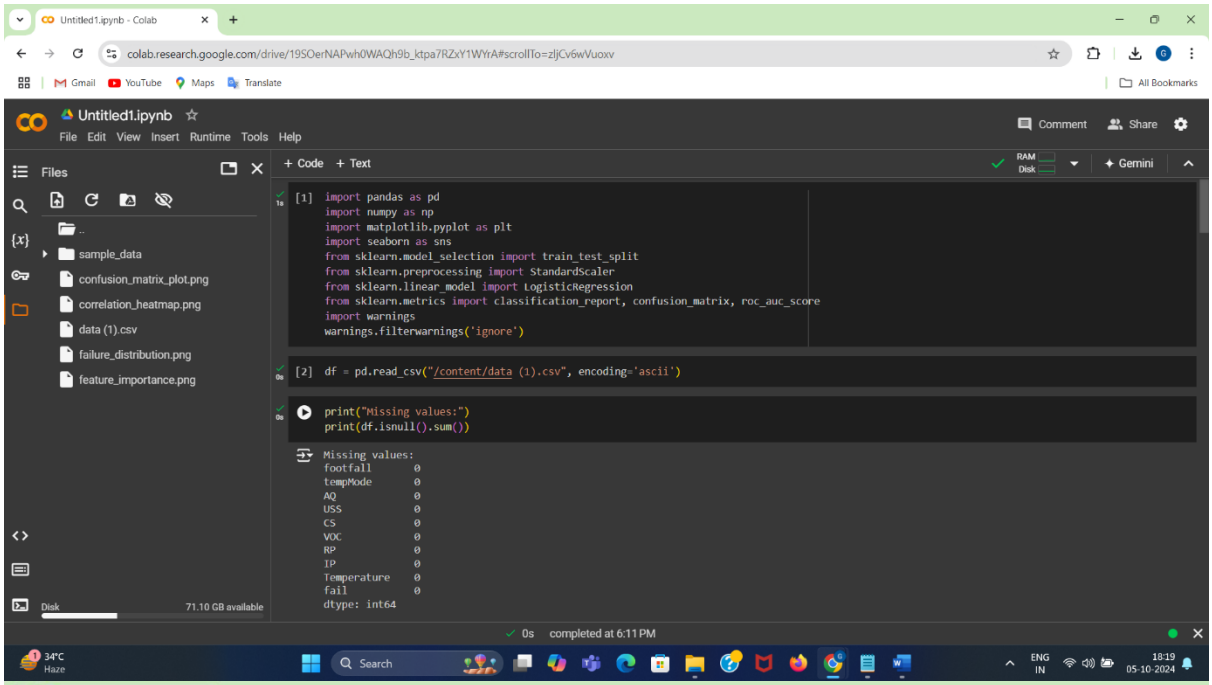
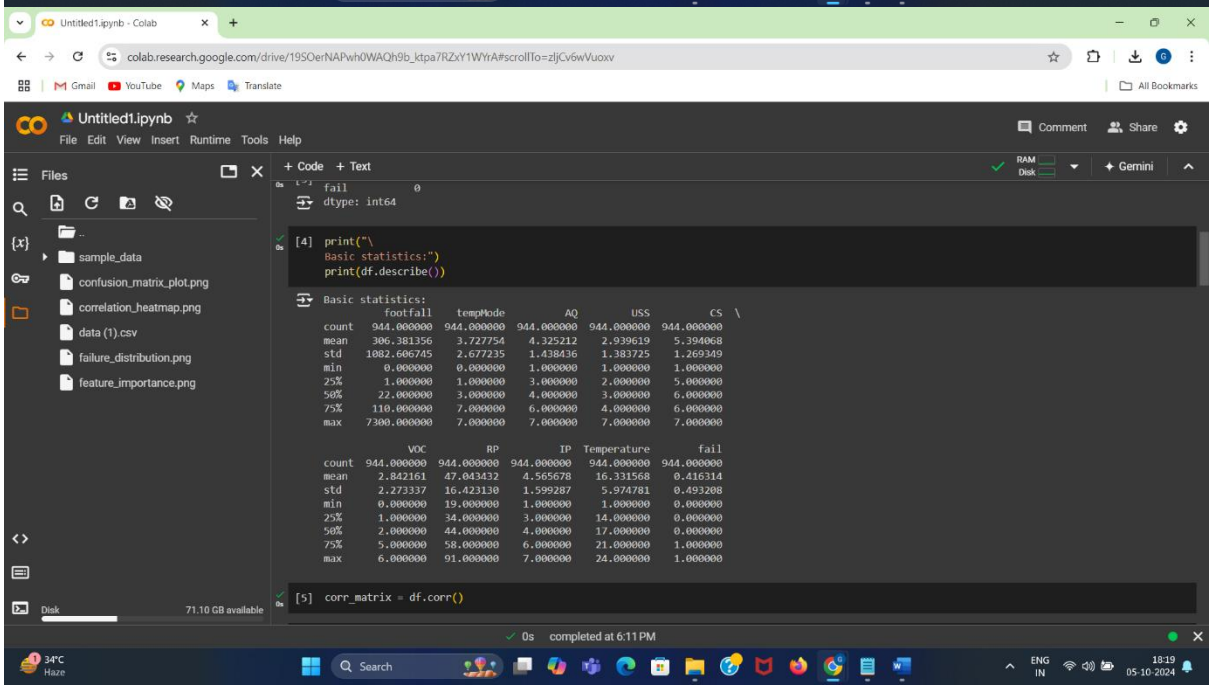
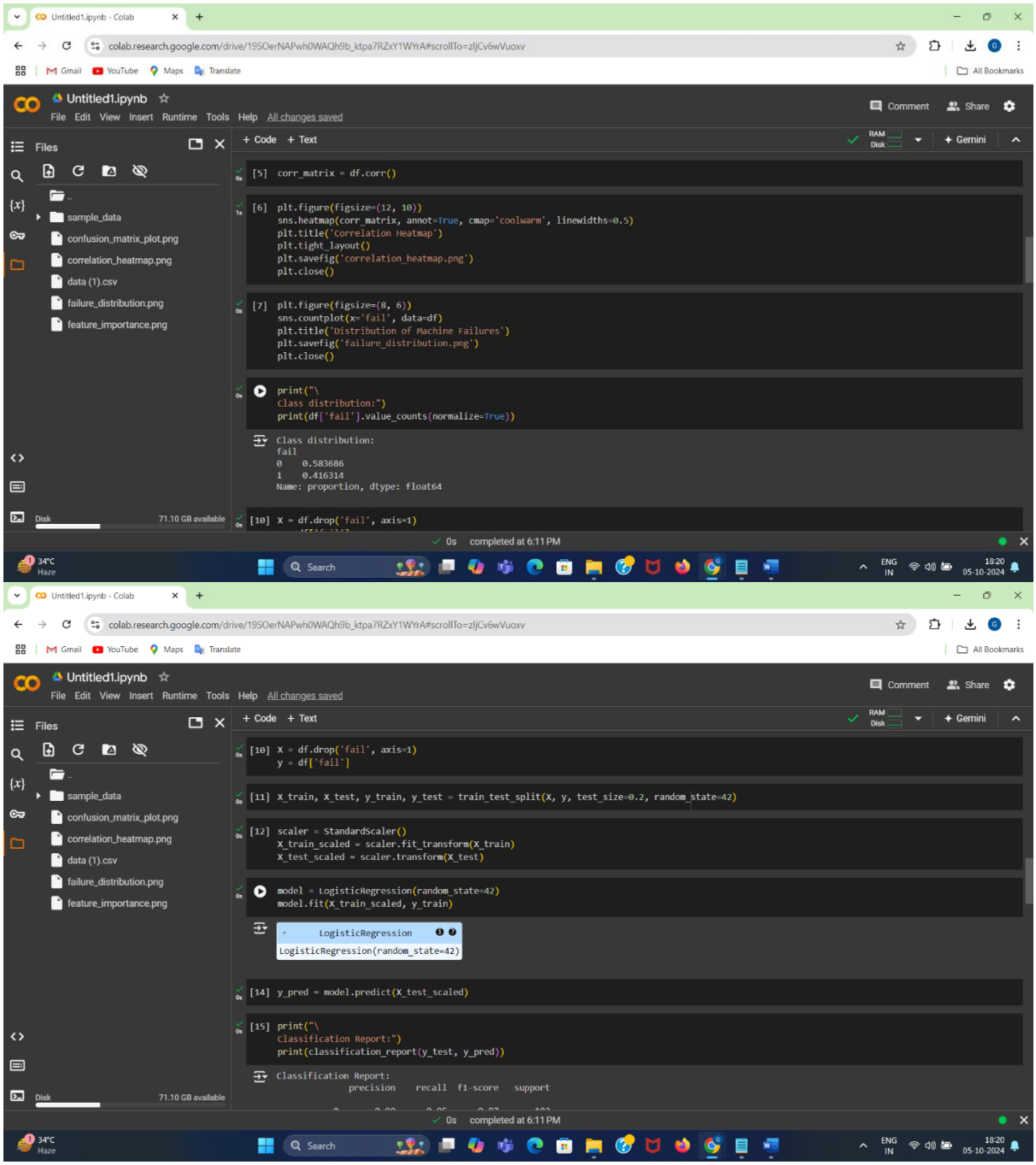
Machine Failure Prediction Project

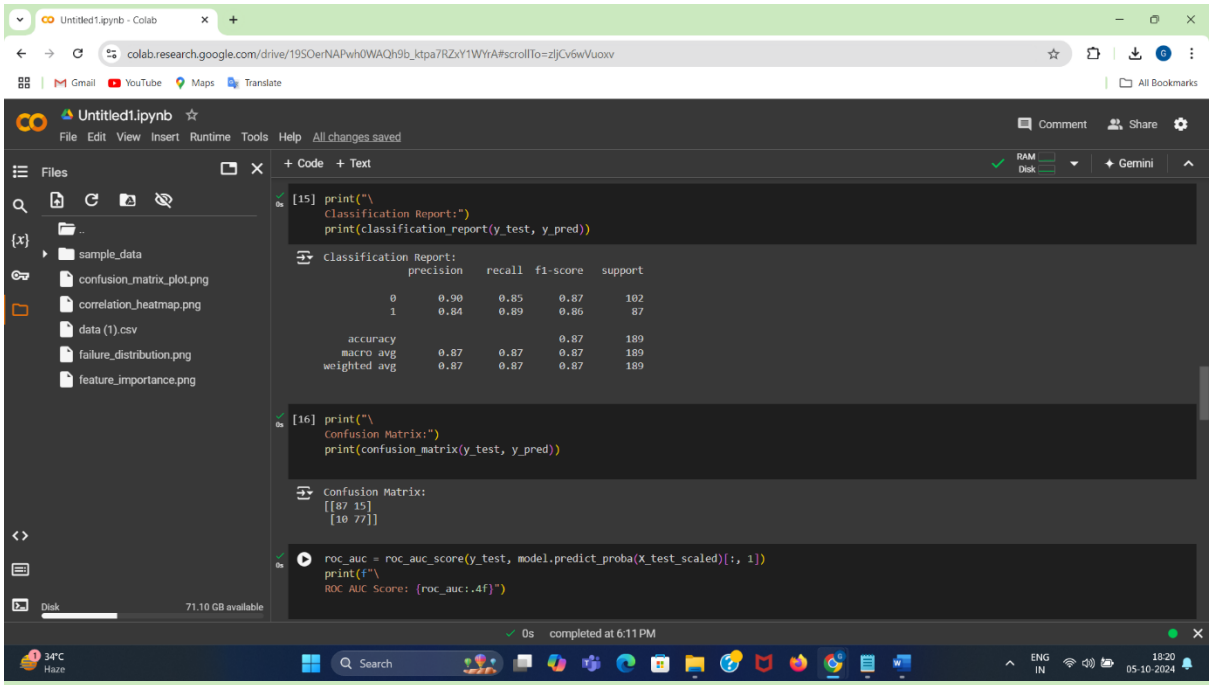
I have used a logistic regression model to analyse the relationship between sensor readings and the likelihood of machine failure.

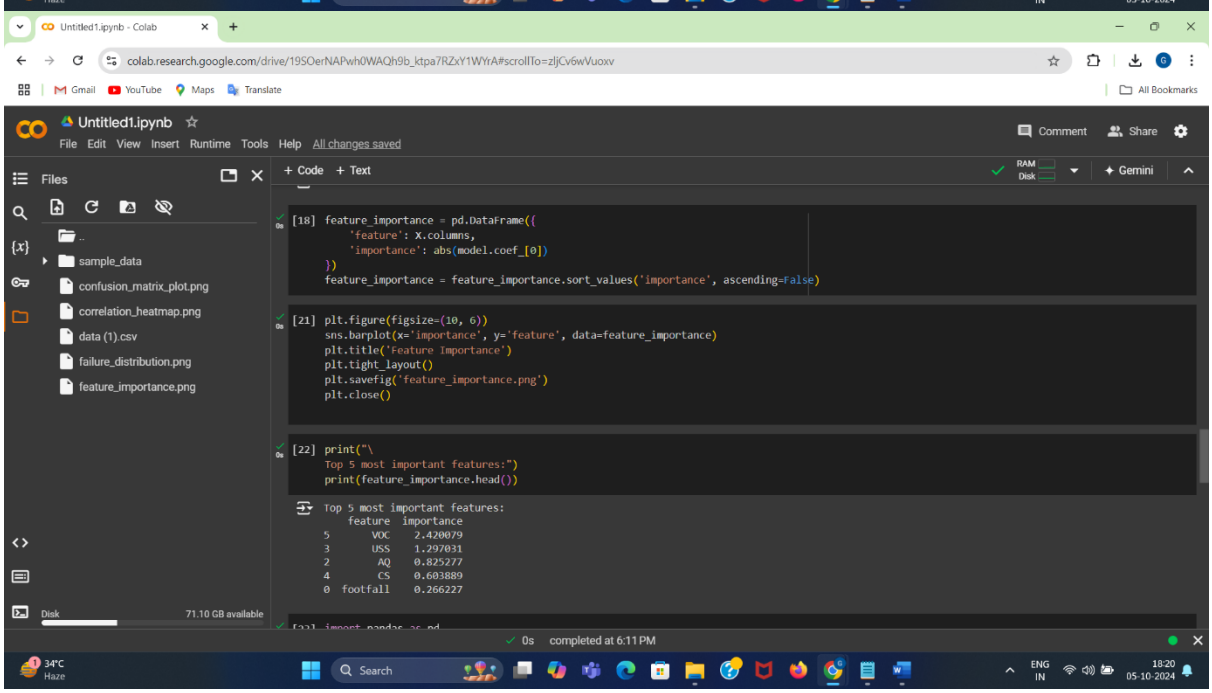


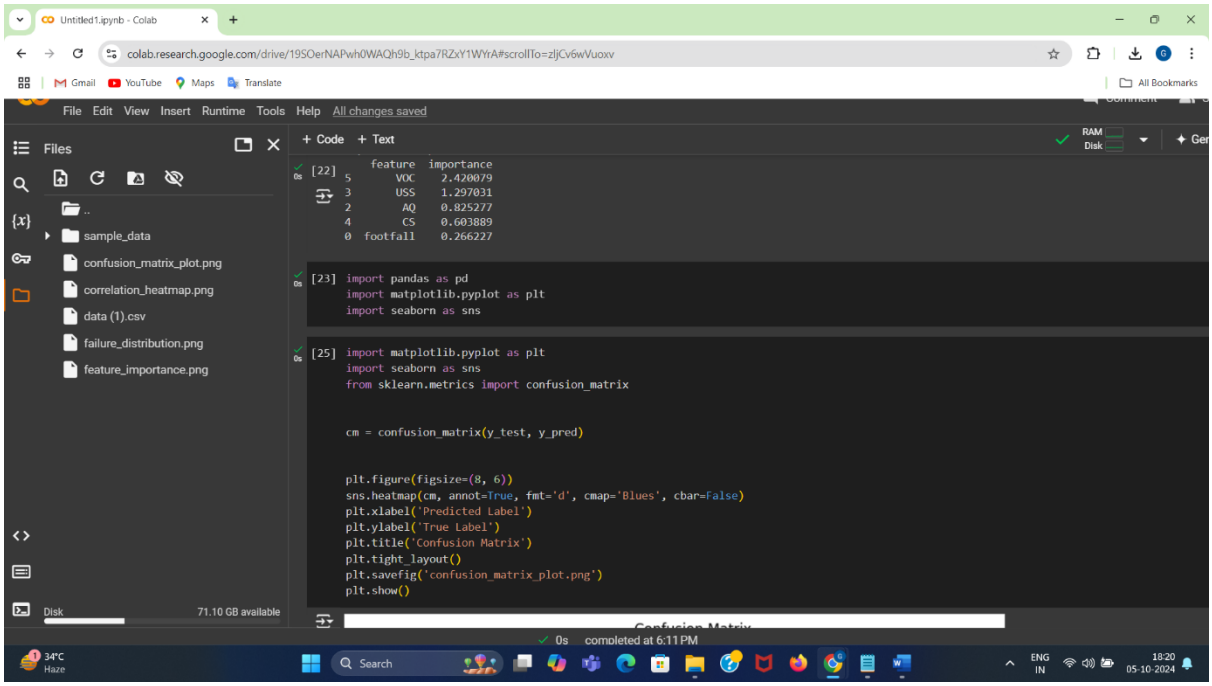



Untitled1.ipynb - Colab
colab.research.google.com/drive/19SOerNAPwh0WAQh9b_ktpa7RZxY1WYrA#scrollTo=zIjCv6wVuoxv
Untitled1.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
Files
+ Code + Text
Gemini
sample_data
confusion_matrix_plot.png
correlation_heatmap.png
data (1).csv
failure_distribution.png
feature_importance.png
Disk 71.10 GB available
[5] corr_matrix = df.corr()
[6] plt.figure(figsize=(12, 10))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Heatmap')
plt.tight_layout()
plt.savefig('correlation_heatmap.png')
plt.close()
[7] plt.figure(figsize=(8, 6))
sns.countplot(x='fail', data=df)
plt.title('Distribution of Machine Failures')
plt.savefig('failure_distribution.png')
plt.close()
print("\
Class distribution:")
print(df['fail'].value_counts(normalize=True))
Class distribution:
fail
0 0.583686
1 0.416314
Name: proportion, dtype: float64
[10] X = df.drop('fail', axis=1)
0s completed at 6:11 PM
34°C Haze
Search
ENG IN
18:20 05-10-2024

Untitled1.ipynb - Colab
colab.research.google.com/drive/19SOerNAPwh0WAQh9b_ktpa7RZxY1WYrA#scrollTo=zIjCv6wVuoxv
Untitled1.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
Files
+ Code + Text
Gemini
sample_data
confusion_matrix_plot.png
correlation_heatmap.png
data (1).csv
failure_distribution.png
feature_importance.png
Disk 71.10 GB available
[10] X = df.drop('fail', axis=1)
y = df['fail']
[11] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
[12] scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
model = LogisticRegression(random_state=42)
model.fit(X_train_scaled, y_train)
LogisticRegression
LogisticRegression(random_state=42)
[14] y_pred = model.predict(X_test_scaled)
[15] print("\
Classification Report:")
print(classification_report(y_test, y_pred))
Classification Report:
precision recall f1-score support
0s completed at 6:11 PM
34°C Haze
Search
ENG IN
18:20 05-10-2024


Untitled1.ipynb - Colab
colab.research.google.com/drive/19SOerNAPwh0WAQh9b_ktpa7RZxY1WYrA#scrollTo=zIjCv6wVuoxv
Gmail YouTube Maps Translate All Bookmarks
Untitled1.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
Files
+ Code + Text
RAM Disk
Gemini
sample_data
confusion_matrix_plot.png
correlation_heatmap.png
data (1).csv
failure_distribution.png
feature_importance.png
Disk 71.10 GB available
[15] print("\
Classification Report:")
print(classification_report(y_test, y_pred))
Classification Report:
precision recall f1-score support
0 0.90 0.85 0.87 102
1 0.84 0.89 0.86 87
accuracy 0.87 189
macro avg 0.87 0.87 0.87 189
weighted avg 0.87 0.87 0.87 189
[16] print("\
Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))
Confusion Matrix:
[[87 15]
[10 77]]
roc_auc = roc_auc_score(y_test, model.predict_proba(X_test_scaled)[:, 1])
print(f"\
ROC AUC Score: {roc_auc:.4f}")
0s completed at 6:11 PM
34°C Haze
Search
ENG IN
18:20 05-10-2024

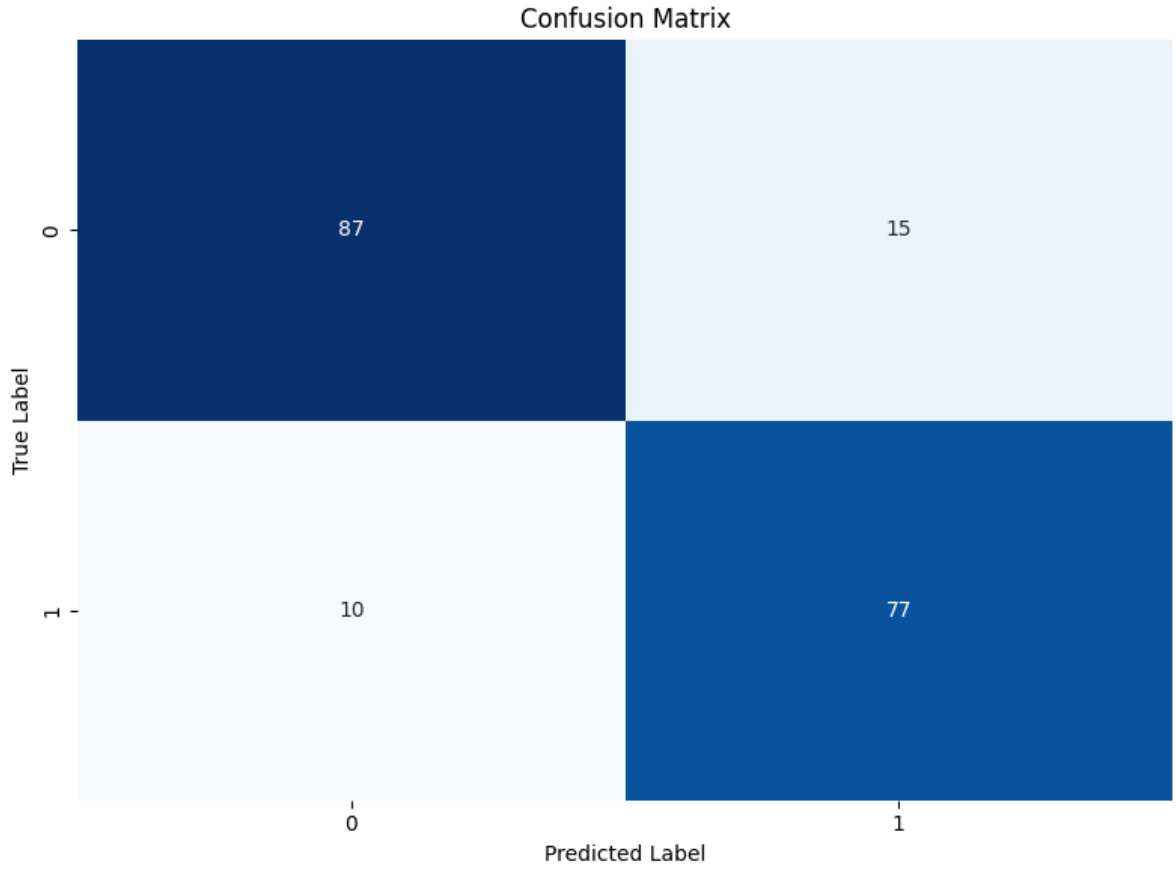

Untitled1.ipynb - Colab
colab.research.google.com/drive/19SOerNAPwh0WAQh9b_ktpa7RZxY1WYrA#scrollTo=zIjCv6wVuoxv
Gmail YouTube Maps Translate All Bookmarks
Untitled1.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
Files
+ Code + Text
RAM Disk
Gemini
sample_data
confusion_matrix_plot.png
correlation_heatmap.png
data (1).csv
failure_distribution.png
feature_importance.png
Disk 71.10 GB available
[18] feature_importance = pd.DataFrame({
'feature': X.columns,
'importance': abs(model.coef_[0])
})
feature_importance = feature_importance.sort_values('importance', ascending=False)
[21] plt.figure(figsize=(10, 6))
sns.barplot(x='importance', y='feature', data=feature_importance)
plt.title('Feature Importance')
plt.tight_layout()
plt.savefig('feature_importance.png')
plt.close()
[22] print("\
Top 5 most important features:")
print(feature_importance.head())
Top 5 most important features:
feature importance
5 VOC 2.420079
3 USS 1.297031
2 AQ 0.825277
4 CS 0.603889
0 footfall 0.266227
0s completed at 6:11 PM
34°C Haze
Search
ENG IN
18:20 05-10-2024



The dataset has no missing values.

Here are the results and visualizations:

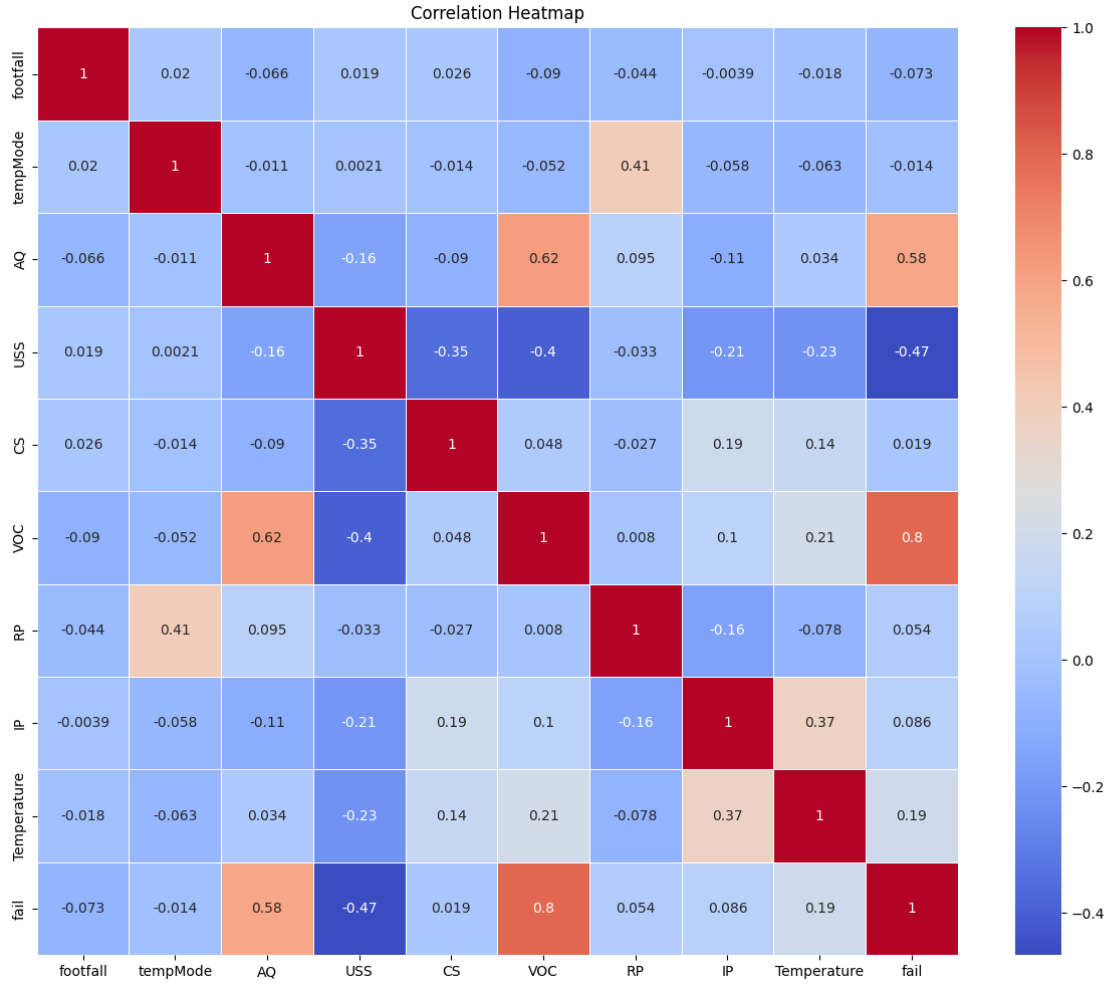
1) Confusion matrix: shows the model's performance.



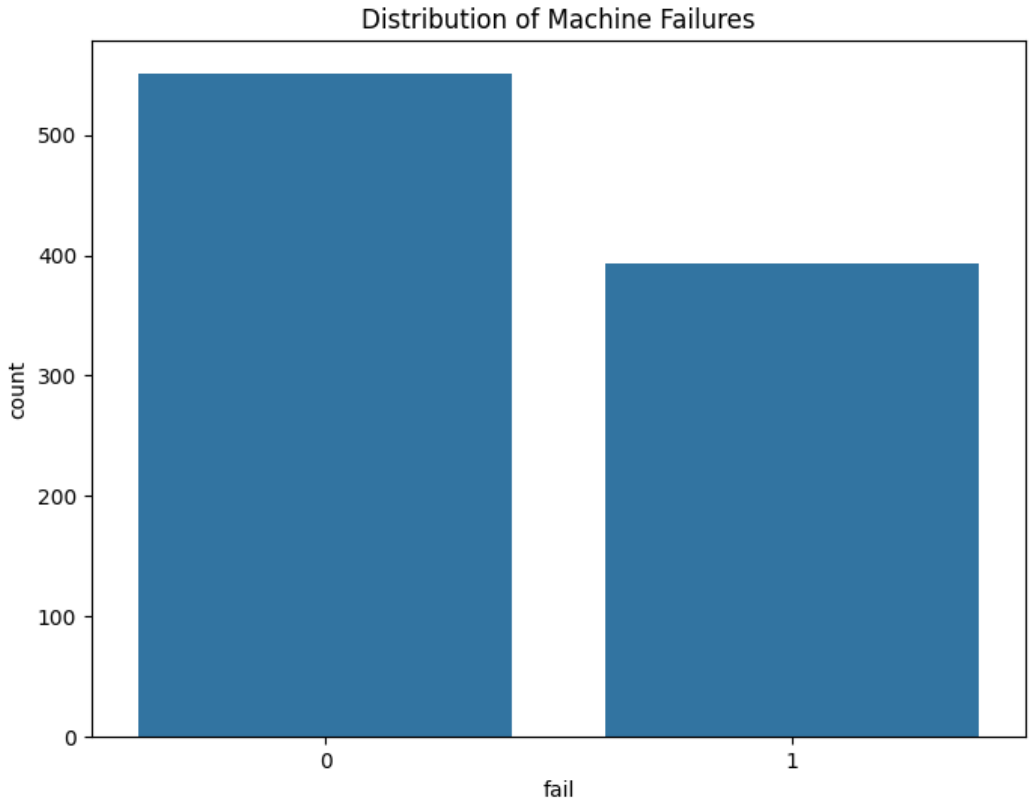
- True Negatives: 87, correctly predicted 87 instances where there was no machine failure.
- False Positives: 15, incorrectly predicted 15 instances as failures when they were actually non-failures.
- False Negatives: 10, missed 10 actual failures, predicting them as non-failures.
- True Positives: 77, correctly identified 77 instances of machine failures.

Model performed well in accurately in identifying failures and non- failures.

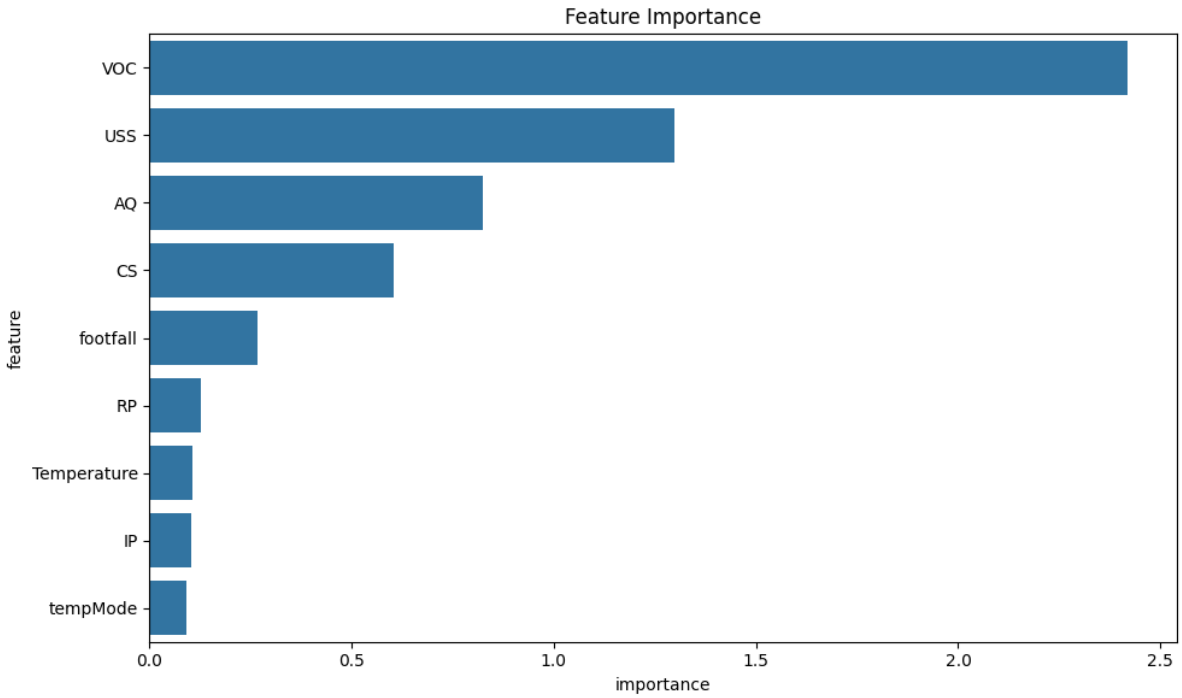
2) Correlation Heatmap: shows relationship between different features.



3) Failure Distribution: shows the distribution of machine failures in the dataset.



4) Feature Importance: highlights the most important features in predicting machine failures.



Other Insights and Results:

1) Top 5 most important features:

VOC 2.420079

USS 1.297031
AQ 0.825277
CS 0.603889
footfall 0.266227

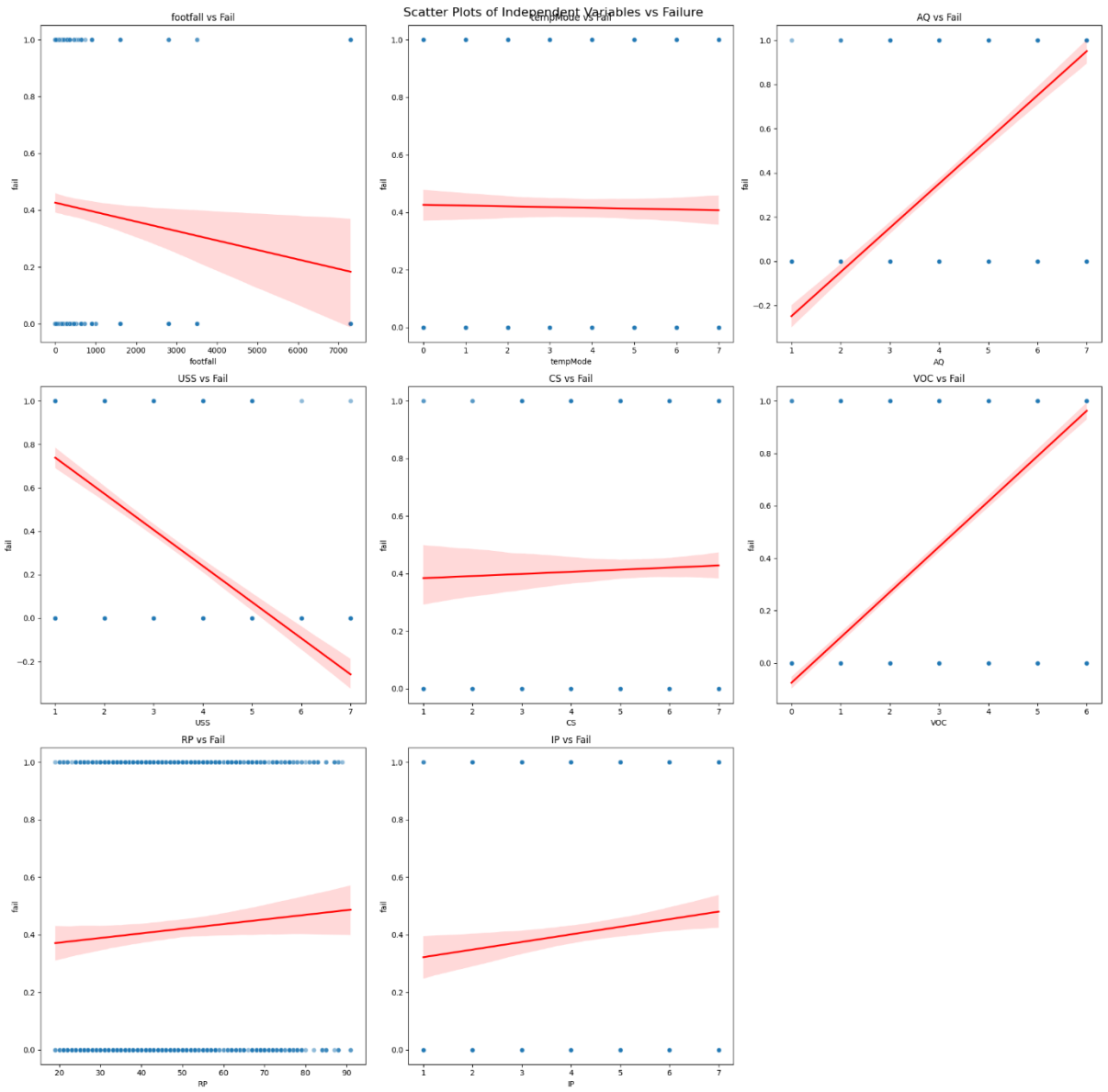
2) ROC AUC Score: 0.9446
3) Classification Report:

	precision	recall	F1 - score	support
0	0.90	0.85	0.87	102
1	0.84	0.89	0.86	87
accuracy			0.87	189
macro avg	0.87	0.87	0.87	189
weight avg	0.87	0.87	0.87	189

4) Class distribution:
fail
0 0.583686
1 0.416314
Name: proportion, dtype: float64

The logistic regression model performed well with a high ROC AUC score, indicating good predictive power. The most important features were VOC, USS, and AQ.

Correlation scatter plots to show the relationship between all features with the dependent variable (fail):



These correlation coefficients show the following observations from the scatter plots:

1. VOC has the strongest positive correlation (0.797) with failure.
2. AQ has the second strongest positive correlation (0.583).
3. USS has a moderate negative correlation (-0.467).
4. Temperature has a weak positive correlation (0.190).
5. Other variables have very weak correlations, either positive or negative.