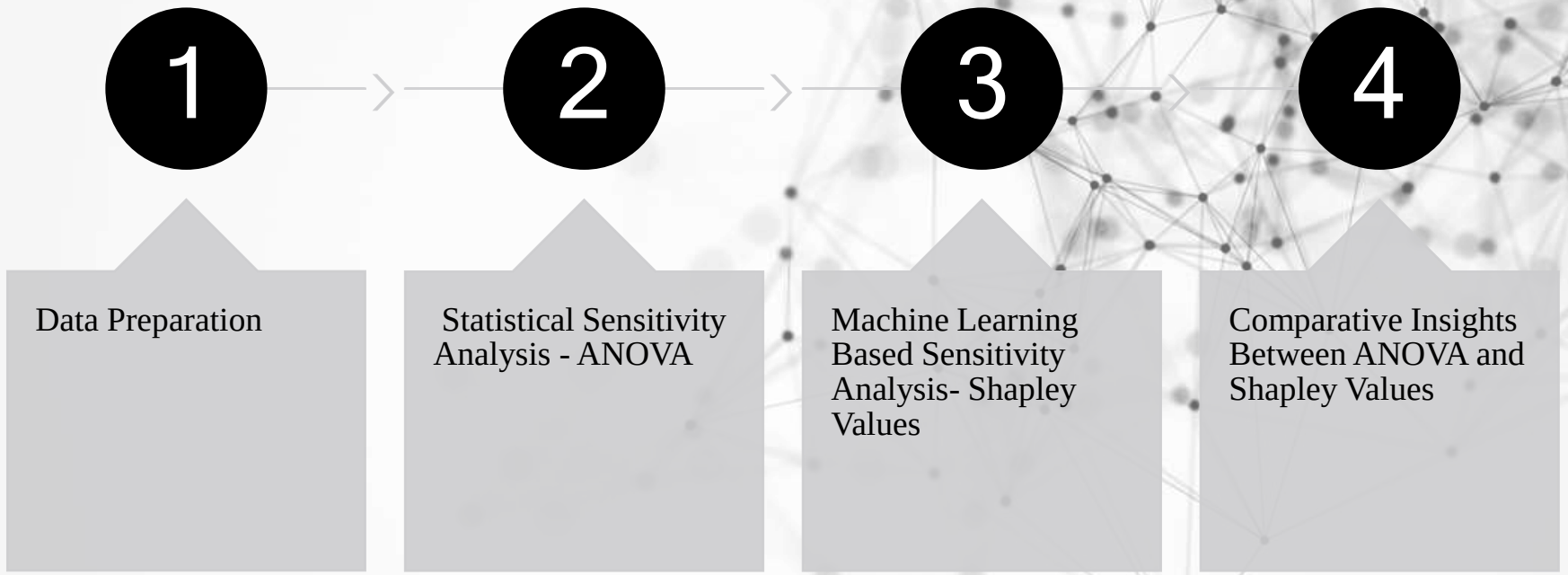
ENS 491
Machine Learning Exercise Part 2 Report

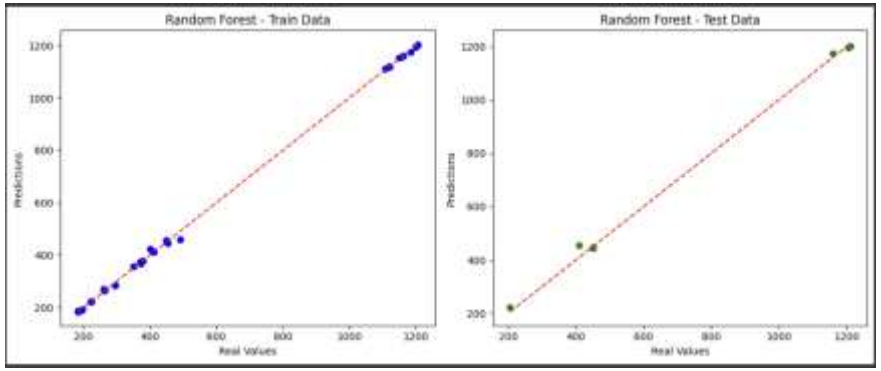
- Zeynep Derya Selçuk (29391)
- Darwin Serdar Dikici (30051)
- Selin Tıraş (2949)
- Efe Çelik (29268)
- Tunahan Arslan (29210)

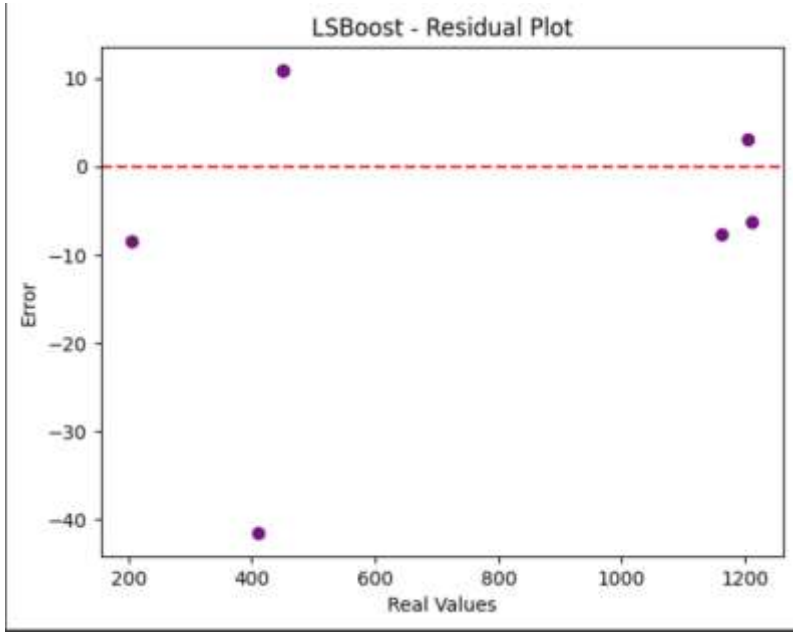

1
2
3
4
Data Preparation
Statistical Sensitivity Analysis - ANOVA
Machine Learning Based Sensitivity Analysis- Shapley Values
Comparative Insights Between ANOVA and Shapley Values

Data Preparation

- Data preprocessing was done in Task 1.
- Normalization was done MinMaxScalar was used.
- LS Boost and Random Forest performance compared.
- You can find the necessary analyzes and steps in the ML exercise part 1 report.
- Necessary additions were made for Part 2, using what was made in Part 1.

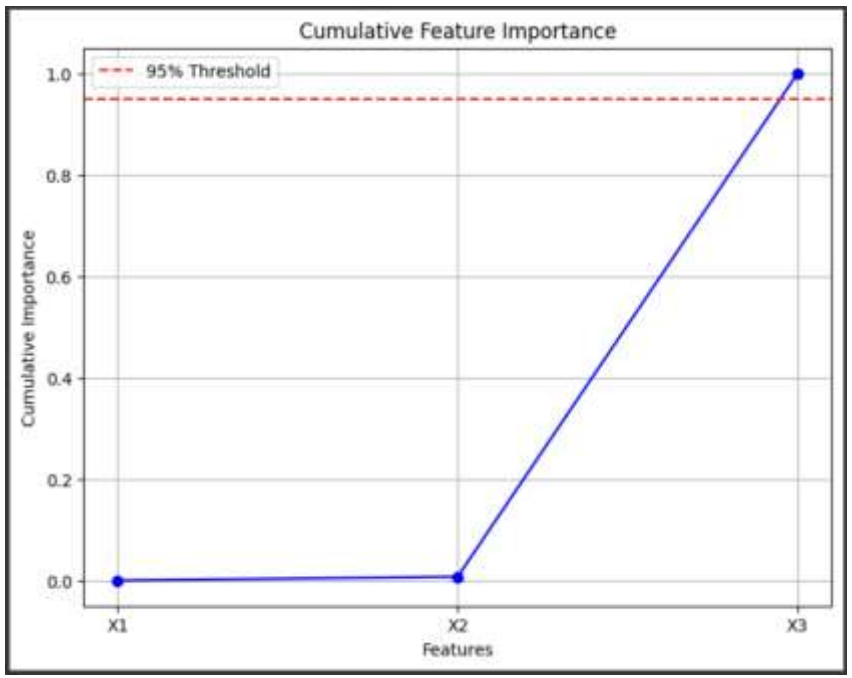
```
Training set size: (38, 3)
Test set size: (7, 3)
```





	Model	R² (Train)	R² (Test)	AME	RMSE
0	Random Forest	0.999636	0.997592	15.078571	20.121620
1	LSBoost	0.999569	0.998180	12.703682	17.494185

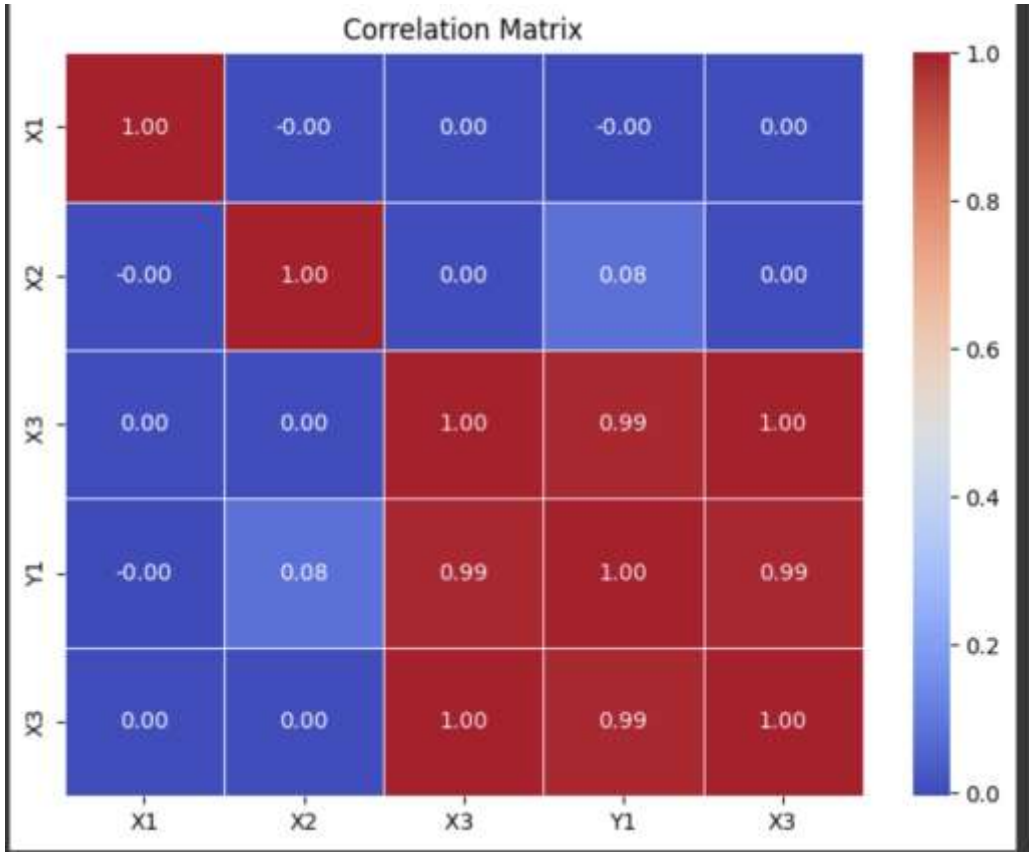
```python

feature_importances = rf_model.feature_importances_


cumulative_importances = feature_importances.cumsum()


features = ['X1', 'X2', 'X3']


plt.figure(figsize=(8, 6))
plt.plot(features, cumulative_importances, marker='o', linestyle='-', color='b')
plt.title("Cumulative Feature Importance")
plt.xlabel("Features")
plt.ylabel("Cumulative Importance")
plt.grid()
plt.axhline(y=0.95, color='r', linestyle='--', label="95% Threshold")
plt.legend()
plt.show()

```

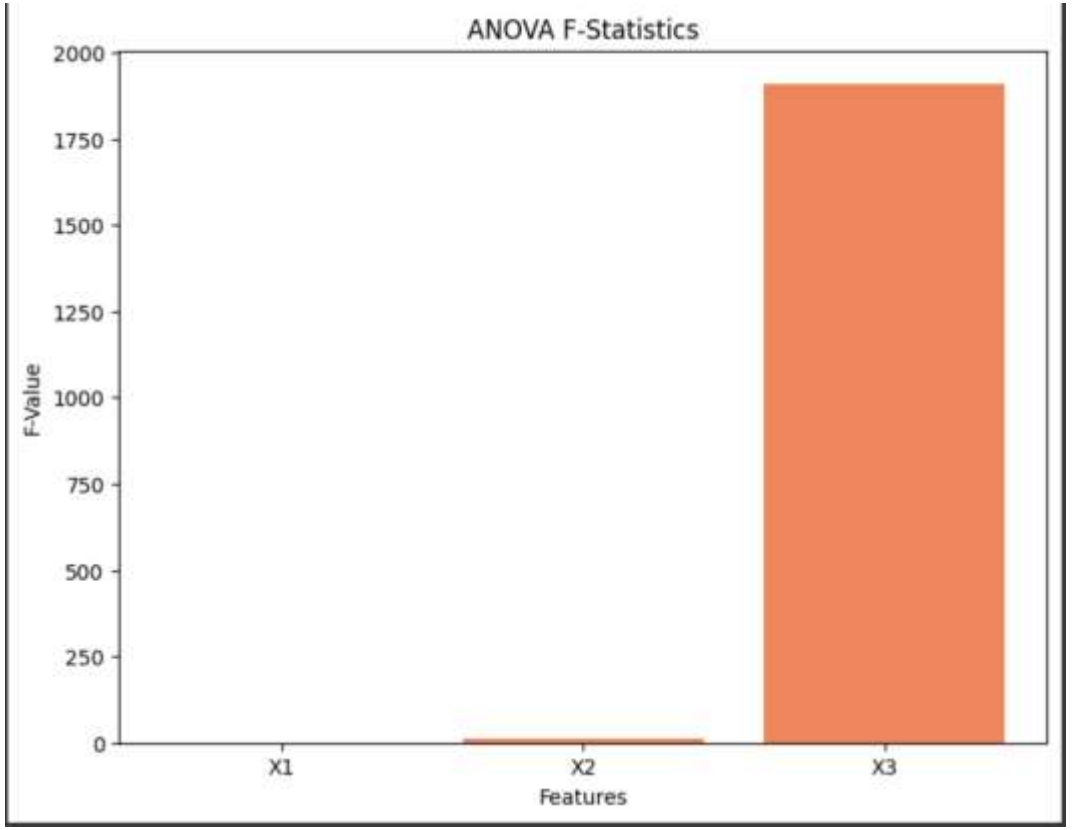


Statistical Sensitivity Analysis - ANOVA

- Conducted an Analysis of Variance (ANOVA) using the Ordinary Least Squares (OLS) method to evaluate the influence of each input feature (X1, X2, X3) on the target output (Y1).
- **Key Findings**: X3 was identified as the most significant feature with the highest F-statistic and lowest P-value.
- X2 showed moderate importance, while X1 had negligible influence.

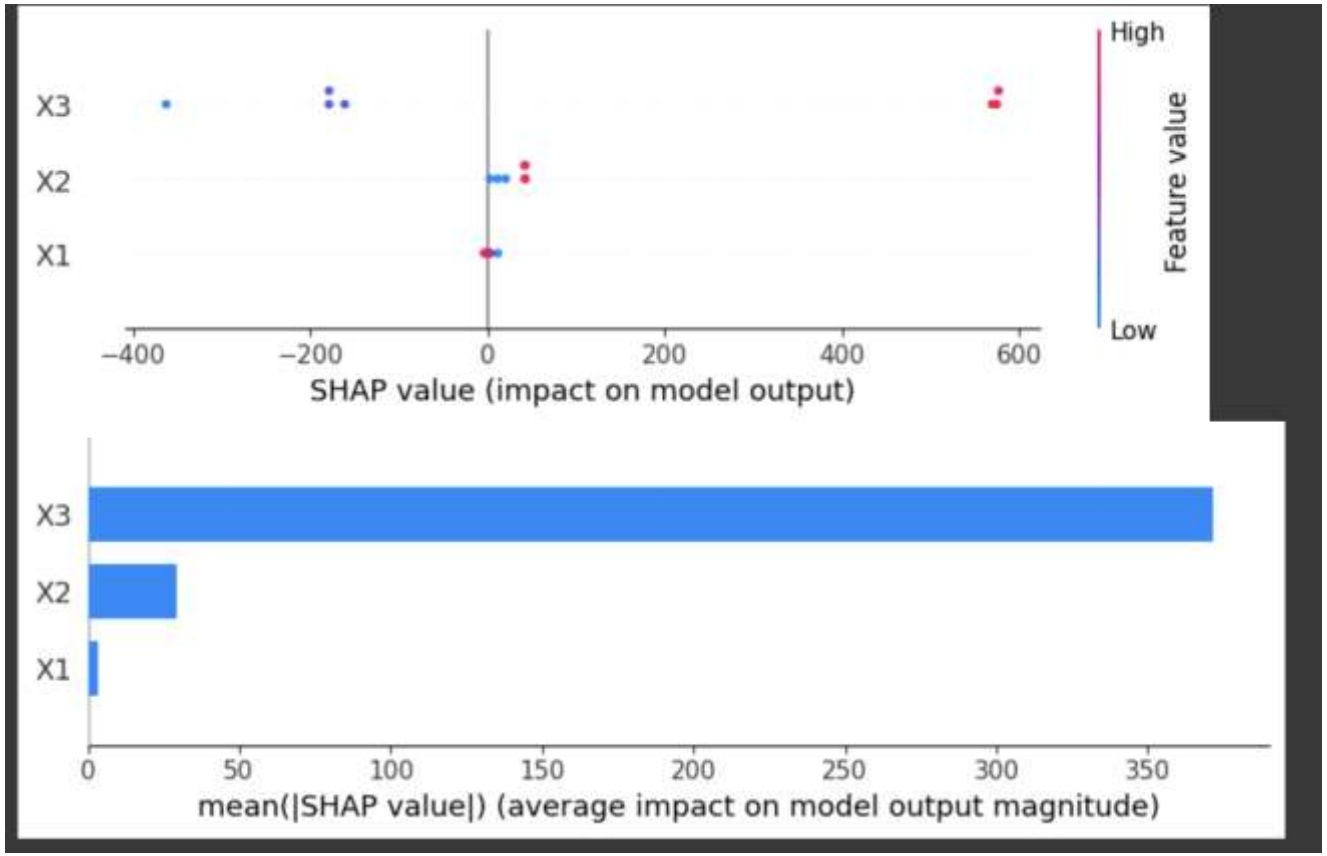
```
ANOVA results:
              df         sum_sq       mean_sq           F        PR(>F)
X1           1.0   1.369000e+02  1.369000e+02    0.036261  8.499172e-01
X2           1.0   5.216670e+04  5.216670e+04   13.817517  6.028574e-04
X3           1.0   7.211158e+06  7.211158e+06 1910.036406  5.115929e-36
Residual    41.0   1.547915e+05  3.775403e+03         NaN           NaN
```

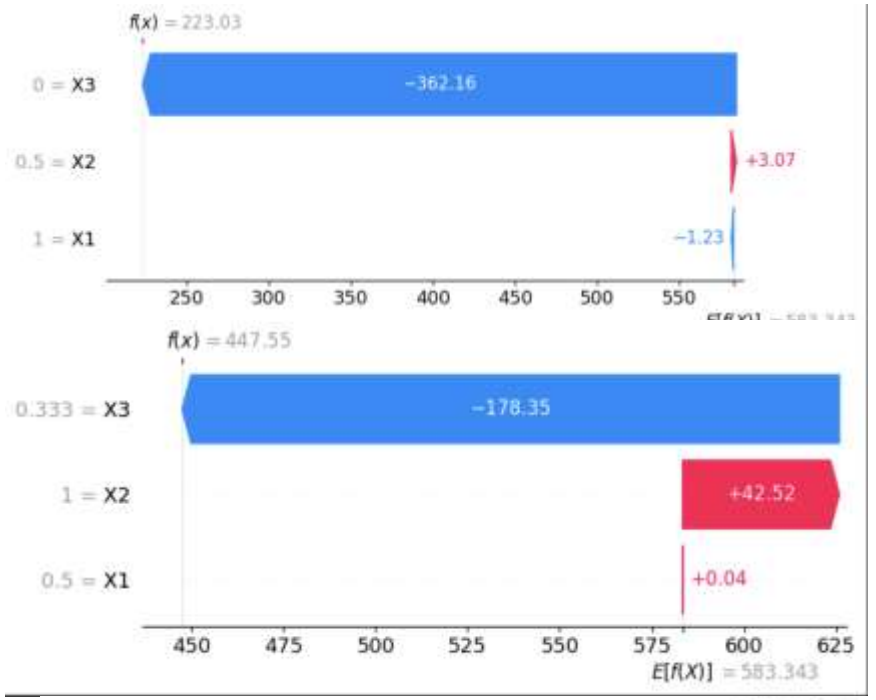


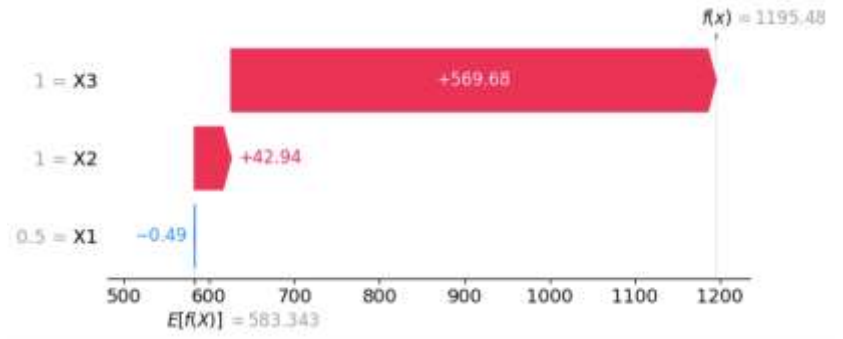


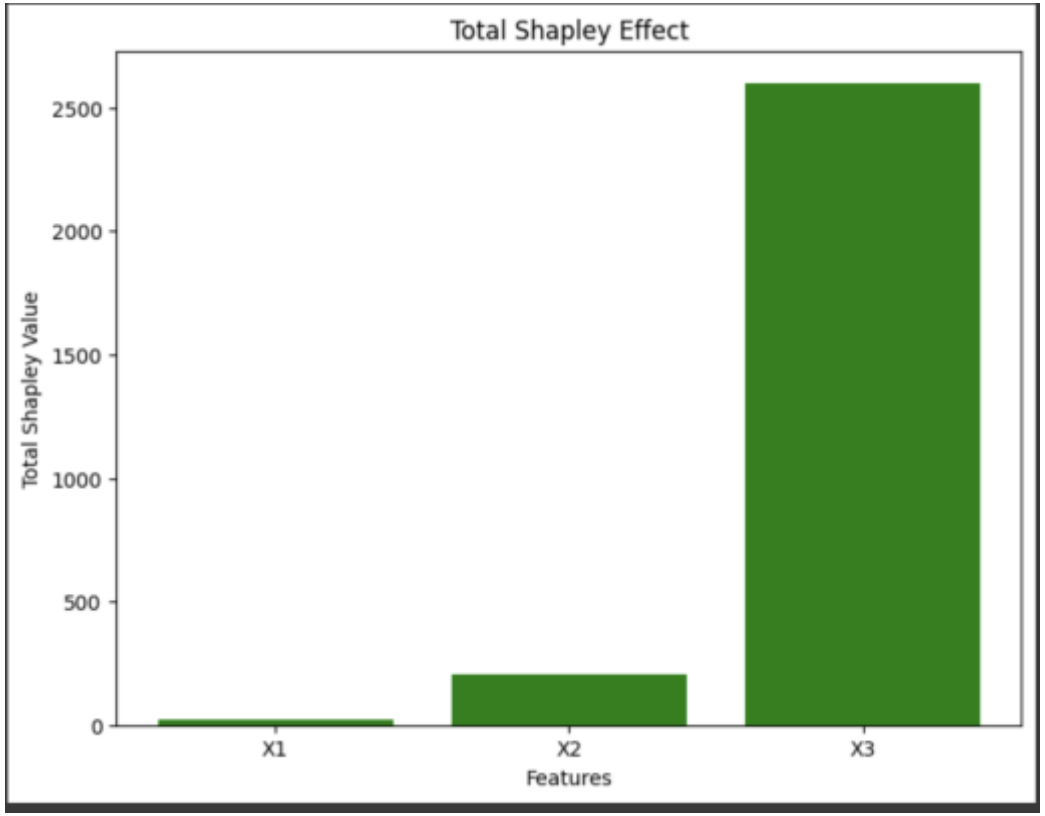
Machine Learning Based Sensitivity Analysis- Shapley Values

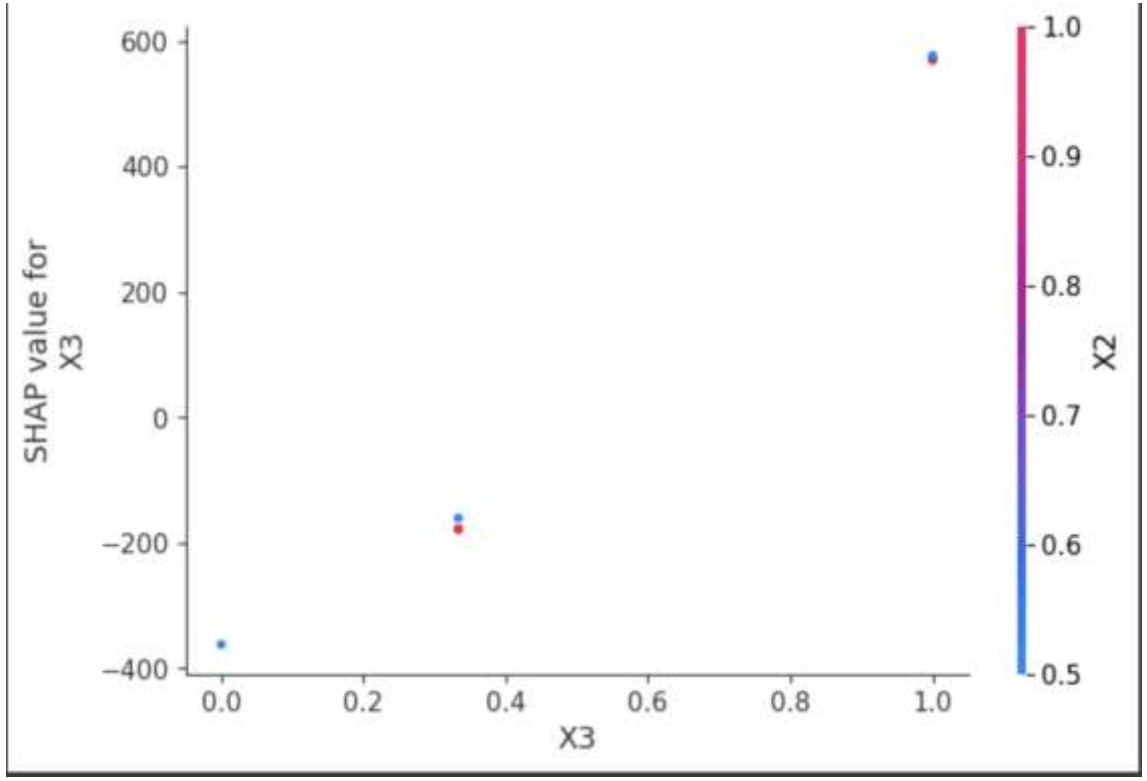
- **Shapley Values**:Used shap.TreeExplainer to calculate Shapley values, providing a machine learning-based perspective on feature importance.
- **Steps**:
  - Calculated Shapley values for the test dataset.
  - Generated summary plots and bar graphs to visualize global feature importance.
  - Highlighted interactions between features (e.g., X3 and X2).
- **Interaction Analysis**:
  - Explored feature interactions using Shapley dependency plots, which revealed a strong relationship between X3 and X2.


X3
X2
X1
−400
−200
0
200
400
600
SHAP value (impact on model output)
High
Feature value
Low
X3
X2
X1
0
50
100
150
200
250
300
350
mean(|SHAP value|) (average impact on model output magnitude)


f(x) = 223.03
0 = X3
−362.16
0.5 = X2
+3.07
1 = X1
−1.23
250
300
350
400
450
500
550
f(x) = 447.55
0.333 = X3
−178.35
1 = X2
+42.52
0.5 = X1
+0.04
450
475
500
525
550
575
600
625
E[f(X)] = 583.343


f(x) = 1195.48
1 = X3
+569.68
1 = X2
+42.94
0.5 = X1
−0.49
500
600
700
800
900
1000
1100
1200
E[f(X)] = 583.343


Total Shapley Effect
Total Shapley Value
2500
2000
1500
1000
500
0
X1
X2
X3
Features

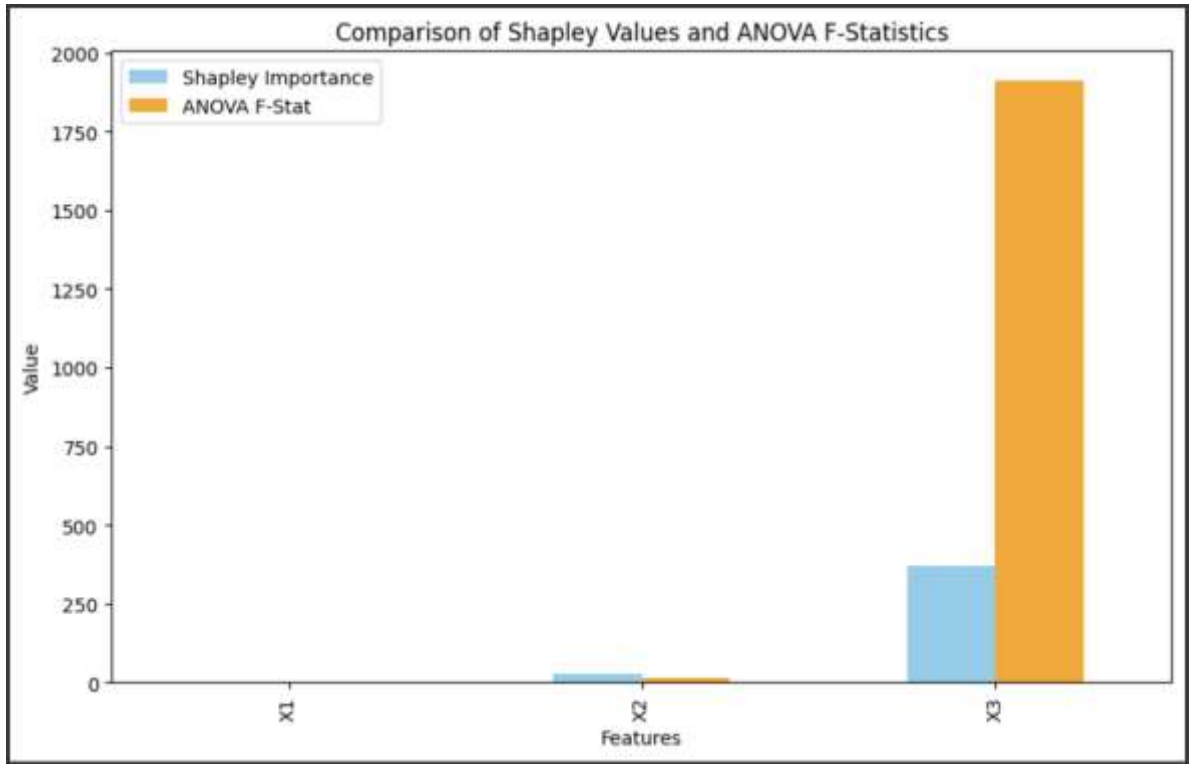

SHAP value for
X3
600
400
200
0
−200
−400
0.0
0.2
0.4
0.6
0.8
1.0
X3
X2
1.0
0.9
0.8
0.7
0.6
0.5

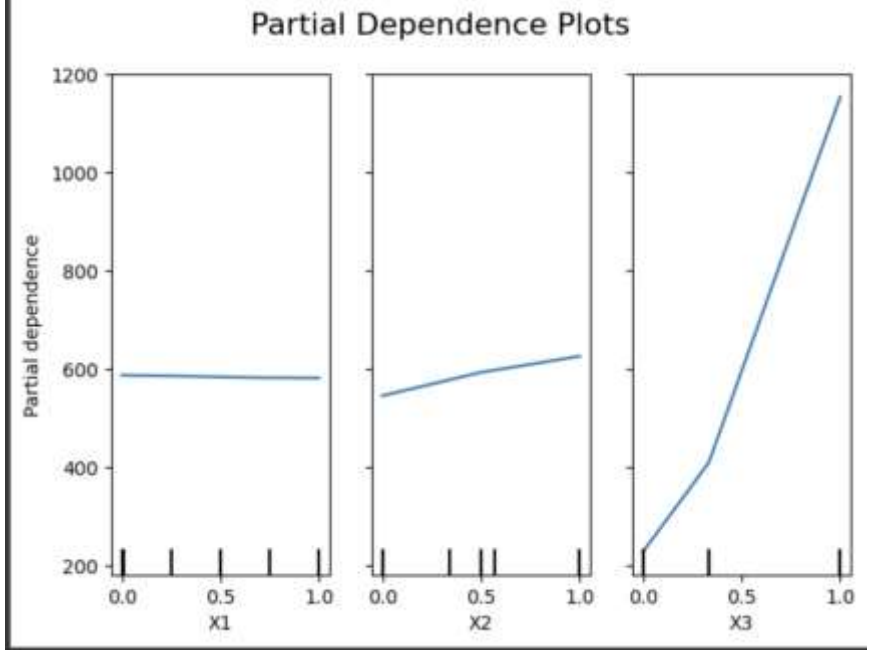
Comparative Insights Between ANOVA and Shapley Values

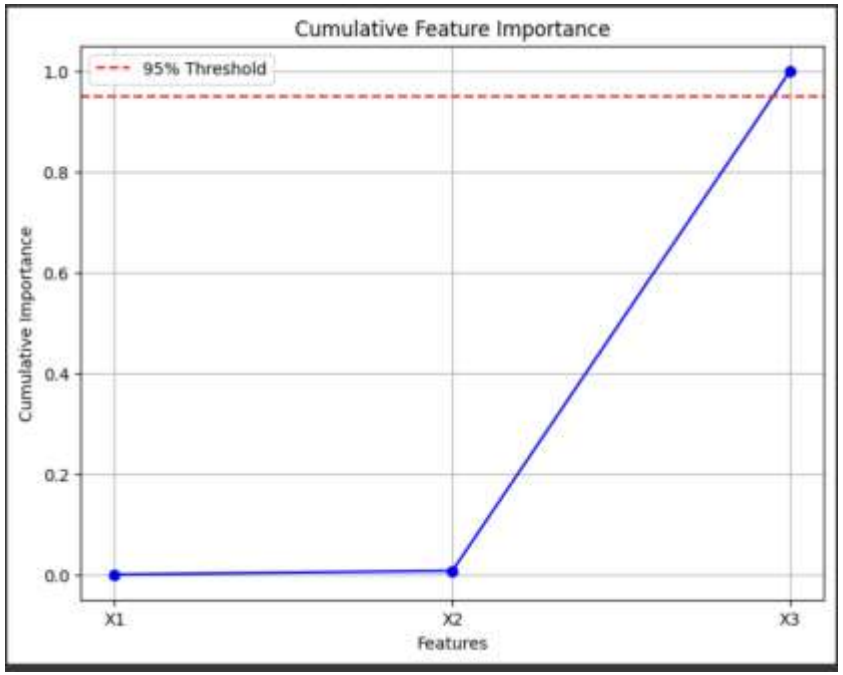
- **Comparison Between ANOVA and Shapley Values**: Both methods confirmed X3 as the most influential feature, followed by X2.
- Shapley analysis provided additional interpretability, capturing feature interactions that ANOVA could not.
- **Conclusion**:X3 is the dominant feature influencing the model's behavior, supported by both statistical and machine learning-based analyses.
- Shapley values provided more granular insights, particularly regarding feature interactions, making it a valuable tool for sensitivity analysis.

Comparison of ANOVA and Shapley Results:

	Feature	ANOVA P-Value	Shapley Mean Importance
0	X1	8.499172e-01	3.209077
1	X2	6.028574e-04	29.489981
2	X3	5.115929e-36	371.504010







END OF THE REPORT