FOREST TYPE CLASSIFIERS

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

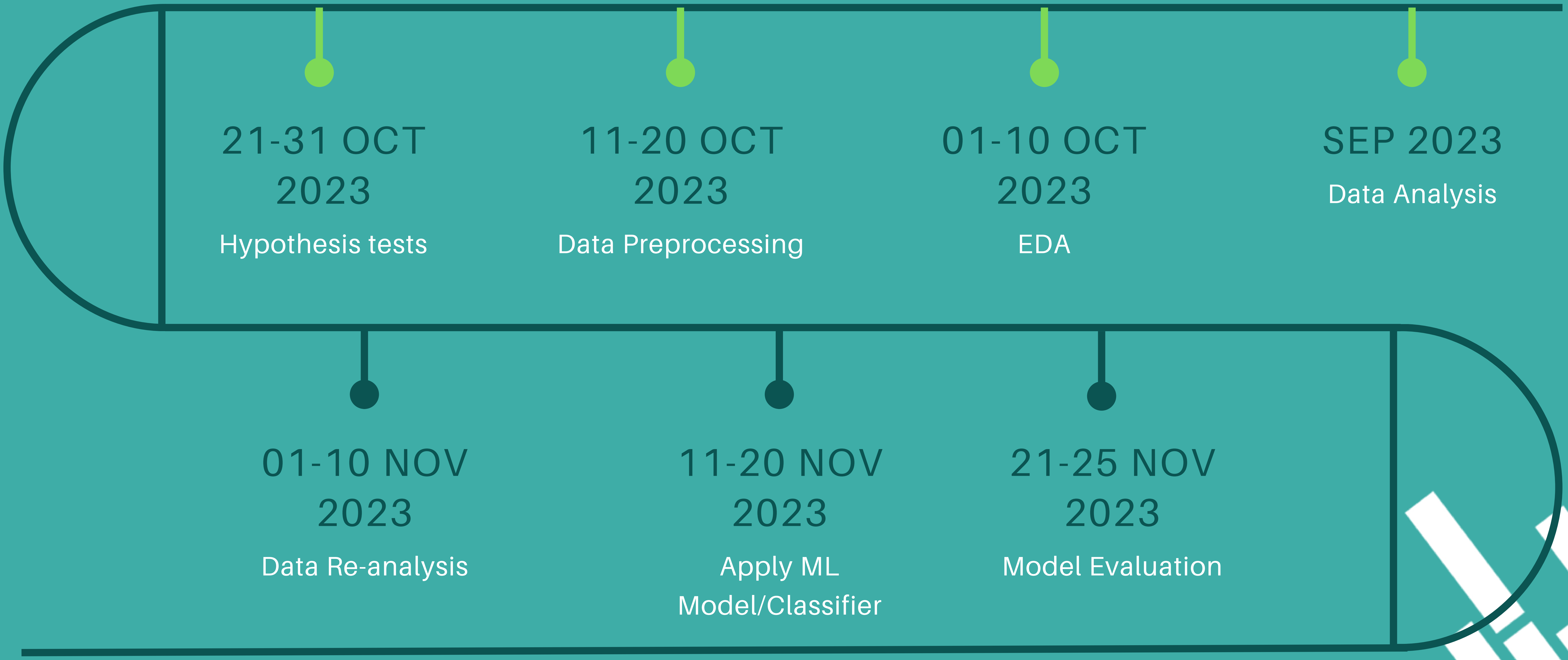
PROBLEM STATEMENT

Given a dataset containing various geographical and environmental features of Forested areas in the United States, the ultimate goal of this project is to analyze the data completely, with its features being engineered through certain techniques, one of them including the dimensionality reduction using the Principal Component Analysis, for the preprocessing step, and finally designing and/or implementing a good Machine Learning/Deep Learning algorithm for us to be able to correctly determine the forest cover type provided the finalized dataset, the description of which has been provided thoroughly in the subsequent section below.

IIIT D

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

TIMELINE



SEP 2023
Data Analysis

01-10 OCT 2023
EDA

11-20 OCT 2023
Data Preprocessing

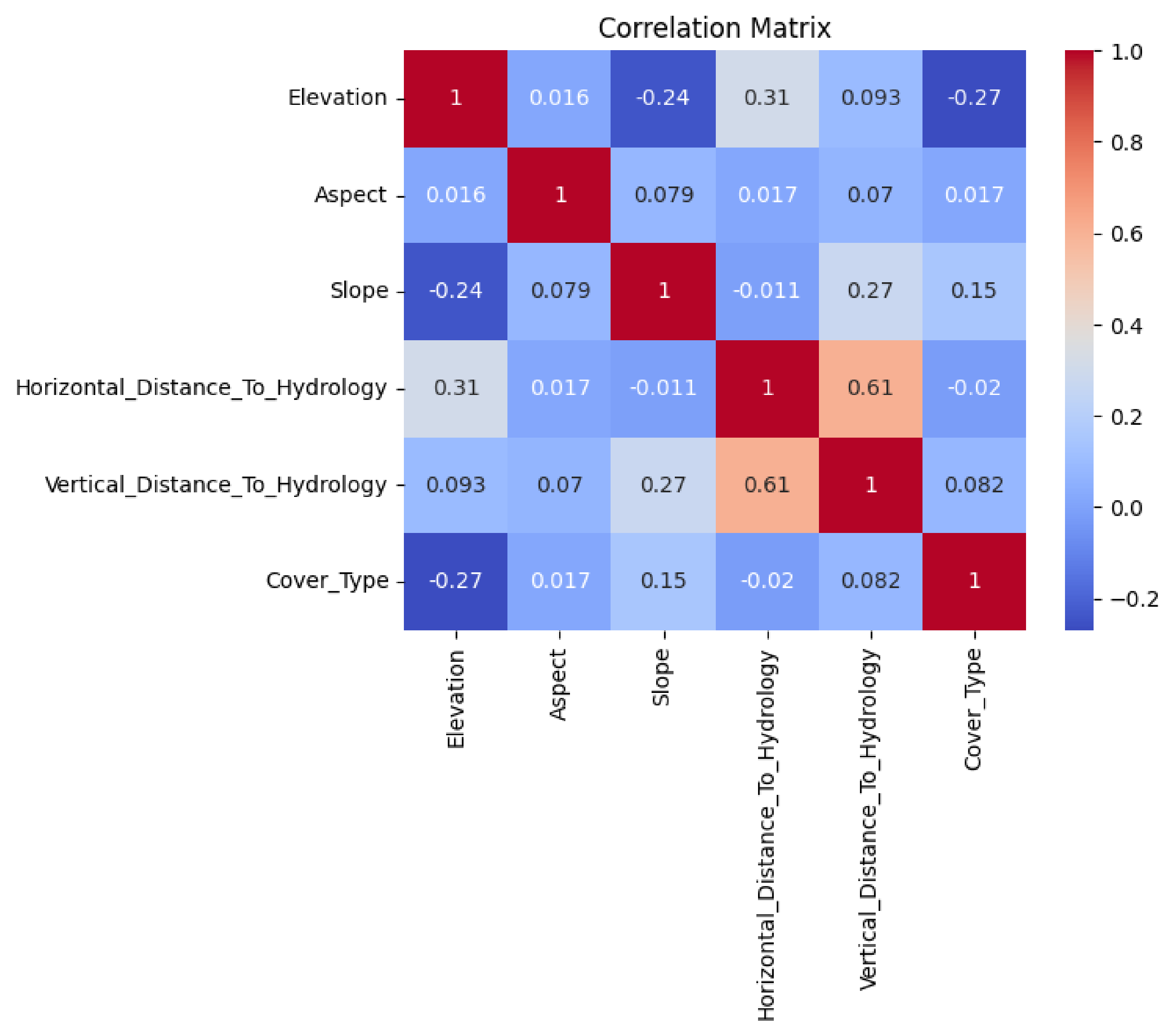
21-31 OCT 2023
Hypothesis tests

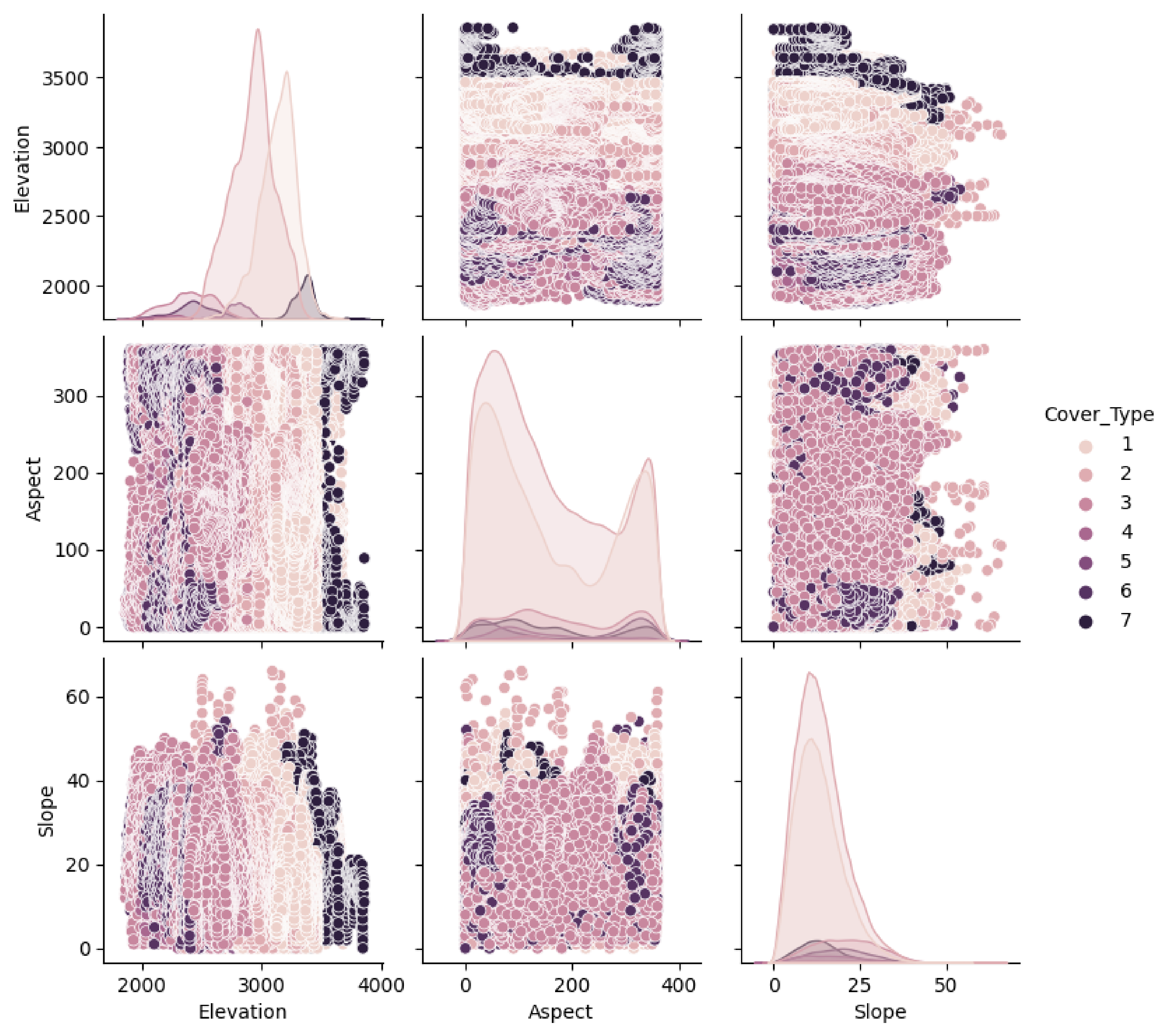
01-10 NOV 2023
Data Re-analysis

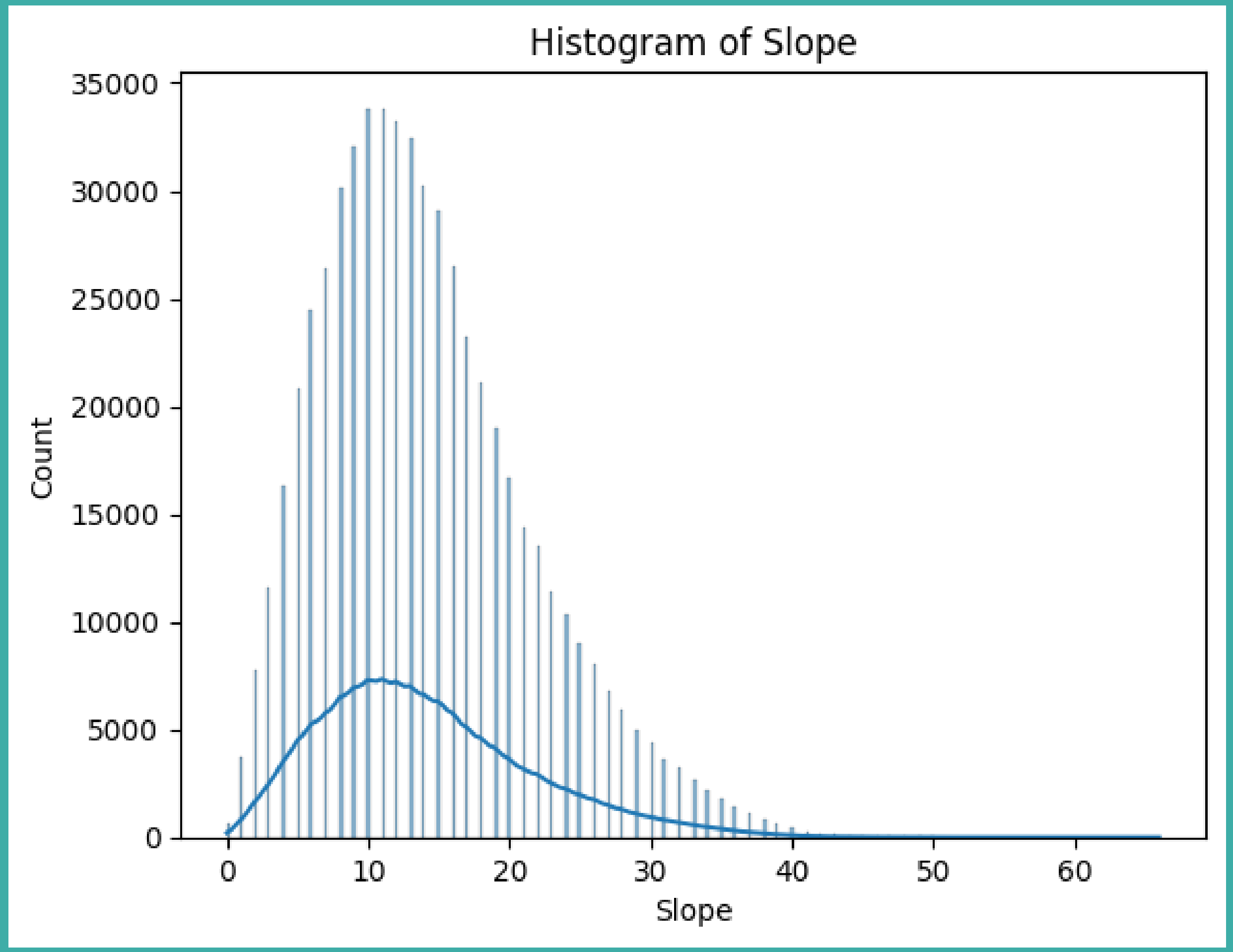
11-20 NOV 2023
Apply ML Model/Classifier

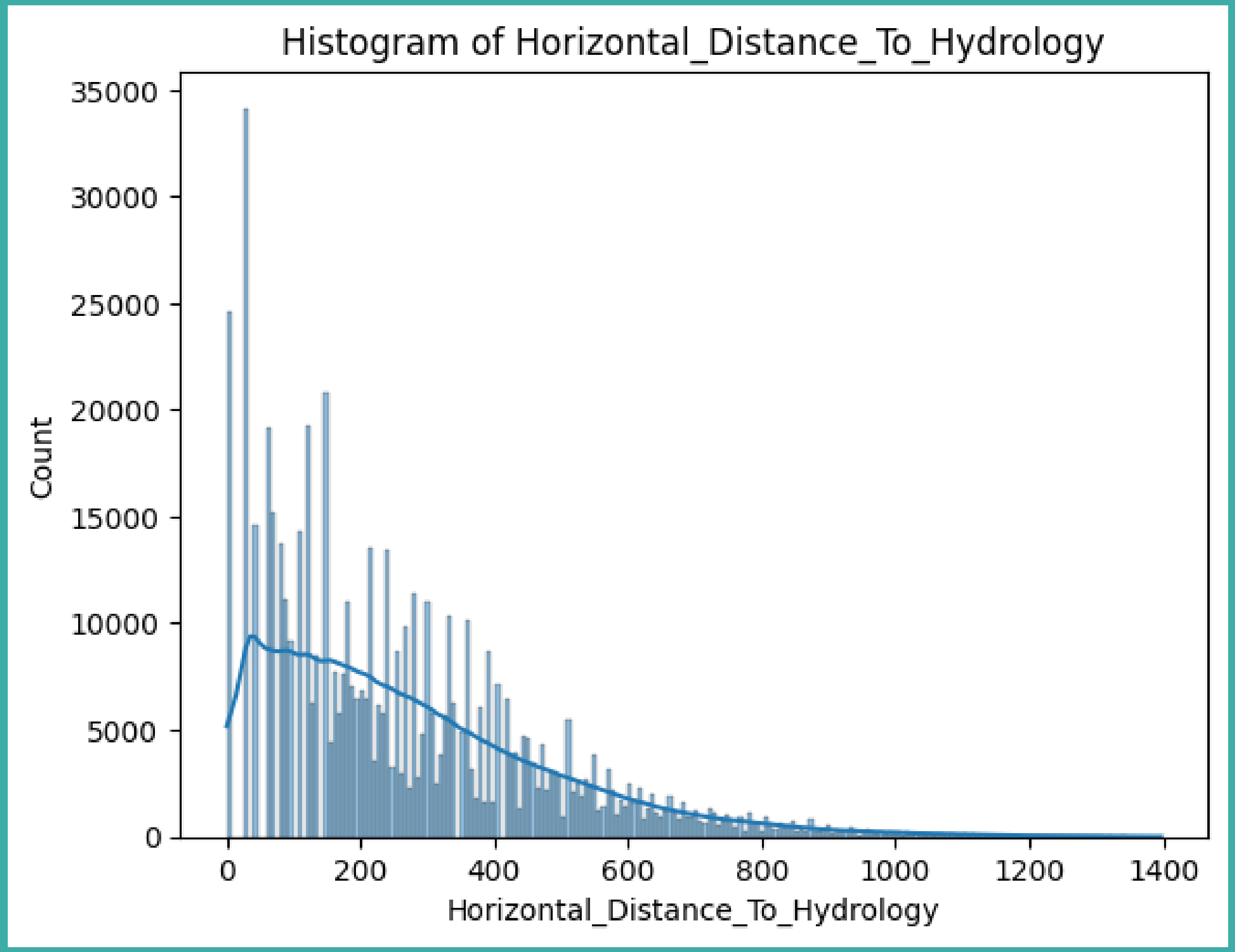
21-25 NOV 2023
Model Evaluation

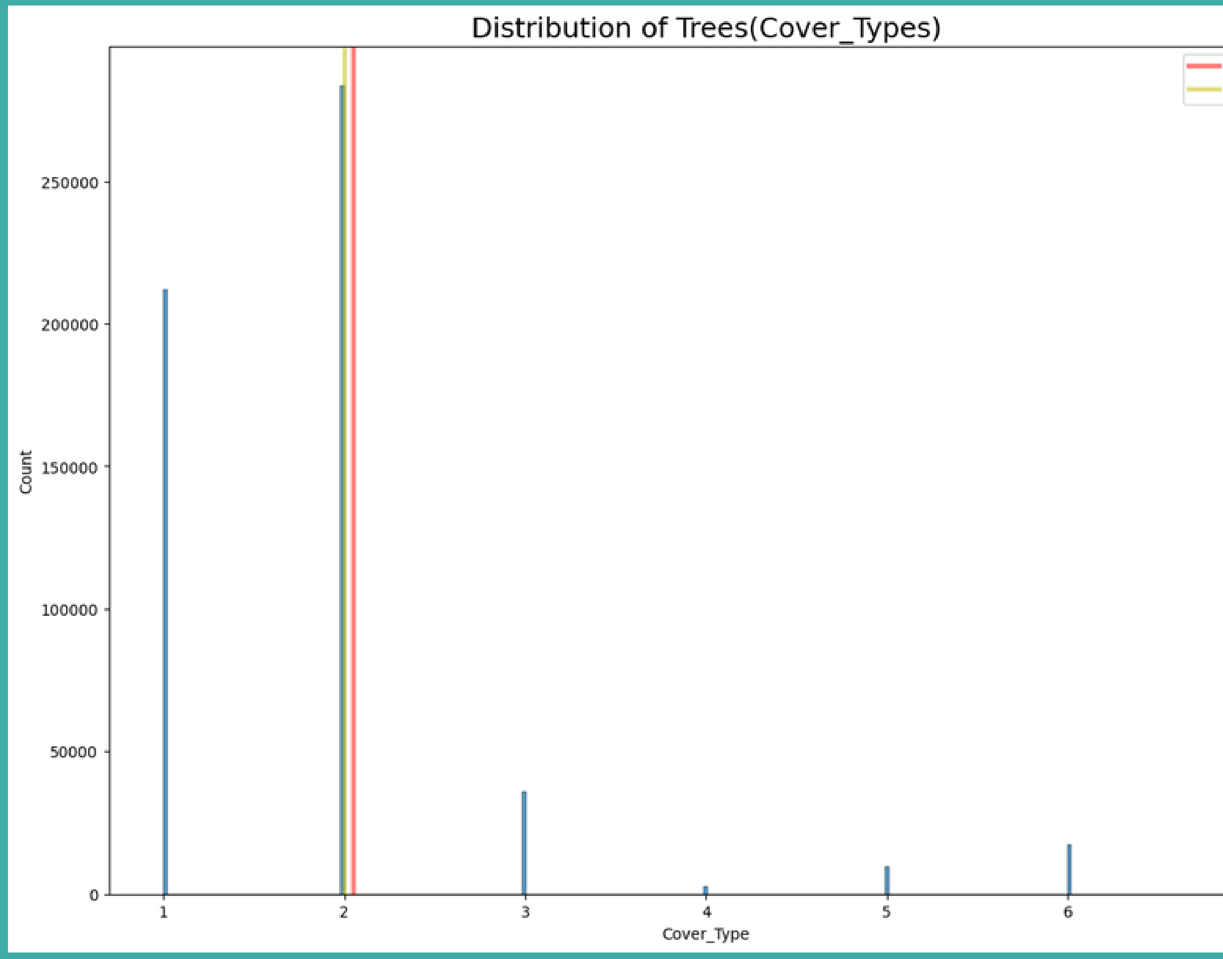
DATASET VISUALIZATION & DISTRIBUTION

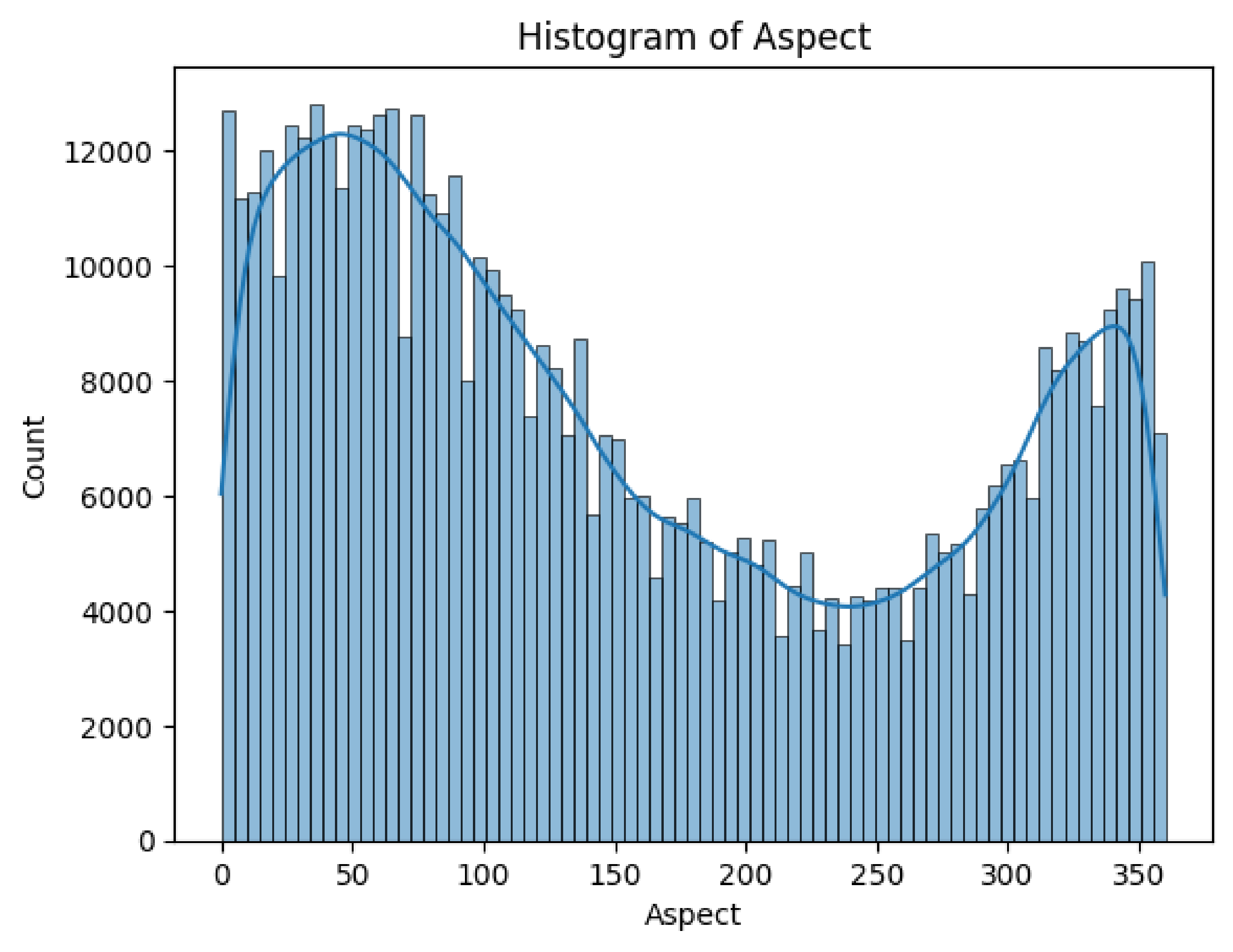


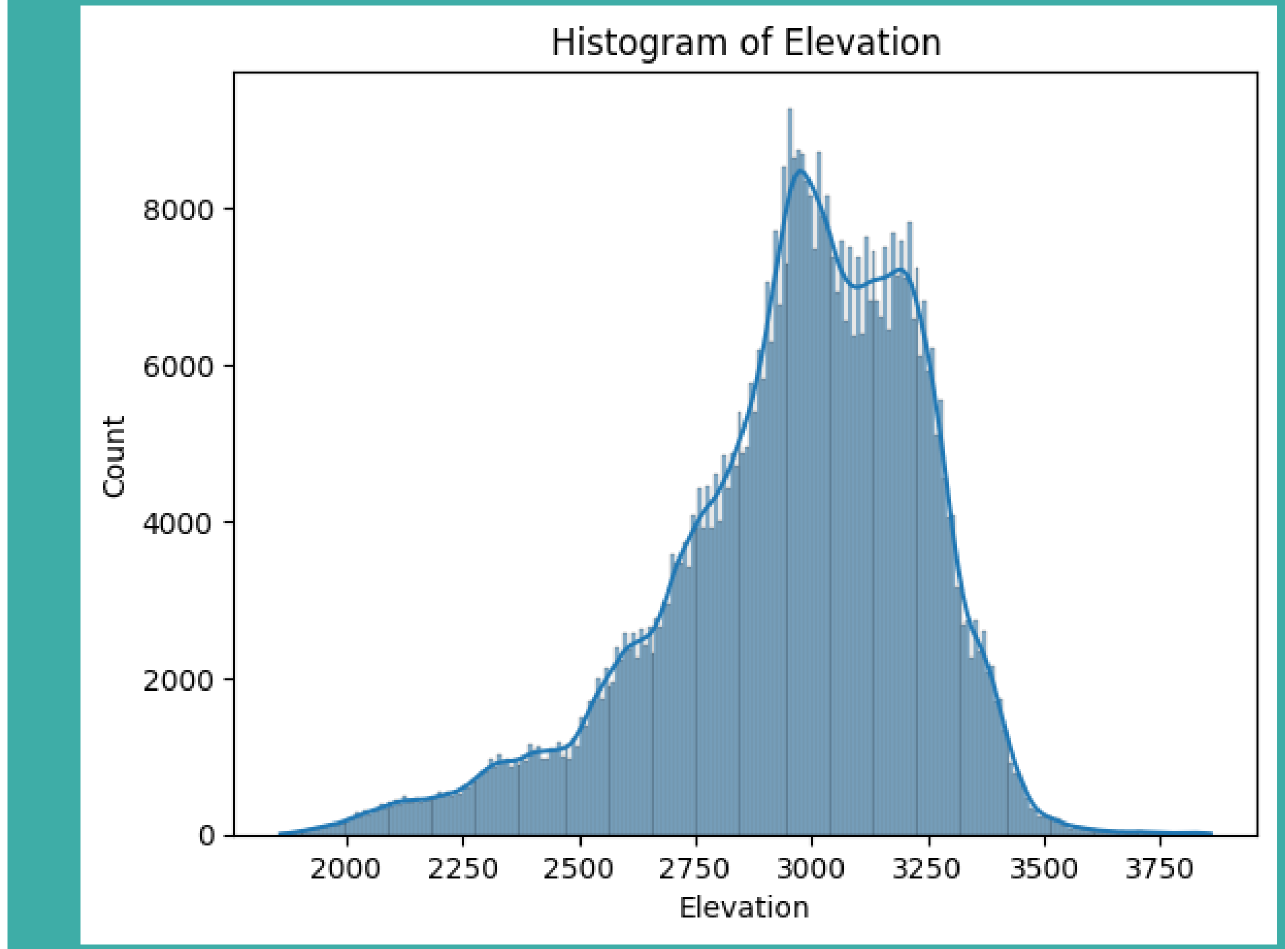


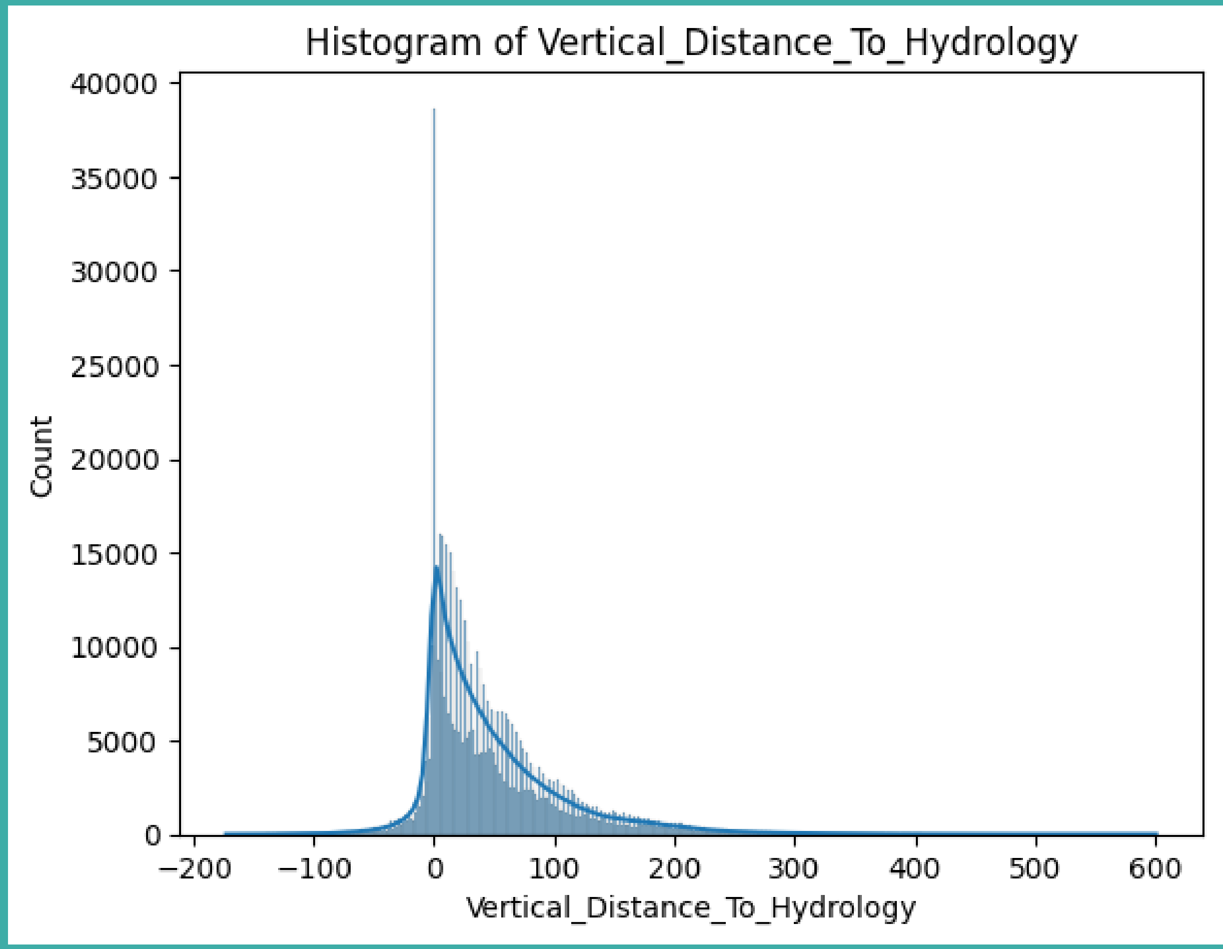

Histogram of Slope
Count
35000
30000
25000
20000
15000
10000
5000
0
0
10
20
30
40
50
60
Slope


Histogram of Horizontal_Distance_To_Hydrology
Count
35000
30000
25000
20000
15000
10000
5000
0
0
200
400
600
800
1000
1200
1400
Horizontal_Distance_To_Hydrology


Distribution of Trees(Cover_Types)
Count
250000
200000
150000
100000
50000
0
1
2
3
4
5
6
Cover_Type


Histogram of Aspect
Count
12000
10000
8000
6000
4000
2000
0
0
50
100
150
200
250
300
350
Aspect


Histogram of Elevation
Count
8000
6000
4000
2000
0
2000
2250
2500
2750
3000
3250
3500
3750
Elevation


Histogram of Vertical_Distance_To_Hydrology
Count
40000
35000
30000
25000
20000
15000
10000
5000
0
−200
−100
0
100
200
300
400
500
600
Vertical_Distance_To_Hydrology

WHY PREPROCESSING IS NEEDED/ CHALLENGES

- Missing data
- Outliers
- Data Scaling
- Categorical Data
- Data Skewness
- Dimensionality
- Class Imbalance
- Data Quality
- Normalization & Encoding
- Feature Engineering

METHODOLOGY

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

MAIN POINTS

The essential steps in this project:

1. Perform Data Pre-processing.
2. Conduct appropriate Hypothesis tests with validation experiments.
3. Analyze the importance of certain attributes.
4. Apply a machine learning classifier model.

PREPROCESSING STEPS

Data Pre-processing includes:

a. Addition of appropriate column names to the data

b. Data Cleaning (removal of null, nan values etc)

c. Handling Outliers

HYPOTHESIS TESTING

Tests Used (sample size=100):

1. t-Test
2. Chi-Squared Test

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

THE T-TEST

Validation Experiments Used:

1. Conduct the same hypothesis test on large number of samples (10000).
2. Validate through actual calculation of mean for samples of a particular attribute.

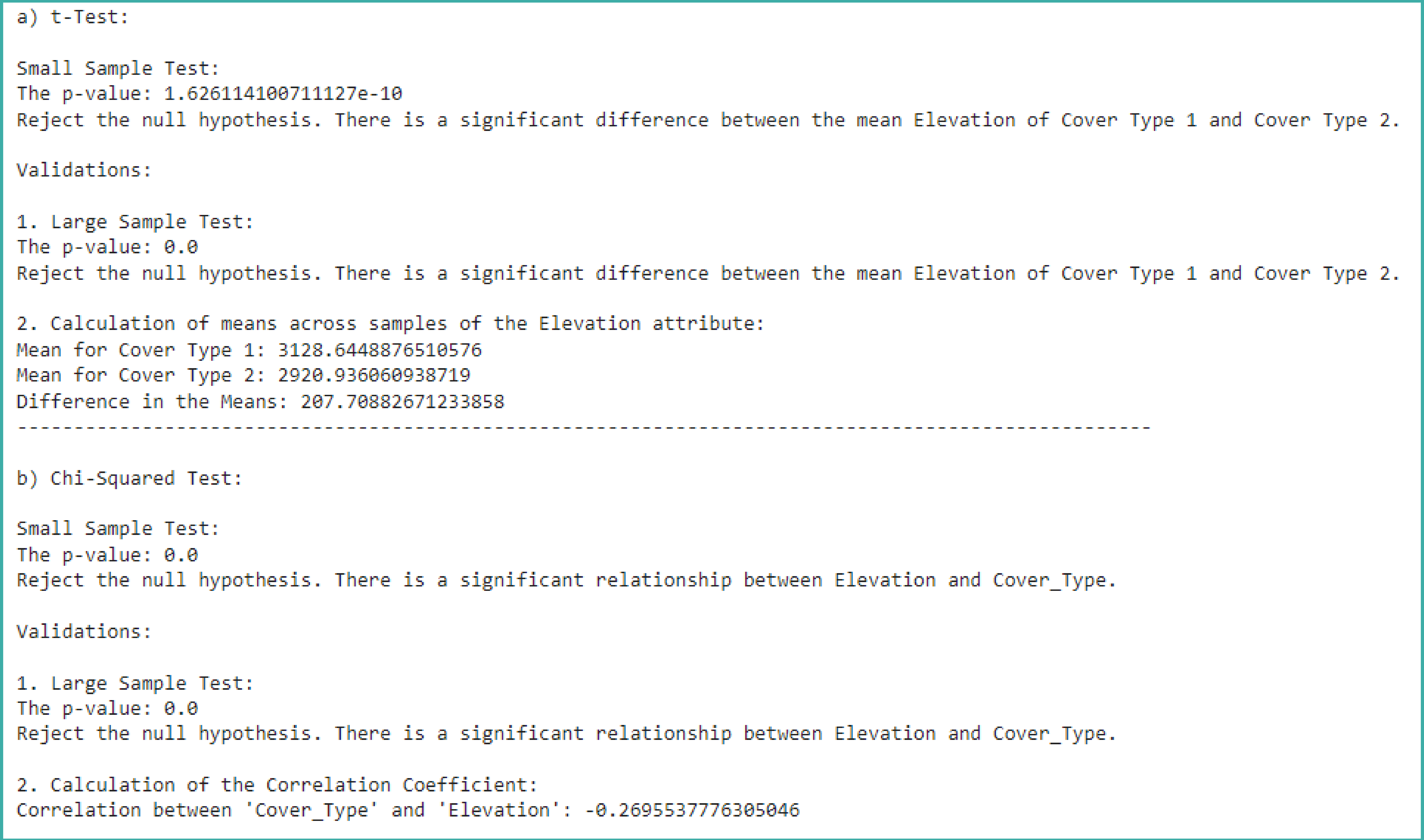
INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

THE CHI-SQUARED-TEST

Validation Experiments Used:

1. Conduct the same hypothesis test on large number of samples (10000).
2. Validate through actual calculation of Correlation between the selected attribute and the Cover_Type attribute.

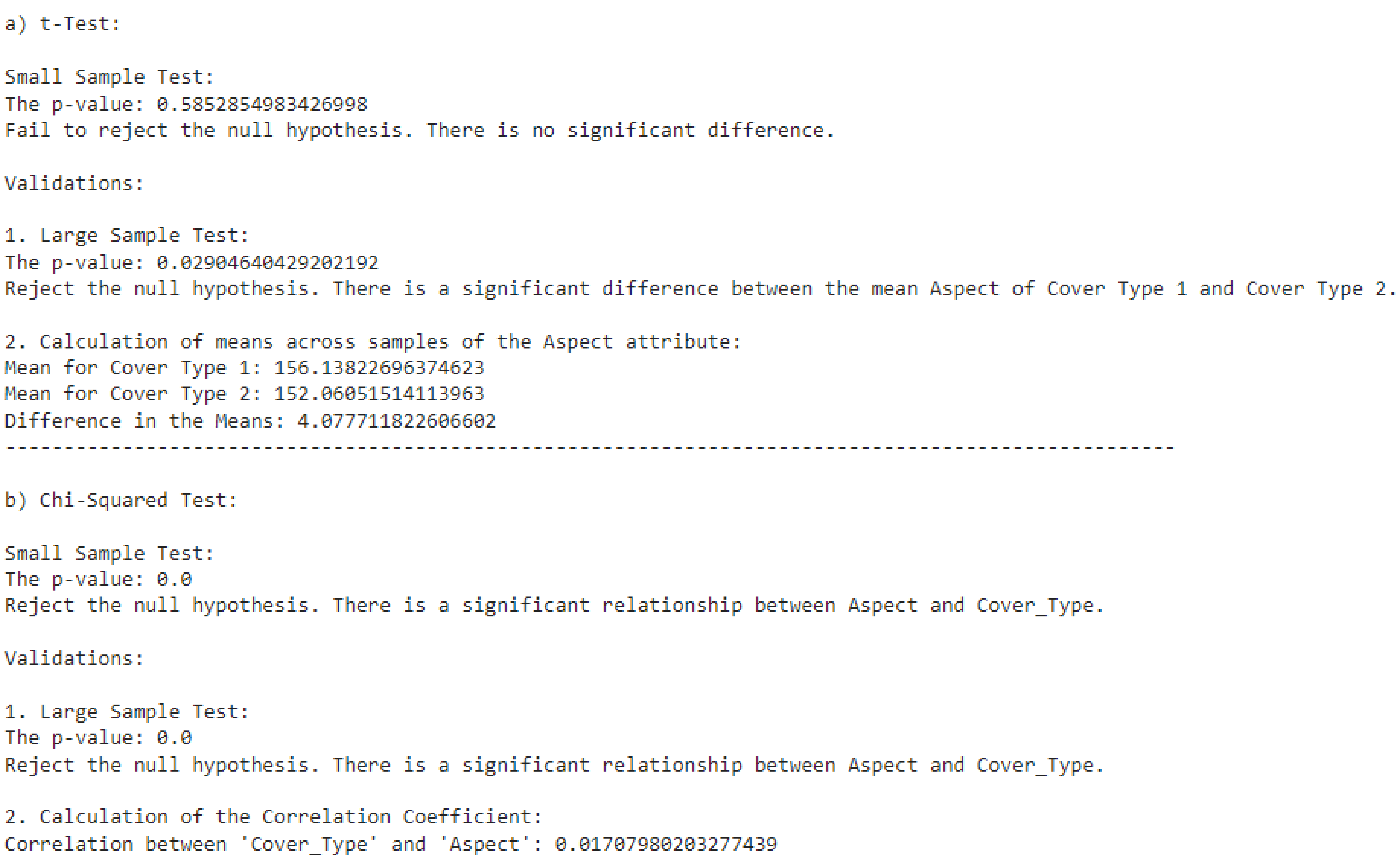
INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI



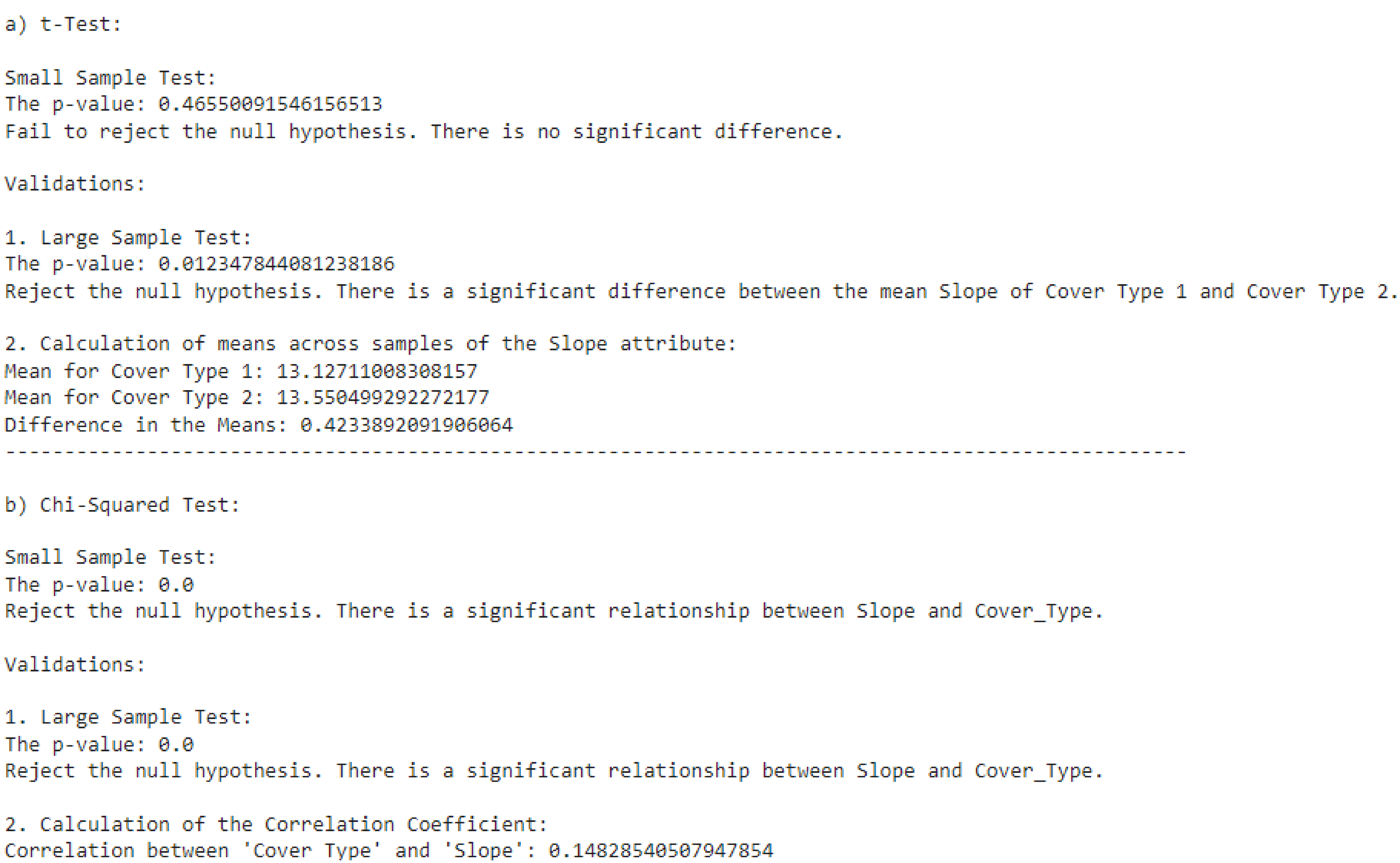
```
a) t-Test:

Small Sample Test:
The p-value: 1.626114100711127e-10
Reject the null hypothesis. There is a significant difference between the mean Elevation of Cover Type 1 and Cover Type 2.

Validations:

1. Large Sample Test:
The p-value: 0.0
Reject the null hypothesis. There is a significant difference between the mean Elevation of Cover Type 1 and Cover Type 2.

2. Calculation of means across samples of the Elevation attribute:
Mean for Cover Type 1: 3128.6448876510576
Mean for Cover Type 2: 2920.936060938719
Difference in the Means: 207.70882671233858
----------------------------------------------------------------------------------------------------

b) Chi-Squared Test:

Small Sample Test:
The p-value: 0.0
Reject the null hypothesis. There is a significant relationship between Elevation and Cover_Type.

Validations:

1. Large Sample Test:
The p-value: 0.0
Reject the null hypothesis. There is a significant relationship between Elevation and Cover_Type.

2. Calculation of the Correlation Coefficient:
Correlation between 'Cover_Type' and 'Elevation': -0.2695537776305046
```

DATA RE-ANALYSIS

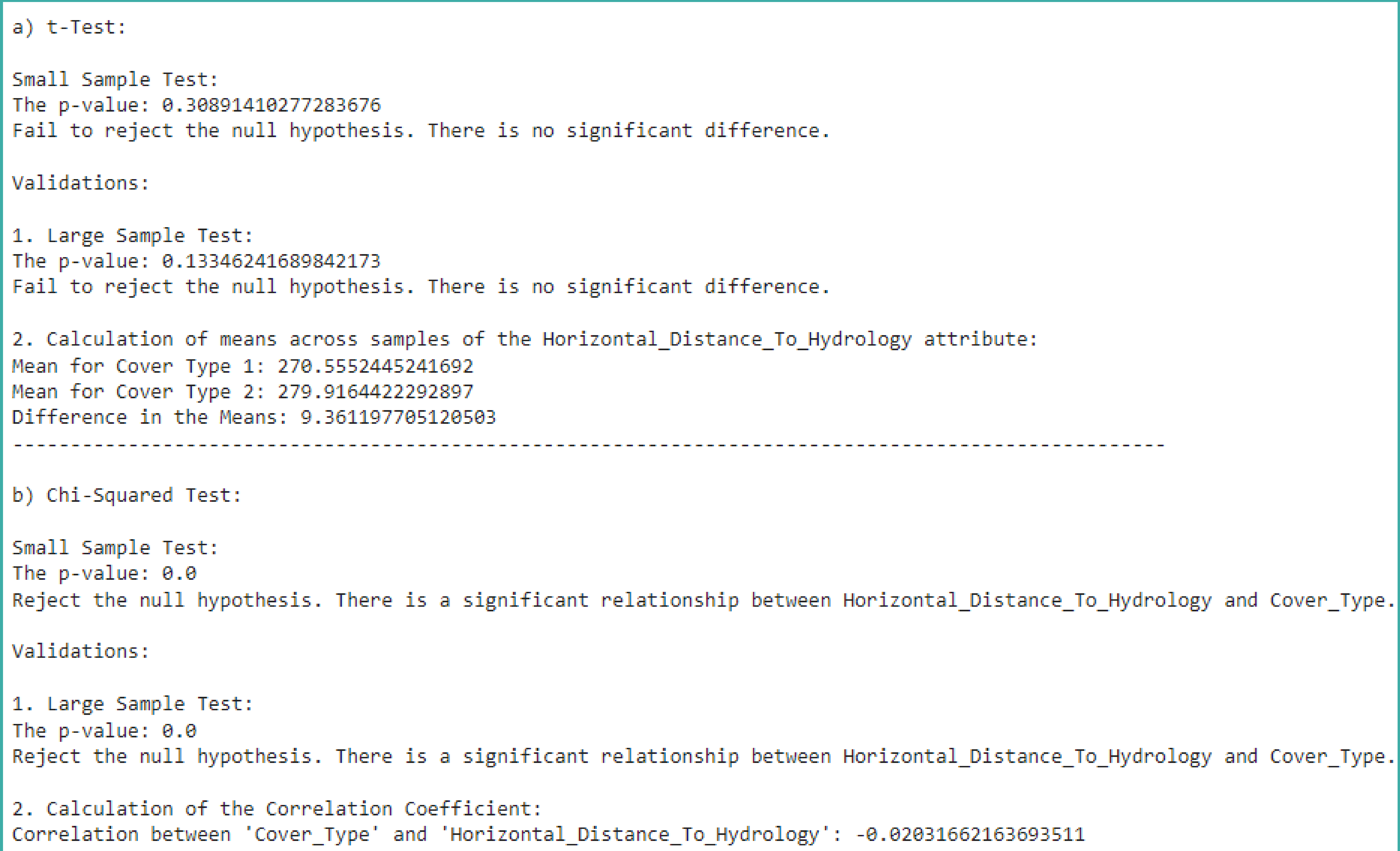
1. Conduct the same hypothesis tests for further attributes.
2. Determine for which attributes the null hypothesis is rejected, i.e, determine which attributes are imprtant for covertype prediction.



```
a) t-Test:

Small Sample Test:
The p-value: 0.5852854983426998
Fail to reject the null hypothesis. There is no significant difference.

Validations:

1. Large Sample Test:
The p-value: 0.02904640429202192
Reject the null hypothesis. There is a significant difference between the mean Aspect of Cover Type 1 and Cover Type 2.

2. Calculation of means across samples of the Aspect attribute:
Mean for Cover Type 1: 156.13822696374623
Mean for Cover Type 2: 152.06051514113963
Difference in the Means: 4.077711822606602
----------------------------------------------------------------------------------------------------

b) Chi-Squared Test:

Small Sample Test:
The p-value: 0.0
Reject the null hypothesis. There is a significant relationship between Aspect and Cover_Type.

Validations:

1. Large Sample Test:
The p-value: 0.0
Reject the null hypothesis. There is a significant relationship between Aspect and Cover_Type.

2. Calculation of the Correlation Coefficient:
Correlation between 'Cover_Type' and 'Aspect': 0.01707980203277439
```



```
a) t-Test:

Small Sample Test:
The p-value: 0.46550091546156513
Fail to reject the null hypothesis. There is no significant difference.

Validations:

1. Large Sample Test:
The p-value: 0.012347844081238186
Reject the null hypothesis. There is a significant difference between the mean Slope of Cover Type 1 and Cover Type 2.

2. Calculation of means across samples of the Slope attribute:
Mean for Cover Type 1: 13.12711008308157
Mean for Cover Type 2: 13.550499292272177
Difference in the Means: 0.4233892091906064
------------------------------------------------------------------------------------------------------------------------


b) Chi-Squared Test:

Small Sample Test:
The p-value: 0.0
Reject the null hypothesis. There is a significant relationship between Slope and Cover_Type.

Validations:

1. Large Sample Test:
The p-value: 0.0
Reject the null hypothesis. There is a significant relationship between Slope and Cover_Type.

2. Calculation of the Correlation Coefficient:
Correlation between 'Cover Type' and 'Slope': 0.14828540507947854
```



```
a) t-Test:

Small Sample Test:
The p-value: 0.30891410277283676
Fail to reject the null hypothesis. There is no significant difference.

Validations:

1. Large Sample Test:
The p-value: 0.13346241689842173
Fail to reject the null hypothesis. There is no significant difference.

2. Calculation of means across samples of the Horizontal_Distance_To_Hydrology attribute:
Mean for Cover Type 1: 270.5552445241692
Mean for Cover Type 2: 279.9164422292897
Difference in the Means: 9.361197705120503
-----------------------------------------------------------------------------------------------------

b) Chi-Squared Test:

Small Sample Test:
The p-value: 0.0
Reject the null hypothesis. There is a significant relationship between Horizontal_Distance_To_Hydrology and Cover_Type.

Validations:

1. Large Sample Test:
The p-value: 0.0
Reject the null hypothesis. There is a significant relationship between Horizontal_Distance_To_Hydrology and Cover_Type.

2. Calculation of the Correlation Coefficient:
Correlation between 'Cover_Type' and 'Horizontal_Distance_To_Hydrology': -0.02031662163693511
```

CONCLUSIONS FROM TESTS

1. Almost all of the attributes are relevant in determination of the cover type label.

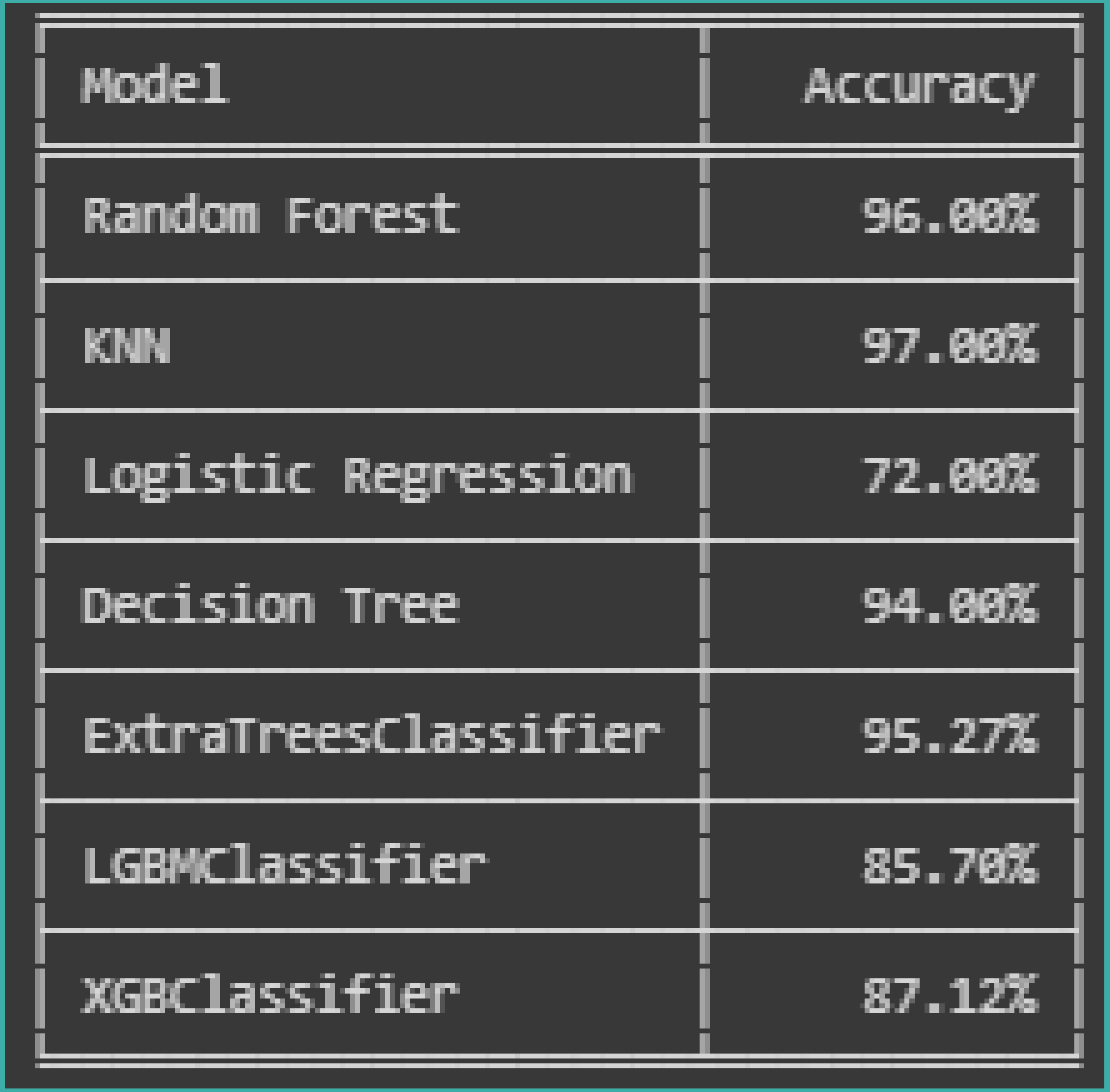
2. Two to three attributes showed contradicting results (different for both t-tests and chi-squared tests), suggesting a possible lack of evidence, or insufficient data.

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

ML CLASSIFIERS

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

CLASSIFIERS

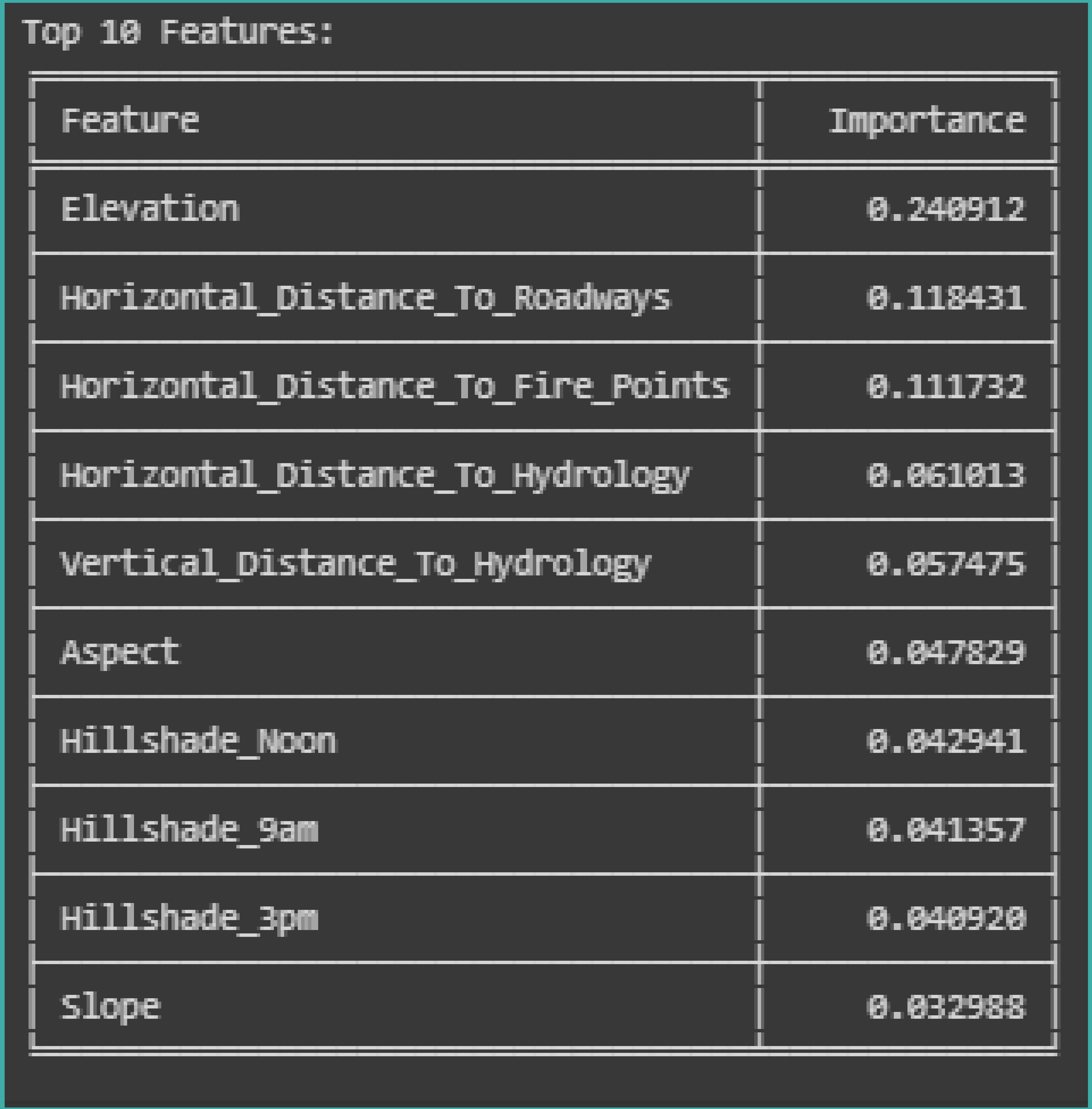


Model	Accuracy
Random Forest	96.00%
KNN	97.00%
Logistic Regression	72.00%
Decision Tree	94.00%
ExtraTreesClassifier	95.27%
LGBMClassifier	85.70%
XGBClassifier	87.12%

RESULTS

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

IMPORTANT FEATURES



```
Top 10 Features:
╒════════════════════════════════════╤══════════════╕
│ Feature                            │   Importance │
╞════════════════════════════════════╪══════════════╡
│ Elevation                          │     0.240912 │
├────────────────────────────────────┼──────────────┤
│ Horizontal_Distance_To_Roadways    │     0.118431 │
├────────────────────────────────────┼──────────────┤
│ Horizontal_Distance_To_Fire_Points │     0.111732 │
├────────────────────────────────────┼──────────────┤
│ Horizontal_Distance_To_Hydrology   │     0.061013 │
├────────────────────────────────────┼──────────────┤
│ Vertical_Distance_To_Hydrology     │     0.057475 │
├────────────────────────────────────┼──────────────┤
│ Aspect                             │     0.047829 │
├────────────────────────────────────┼──────────────┤
│ Hillshade_Noon                     │     0.042941 │
├────────────────────────────────────┼──────────────┤
│ Hillshade_9am                      │     0.041357 │
├────────────────────────────────────┼──────────────┤
│ Hillshade_3pm                      │     0.040920 │
├────────────────────────────────────┼──────────────┤
│ Slope                              │     0.032988 │
╘════════════════════════════════════╧══════════════╛
```

CONCLUSION

- **Data Preprocessing:**
  - one-hot encoding, outliers, select imp features
- **Models Implemented:**
  - Random Forest, KNN, Logistic regression, Decision Tree, Extra Trees, LGBM, and XGBoost.
- **Model Performance:**
  - KNN (accuracy 97%), Random Forest(96%)
- **Feature Importance:**
  - Elevation and Horizontal distance to roadways and fire points

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

CONCLUSION

- **Insights for Conservation:**
  - Understanding influential features aids conservation
  - Provides insights into factors affecting different forest cover types.
- **Areas for Improvement:**
  - faster training model for such a large dataset
  - explore feature engineering and ensemble methods.