Landslide Prediction Using Different Machine Learning Models

Wasifa Tasnim Mrittika
ID: 21-45108-2
Dept Name: CSE
Institute Name: AIUB
Dhala, Bangladesh.
21-45108-2@student.aiub.edu

MD Tahsin Tasnim Aurin
ID: 21-45213-2
Dept Name: CSE
Institute Name: AIUB
Dhala, Bangladesh.
21-45213-2@student.aiub.edu

MD.Ibrahim Khalil Lajim
ID: 21-45123-2
Dept Name: CSE
Institute Name: AIUB
Dhala, Bangladesh.
21-45123-2@student.aiub.edu

MD Minhazur Rahman
ID: 21-45127-2
Dept Name: CSE
Institute Name: AIUB
Dhala, Bangladesh.
21-45127-2@student.aiub.edu

Abstract— This project focuses on landslide prediction using machine learning models such as Random Forest, Decision Trees, Support Vector Machines (SVM), and Gradient Boosting Classifier. We employ performance metrics like confusion matrix, precision, recall, accuracy, F1 curves, and ROC-AUC curves to evaluate the models. The dataset comprises 18 columns, with 17 features and 1 target column, all containing numerical data. The model with the highest accuracy among the tested algorithms will be selected for landslide prediction.

I. INTRODUCTION

Landslides pose a significant threat to communities and infrastructure, necessitating effective prediction models for early warning systems. This project explores the application of machine learning techniques to predict landslides, aiming to enhance our ability to anticipate and mitigate the impact of these natural disasters.

II. MOTIVATION OF THE PROJECT

The motivation behind undertaking this project lies in the imperative need for accurate landslide prediction. By leveraging machine learning models, we aim to provide a tool that not only aids in safeguarding human lives and infrastructure but also contributes to a better understanding of the factors influencing landslides. The societal benefits of an effective landslide prediction system include improved disaster preparedness and response.

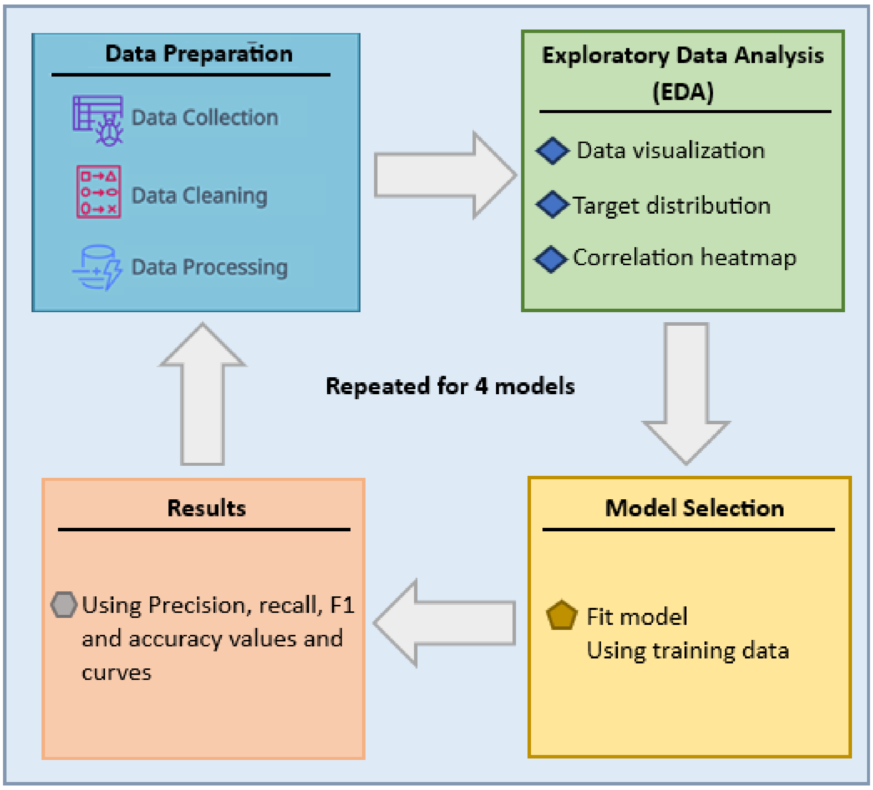
III. OBJECTIVE OF THE PROJECT

The primary objective of this project is to develop a reliable landslide prediction model using machine learning algorithms. Specifically, we aim to:

- Evaluate the performance of four machine learning models: Random Forest, Decision Trees, SVM, and Gradient Boosting Classifier.
- Utilize various performance metrics such as confusion matrix, precision, recall, accuracy, F1 curves, and ROC-AUC curves for model evaluation.
- Identify the model with the highest accuracy among the tested algorithms to serve as the optimal landslide prediction tool.

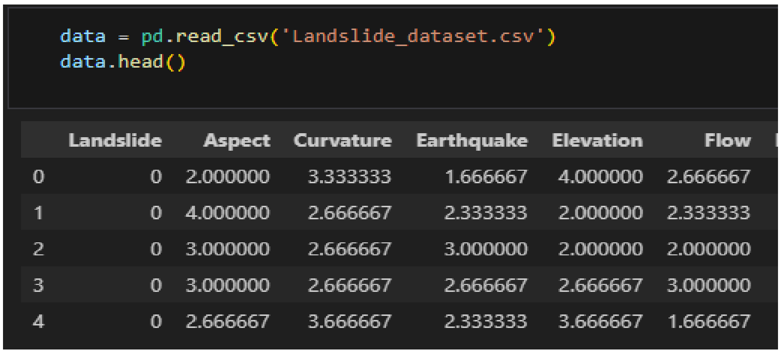
IV. METHODOLOGY

The methodology involves a systematic approach to landslide prediction using machine learning models. In the initial phase of our project, we commenced importing the necessary libraries to facilitate our analysis. Subsequently, we engaged in meticulous data preparation, encompassing data collection, cleaning, and processing to ensure the dataset's readiness for modeling. Following this, we delved into Exploratory Data Analysis (EDA), employing techniques such as visualizing the distribution of the target variable and constructing a correlation heatmap to gain insights into the dataset's structure and relationships. The subsequent step involved model selection, where we opted for four distinct machine learning models: Random Forest, Decision Trees, Support Vector Machines (SVM), and Gradient Boosting Classifier. To understand the significance of each feature, we performed feature importance analysis using each model. Subsequently, we evaluated the models' performance by calculating crucial metrics like confusion matrix, precision, recall, accuracy, and F1 values. Visual representations, such as curves plotting actual versus predicted results, were generated for each metric. Finally, we quantified the Mean Squared Error (MSE) to provide a comprehensive assessment of the models' predictive capabilities. This holistic process ensures a thorough understanding of the dataset, model selection, and comprehensive evaluation of predictive performance.



A. *Data Collection*

For every model, the dataset Landslide_dataset.csv, was loaded using the Pandas library. This dataset serves as the foundation for training and evaluating machine learning models for landslide prediction. This dataset was found in the Kaggle website.

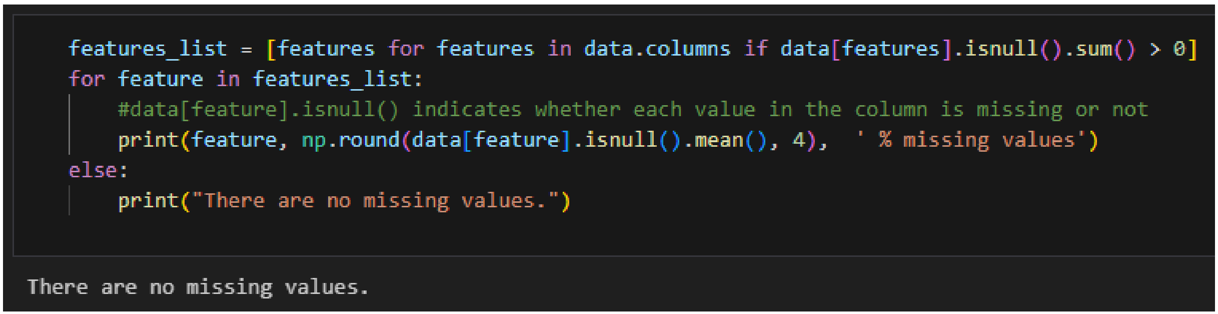


```
data = pd.read_csv('Landslide_dataset.csv')
data.head()
```

```
   Landslide    Aspect  Curvature  Earthquake  Elevation      Flow
0          0  2.000000   3.333333    1.666667   4.000000  2.666667
1          0  4.000000   2.666667    2.333333   2.000000  2.333333
2          0  3.000000   2.666667    3.000000   2.000000  2.000000
3          0  3.000000   2.666667    2.666667   2.666667  3.000000
4          0  2.666667   3.666667    2.333333   3.666667  1.666667
```

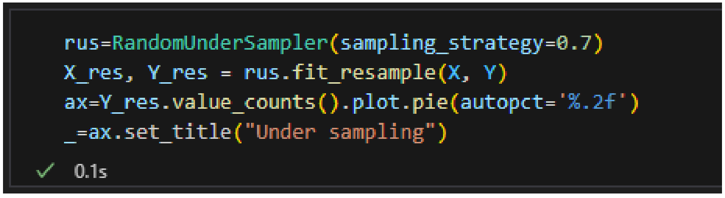
B. *Data processing*

Data preparation involved identifying and handling missing values. The absence of missing values in any feature was confirmed, ensuring a robust dataset for subsequent analysis. Then feature and target variables were separated and separated the dataset into training and testing set. There were no missing values found in the dataset. However, the dataset was imbalanced having more values of 0 than 1 so under sampling was done to fix this problem.



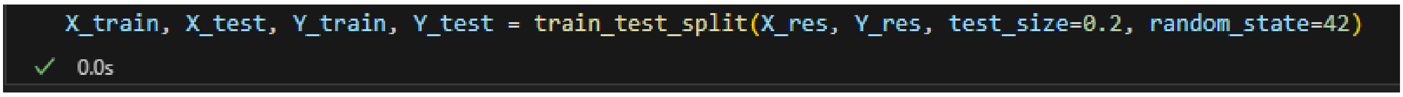
```
features_list = [features for features in data.columns if data[features].isnull().sum() > 0]
for feature in features_list:
    #data[feature].isnull() indicates whether each value in the column is missing or not
    print(feature, np.round(data[feature].isnull().mean(), 4),  ' % missing values')
else:
    print("There are no missing values.")
```

```
There are no missing values.
```



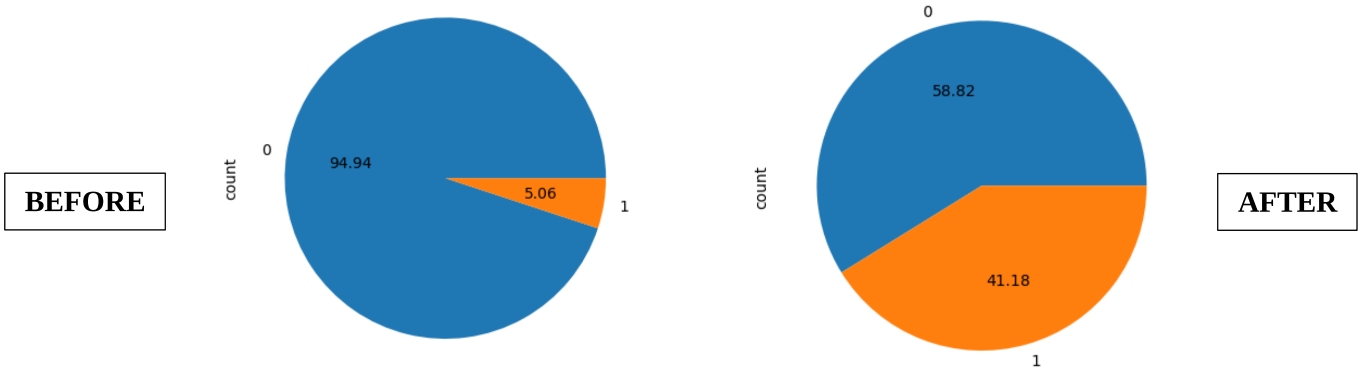
```
rus=RandomUnderSampler(sampling_strategy=0.7)
X_res, Y_res = rus.fit_resample(X, Y)
ax=Y_res.value_counts().plot.pie(autopct='%.2f')
_=ax.set_title("Under sampling")
```

✓ 0.1s



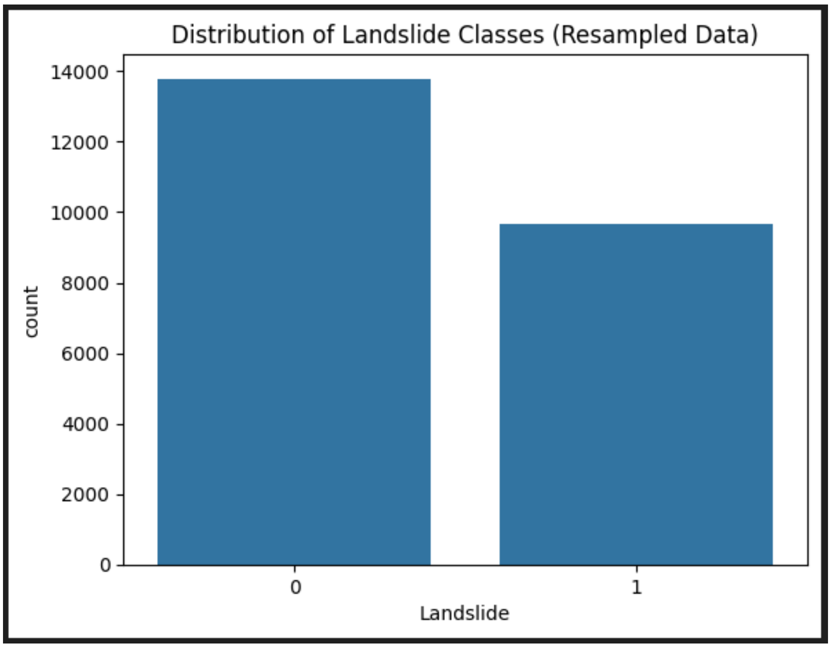
```
X_train, X_test, Y_train, Y_test = train_test_split(X_res, Y_res, test_size=0.2, random_state=42)
```

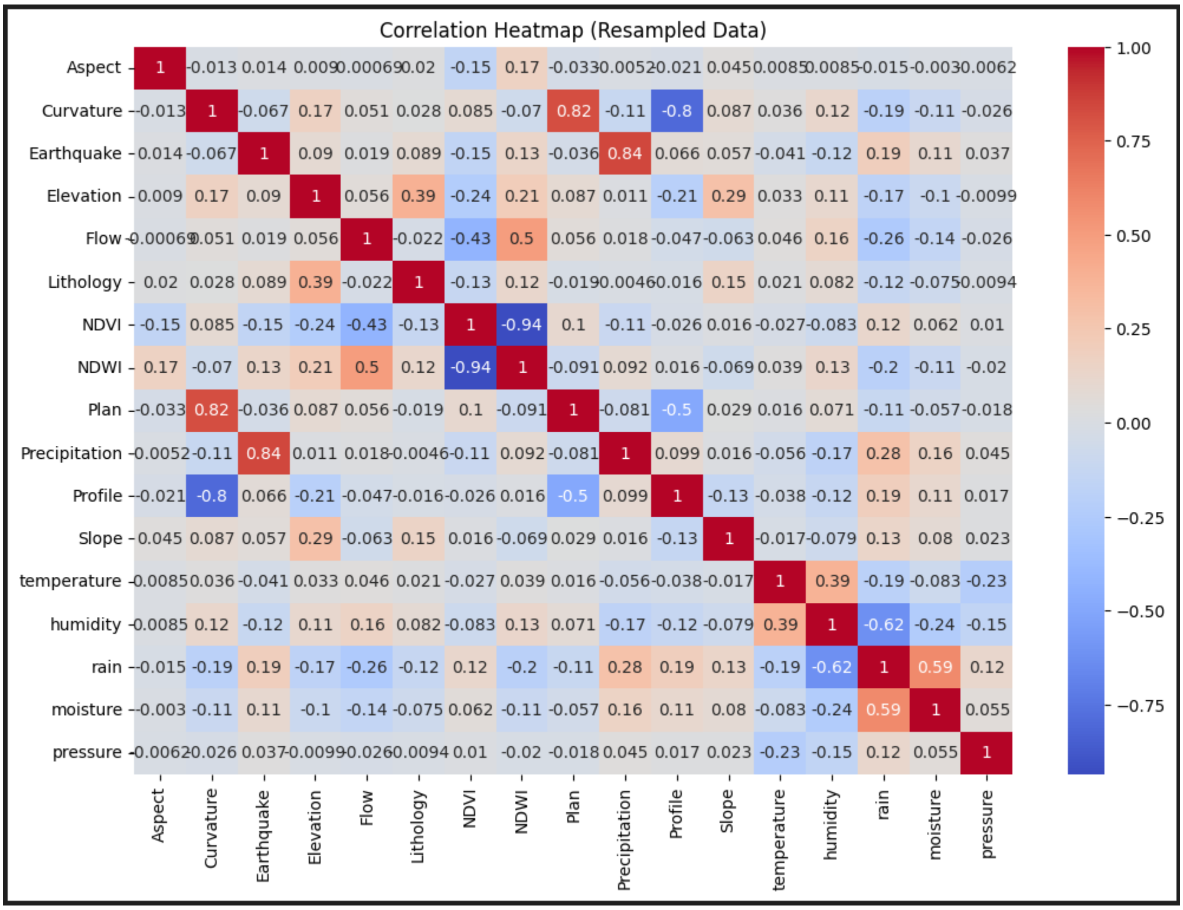
✓ 0.0s



C. *Dataset description*

The dataset comprises both features and a target variable. Features include various numerical attributes related to landslide prediction, while the target variable, labeled 'Landslide,' represents the class to be predicted. There are a total of 17 features in the dataset named - Aspect, Curvature, Earthquake, Elevation, Flow, Lithology, NDVI, NDWI, Plan, Precipitation, Profile, Slope, temperature, humidity, rain, moisture, and pressure. Exploratory Data Analysis (EDA) methods were employed to visualize the distribution of landslide classes and generate a correlation heatmap to uncover relationships within the dataset.



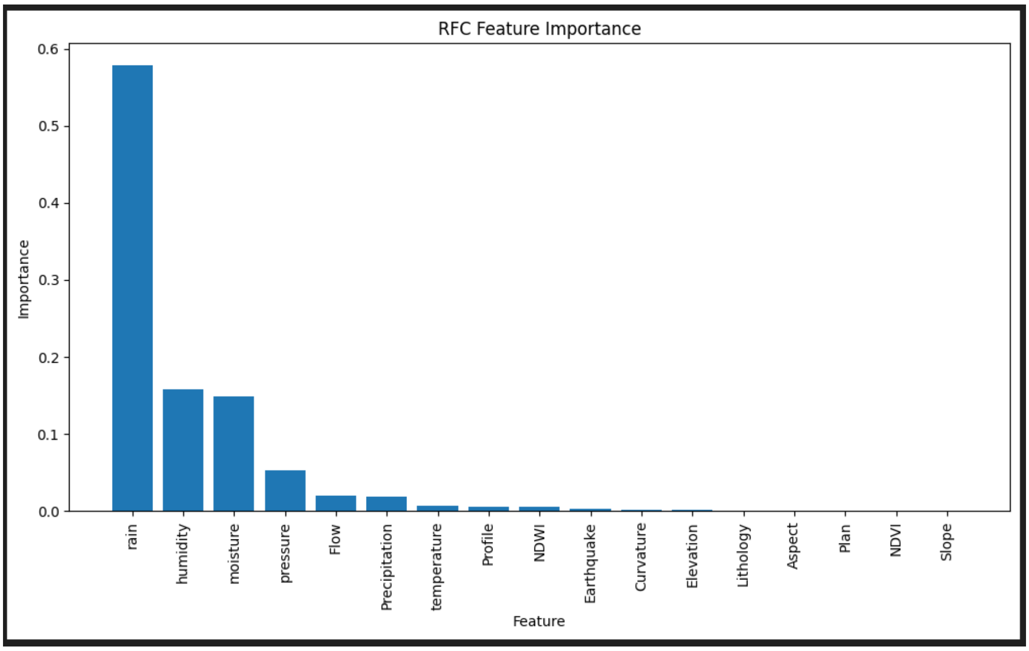


	Aspect	Curvature	Earthquake	Elevation	Flow	Lithology	NDVI	NDWI	Plan	Precipitation	Profile	Slope	temperature	humidity	rain	moisture	pressure
Aspect	1	-0.013	0.014	0.009	0.00069	0.02	-0.15	0.17	-0.033	0.0052	-0.021	0.045	0.0085	0.0085	-0.015	-0.003	0.0062
Curvature	-0.013	1	-0.067	0.17	0.051	0.028	0.085	-0.07	0.82	-0.11	-0.8	0.087	0.036	0.12	-0.19	-0.11	-0.026
Earthquake	0.014	-0.067	1	0.09	0.019	0.089	-0.15	0.13	-0.036	0.84	0.066	0.057	-0.041	-0.12	0.19	0.11	0.037
Elevation	0.009	0.17	0.09	1	0.056	0.39	-0.24	0.21	0.087	0.011	-0.21	0.29	0.033	0.11	-0.17	-0.1	-0.0099
Flow	0.00069	0.051	0.019	0.056	1	-0.022	-0.43	0.5	0.056	0.018	-0.047	-0.063	0.046	0.16	-0.26	-0.14	-0.026
Lithology	0.02	0.028	0.089	0.39	-0.022	1	-0.13	0.12	-0.019	-0.0046	-0.016	0.15	0.021	0.082	-0.12	-0.075	0.0094
NDVI	-0.15	0.085	-0.15	-0.24	-0.43	-0.13	1	-0.94	0.1	-0.11	-0.026	0.016	-0.027	-0.083	0.12	0.062	0.01
NDWI	0.17	-0.07	0.13	0.21	0.5	0.12	-0.94	1	-0.091	0.092	0.016	-0.069	0.039	0.13	-0.2	-0.11	-0.02
Plan	-0.033	0.82	-0.036	0.087	0.056	-0.019	0.1	-0.091	1	-0.081	-0.5	0.029	0.016	0.071	-0.11	-0.057	-0.018
Precipitation	0.0052	-0.11	0.84	0.011	0.018	-0.0046	-0.11	0.092	-0.081	1	0.099	0.016	-0.056	-0.17	0.28	0.16	0.045
Profile	-0.021	-0.8	0.066	-0.21	-0.047	-0.016	-0.026	0.016	-0.5	0.099	1	-0.13	-0.038	-0.12	0.19	0.11	0.017
Slope	0.045	0.087	0.057	0.29	-0.063	0.15	0.016	-0.069	0.029	0.016	-0.13	1	-0.017	-0.079	0.13	0.08	0.023
temperature	0.0085	0.036	-0.041	0.033	0.046	0.021	-0.027	0.039	0.016	-0.056	-0.038	-0.017	1	0.39	-0.19	-0.083	-0.23
humidity	0.0085	0.12	-0.12	0.11	0.16	0.082	-0.083	0.13	0.071	-0.17	-0.12	-0.079	0.39	1	-0.62	-0.24	-0.15
rain	-0.015	-0.19	0.19	-0.17	-0.26	-0.12	0.12	-0.2	-0.11	0.28	0.19	0.13	-0.19	-0.62	1	0.59	0.12
moisture	-0.003	-0.11	0.11	-0.1	-0.14	-0.075	0.062	-0.11	-0.057	0.16	0.11	0.08	-0.083	-0.24	0.59	1	0.055
pressure	0.0062	-0.026	0.037	-0.0099	-0.026	0.0094	0.01	-0.02	-0.018	0.045	0.017	0.023	-0.23	-0.15	0.12	0.055	1

D. ***Machine Learning model development and evaluation***

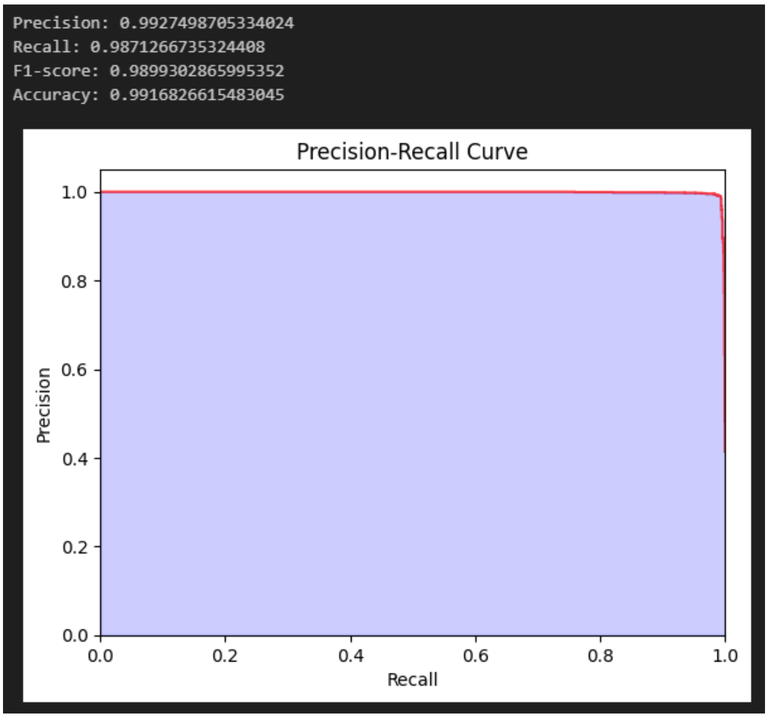
1. <u>Random Forest Classifier:</u>

- A Random Forest Classifier was chosen for its versatility and ability to handle complex relationships in the data. It can provide insights into feature importance, which can be valuable for understanding the factors contributing to landslides.
- Hyperparameters were tuned with values such as criterion='entropy,' max_depth=3, max_features='sqrt,' and n_estimators=100.
- The model was trained on the training set (80% of the data), and predictions were made on the test set (20% of the data).
- Feature importance was analyzed, revealing insights into the significance of each feature for landslide prediction.

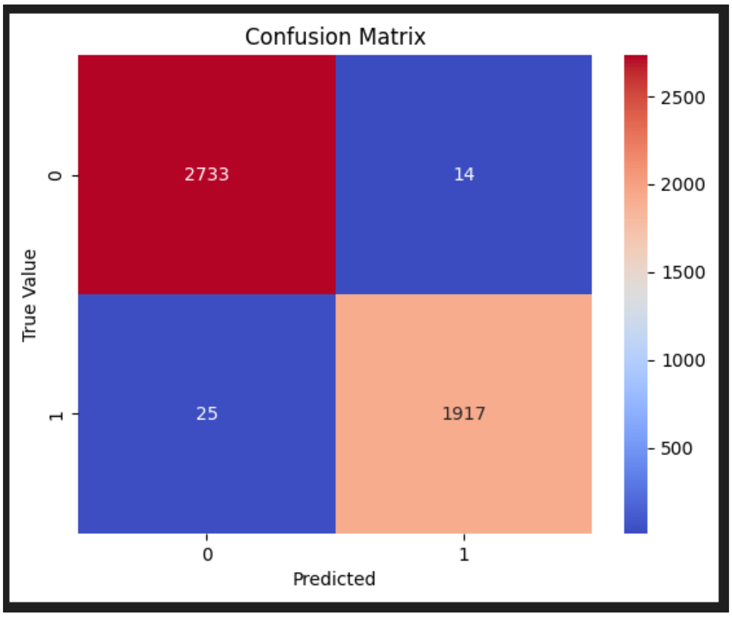


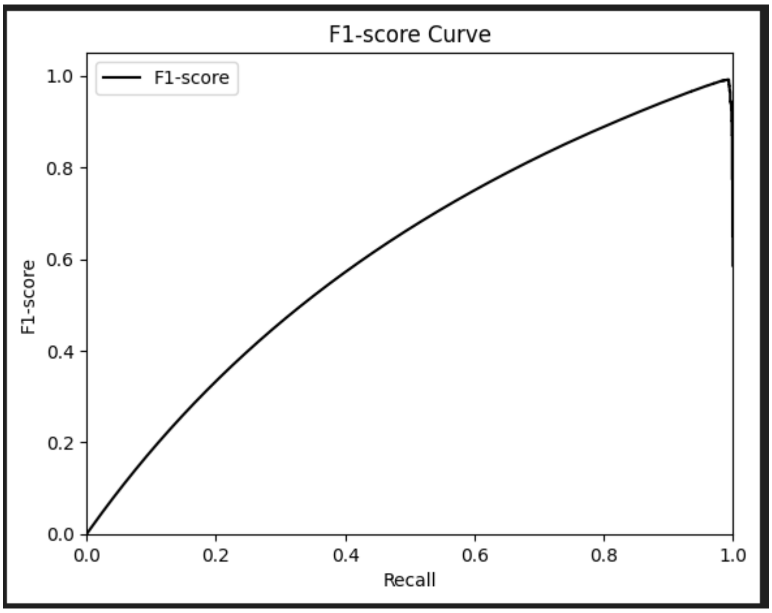
Model Evaluation:

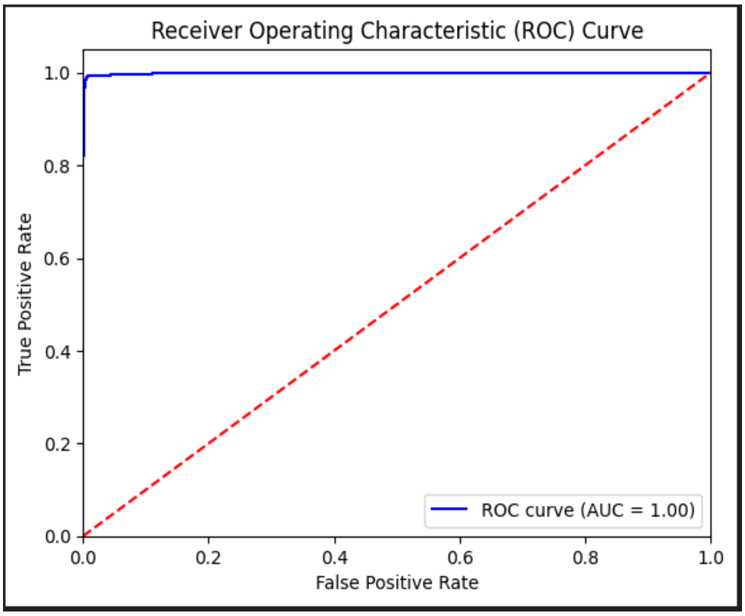
- Key metrics including precision, recall, F1-score, and accuracy were calculated for the test set.
- Precision-Recall and F1-score curves were plotted to visualize the trade-off between precision and recall.
- A confusion matrix provided a detailed breakdown of the model's performance.
- The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were employed to assess the model's ability to discriminate between classes.



```
Precision: 0.9927498705334024
Recall: 0.9871266735324408
F1-score: 0.9899302865995352
Accuracy: 0.9916826615483045
```

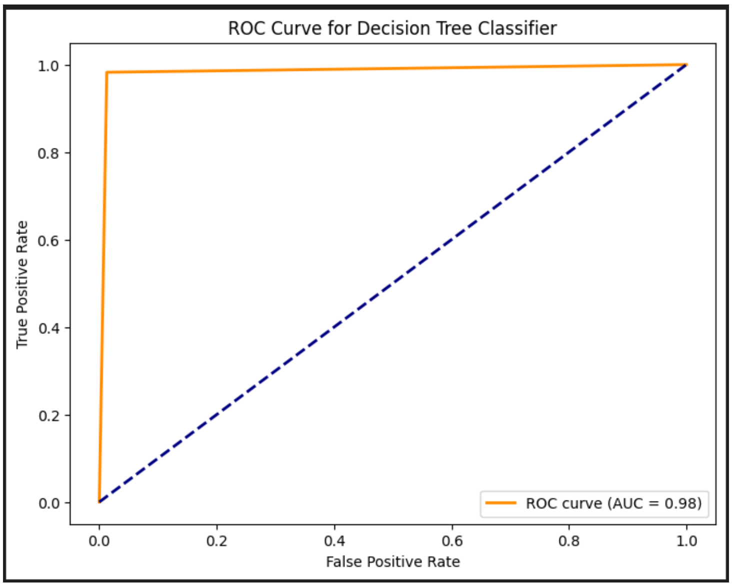


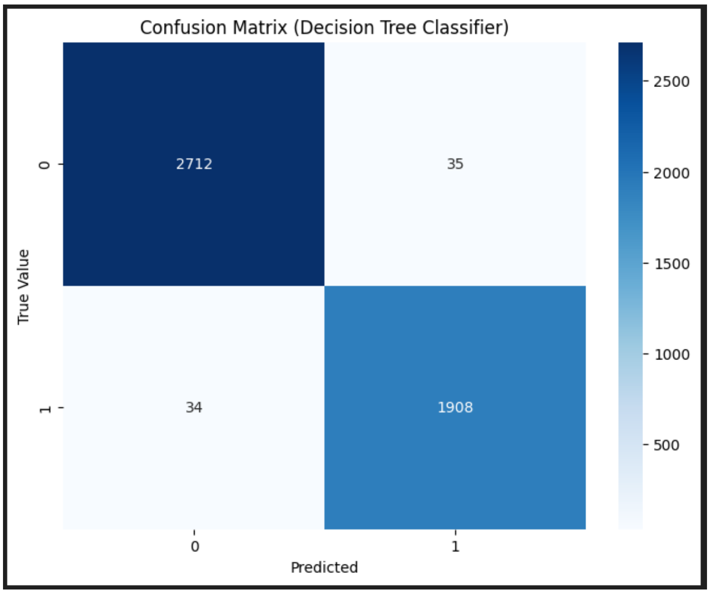




2. **Decision Tree Classifier:**

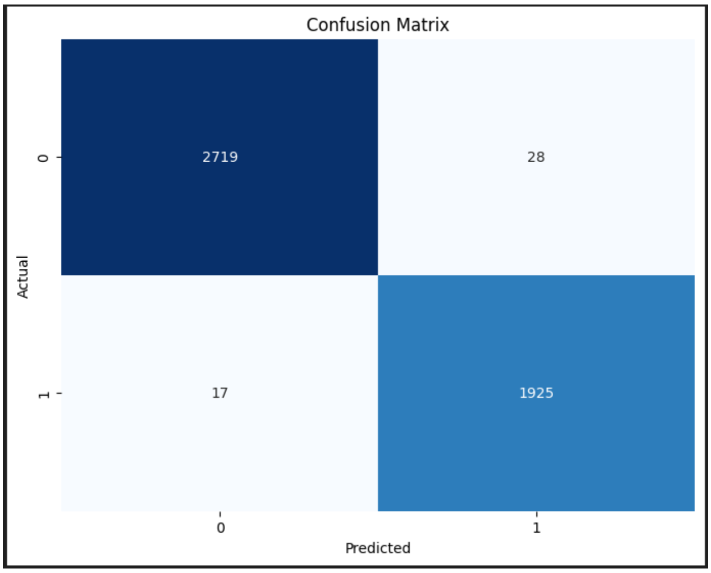
- A Decision Tree Classifier was chosen for its interpretability and since it can capture non-linear relationships in the data, which might be present in a dataset with numerical features.
- The model was trained on 80% of the data and evaluated on the remaining 20%.
- Evaluation metrics, such as precision, recall, F1-score, and accuracy, were calculated for the test set.
- Precision-Recall and F1-score curves provided insights into the trade-off between precision and recall.
- A confusion matrix detailed the model's performance in predicting landslide classes.
- The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) assessed the model's discrimination capabilities.

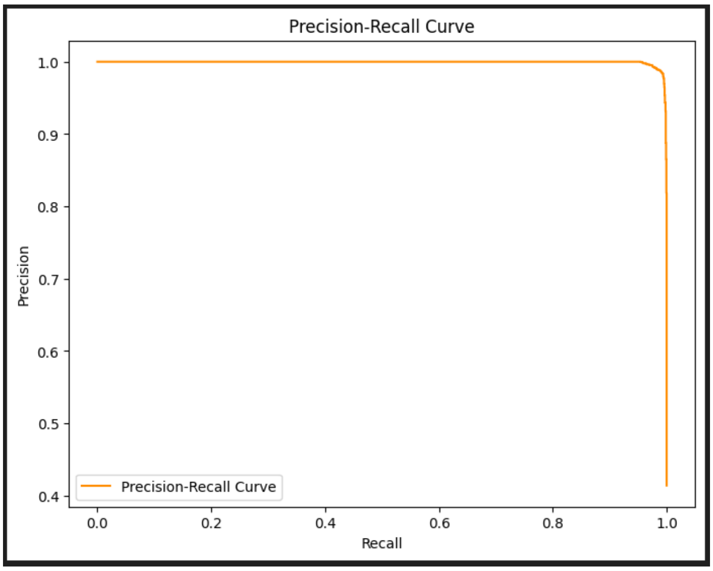


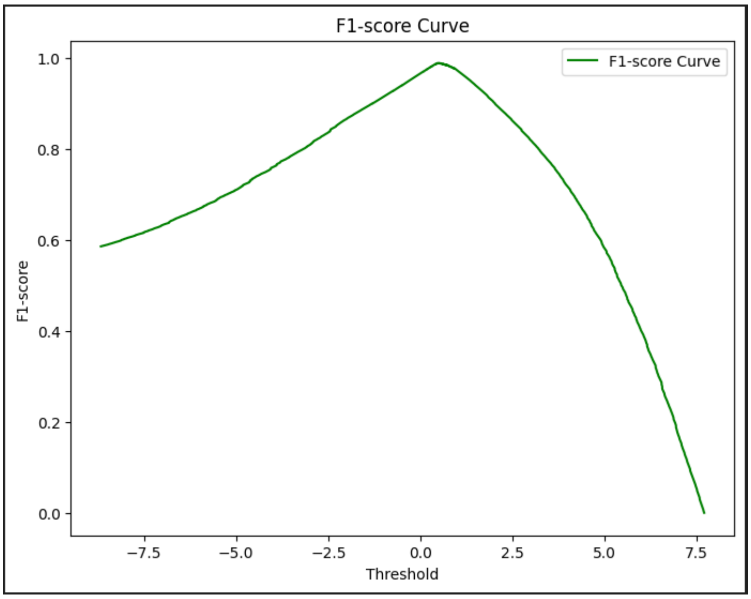


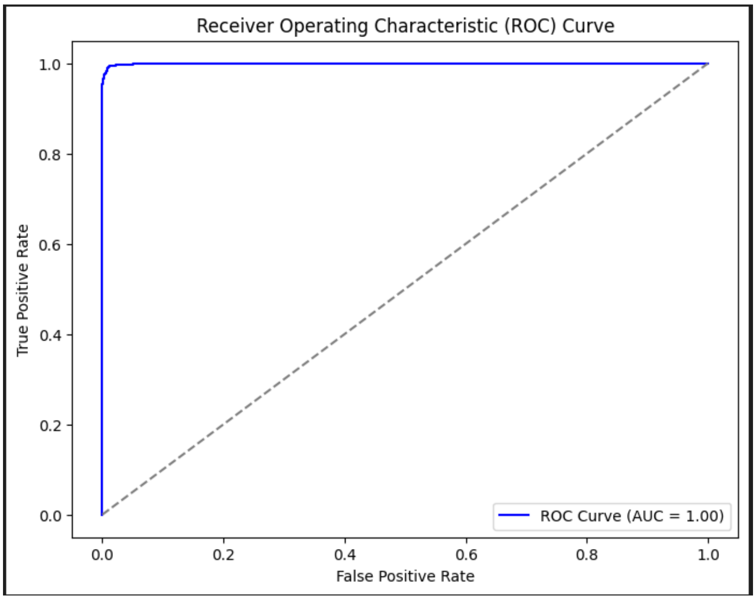
3. **Gradient Boosting Classifier:**

- A Gradient Boosting Classifier (GBC) was selected for its ability to handle complex relationships in the data and improve predictive performance.
- The model was trained on under sampled data to address class imbalance and evaluated for accuracy.
- Key metrics, including precision, recall, F1-score, and accuracy, were computed to gauge model performance.
- Precision-Recall and F1-score curves illuminated the precision-recall trade-off.
- The confusion matrix provided a detailed breakdown of the model's predictive outcomes.
- The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) assessed the model's discrimination capabilities.



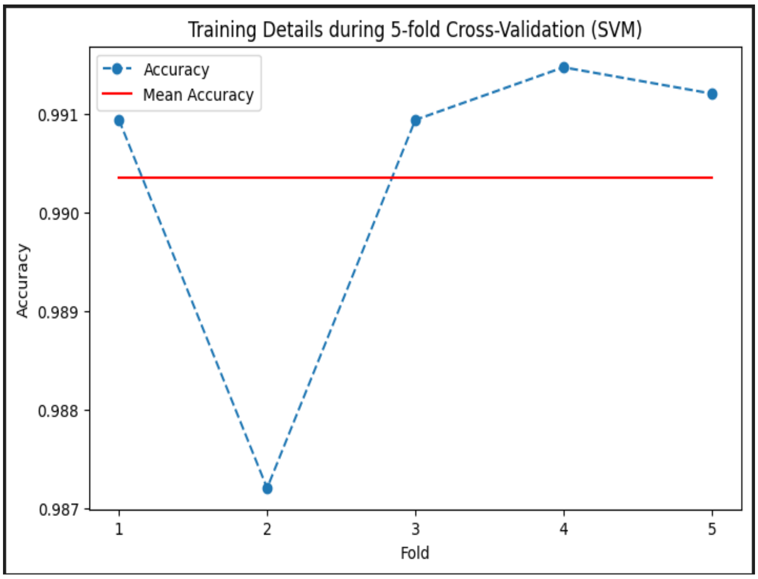


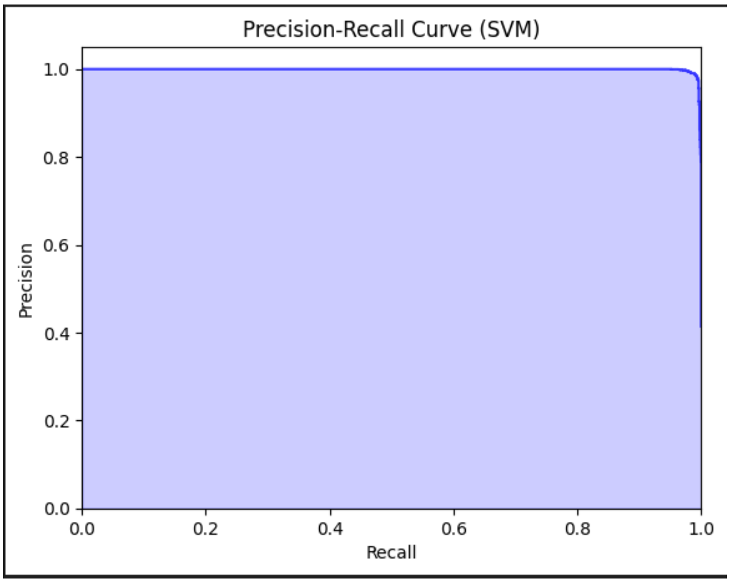


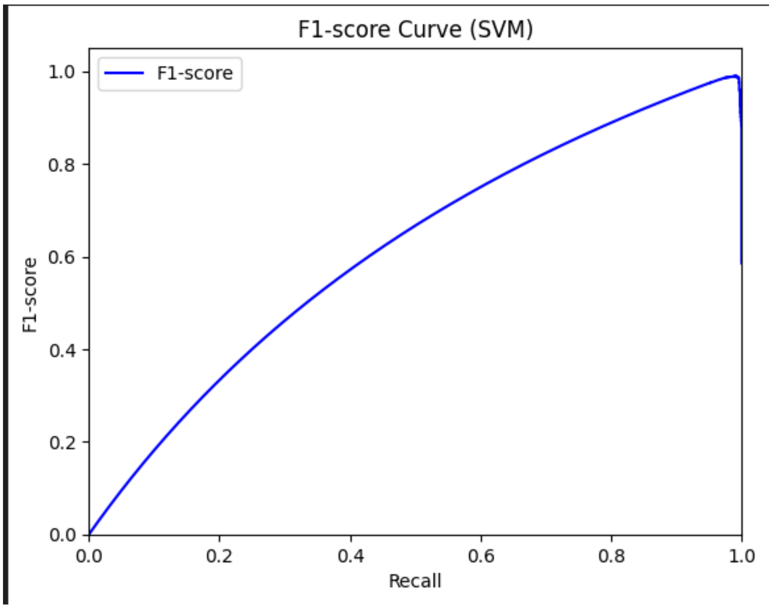


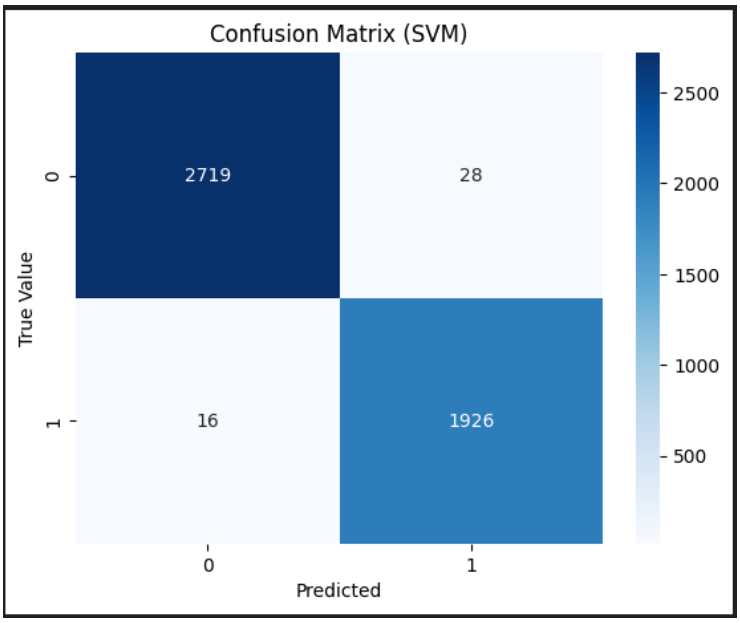
4. Support Vector Machine (SVM):

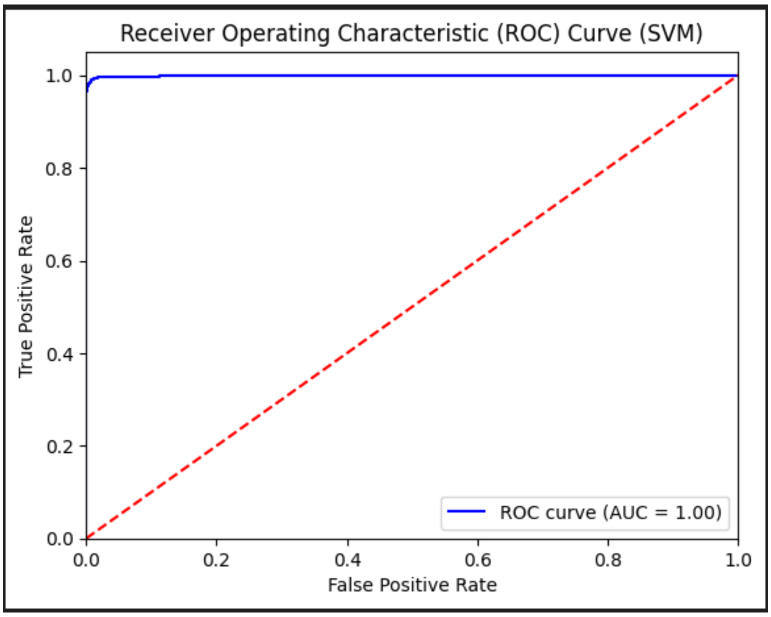
- A Support Vector Machine (SVM) model, utilizing a linear kernel, was employed for its interpretability and robustness. It can be robust against overfitting, especially in high-dimensional space.
- The dataset underwent preprocessing, involving label encoding of categorical variables and addressing class imbalance through random under-sampling.
- The SVM model was trained on the under-sampled data and evaluated for accuracy on the test set.
- Evaluation metrics, including precision, recall, F1-score, and accuracy, were calculated, providing a comprehensive overview of the model's performance.
- Precision-Recall and F1-score curves were plotted to visualize the trade-off between precision and recall.
- The confusion matrix offered a detailed breakdown of the model's predictive outcomes, highlighting true positive, true negative, false positive, and false negative results.
- The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were utilized to assess the SVM model's discrimination capabilities, demonstrating its ability to distinguish between classes.











V. Results

a. Precision-Recall Curve:

The Precision-Recall curve illustrates the trade-off between precision and recall. The area under the curve (AUC) quantifies the classifier's ability to balance these two metrics.

b. F1-score Curve:

The F1-score curve visualizes the balance between precision and recall across different thresholds.

c. Accuracy Curve:

Accuracy is a measure of the overall correctness of the model. It is calculated as the ratio of correctly predicted instances to the total instances.

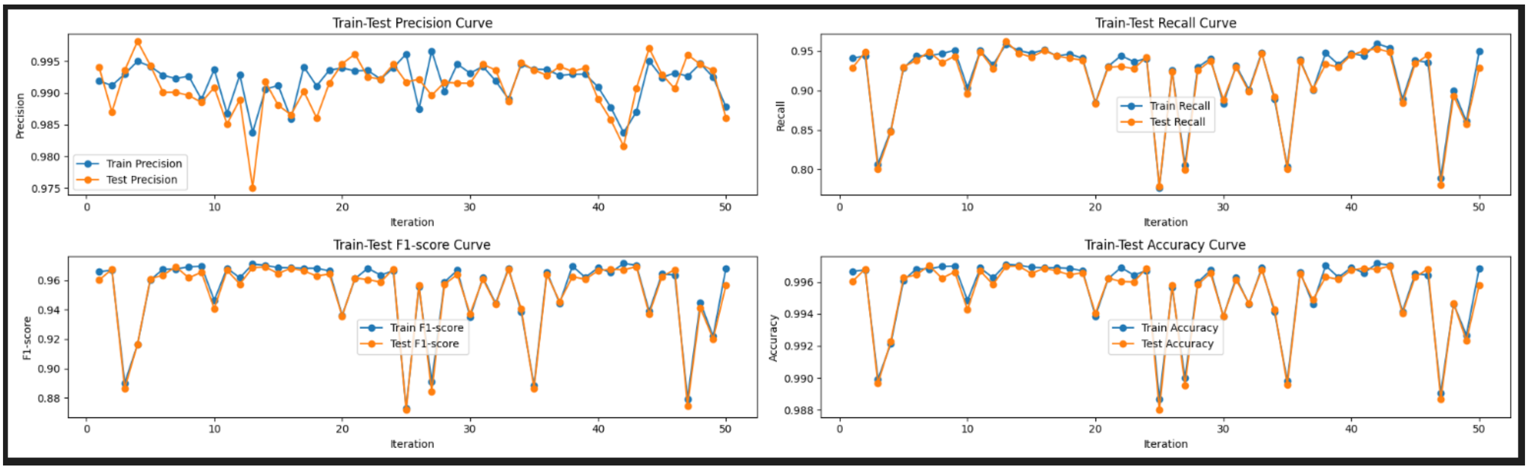
d. ROC-AUC Curve:

The Receiver Operating Characteristic (ROC) curve evaluates the classifier's performance across various discrimination thresholds. The AUC provides a summary measure of the ROC curve.

To ensure the model's consistency, multiple iterations were performed, each involving a different train-test split. The curves are illustrated below.

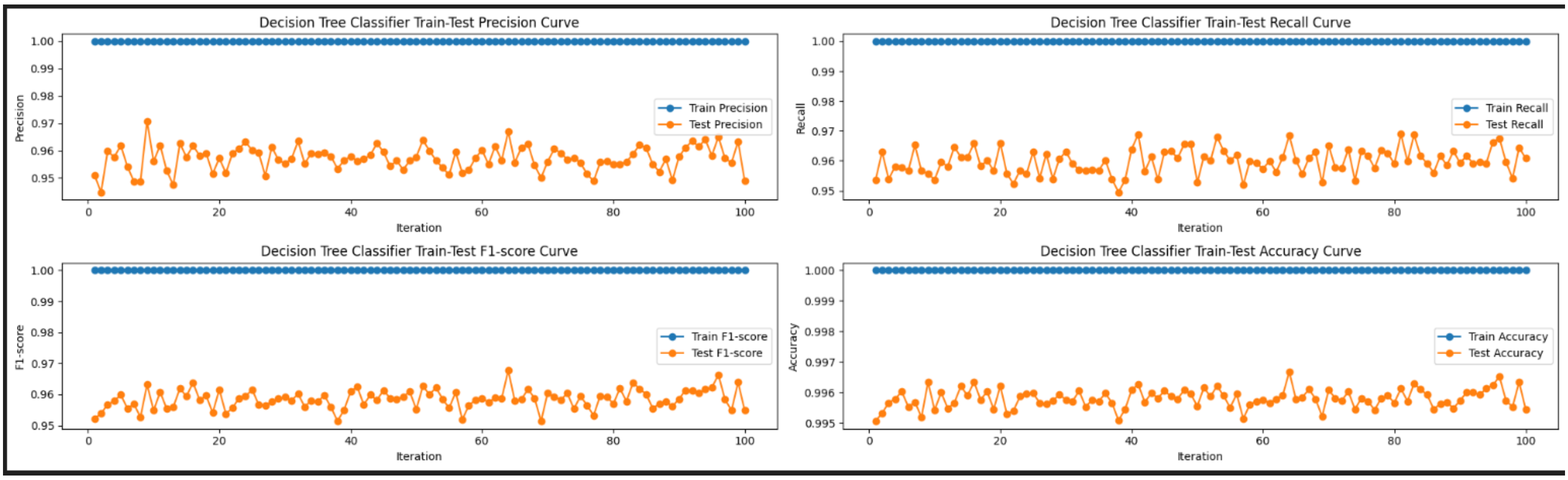
1. Random Forest Classifier Results:

Metric	Value
Precision	0.9927
Recall	0.9871
F1-score	0.9899
Accuracy	0.9917



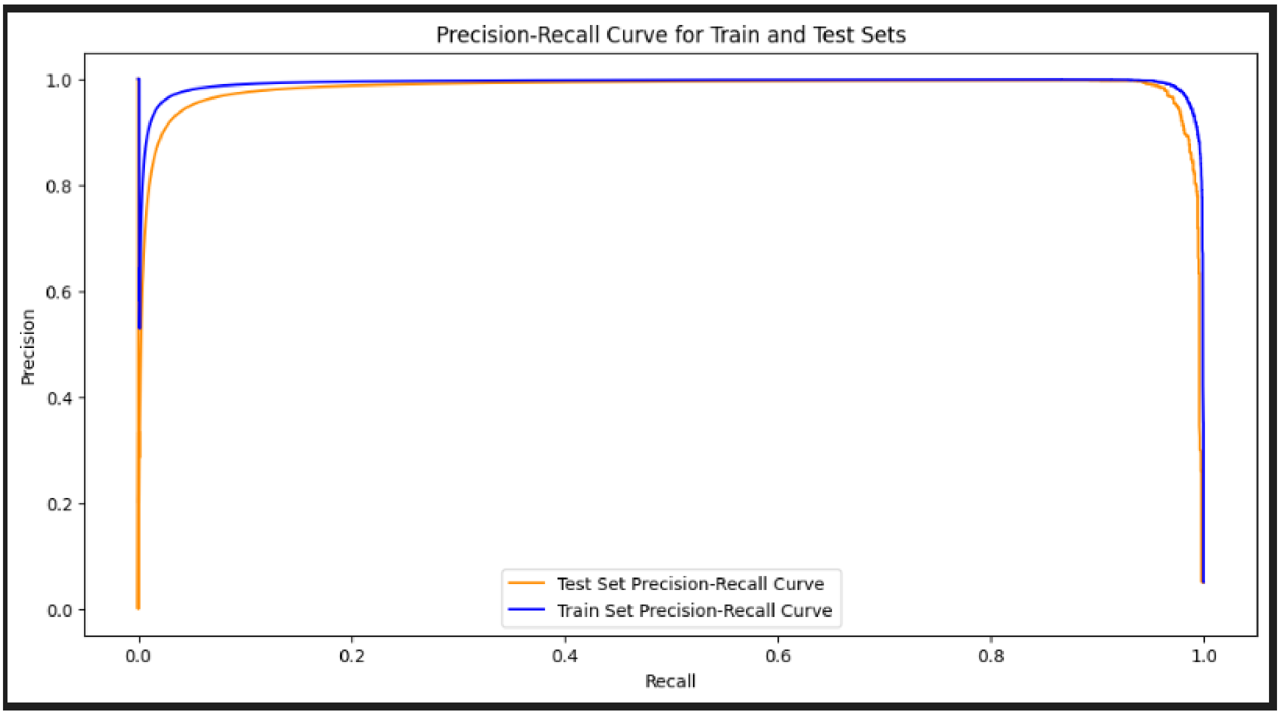
2. Decision Tree Classifier Results:

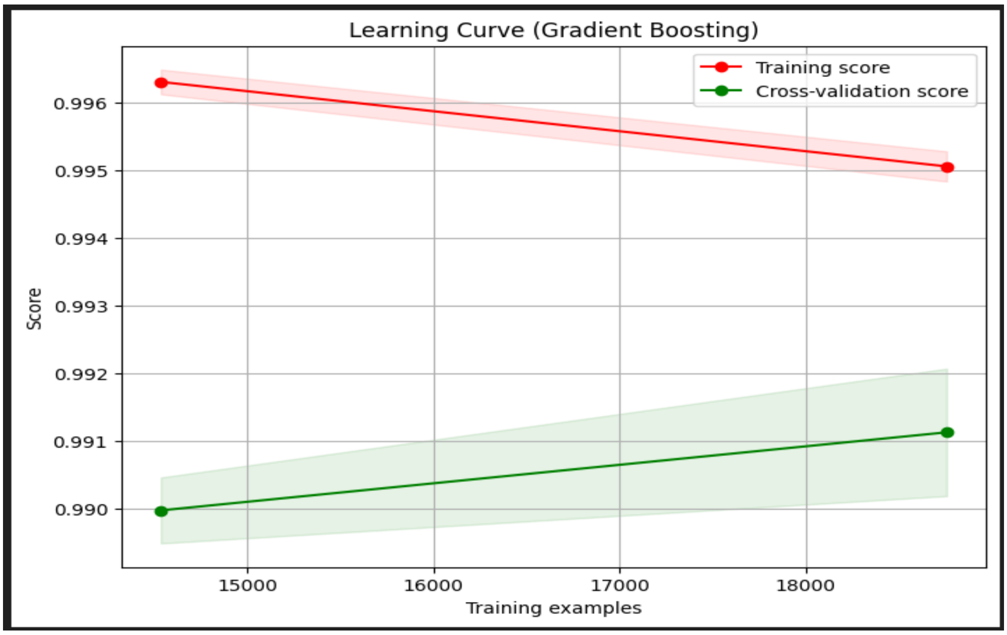
Metric	Value
Precision	0.9819
Recall	0.9824
F1-score	0.9822
Accuracy	0.9852



3. Gradient Boosting Classifier Results:

Metric	Value
Precision	0.9849
Recall	0.9582
F1-score	0.9714
Accuracy	0.9971





4. Support Vector Machine (SVM) Results:

Metric	Value
Precision	0.9856
Recall	0.9917
F1-score	0.9887
Accuracy	0.9906

In conclusion, the performance of each model was evaluated based on key metrics, including precision, recall, F1-score, and accuracy, to determine their efficacy in predicting landslides.

The Random Forest Classifier exhibited exceptional precision (0.9927), recall (0.9871), and F1-score (0.9899), resulting in an impressive accuracy of 99.17%. Similarly, the Decision Tree Classifier demonstrated high precision (0.9819), recall (0.9824), and F1-score (0.9822), with an accuracy of 98.52%. The Gradient Boosting Classifier achieved a commendable precision (0.9849), though with a slightly lower recall (0.9582), resulting in an overall F1-score of 0.9714. Notably, it achieved the highest accuracy among the models at 99.71%. The Support Vector Machine (SVM) exhibited a balanced performance with precision (0.9856), recall (0.9917), and F1-score (0.9887), contributing to an accuracy of 99.06%.

Considering these results, the **Gradient Boosting Classifier** stands out as the most accurate model for landslide prediction, closely followed by the Random Forest Classifier, while the Decision Tree Classifier and SVM also exhibit robust performance. However, the choice of the most suitable model may depend on specific priorities, such as the emphasis on precision, recall, or a balance of both, depending on the application's requirements.

VI. References

[1] Kaggle: https://www.kaggle.com/datasets/vigneshwarchandran/landslide-data-large-numeric

[2] Scikit-learn: https://scikit-learn.org/stable/documentation.html

[3] Youtube: https://youtu.be/4SivdTLIwHc?si=vO19VXfX95hnT2e4