

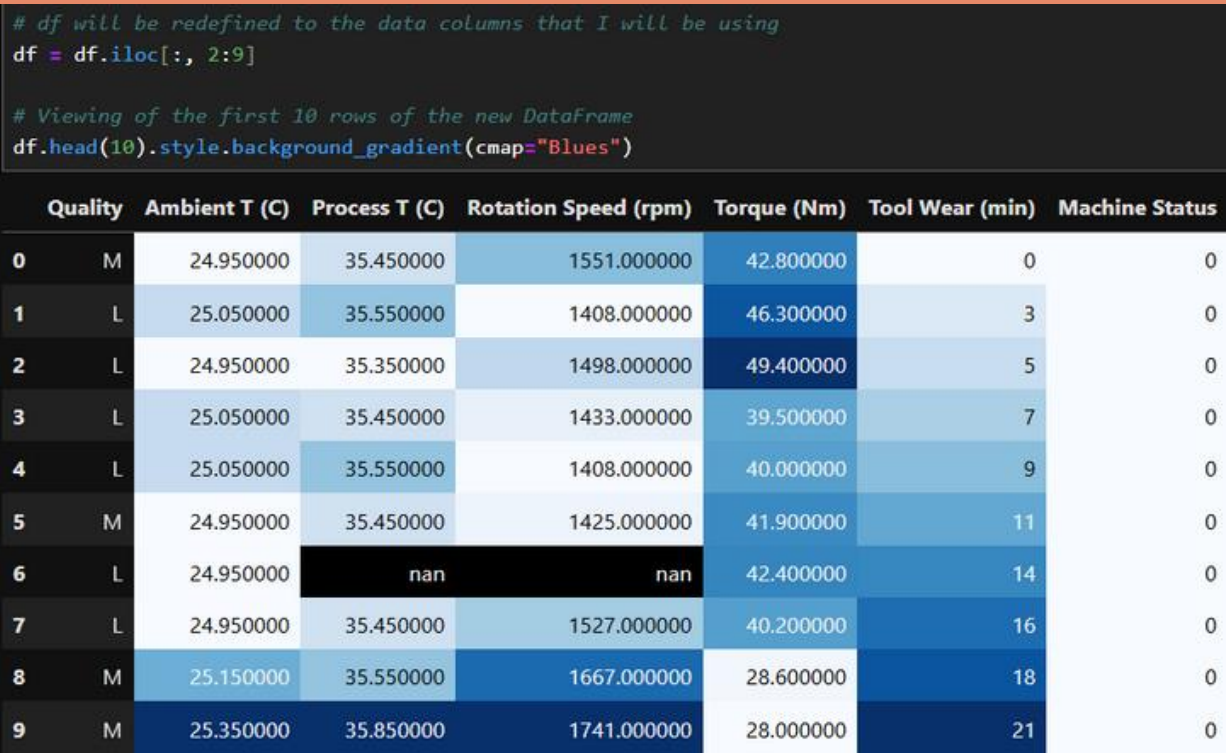
Factory Machine Prediction

GOH KUN MING
DAAA/FT/02/1B

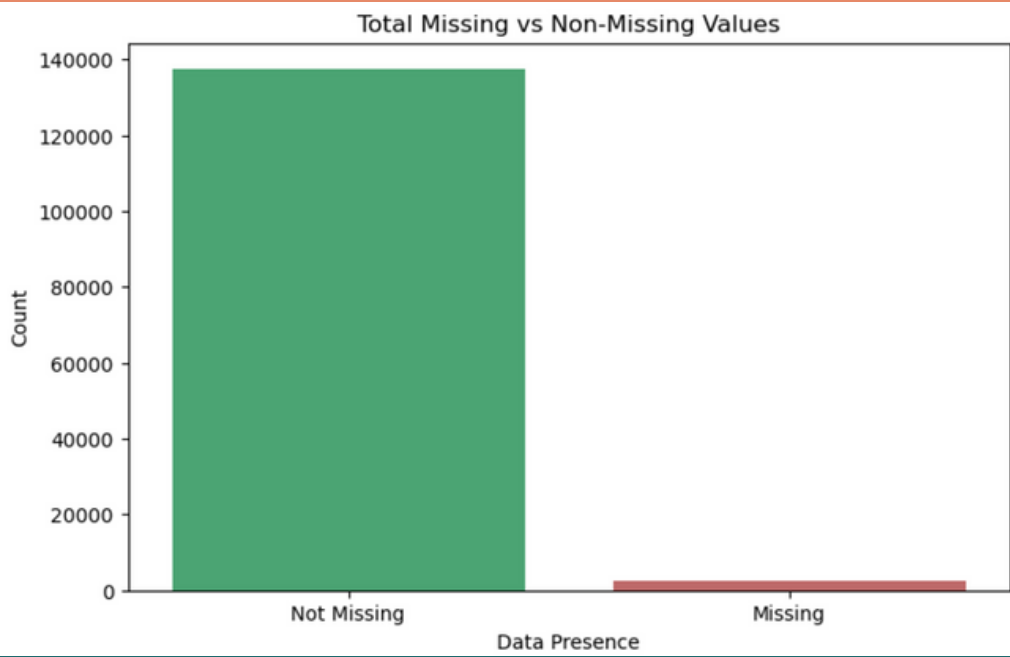
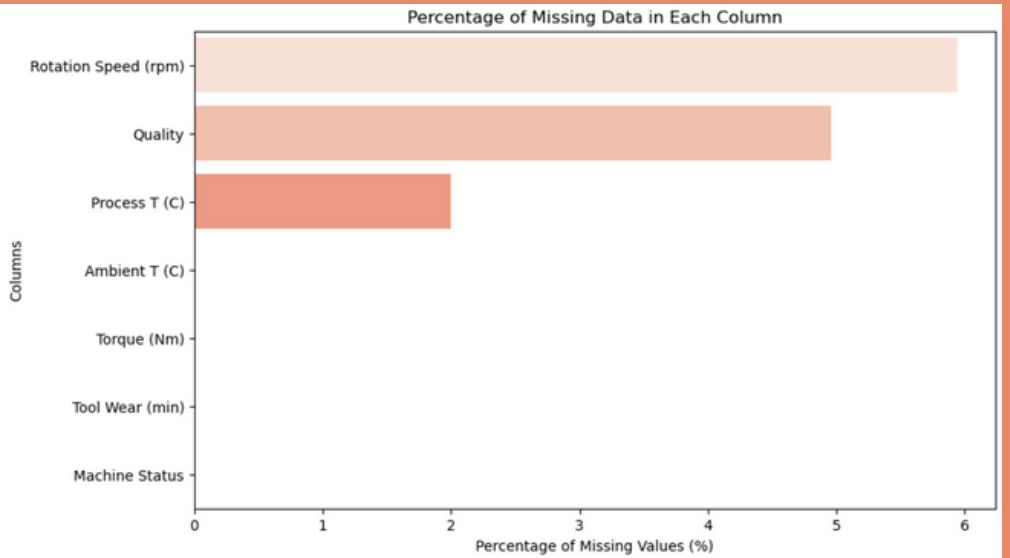
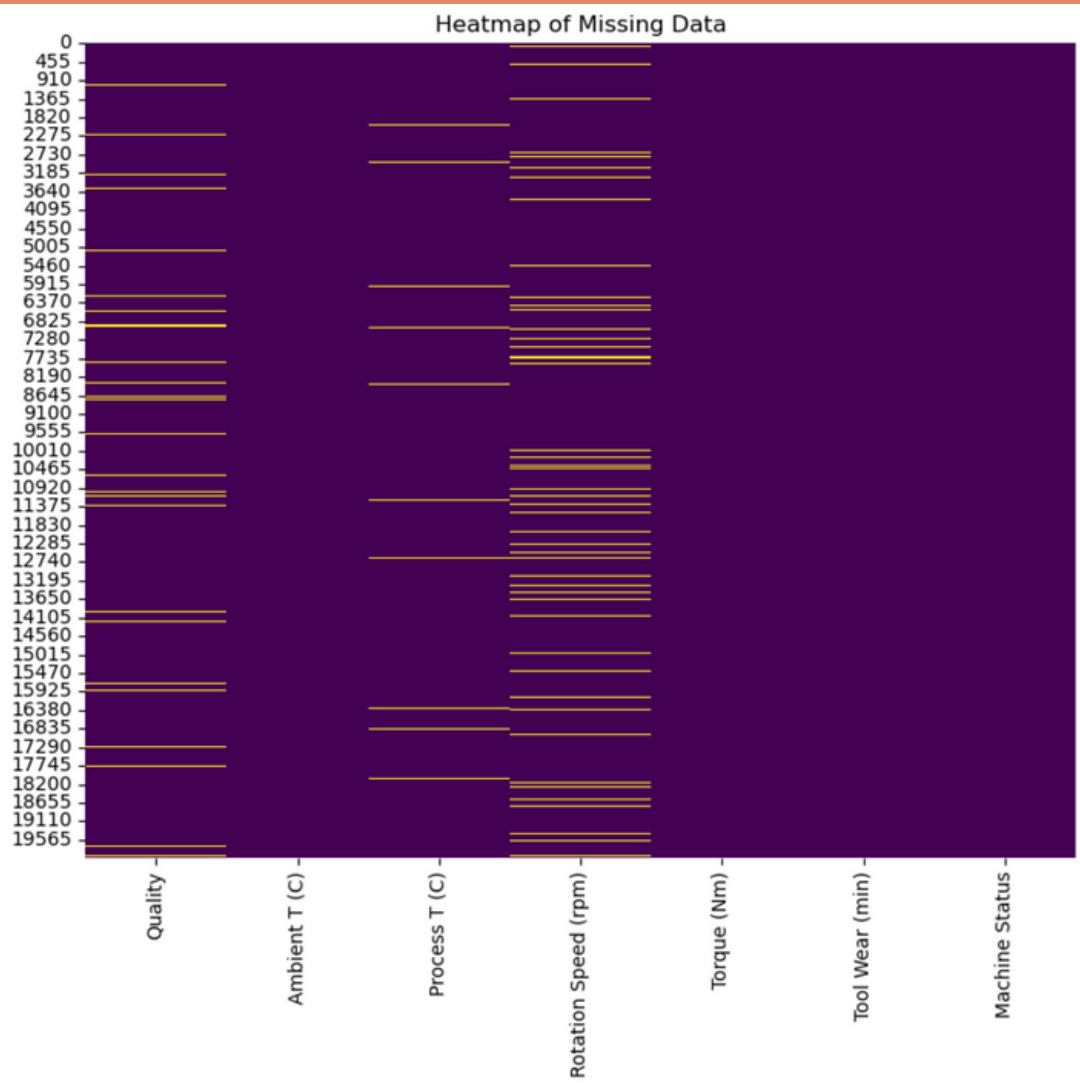


Data Cleaning and Analysis

Dropping Exclusive Data



Missing Data Visualisations



Checking Missing Data

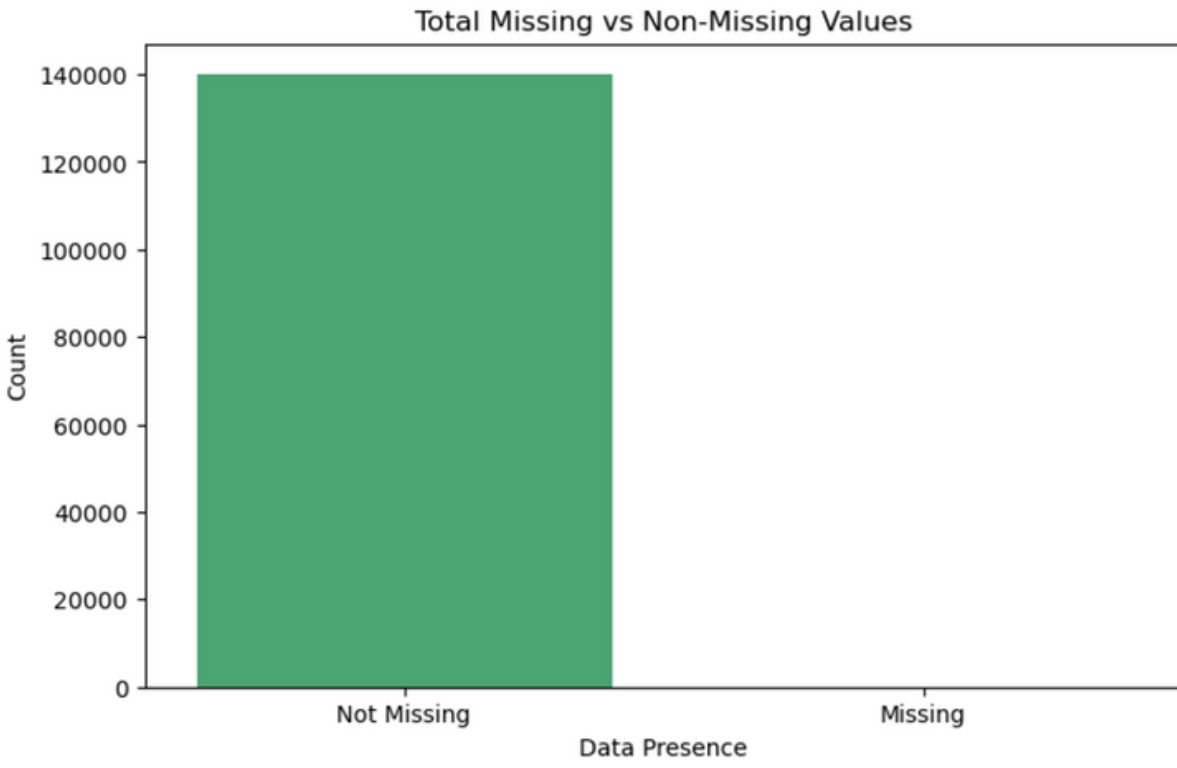
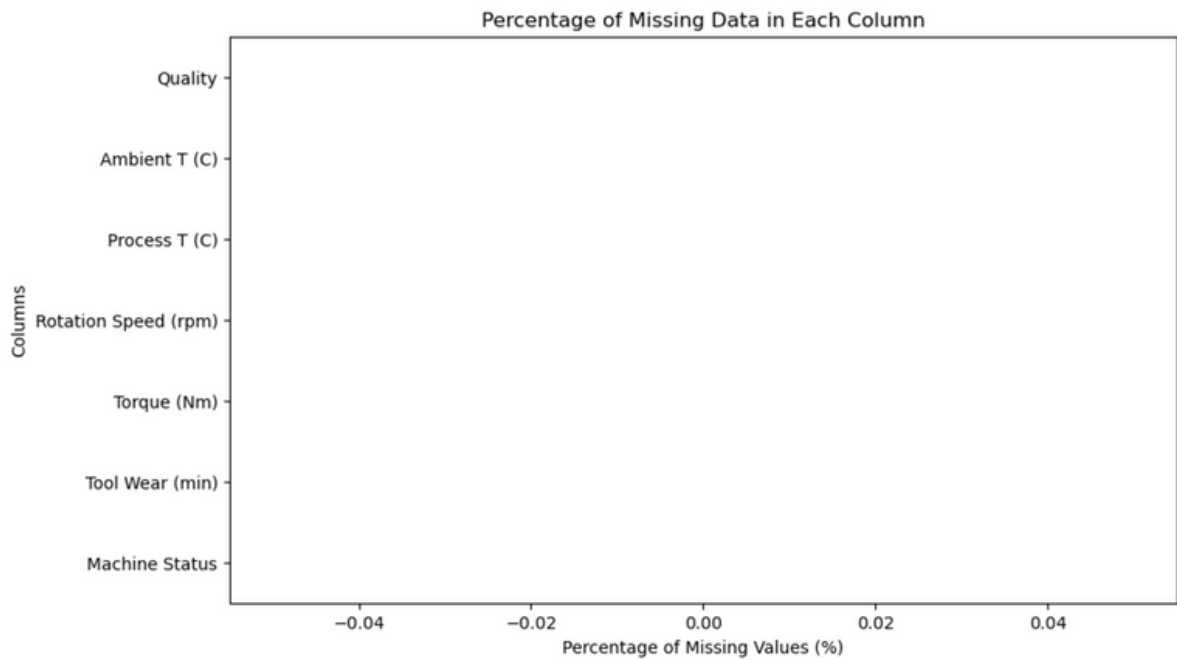
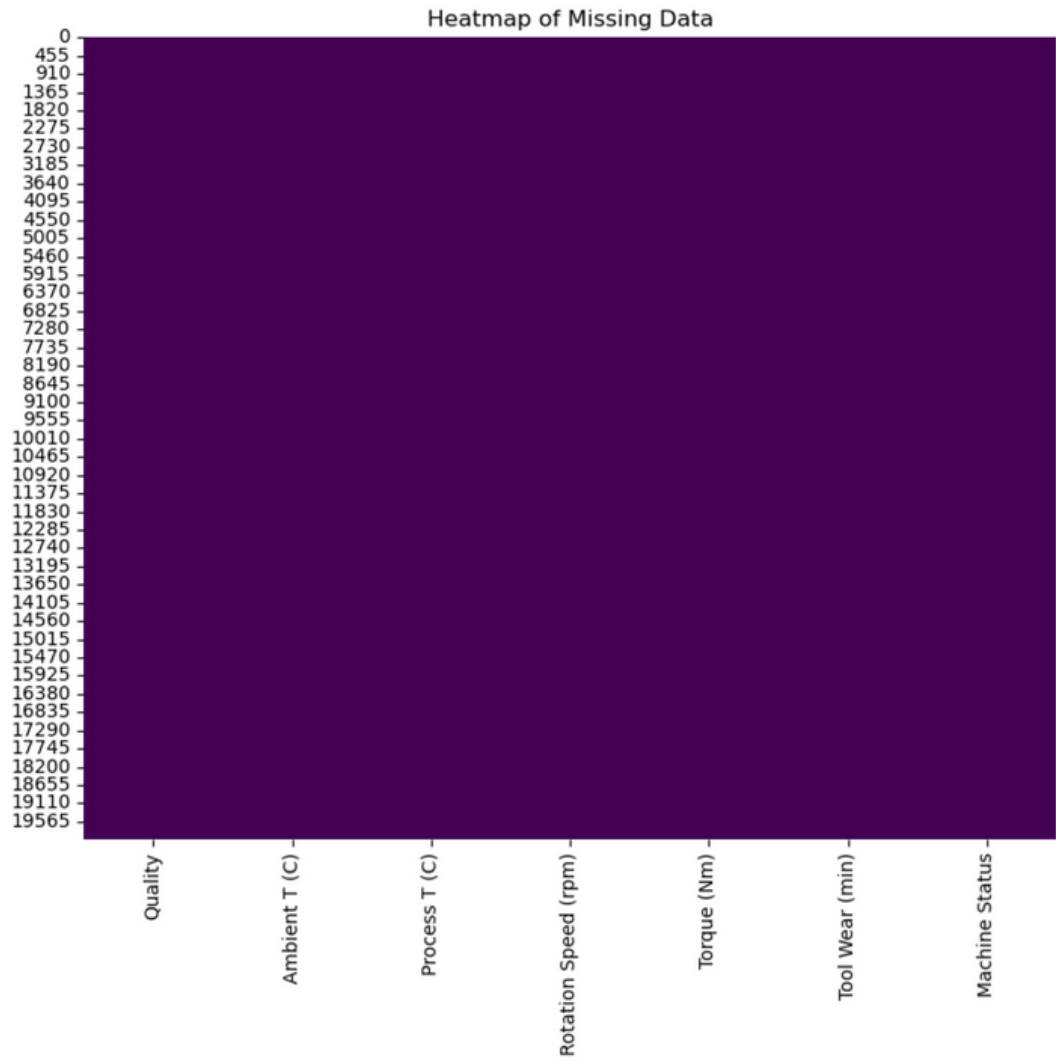
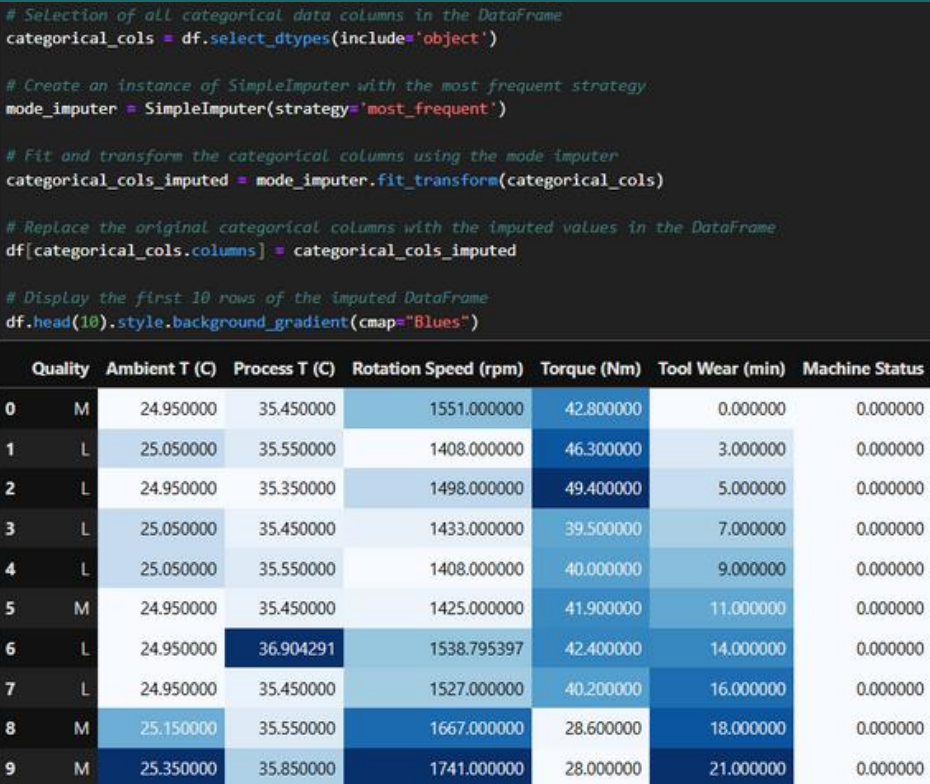
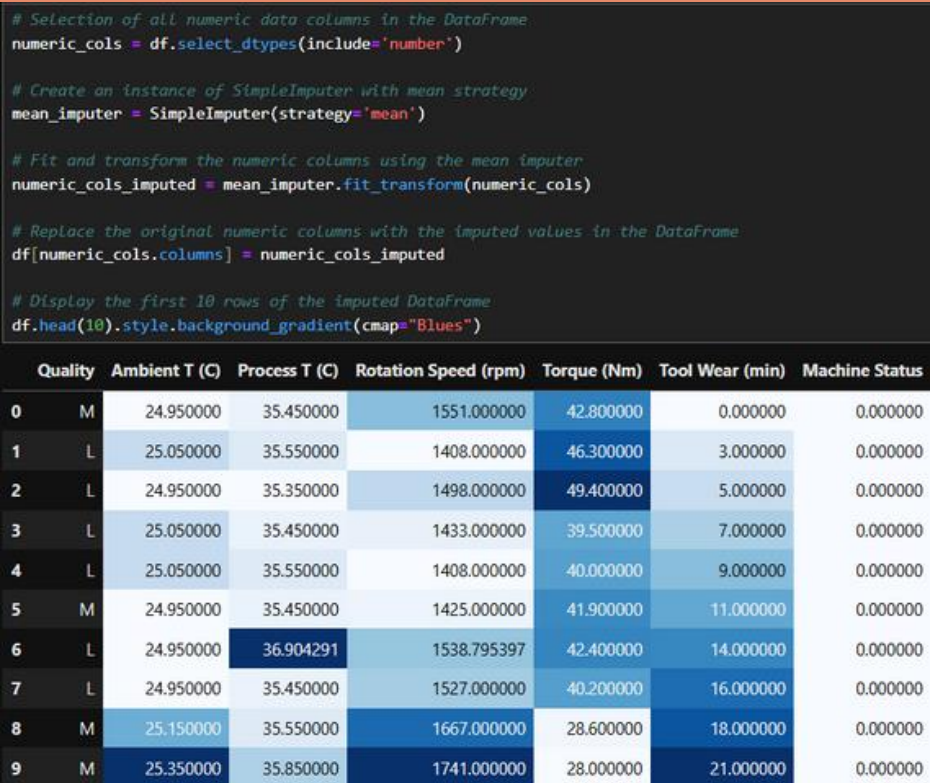
```
# Printing of total data that
df.isnull().sum()

Quality          991
Ambient T (C)    0
Process T (C)    400
Rotation Speed (rpm) 1188
Torque (Nm)      0
Tool Wear (min)  0
Machine Status   0
dtype: int64
```

Data Cleaning and Analysis

Imputing Missing Data

Verifying Data's Presence



Checking Duplicated Data

```
# Count number of duplicated data rows
num_duplicates = df.duplicated().sum()

# Printing number of duplicated data rows
print(f"Number of duplicated rows: {num_duplicates}")

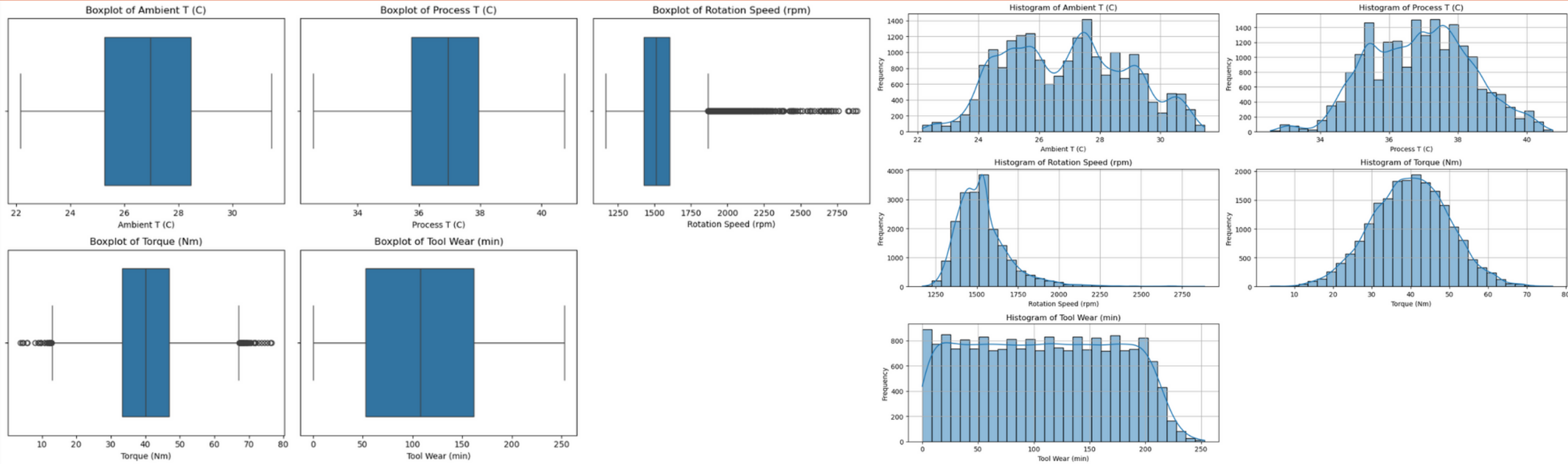
Number of duplicated rows: 0
```

Feature Engineering

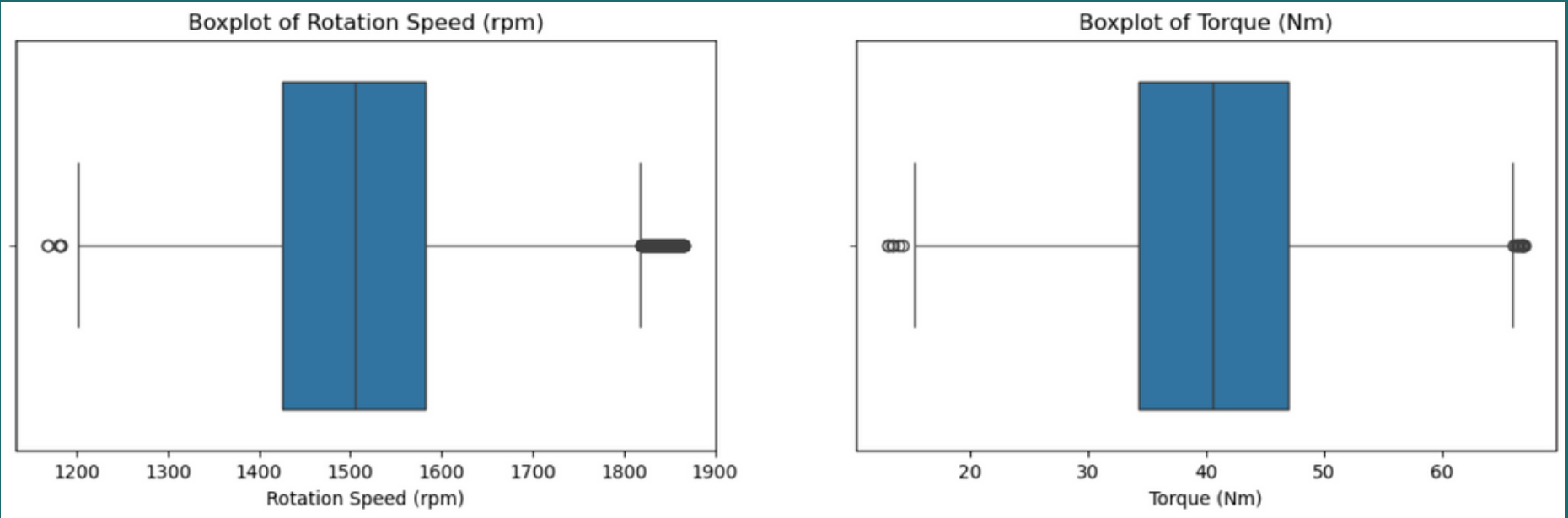
Dummy Variables
(For Quality)

Quality_L	Quality_M
0	1
1	0
1	0
1	0
1	0
0	1
1	0
1	0
0	1
0	1
1	0
0	0

Verifying Outliers



Verifying Removal of Some Outliers



New Features

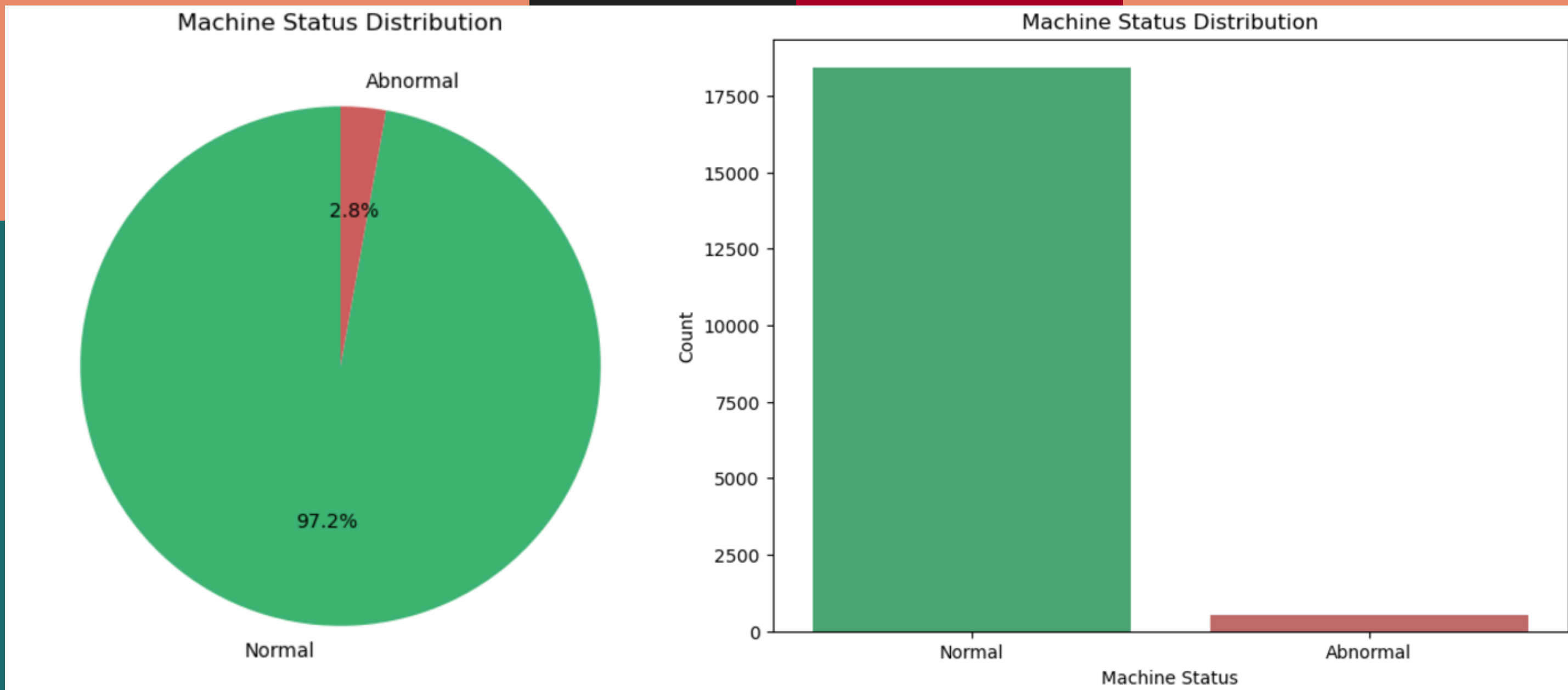
Temperature Gradient = $T_{\text{process}} - T_{\text{ambient}}$

Power Indicator = $\frac{\text{Torque}}{\text{Rotational Speed}}$

Temperature Gradient	Power Indicator
10.500000	0.027595
10.500000	0.032884
10.400000	0.032977

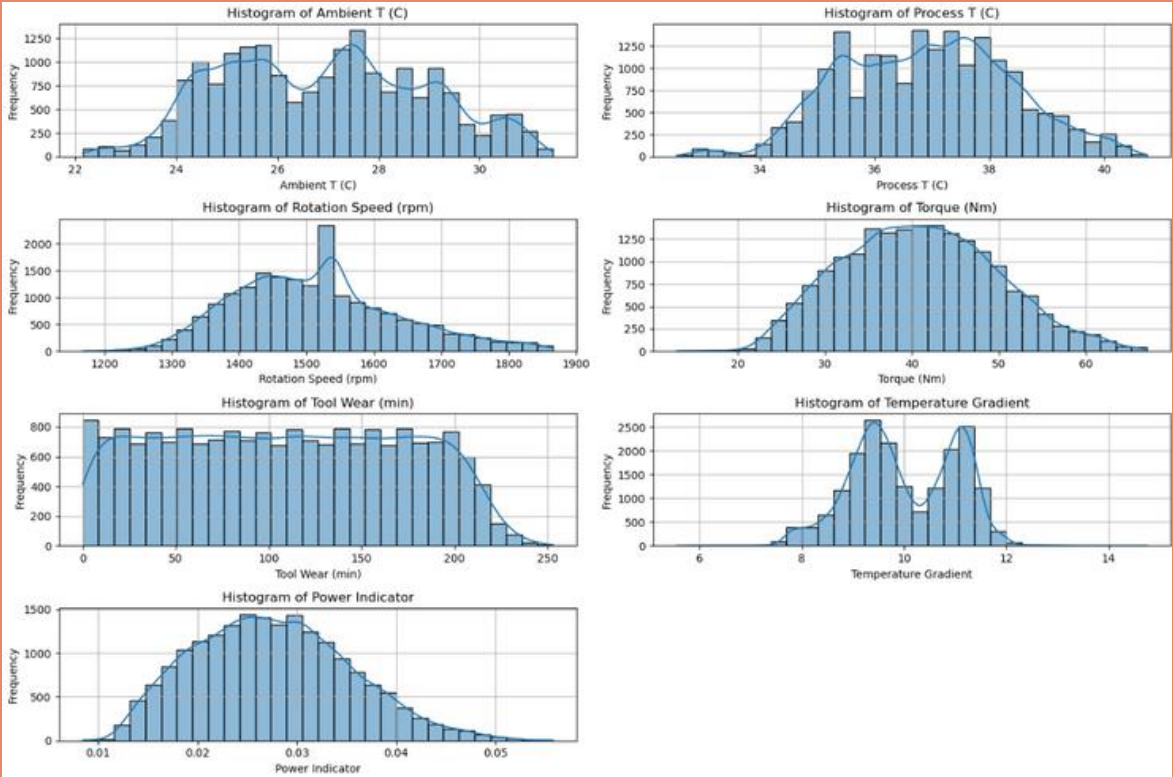
Data Visualisation (Machine Status)

	Machine Status	Count	Percentage (%)
1	Normal	18428	97.194093
2	Abnormal	532	2.805907

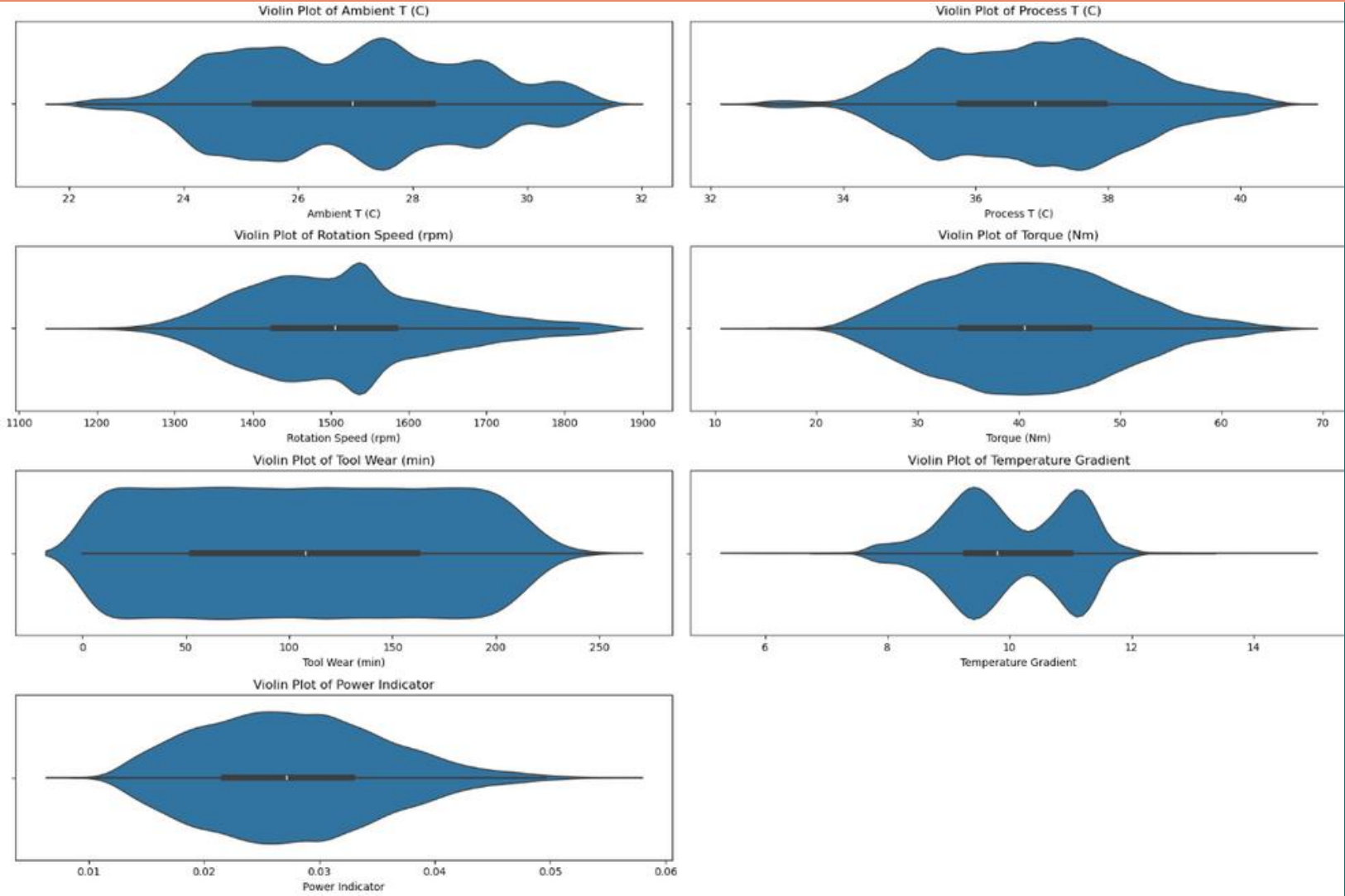


Data Visualisation (Univariate Analysis)

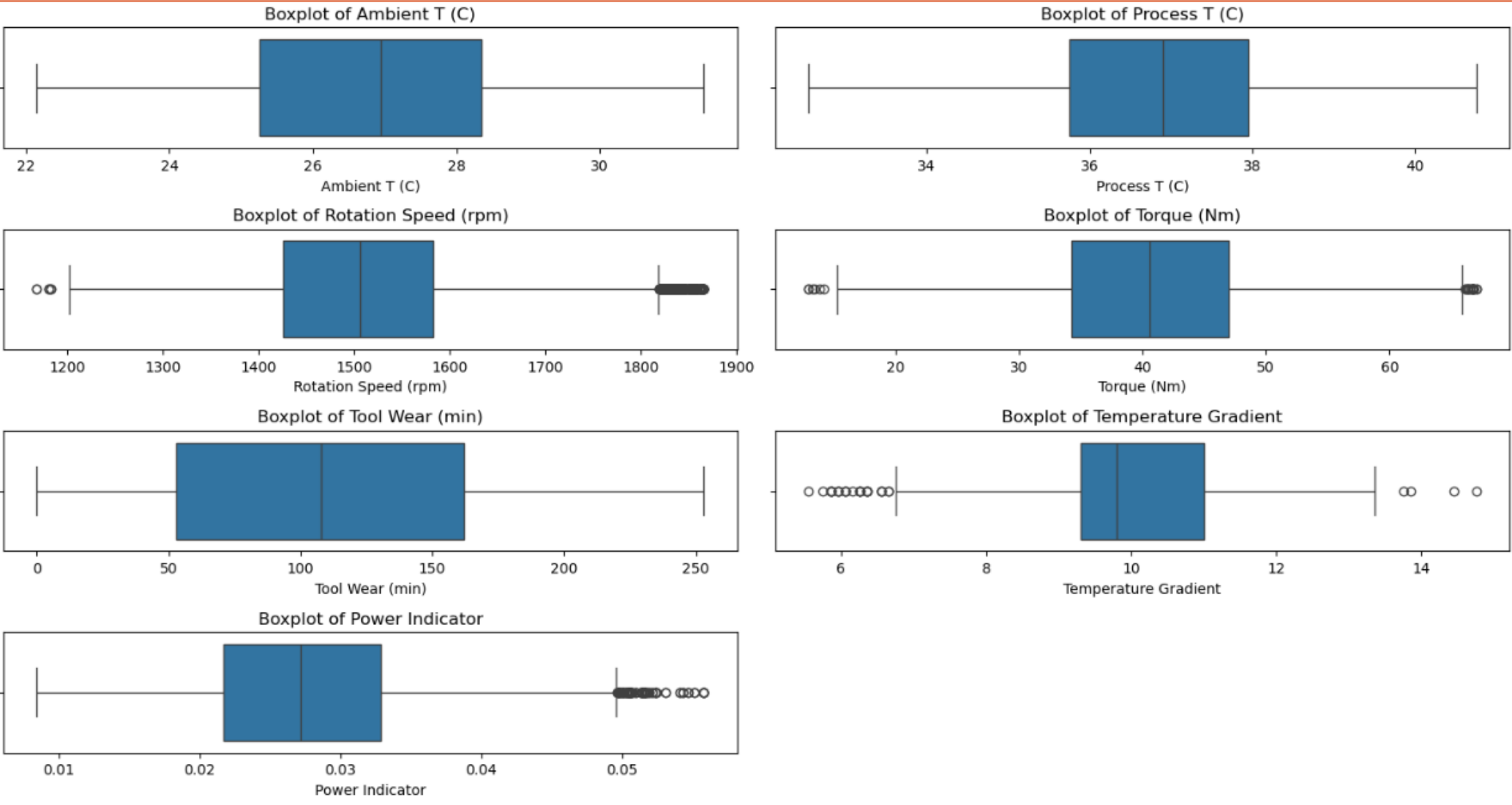
Histograms



Violin Plots



Box Plots

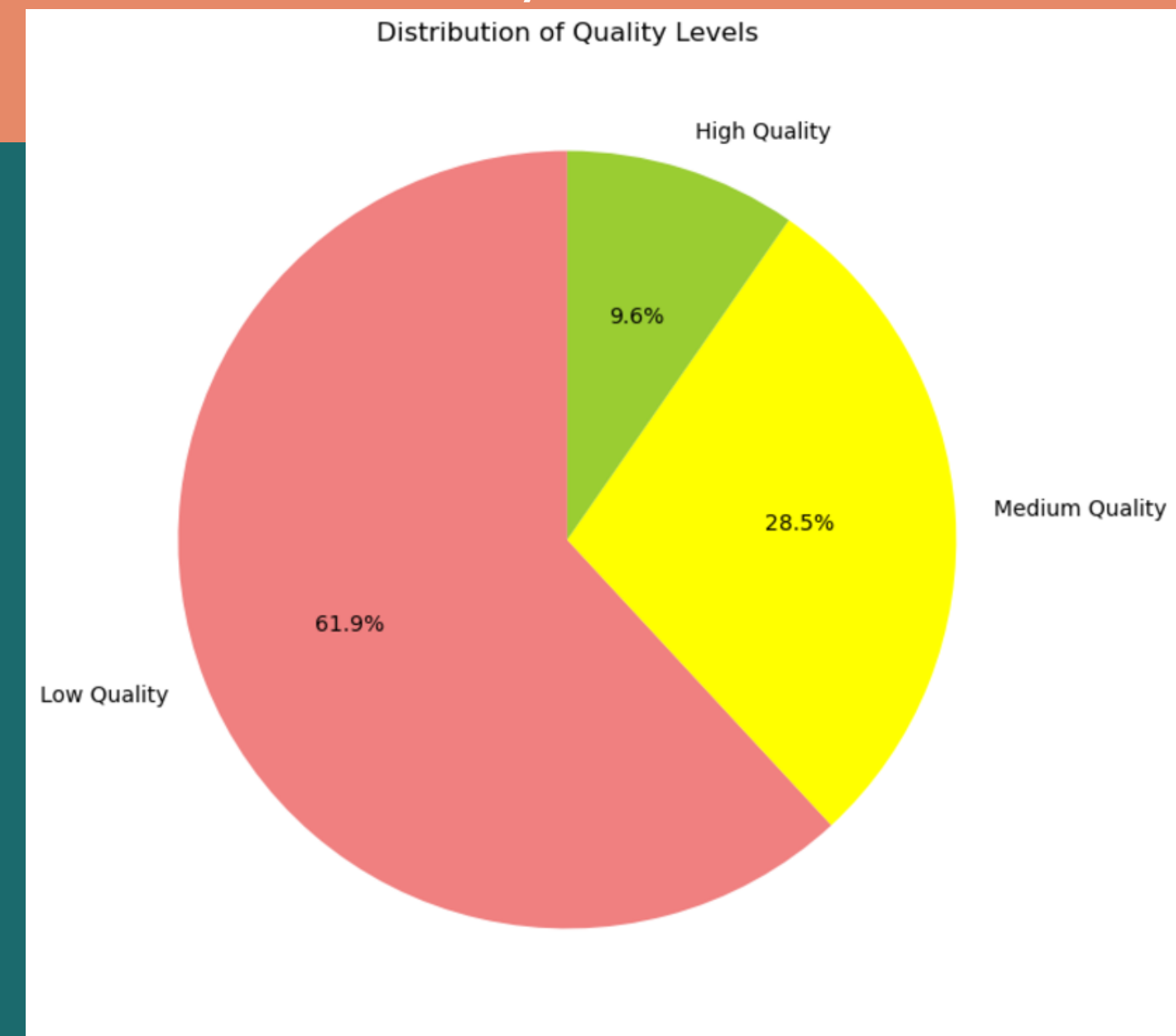


Data Visualisation (Univariate Analysis)

Quality Stacked Bar Chart

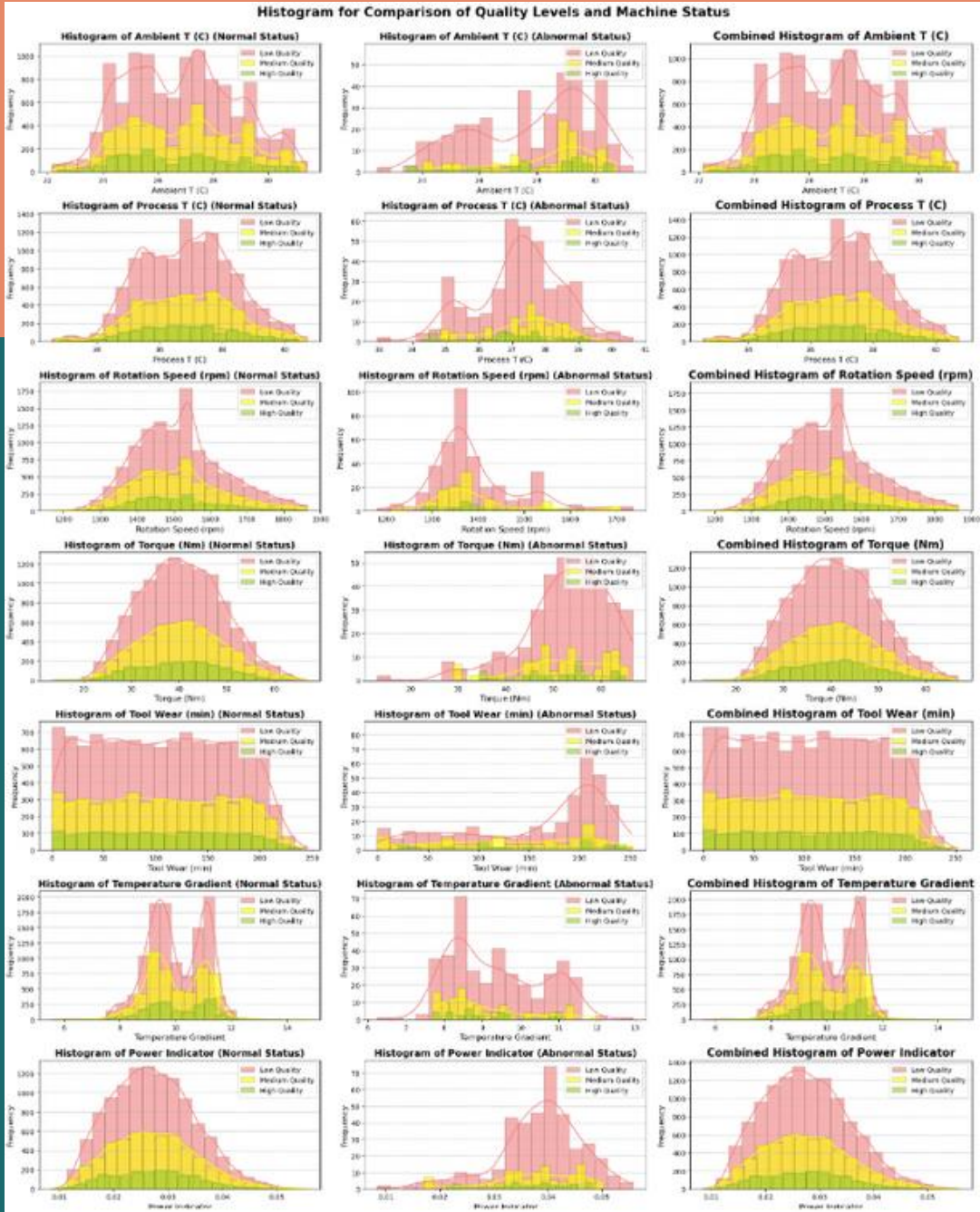


Quality Pie Chart

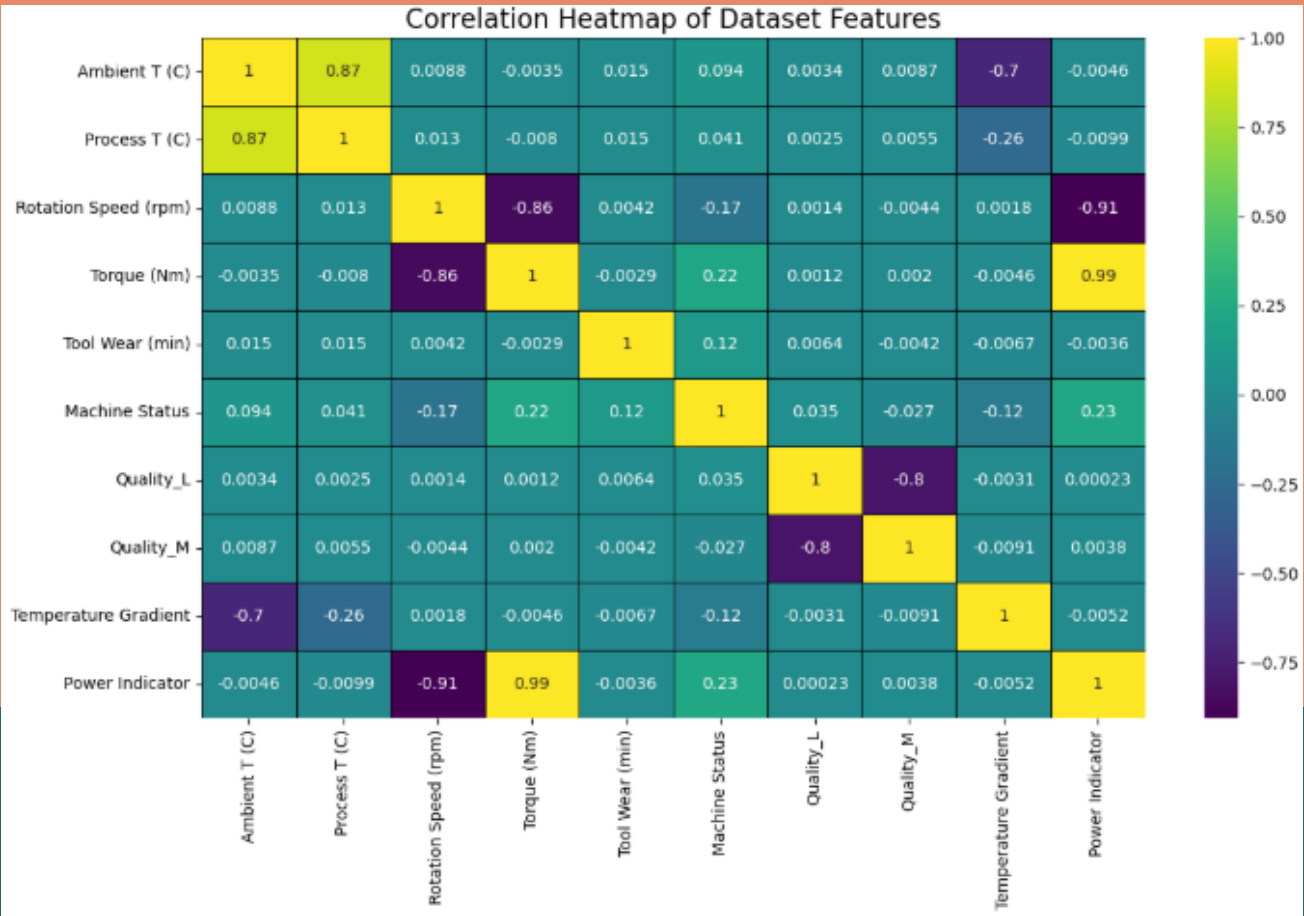


Data Visualisation (Bivariate Analysis)

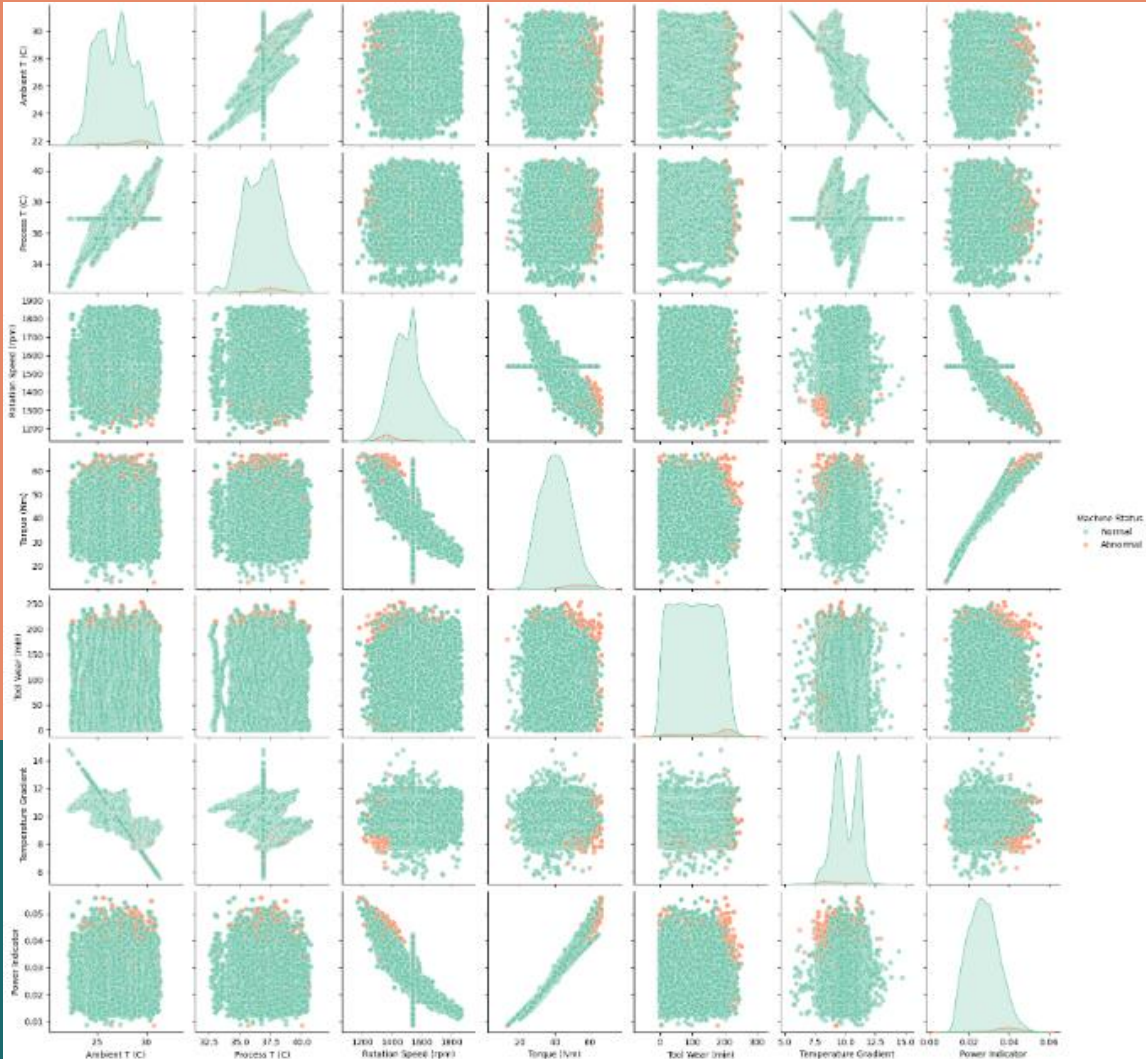
Different Quality
Relationship with
Machine Status



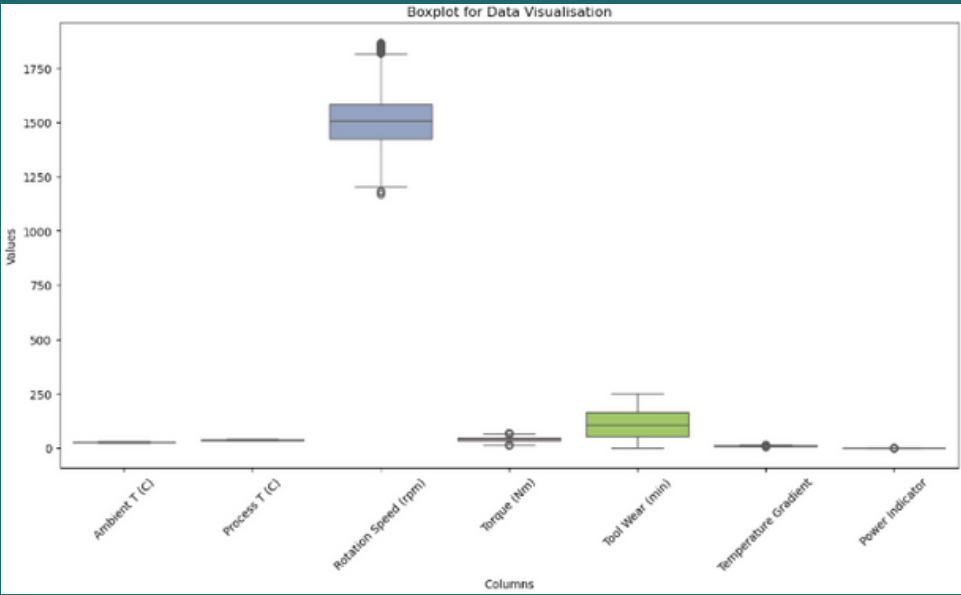
Correlation Heatmap



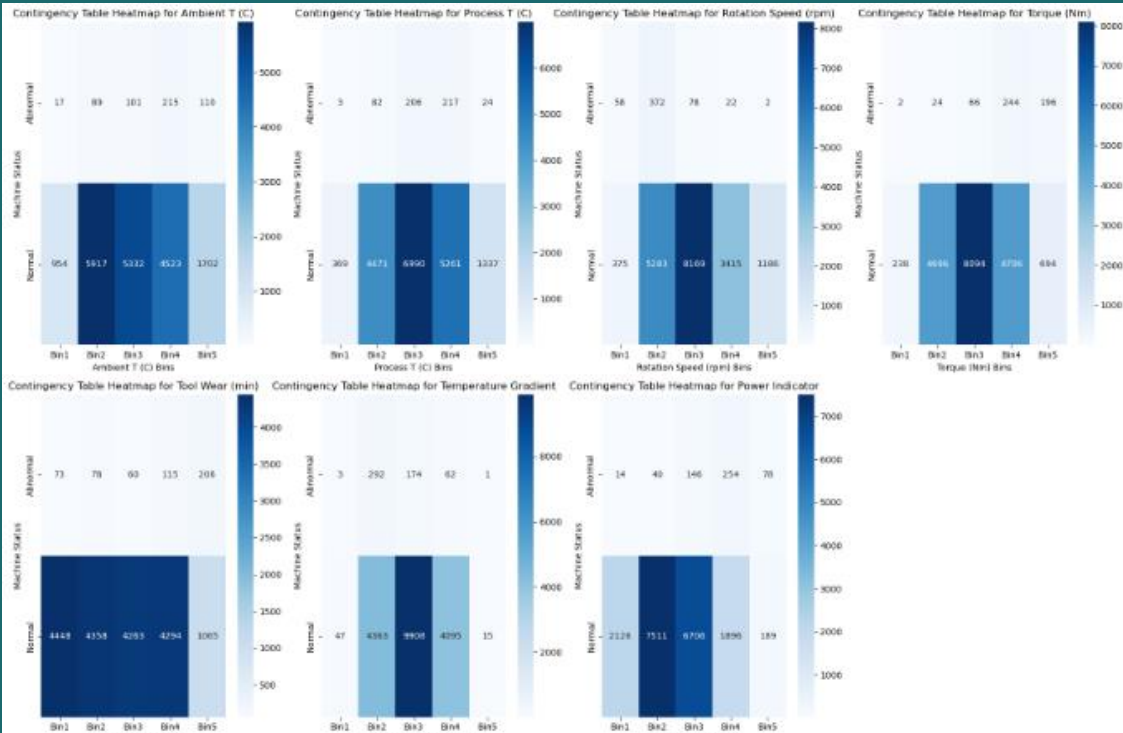
Pairplot



Box Plot



Contingency Table



Data Preprocessing and Dummy Baseline

Separation of Feature and Target

```
# X = Features
X = df.drop("Machine Status", axis = 1)

# Y = Target
Y = df["Machine Status"]

# Execution Confirmation
print("Feature and Target Separation Successfully Executed")
```

Feature and Target Separation Successfully Executed

Train-Test Split

```
# Splitting of Trianing Data and Testing Data
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, shuffle = True, random_state = 28)

# Execution Confirmation
print("Train Test Split Successfully Executed")
```

Train Test Split Successfully Executed

Dummy Baseline

	Accuracy	Balanced Accuracy	Recall	F1 Score	ROC AUC
DummyClassifier	0.500791	0.460673	0.500791	0.646458	0.5

Model Training



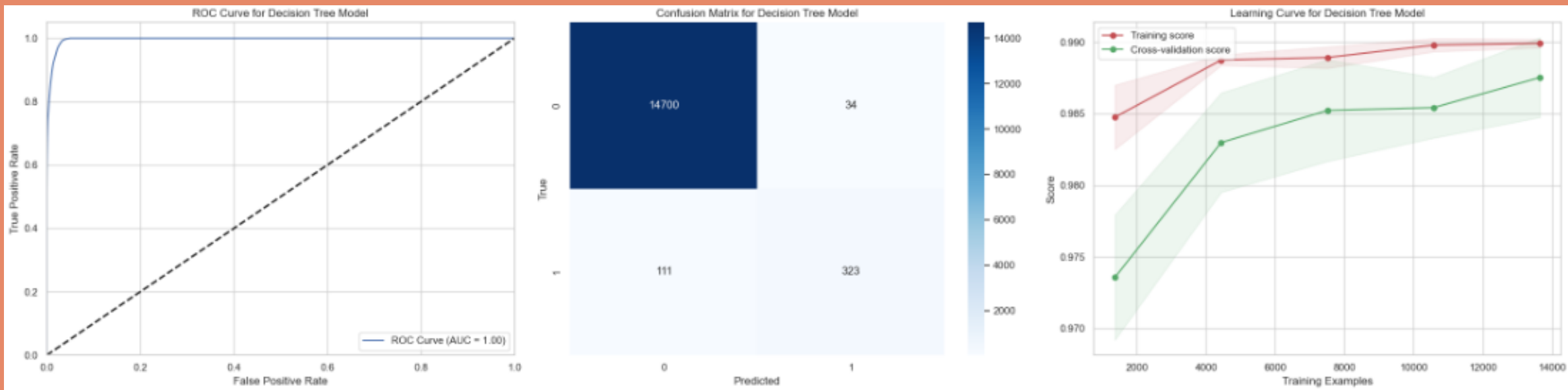
Model Training

Training Scores for All Models

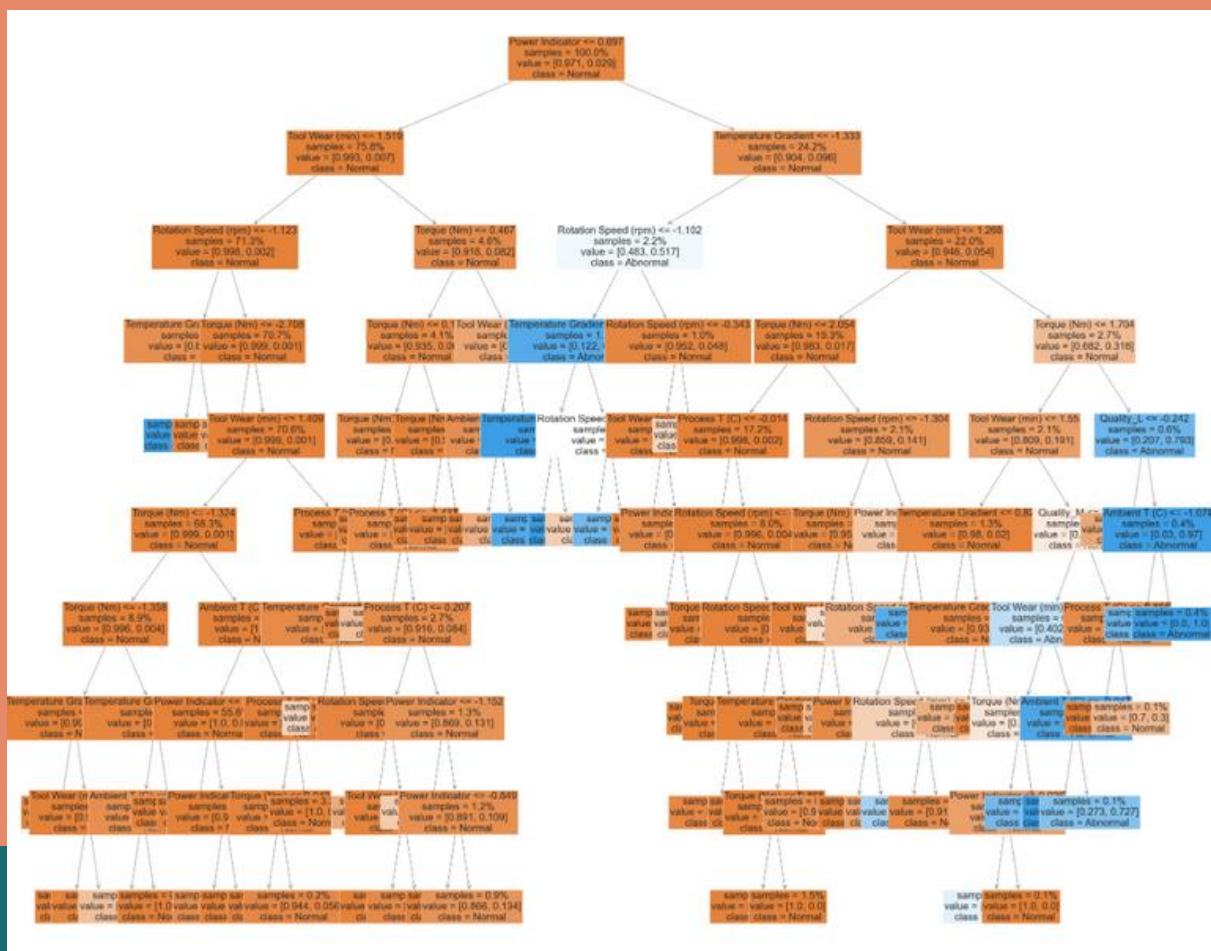
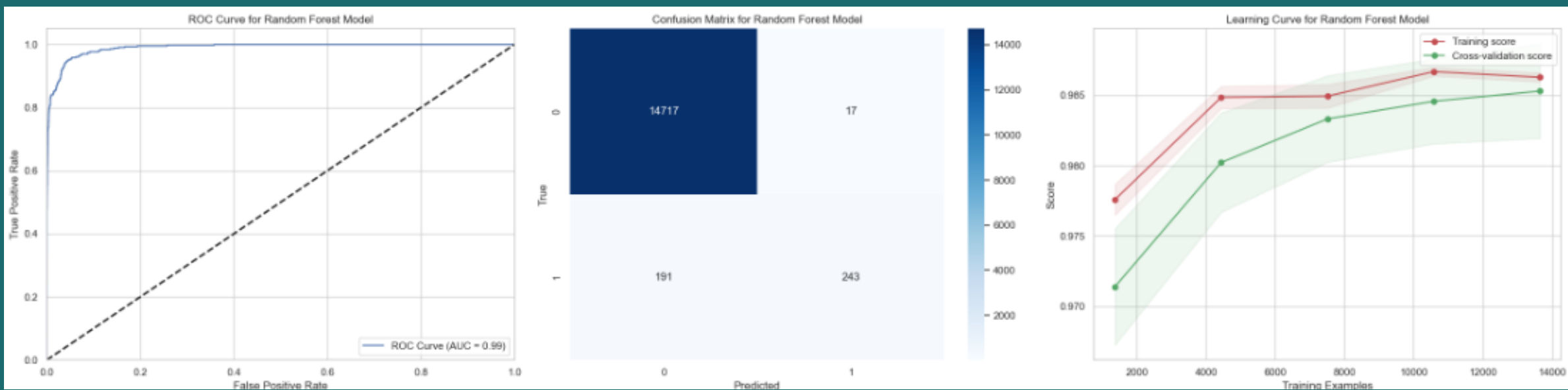
	fit_time	score_time	test_accuracy	train_accuracy	test_balanced_accuracy	train_balanced_accuracy	test_recall	train_recall	test_f1	train_f1	test_roc_auc	train_roc_auc
model												
LogisticRegression	0.158233	0.022023	0.973336	0.973603	0.551721	0.561400	0.551721	0.561400	0.583954	0.599504	0.551721	0.561400
KNeighborsClassifier	0.055355	0.318681	0.974984	0.979875	0.631803	0.692757	0.631803	0.692757	0.682018	0.757445	0.631803	0.692757
SVC	1.419350	0.579905	0.974192	0.975490	0.568366	0.587650	0.568366	0.587650	0.608972	0.639525	0.568366	0.587650
DecisionTreeClassifier	0.090754	0.039160	0.990211	1.000000	0.913817	1.000000	0.913817	1.000000	0.910928	1.000000	0.913817	1.000000
RandomForestClassifier	2.618238	0.056462	0.992650	1.000000	0.887574	1.000000	0.887574	1.000000	0.926157	1.000000	0.887574	1.000000
AdaBoostClassifier	1.167915	0.061291	0.977192	0.977897	0.704793	0.716097	0.704793	0.716097	0.748158	0.760250	0.704793	0.716097
GaussianNB	0.007405	0.013698	0.947034	0.947651	0.750493	0.748490	0.750493	0.748490	0.669600	0.671706	0.750493	0.748490
GradientBoostingClassifier	3.325692	0.021656	0.990771	0.993959	0.866947	0.908955	0.866947	0.908955	0.906710	0.941401	0.866947	0.908955
Perceptron	0.014942	0.017935	0.968721	0.968147	0.530349	0.529974	0.530349	0.529974	0.521684	0.522870	0.530349	0.529974
ExtraTreesClassifier	1.173980	0.102222	0.992683	1.000000	0.880084	1.000000	0.880084	1.000000	0.925538	1.000000	0.880084	1.000000
RidgeClassifier	0.011600	0.013115	0.971688	0.971312	0.500000	0.500000	0.500000	0.500000	0.492819	0.492724	0.500000	0.500000
RidgeClassifierCV	0.016845	0.012785	0.971688	0.971312	0.500000	0.500000	0.500000	0.500000	0.492819	0.492724	0.500000	0.500000
SGDClassifier	0.131574	0.033714	0.952802	0.950247	0.546921	0.542923	0.546921	0.542923	0.524736	0.520861	0.546921	0.542923
Baseline	0.002449	0.013055	0.505307	0.500742	0.512566	0.507744	0.512566	0.507744	0.360501	0.358254	0.512566	0.507744

Model Comparison (RFC and DT)

DT:



RFC:



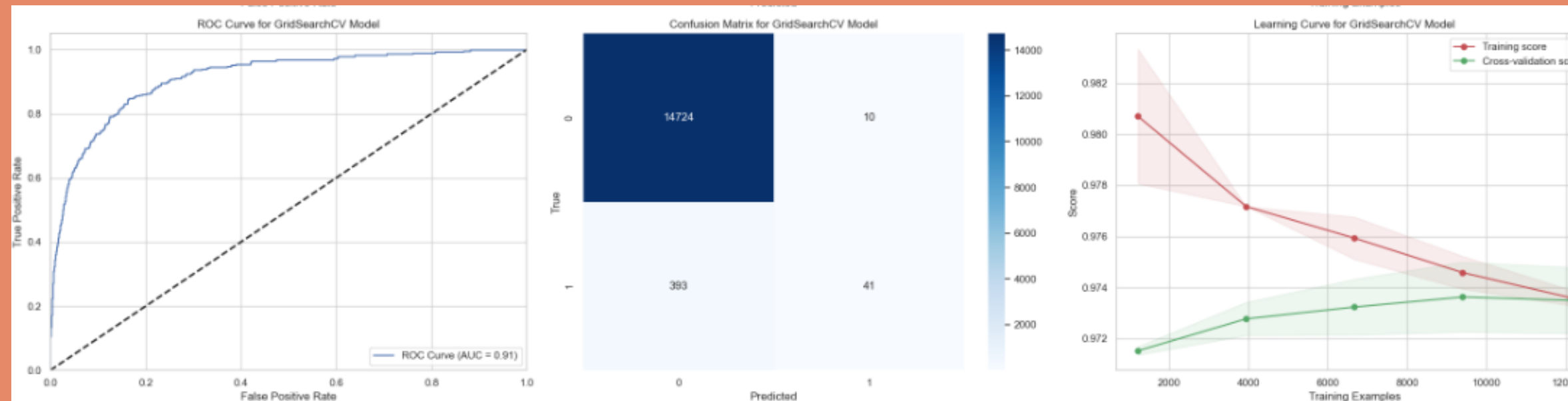
Decision Tree
Incorrect Data

	Model	Accuracy	Balanced Accuracy	Recall	F1 Score	ROC AUC
0	Decision Tree	0.990440	0.870966	0.990440	0.989988	0.996741
1	Random Forest	0.986287	0.779377	0.986287	0.984608	0.990057

Incorrect Predictions on Training Data:											
	Ambient T (C)	Process T (C)	Rotation Speed (rpm)	Torque (Nm)	Tool Wear (min)	Quality_L	Quality_M	Temperature Gradient	Power Indicator	Predicted	Actual
63	29.55	39.25	1454.000000	54.8	253.0	0	1	9.7	0.037689	0.0	1.0
229	28.75	37.35	1340.000000	55.7	204.0	1	0	8.6	0.041567	0.0	1.0
603	30.55	39.15	1375.000000	46.2	98.0	1	0	8.6	0.033600	1.0	0.0
754	28.95	37.45	1538.795397	60.6	90.0	1	0	8.5	0.039381	0.0	1.0
763	29.55	37.35	1446.000000	40.9	211.0	1	0	7.8	0.028285	0.0	1.0
778	30.25	38.65	1538.795397	47.0	210.0	0	1	8.4	0.030543	0.0	1.0
886	25.95	37.35	1389.000000	61.4	64.0	1	0	11.4	0.044204	1.0	0.0
1199	28.95	37.55	1314.000000	52.8	8.0	1	0	8.6	0.040183	0.0	1.0
1443	26.05	37.45	1389.000000	61.4	64.0	1	0	11.4	0.044204	1.0	0.0
1468	25.75	35.85	1455.000000	41.3	208.0	1	0	10.1	0.028385	0.0	1.0

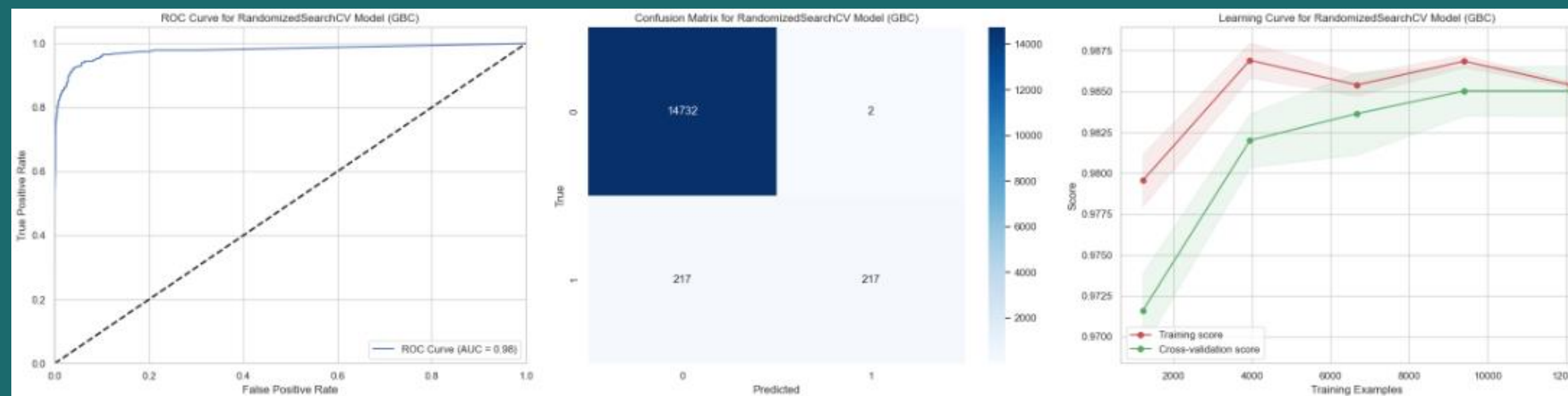
Models Selection and Improvement

Logistics Regression



	Accuracy	Balanced Accuracy	Recall	F1 Score	ROC AUC
RandomizedSearchCV	0.973431	0.560313	0.973431	0.964221	0.910909
GridSearchCV	0.973431	0.546896	0.973431	0.963111	0.910915

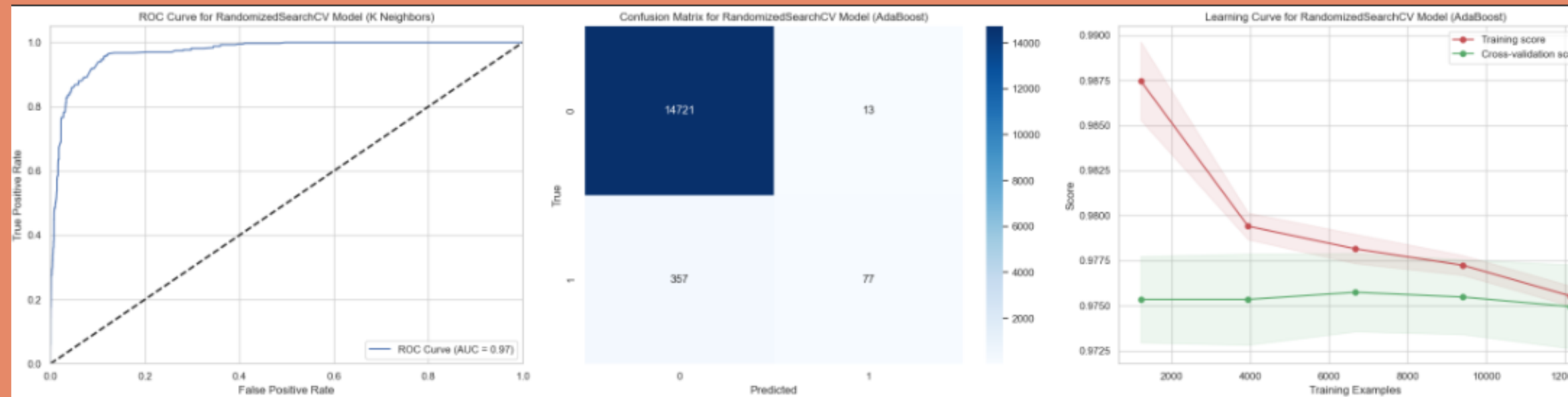
Gradient Booster Classifier



	Accuracy	Balanced Accuracy	Recall	F1 Score	ROC AUC
RandomizedSearchCV (GBC)	0.985562	0.749932	0.985562	0.983237	0.977823
GridSearchCV (GBC)	0.985628	0.751084	0.985628	0.983328	0.978034

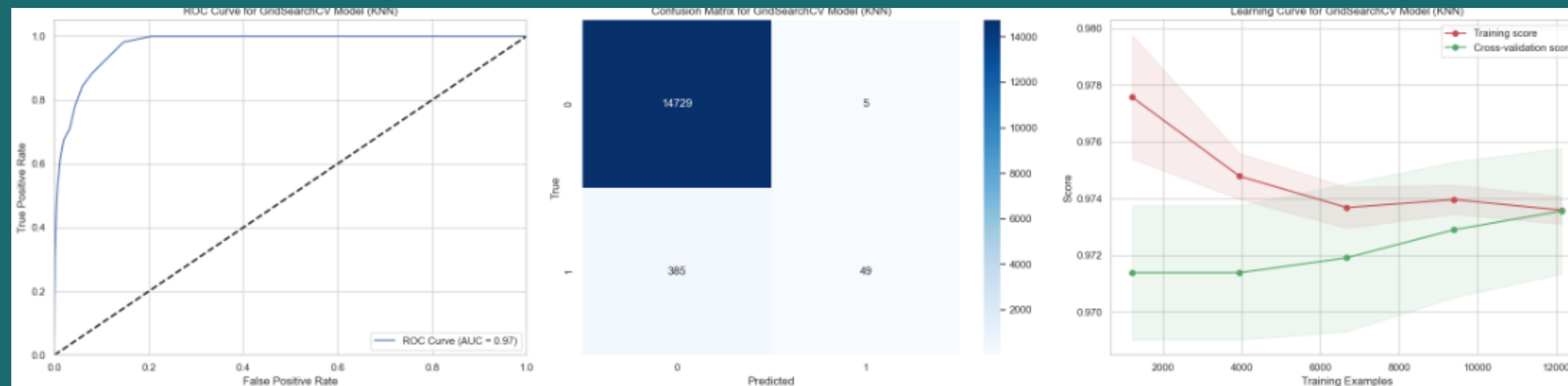
Models Selection and Improvement

AdaBoost Classifier



	Accuracy	Balanced Accuracy	Recall	F1 Score	ROC AUC
RandomizedSearchCV (AdaBoost)	0.975607	0.588269	0.975607	0.96774	0.970166
GridSearchCV (AdaBoost)	0.975607	0.588269	0.975607	0.96774	0.970166

K Neighbors Classifier



	Accuracy	Balanced Accuracy	Recall	F1 Score	ROC AUC
RandomizedSearchCV (KNN)	1.000000	1.000000	1.000000	1.000000	1.000000
GridSearchCV (KNN)	0.974288	0.556282	0.974288	0.964441	0.973147

Voting Classifier

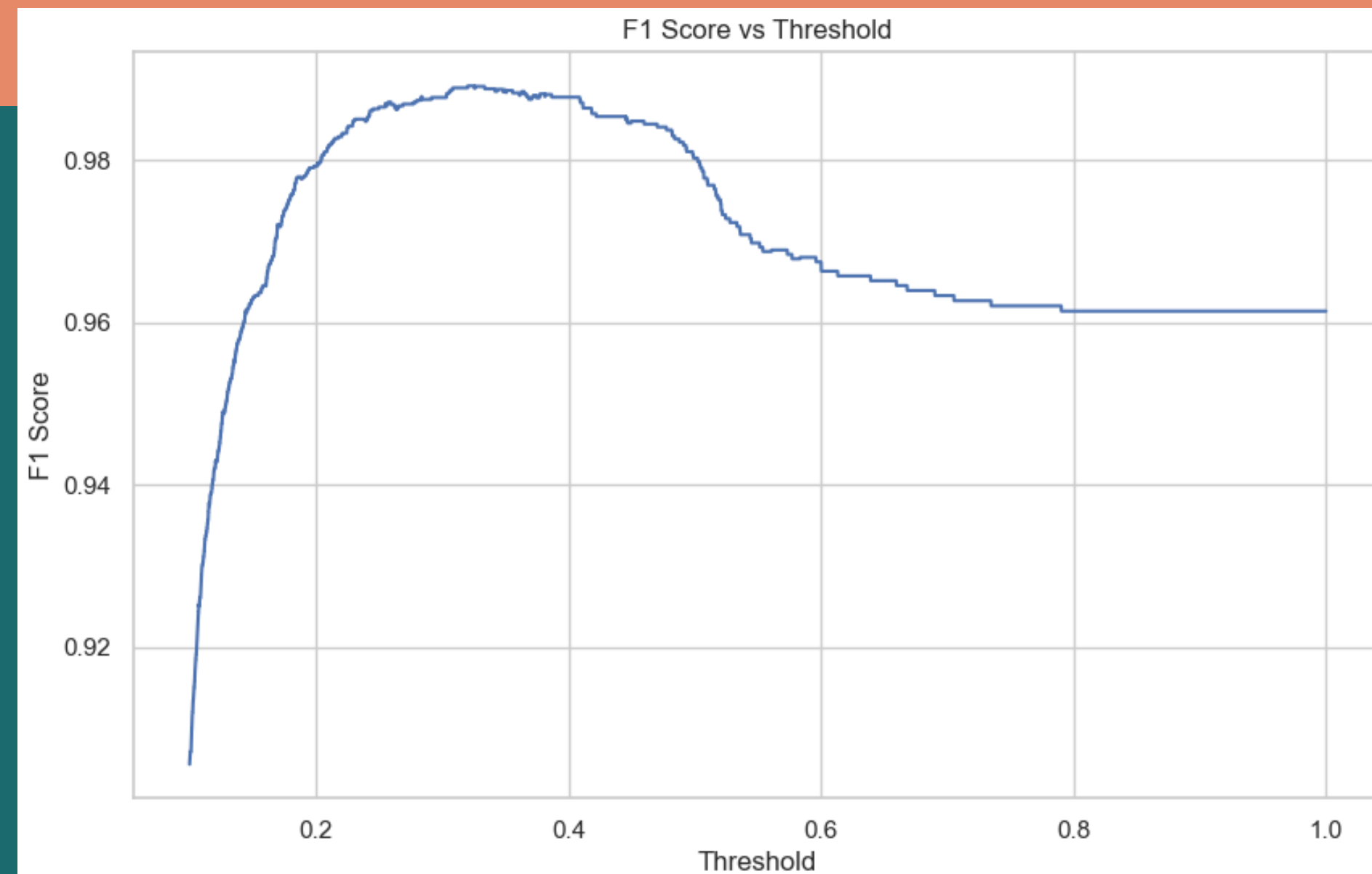
Initial Test Results

	Accuracy	Balanced Accuracy	Recall	F1 Score
VotingClassifier	0.98365	0.693607	0.98365	0.980278

Initial Confusion Matrix



Threshold Improvement

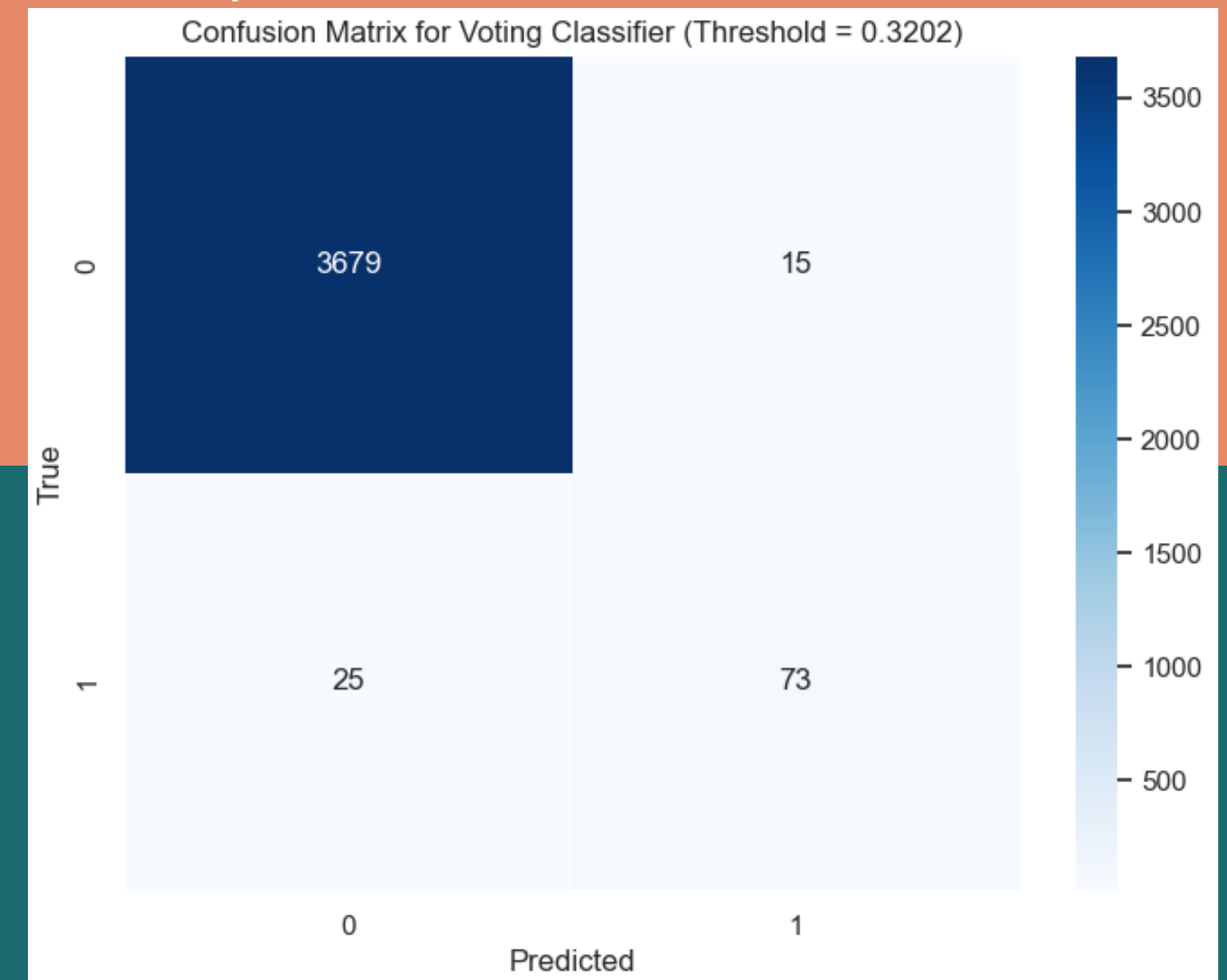


Voting Classifier

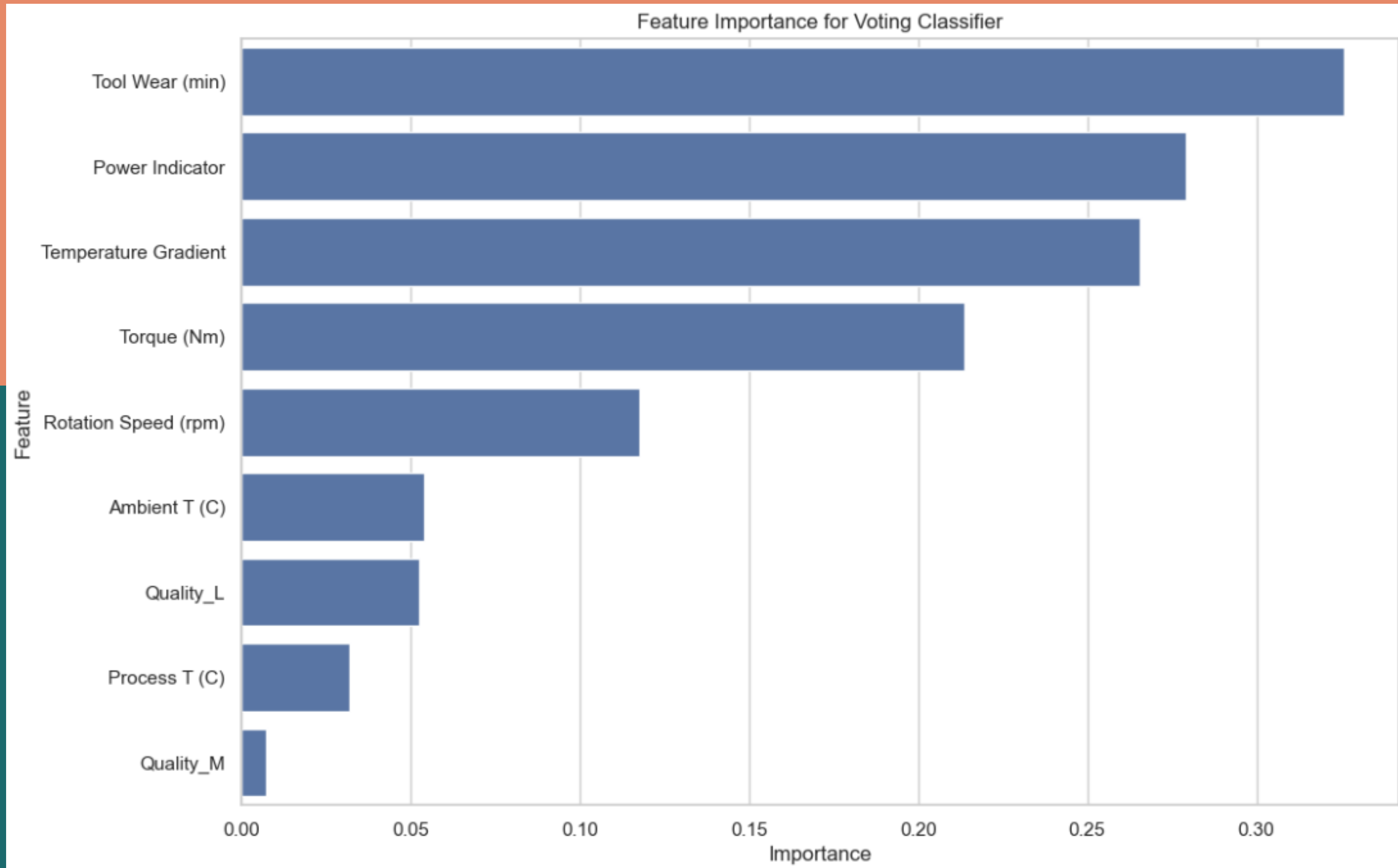
Improved Test Result

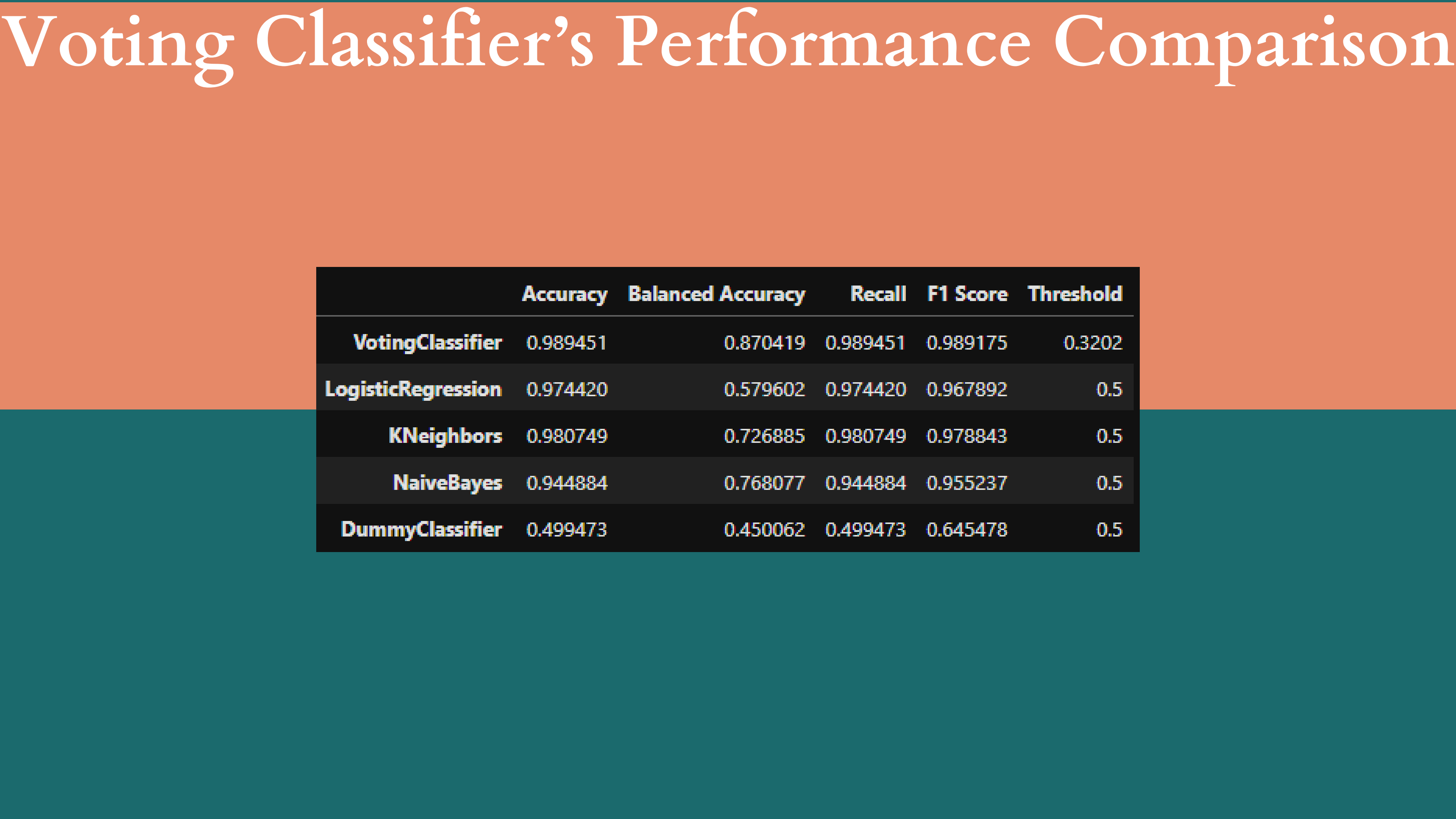
	Accuracy	Balanced Accuracy	Recall	F1 Score	Threshold
VotingClassifier	0.989451	0.870419	0.989451	0.989175	0.3202

Improved Confusion Matrix



Voting Classifier's Feature Importance





Saving of Voting Classifier

```
# Saving of Voting Classifier  
joblib.dump(voting_clf, 'CA1-Saved-Models/final-model_A.pkl')
```

```
# Code Execution Confirmation  
print("Voting Classifier Model Saved Successfully")
```

```
Voting Classifier Model Saved Successfully
```

A close-up, dimly lit photograph of a person's hands working on a small electronic circuit board. The person is using a green-handled soldering iron to solder components on the board, which is held in place by a metal clamp. The work is taking place on a wooden surface. In the background, a white box with some red text is visible, and the overall scene is softly focused, emphasizing the hands and the soldering process.

Thank you