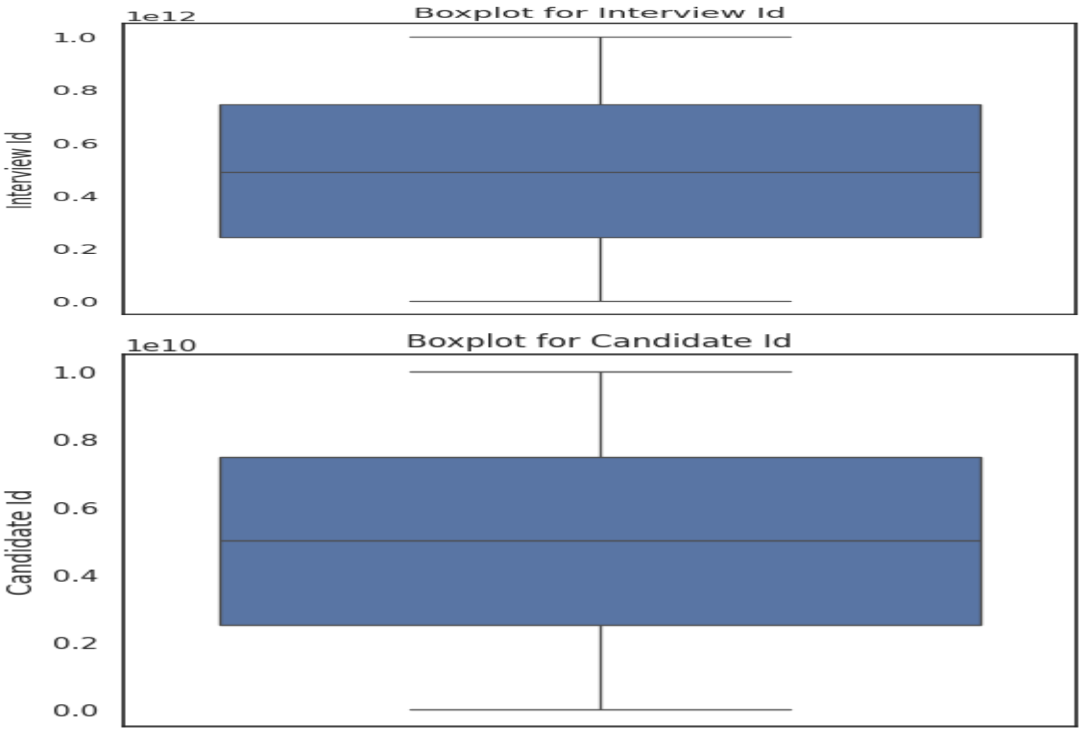


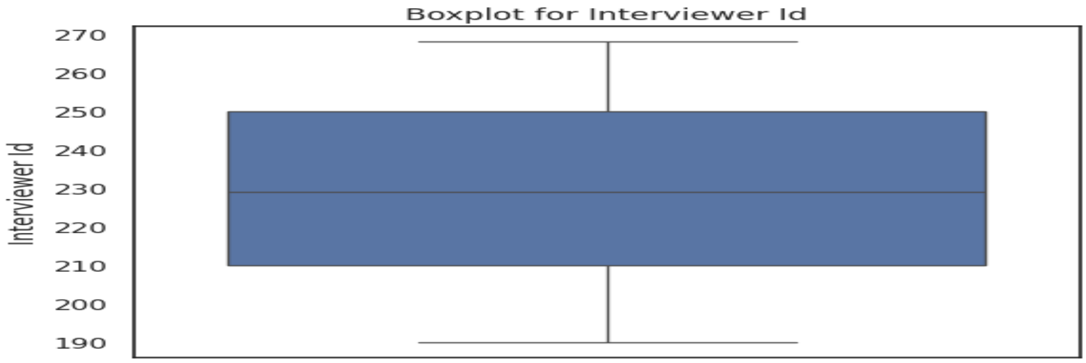


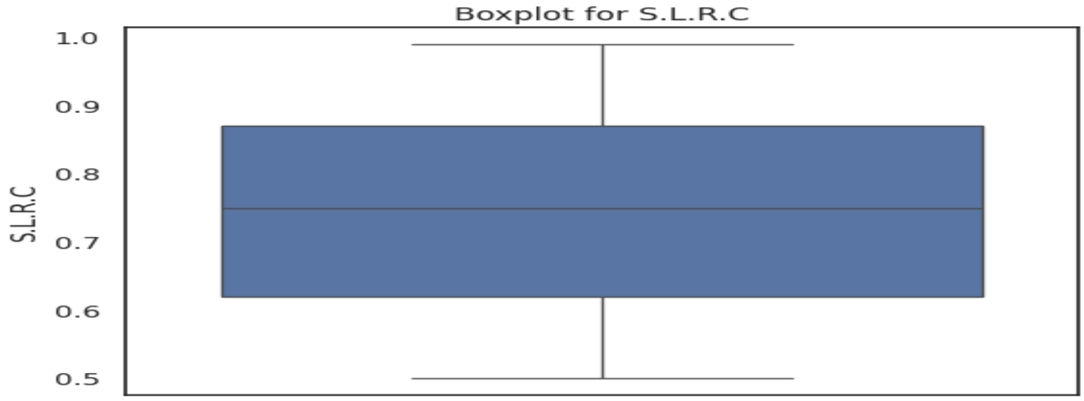


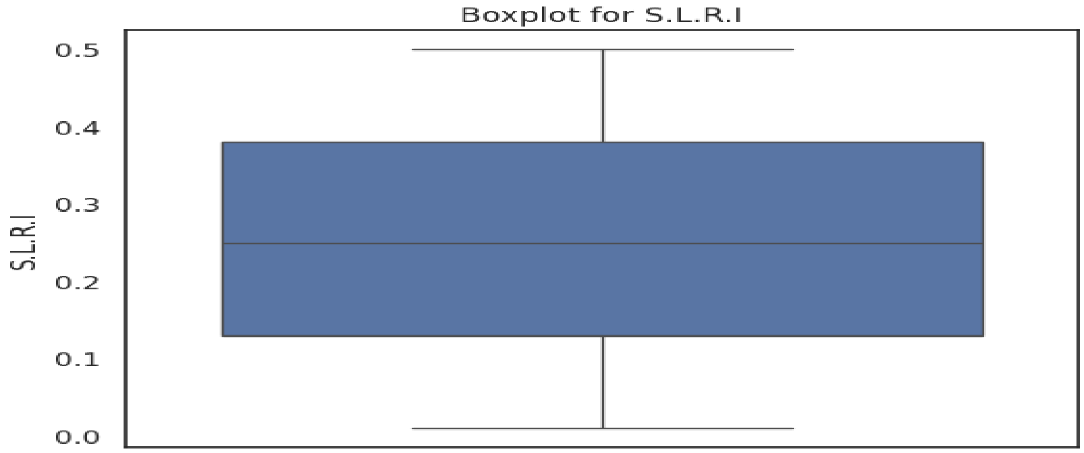


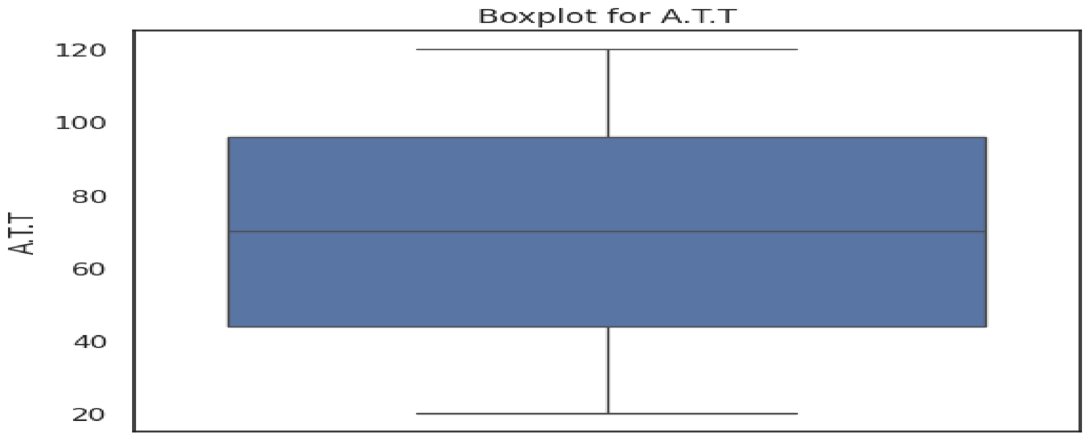
B- Identified outliers for Numerical columns:

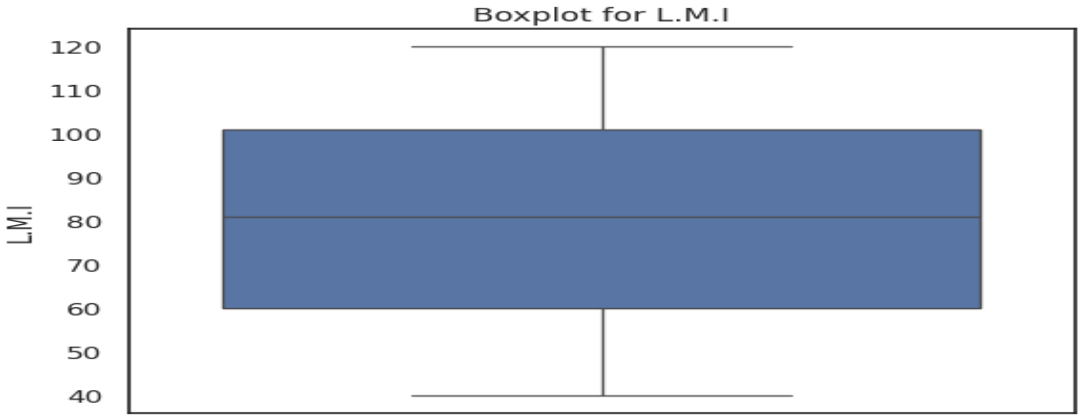


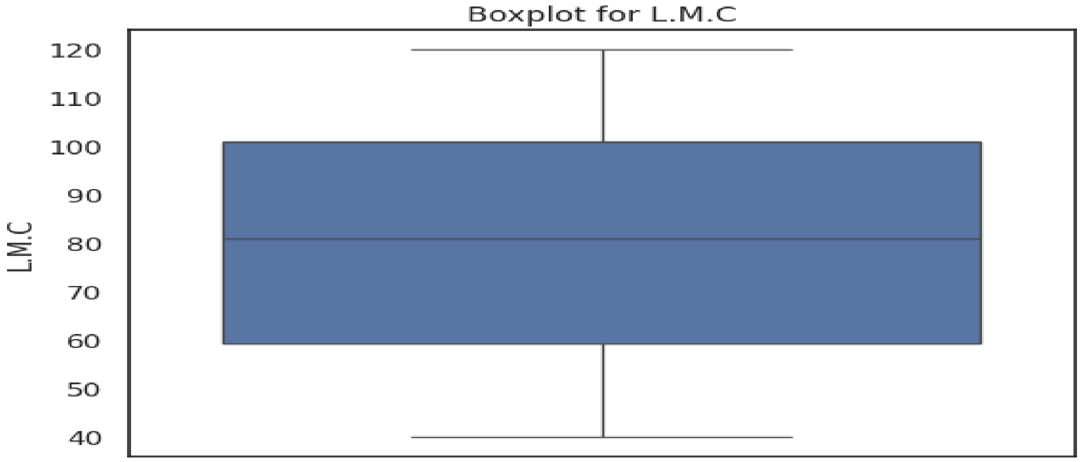

Boxplot for Interviewer Id

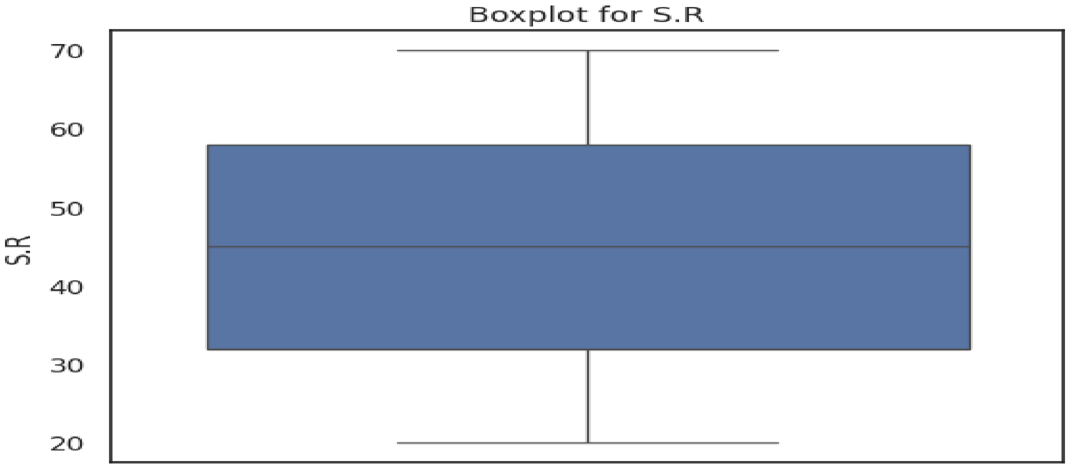

Boxplot for S.L.R.C

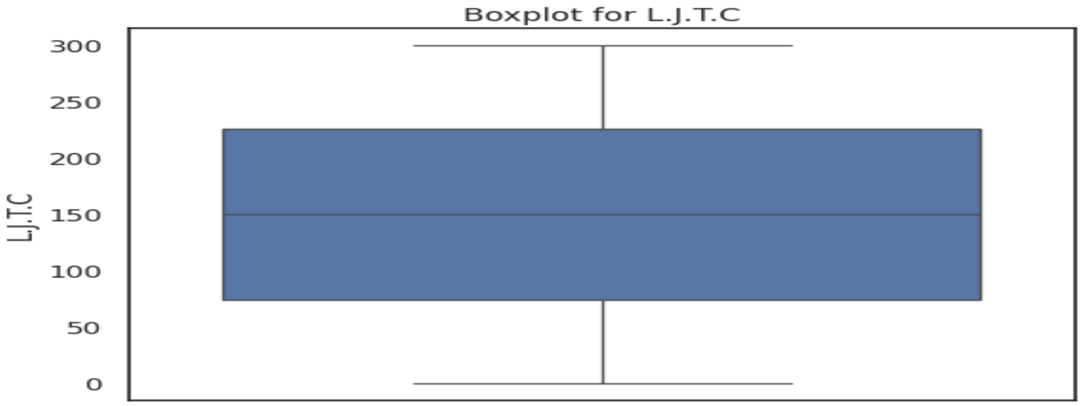

Boxplot for S.L.R.I

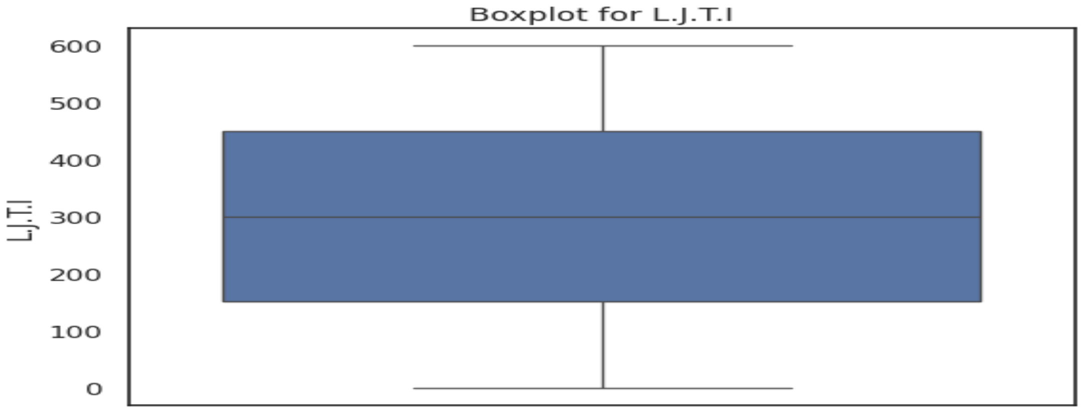

Boxplot for A.T.T

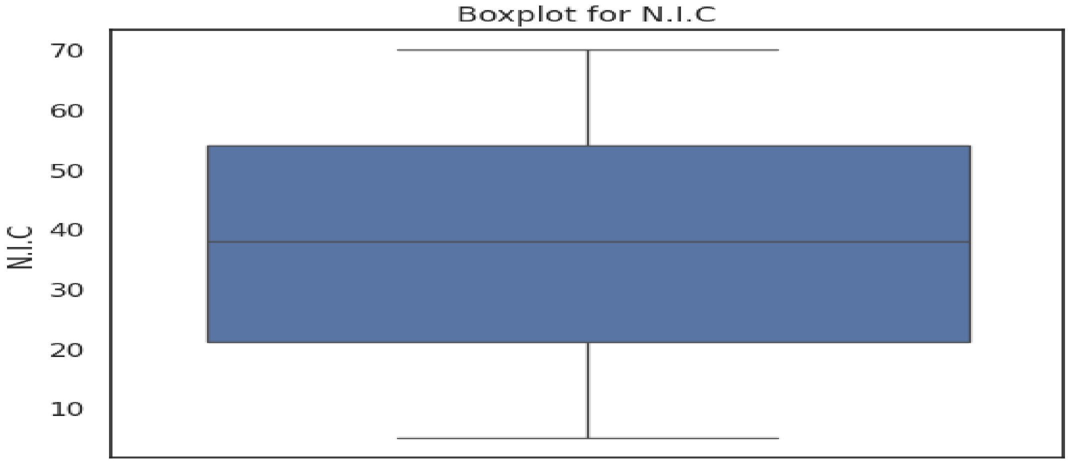

Boxplot for L.M.I
L.M.I
120
110
100
90
80
70
60
50
40

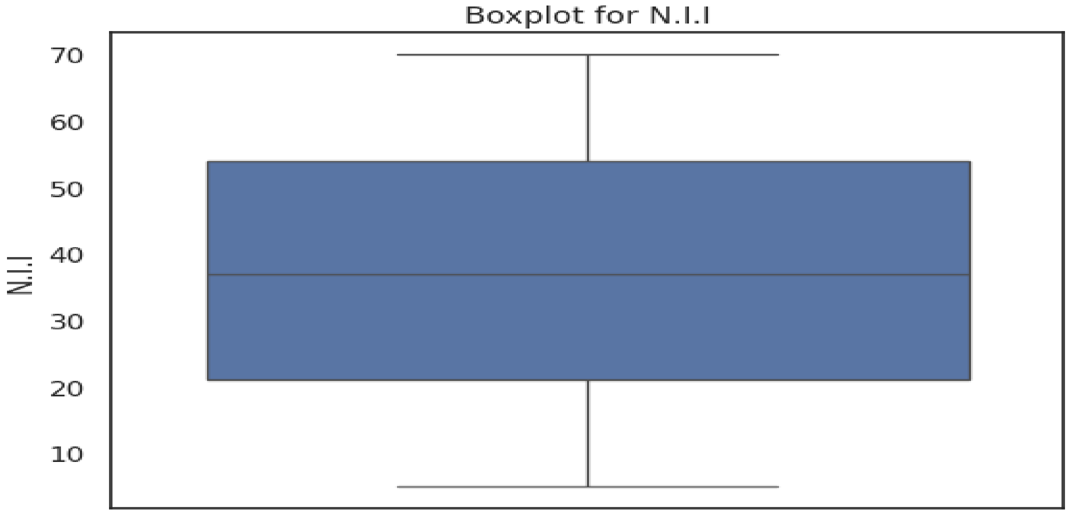

Boxplot for L.M.C
L.M.C
120
110
100
90
80
70
60
50
40

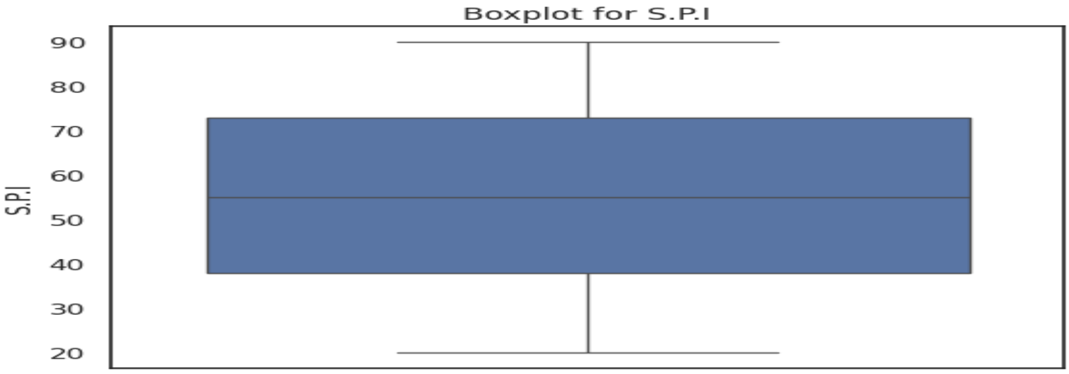

Boxplot for S.R
S.R
70
60
50
40
30
20

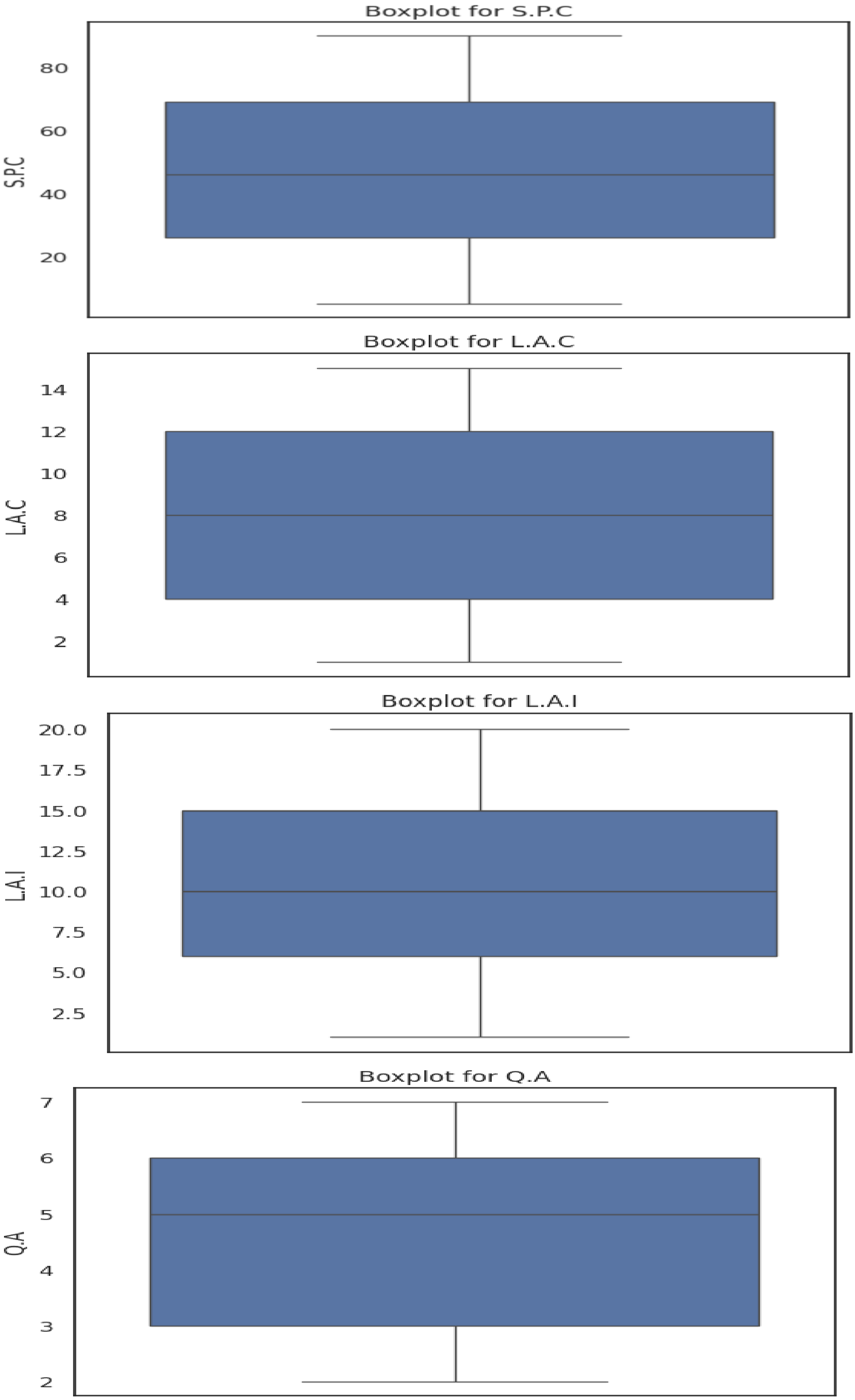

Boxplot for L.J.T.C
L.J.T.C
300
250
200
150
100
50
0

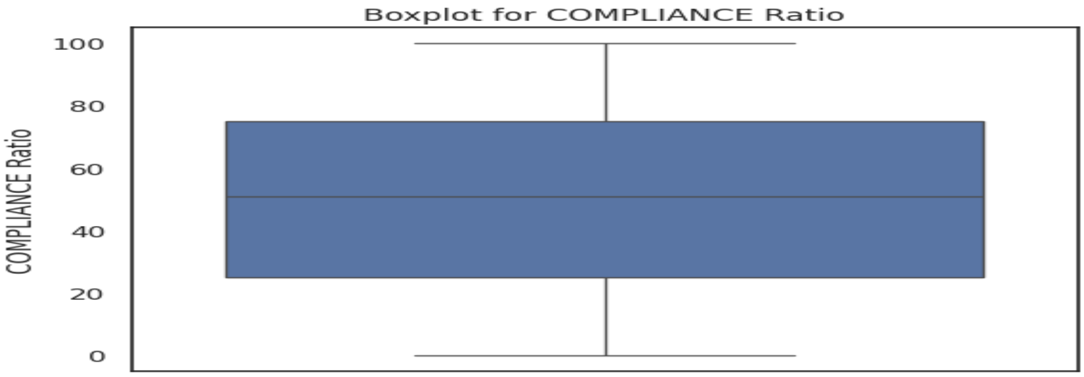

Boxplot for L.J.T.I
L.J.T.I
600
500
400
300
200
100
0

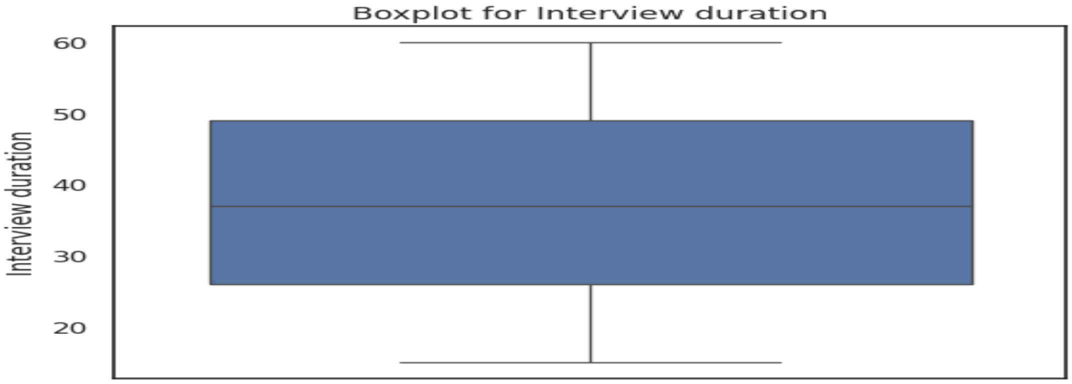

Boxplot for N.I.C
N.I.C
70
60
50
40
30
20
10


Boxplot for N.I.I
N.I.I
70
60
50
40
30
20
10


Boxplot for S.P.I
S.P.I
90
80
70
60
50
40
30
20


Boxplot for S.P.C
S.P.C
80
60
40
20
Boxplot for L.A.C
L.A.C
14
12
10
8
6
4
2
Boxplot for L.A.I
L.A.I
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
Boxplot for Q.A
Q.A
7
6
5
4
3
2

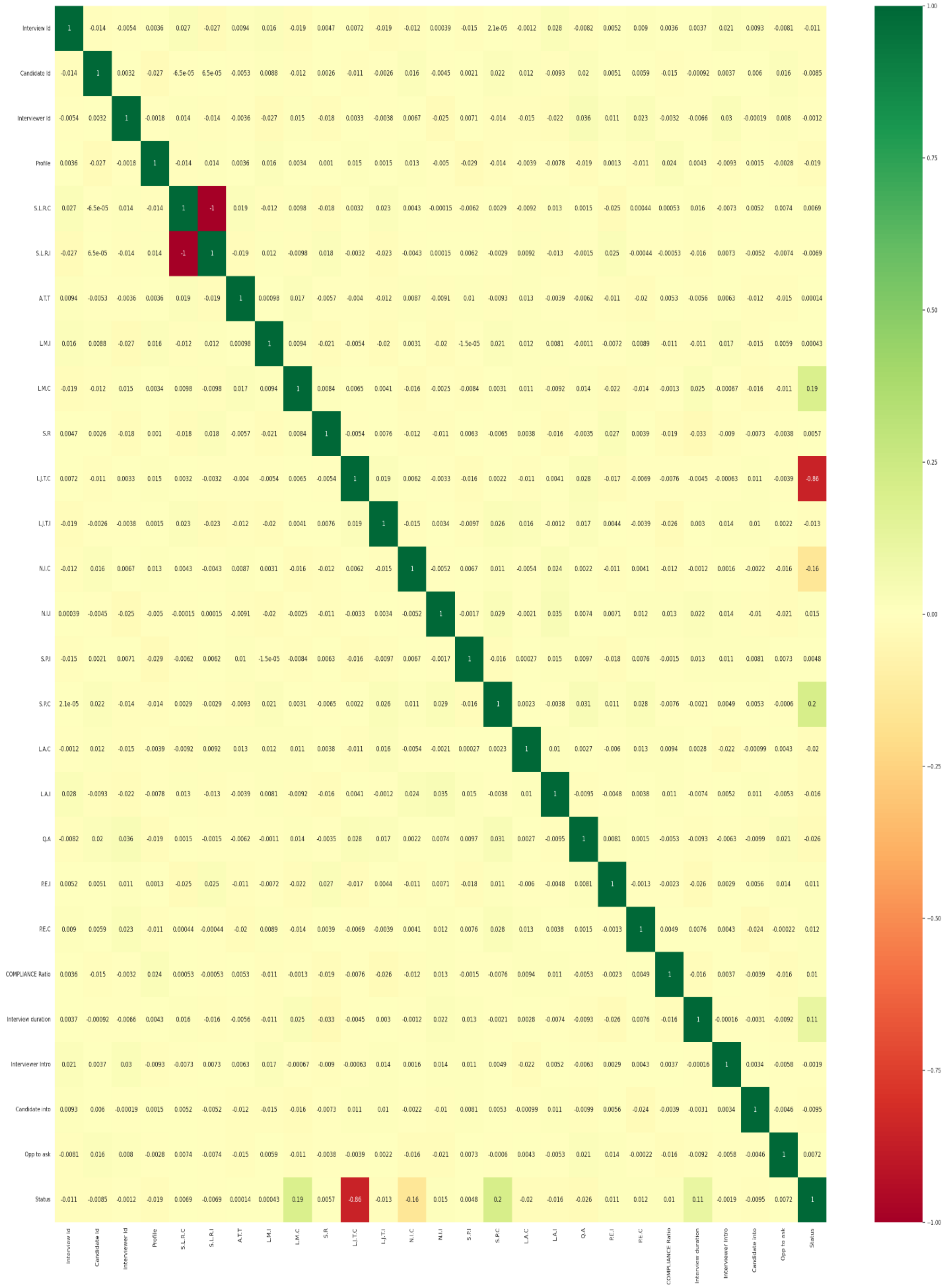




We found that **Interview duration:**

- Mean: **37.3** minutes
- Standard deviation: **13.2** minutes.
- Shortest interview: **15** minutes
- Longest Interview: **60** minutes

C- Checked correlations between features:



	Interview Id	Candidate Id	Interviewer Id	Profile	S.L.R.C	S.L.R.I	A.T.T	L.M.I	L.M.C	S.R	L.J.T.C	L.J.T.I	N.I.C	N.I.I	S.P.I	S.P.C	L.A.C	L.A.I	Q.A	P.E.I	P.E.C	COMPLIANCE Ratio	Interview duration	Interviewer Intro	Candidate into	Opp to ask	Status
Interview Id	1	-0.014	-0.0054	0.0036	0.027	-0.027	0.0094	0.016	-0.019	0.0047	0.0072	-0.019	-0.012	0.00039	-0.015	2.1e-05	-0.0012	0.028	-0.0082	0.0052	0.009	0.0036	0.0037	0.021	0.0093	-0.0081	-0.011
Candidate Id	-0.014	1	0.0032	-0.027	-6.5e-05	6.5e-05	-0.0053	0.0088	-0.012	0.0026	-0.011	-0.0026	0.016	-0.0045	0.0021	0.022	0.012	-0.0093	0.02	0.0051	0.0059	-0.015	-0.00092	0.0037	0.006	0.016	-0.0085
Interviewer Id	-0.0054	0.0032	1	-0.0018	0.014	-0.014	-0.0036	-0.027	0.015	-0.018	0.0033	-0.0038	0.0067	-0.025	0.0071	-0.014	-0.015	-0.022	0.036	0.011	0.023	-0.0032	-0.0066	0.03	-0.00019	0.008	-0.0012
Profile	0.0036	-0.027	-0.0018	1	-0.014	0.014	0.0036	0.016	0.0034	0.001	0.015	0.0015	0.013	-0.005	-0.029	-0.014	-0.0039	-0.0078	-0.019	0.0013	-0.011	0.024	0.0043	-0.0093	0.0015	-0.0028	-0.019
S.L.R.C	0.027	-6.5e-05	0.014	-0.014	1	-1	0.019	-0.012	0.0098	-0.018	0.0032	0.023	0.0043	-0.00015	-0.0062	0.0029	-0.0092	0.013	0.0015	-0.025	0.00044	0.00053	0.016	-0.0073	0.0052	0.0074	0.0069
S.L.R.I	-0.027	6.5e-05	-0.014	0.014	-1	1	-0.019	0.012	-0.0098	0.018	-0.0032	-0.023	-0.0043	0.00015	0.0062	-0.0029	0.0092	-0.013	-0.0015	0.025	-0.00044	-0.00053	-0.016	0.0073	-0.0052	-0.0074	-0.0069
A.T.T	0.0094	-0.0053	-0.0036	0.0036	0.019	-0.019	1	0.00098	0.017	-0.0057	-0.004	-0.012	0.0087	-0.0091	0.01	-0.0093	0.013	-0.0039	-0.0062	-0.011	-0.02	0.0053	-0.0056	0.0063	-0.012	-0.015	0.00014
L.M.I	0.016	0.0088	-0.027	0.016	-0.012	0.012	0.00098	1	0.0094	-0.021	-0.0054	-0.02	0.0031	-0.02	-1.5e-05	0.021	0.012	0.0081	-0.0011	-0.0072	0.0089	-0.011	-0.011	0.017	-0.015	0.0059	0.00043
L.M.C	-0.019	-0.012	0.015	0.0034	0.0098	-0.0098	0.017	0.0094	1	0.0084	0.0065	0.0041	-0.016	-0.0025	-0.0084	0.0031	0.011	-0.0092	0.014	-0.022	-0.014	-0.0013	0.025	-0.00067	-0.016	-0.011	0.19
S.R	0.0047	0.0026	-0.018	0.001	-0.018	0.018	-0.0057	-0.021	0.0084	1	-0.0054	0.0076	-0.012	-0.011	0.0063	-0.0065	0.0038	-0.016	-0.0035	0.027	0.0039	-0.019	-0.033	-0.009	-0.0073	-0.0038	0.0057
L.J.T.C	0.0072	-0.011	0.0033	0.015	0.0032	-0.0032	-0.004	-0.0054	0.0065	-0.0054	1	0.019	0.0062	-0.0033	-0.016	0.0022	-0.011	0.0041	0.028	-0.017	-0.0069	-0.0076	-0.0045	-0.00063	0.011	-0.0039	-0.86
L.J.T.I	-0.019	-0.0026	-0.0038	0.0015	0.023	-0.023	-0.012	-0.02	0.0041	0.0076	0.019	1	-0.015	0.0034	-0.0097	0.026	0.016	-0.0012	0.017	0.0044	-0.0039	-0.026	0.003	0.014	0.01	0.0022	-0.013
N.I.C	-0.012	0.016	0.0067	0.013	0.0043	-0.0043	0.0087	0.0031	-0.016	-0.012	0.0062	-0.015	1	-0.0052	0.0067	0.011	-0.0054	0.024	0.0022	-0.011	0.0041	-0.012	-0.0012	0.0016	-0.0022	-0.016	-0.16
N.I.I	0.00039	-0.0045	-0.025	-0.005	-0.00015	0.00015	-0.0091	-0.02	-0.0025	-0.011	-0.0033	0.0034	-0.0052	1	-0.0017	0.029	-0.0021	0.035	0.0074	0.0071	0.012	0.013	0.022	0.014	-0.01	-0.021	0.015
S.P.I	-0.015	0.0021	0.0071	-0.029	-0.0062	0.0062	0.01	-1.5e-05	-0.0084	0.0063	-0.016	-0.0097	0.0067	-0.0017	1	-0.016	0.00027	0.015	0.0097	-0.018	0.0076	-0.0015	0.013	0.011	0.0081	0.0073	0.0048
S.P.C	2.1e-05	0.022	-0.014	-0.014	0.0029	-0.0029	-0.0093	0.021	0.0031	-0.0065	0.0022	0.026	0.011	0.029	-0.016	1	0.0023	-0.0038	0.031	0.011	0.028	-0.0076	-0.0021	0.0049	0.0053	-0.0006	0.2
L.A.C	-0.0012	0.012	-0.015	-0.0039	-0.0092	0.0092	0.013	0.012	0.011	0.0038	-0.011	0.016	-0.0054	-0.0021	0.00027	0.0023	1	0.01	0.0027	-0.006	0.013	0.0094	0.0028	-0.022	-0.00099	0.0043	-0.02
L.A.I	0.028	-0.0093	-0.022	-0.0078	0.013	-0.013	-0.0039	0.0081	-0.0092	-0.016	0.0041	-0.0012	0.024	0.035	0.015	-0.0038	0.01	1	-0.0095	-0.0048	0.0038	0.011	-0.0074	0.0052	0.011	-0.0053	-0.016
Q.A	-0.0082	0.02	0.036	-0.019	0.0015	-0.0015	-0.0062	-0.0011	0.014	-0.0035	0.028	0.017	0.0022	0.0074	0.0097	0.031	0.0027	-0.0095	1	0.0081	0.0015	-0.0053	-0.0093	-0.0063	-0.0099	0.021	-0.026
P.E.I	0.0052	0.0051	0.011	0.0013	-0.025	0.025	-0.011	-0.0072	-0.022	0.027	-0.017	0.0044	-0.011	0.0071	-0.018	0.011	-0.006	-0.0048	0.0081	1	-0.0013	-0.0023	-0.026	0.0029	0.0056	0.014	0.011
P.E.C	0.009	0.0059	0.023	-0.011	0.00044	-0.00044	-0.02	0.0089	-0.014	0.0039	-0.0069	-0.0039	0.0041	0.012	0.0076	0.028	0.013	0.0038	0.0015	-0.0013	1	0.0049	0.0076	0.0043	-0.024	-0.00022	0.012
COMPLIANCE Ratio	0.0036	-0.015	-0.0032	0.024	0.00053	-0.00053	0.0053	-0.011	-0.0013	-0.019	-0.0076	-0.026	-0.012	0.013	-0.0015	-0.0076	0.0094	0.011	-0.0053	-0.0023	0.0049	1	-0.016	0.0037	-0.0039	-0.016	0.01
Interview duration	0.0037	-0.00092	-0.0066	0.0043	0.016	-0.016	-0.0056	-0.011	0.025	-0.033	-0.0045	0.003	-0.0012	0.022	0.013	-0.0021	0.0028	-0.0074	-0.0093	-0.026	0.0076	-0.016	1	-0.00016	-0.0031	-0.0092	0.11
Interviewer Intro	0.021	0.0037	0.03	-0.0093	-0.0073	0.0073	0.0063	0.017	-0.00067	-0.009	-0.00063	0.014	0.0016	0.014	0.011	0.0049	-0.022	0.0052	-0.0063	0.0029	0.0043	0.0037	-0.00016	1	0.0034	-0.0058	-0.0019
Candidate into	0.0093	0.006	-0.00019	0.0015	0.0052	-0.0052	-0.012	-0.015	-0.016	-0.0073	0.011	0.01	-0.0022	-0.01	0.0081	0.0053	-0.00099	0.011	-0.0099	0.0056	-0.024	-0.0039	-0.0031	0.0034	1	-0.0046	-0.0095
Opp to ask	-0.0081	0.016	0.008	-0.0028	0.0074	-0.0074	-0.015	0.0059	-0.011	-0.0038	-0.0039	0.0022	-0.016	-0.021	0.0073	-0.0006	0.0043	-0.0053	0.021	0.014	-0.00022	-0.016	-0.0092	-0.0058	-0.0046	1	0.0072
Status	-0.011	-0.0085	-0.0012	-0.019	0.0069	-0.0069	0.00014	0.00043	0.19	0.0057	-0.86	-0.013	-0.16	0.015	0.0048	0.2	-0.02	-0.016	-0.026	0.011	0.012	0.01	0.11	-0.0019	-0.0095	0.0072	1

We found that five columns in training data are important for the target column status.

Modeling:

The Selected Model Name	The Model Description	hyperparameter tuning performed

Decision Tree	A decision tree is a flowchart-like tree structure where an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome.	criterion='gini',
		max_depth=12,
		min_samples_split=15

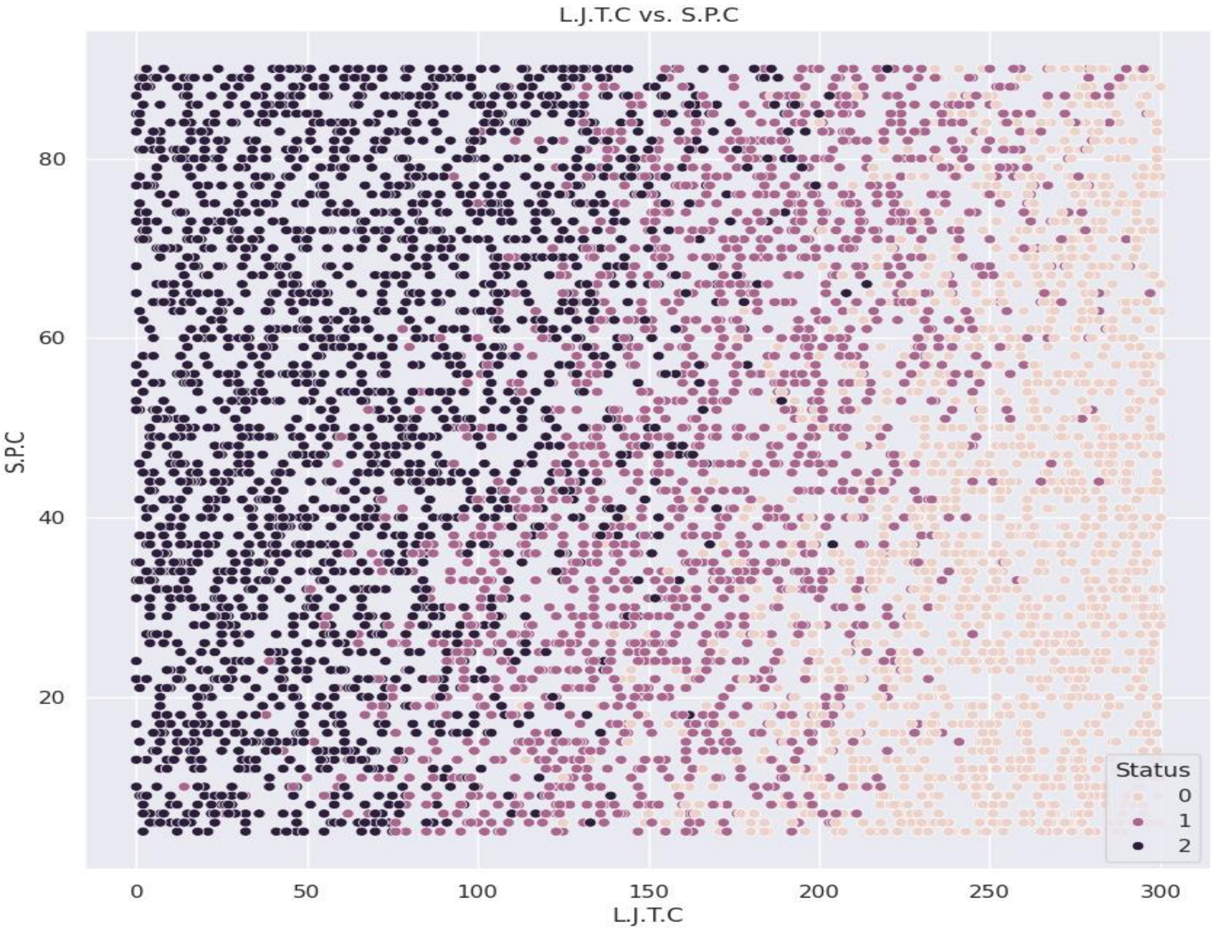
Random Forest	An ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest.	n_estimators=80,
		criterion='gini',
		max_depth=12,
		min_samples_split=15

AdaBoost	Mainly used for classification, and the base learner (the machine learning algorithm that is boosted) is usually a decision tree with only one level, also called as stumps. It makes use of weighted errors to build a strong classifier from a series of weak classifiers.	n_estimators=1000

Gradient Boosting	Based on boosting in a functional space, where the target is pseudo-residuals rather than the typical residuals used in traditional boosting. It gives a prediction model in the form of an ensemble of weak prediction models, i.e., models that make very few assumptions about the data, which are typically simple decision trees.	n_estimators=200,
		learning_rate=0.05,
		random_state=8,
		max_features=15

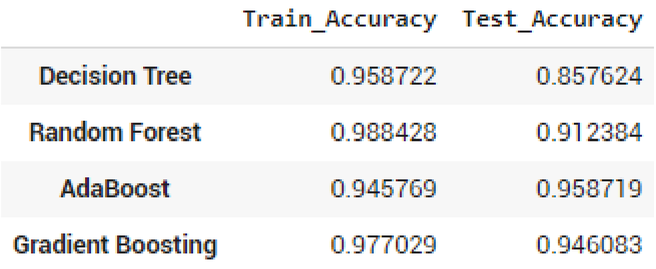
Results and Evaluation:

The visualizing of data:



The evaluation metrics for each model:

The detail of evaluation metrics for each model is shown in attached code file. The summary of the evaluation metrics as the following:



	Train_Accuracy	Test_Accuracy
Decision Tree	0.958722	0.857624
Random Forest	0.988428	0.912384
AdaBoost	0.945769	0.958719
Gradient Boosting	0.977029	0.946083

The result shows that **AdaBoost algorithm performed the best** among the tested models for interview status prediction.

Conclusion:

- There is no noticeable difference when removing the null values and when processing them with the mean and median due the dataset contains a small number of missing values
- AdaBoost algorithm performed the best for interview status prediction.
- This tool can be highly valuable for recruitment processes, helping HR professionals streamline their efforts and make data-driven decisions. This method helps avoid biases, saves time for recruiters, and ensures fair hiring.

Attached files:

interview_status_prediction_by_ml:
https://colab.research.google.com/drive/1xgwAt0L6EBOlwe8AZ3wC2jHWMG4OVu-t#scrollTo=IoLhHh0_Fd4F