Indira Gandhi Delhi Technical University for Women
Kashmere Gate, Delhi, 110006

Internship Project Presentation
on
EMPLOYEE ATTRITION PREDICTION & COMPARITIVE ANALYSIS OF RESAMPLING TECHNIQUES

Guided By:

Dr. Nidhi Grover

Dr. Ritu Rani

Presented By:

Mehak Arora

03601192023

Branch- AI/ML

INTRODUCTION

PROBLEM OVERVIEW:

Machine learning is extensively used across different domains to make data-driven decisions; one such domain is HR Analytics.

Employee attrition refers to the departure of employees for both voluntary and involuntary reasons. Employee attrition becomes a significant concern for organizations because considerable costs and risks are involved in recruiting new employees and it is efficient to retain the trained, experienced employees.

Predicting the employees who are prone to attrition allows the company to take proactive measures to improve retention.

Additionally, datasets used for machine learning models are often unbalanced and it becomes crucial to adopt correct balancing techniques for optimal results.

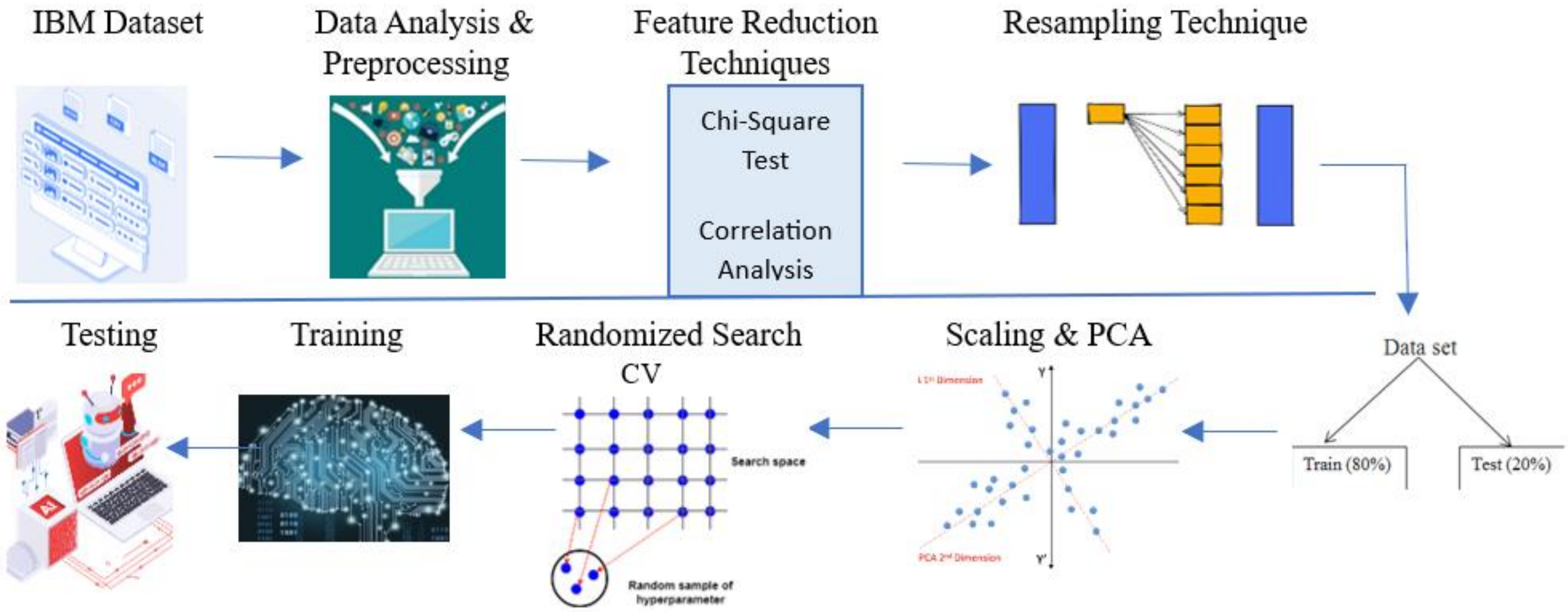
OBJECTIVE:

- Comparative analysis of various resampling techniques used for balancing datasets.
- Apply them with various supervised learning models to conclude the best predicting model for attrition.

DATASET USED: IBM HR Analytics dataset

WORK DESCRIPTION

Methodology followed:



IBM Dataset

Data Analysis & Preprocessing

Feature Reduction Techniques

Chi-Square Test

Correlation Analysis

Resampling Technique

Data set

Train (80%)

Test (20%)

Scaling & PCA

Randomized Search CV

Training

Testing

(i) DATASET DESCRIPTION & ANALYSIS: The dataset used in this work is IBM HR Employee Attrition. The dataset contains 35 features and 1470 data entries. Out of these 35 features, 9 features are categorical features and remaining 26 are numerical features. The dataset used in this work is IBM HR Employee Attrition [1]. The dataset contains 35 features and 1470 data entries. Out of these 35 features, 9 features are categorical features and remaining 26 are numerical features.

(ii) DATA PREPROCESSING AND FEATURE ENGINEERING:

I. Chi Square statistical test was used to study the relevance of categorical features in predicting the target variable.

II. Correlation analysis was done for feature reduction.

III. The categorical features were **Label encoded.**

IV. Standard scaling was applied to the dataset. **Principal component analysis (PCA)** technique was implemented as a part of feature engineering for dimensionality reduction.

(iii) RESAMPLING METHODS: The 8 resampling methods studied under this work includes,

I. UNDER SAMPLING -- Random Under Sampling, Tomek Links, ENN

II. OVER SAMPLING – Random Oversampling, SMOTE, ADASYN

III. HYBRID – SMOTE-Tomek Links, SMOTE-ENN

(iv) EXPERIMENTAL SETUP: The models we used for comparative analysis of each resampling techniques to find the best performing model are **Logistic Regression, Random Forest Classifier, KNN, XGB Classifier, SVC, Decision Tree Classifier.** The balanced dataset was divided in the ratio **80:20 for training and testing.**
Cross-validation, a technique used to assess a model's performance and ensure its generalizability across different subsets of data was applied. **Randomized Search CV** was utilized to efficiently explore a wide range of hyperparameters.

(v) EVALUATION MATRICES: A **confusion matrix, classification report, accuracy score, AUC-ROC** score along with **ROC curve** was **generated for every combination of resampling techniques and classifiers** to gather the performance metrics overview.

WORK OUTCOME

Table 1 and 2 illustrate the results of Accuracy and AUC ROC scores for predictions done based upon the imbalanced(raw) dataset as well as that of various combinations of classifiers and resampling techniques for comparison.

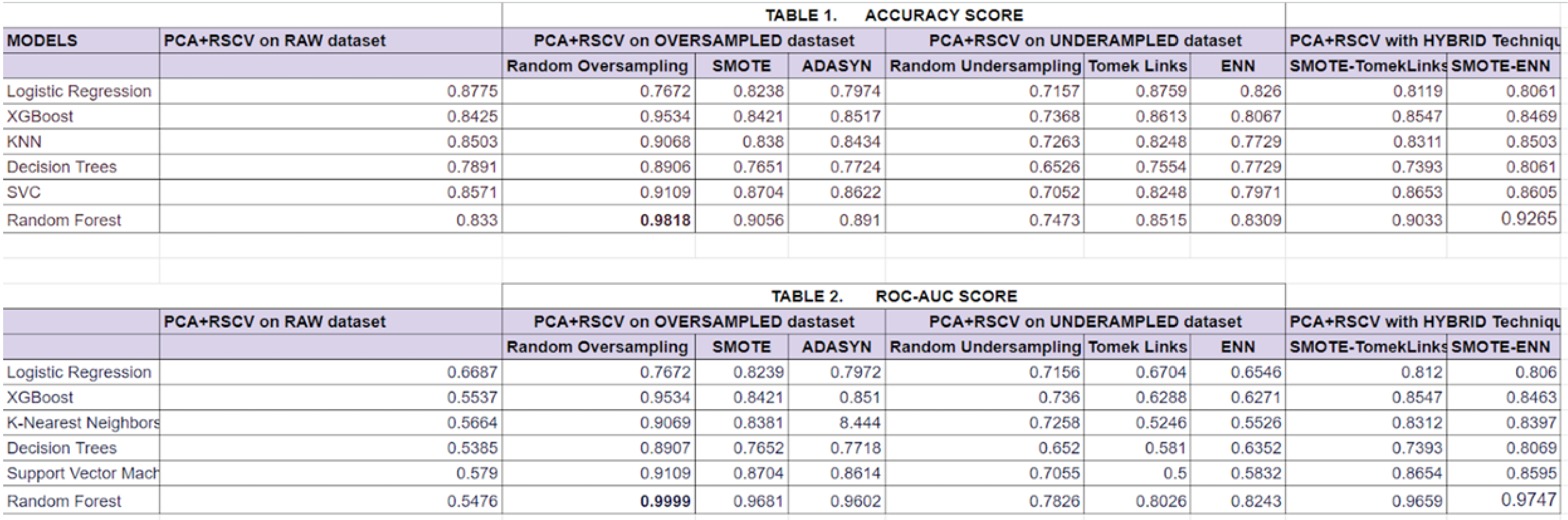


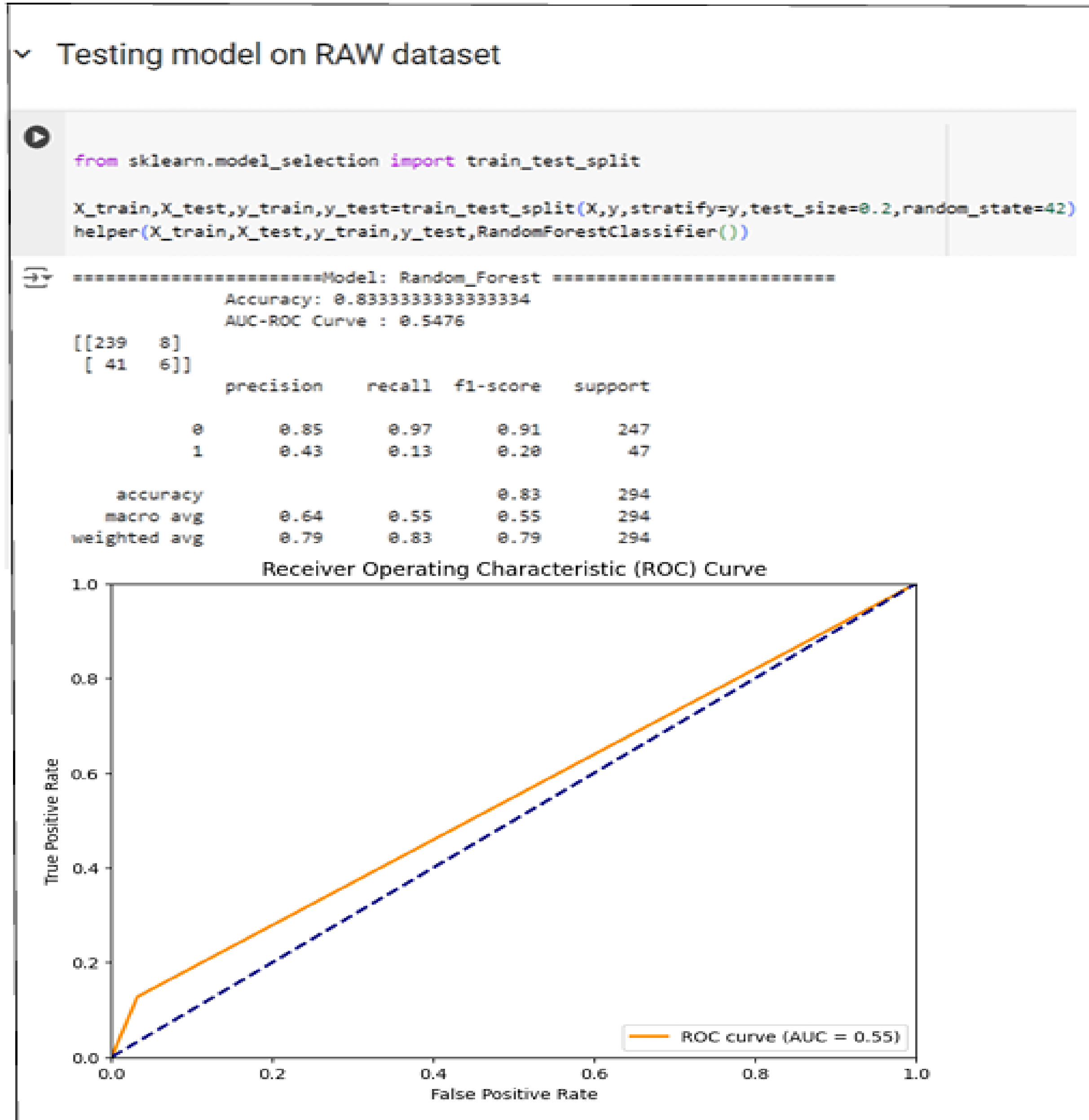
TABLE 1. ACCURACY SCORE

MODELS	PCA+RSCV on RAW dataset	PCA+RSCV on OVERSAMPLED dastaset			PCA+RSCV on UNDERAMPLED dataset			PCA+RSCV with HYBRID Techniqu	
		Random Oversampling	SMOTE	ADASYN	Random Undersampling	Tomek Links	ENN	SMOTE-TomekLinks	SMOTE-ENN
Logistic Regression	0.8775	0.7672	0.8238	0.7974	0.7157	0.8759	0.826	0.8119	0.8061
XGBoost	0.8425	0.9534	0.8421	0.8517	0.7368	0.8613	0.8067	0.8547	0.8469
KNN	0.8503	0.9068	0.838	0.8434	0.7263	0.8248	0.7729	0.8311	0.8503
Decision Trees	0.7891	0.8906	0.7651	0.7724	0.6526	0.7554	0.7729	0.7393	0.8061
SVC	0.8571	0.9109	0.8704	0.8622	0.7052	0.8248	0.7971	0.8653	0.8605
Random Forest	0.833	**0.9818**	0.9056	0.891	0.7473	0.8515	0.8309	0.9033	0.9265

TABLE 2. ROC-AUC SCORE

	PCA+RSCV on RAW dataset	PCA+RSCV on OVERSAMPLED dastaset			PCA+RSCV on UNDERAMPLED dataset			PCA+RSCV with HYBRID Techniqu	
		Random Oversampling	SMOTE	ADASYN	Random Undersampling	Tomek Links	ENN	SMOTE-TomekLinks	SMOTE-ENN
Logistic Regression	0.6687	0.7672	0.8239	0.7972	0.7156	0.6704	0.6546	0.812	0.806
XGBoost	0.5537	0.9534	0.8421	0.851	0.736	0.6288	0.6271	0.8547	0.8463
K-Nearest Neighbors	0.5664	0.9069	0.8381	8.444	0.7258	0.5246	0.5526	0.8312	0.8397
Decision Trees	0.5385	0.8907	0.7652	0.7718	0.652	0.581	0.6352	0.7393	0.8069
Support Vector Mach	0.579	0.9109	0.8704	0.8614	0.7055	0.5	0.5832	0.8654	0.8595
Random Forest	0.5476	**0.9999**	0.9681	0.9602	0.7826	0.8026	0.8243	0.9659	0.9747

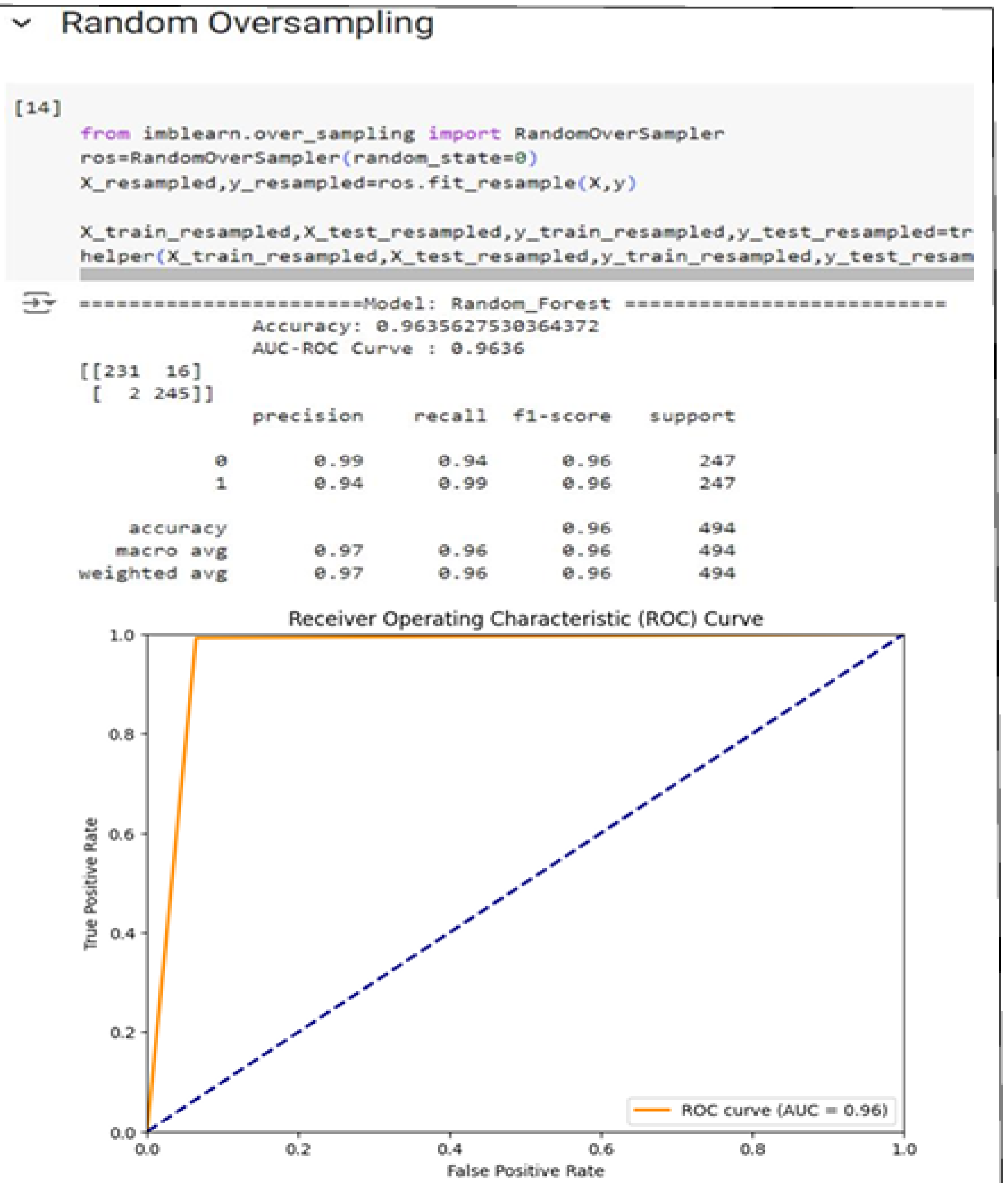
Overall, **Random Forest model performed best** with an **accuracy of 98.18%** and **AUC-ROC score of 99.99%** when cross validation was performed after **using Random Oversampling for balancing.**



Testing model on RAW dataset

```
from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test=train_test_split(X,y,stratify=y,test_size=0.2,random_state=42)
helper(X_train,X_test,y_train,y_test,RandomForestClassifier())
```

```
========================Model: Random_Forest ==========================
              Accuracy: 0.8333333333333334
              AUC-ROC Curve : 0.5476
[[239   8]
 [ 41   6]]
              precision    recall  f1-score   support

           0       0.85      0.97      0.91       247
           1       0.43      0.13      0.20        47

    accuracy                           0.83       294
   macro avg       0.64      0.55      0.55       294
weighted avg       0.79      0.83      0.79       294
```

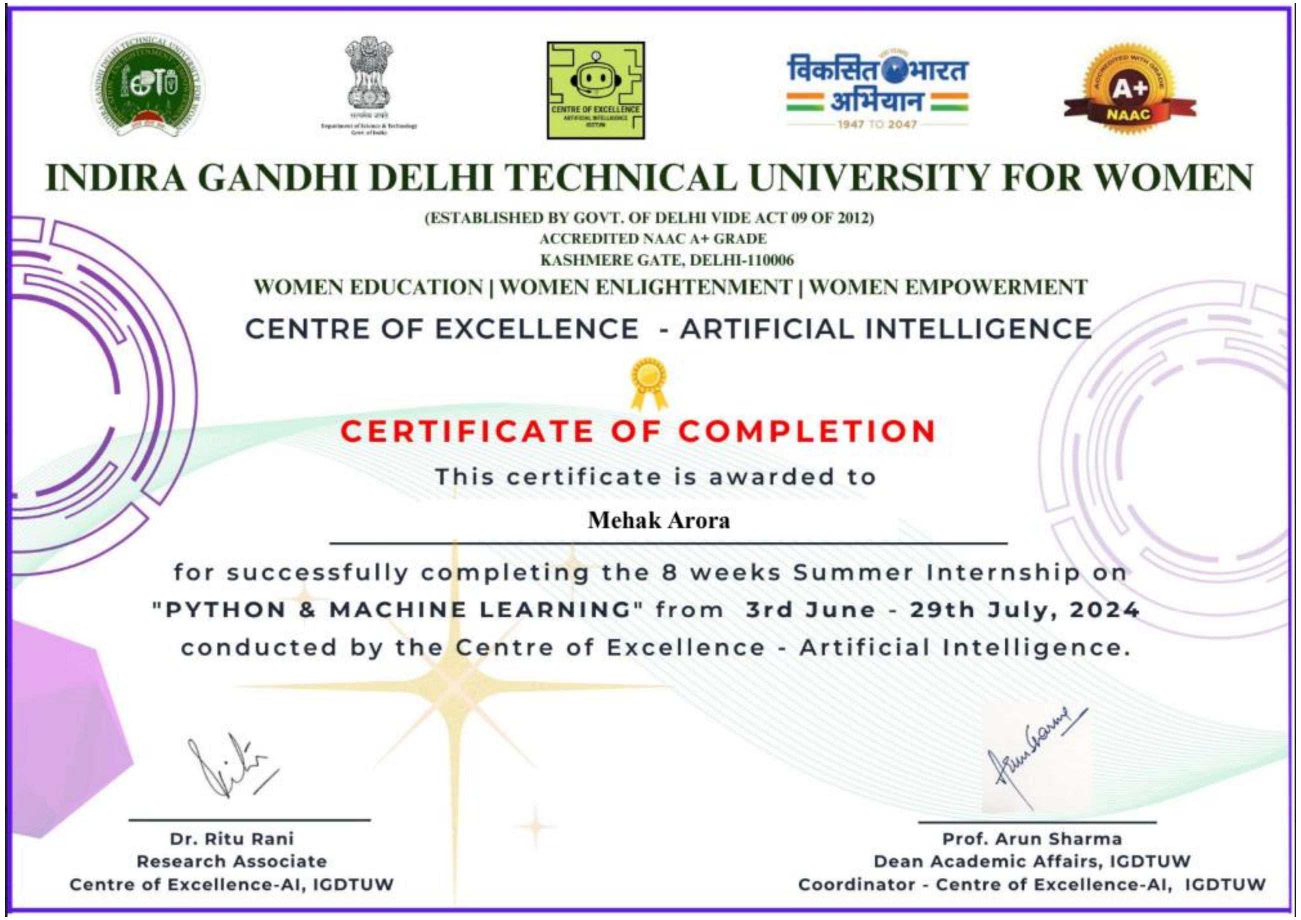


Random Oversampling

```
[14]
from imblearn.over_sampling import RandomOverSampler
ros=RandomOverSampler(random_state=0)
X_resampled,y_resampled=ros.fit_resample(X,y)

X_train_resampled,X_test_resampled,y_train_resampled,y_test_resampled=tr
helper(X_train_resampled,X_test_resampled,y_train_resampled,y_test_resam
```

```
========================Model: Random_Forest ==========================
              Accuracy: 0.9635627530364372
              AUC-ROC Curve : 0.9636
[[231  16]
 [  2 245]]
              precision    recall  f1-score   support

           0       0.99      0.94      0.96       247
           1       0.94      0.99      0.96       247

    accuracy                           0.96       494
   macro avg       0.97      0.96      0.96       494
weighted avg       0.97      0.96      0.96       494
```

INTERNSHIP CERTIFICATE



INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN

(ESTABLISHED BY GOVT. OF DELHI VIDE ACT 09 OF 2012)
ACCREDITED NAAC A+ GRADE
KASHMERE GATE, DELHI-110006

WOMEN EDUCATION | WOMEN ENLIGHTENMENT | WOMEN EMPOWERMENT

CENTRE OF EXCELLENCE - ARTIFICIAL INTELLIGENCE

CERTIFICATE OF COMPLETION

This certificate is awarded to

Mehak Arora

for successfully completing the 8 weeks Summer Internship on "PYTHON & MACHINE LEARNING" from 3rd June - 29th July, 2024 conducted by the Centre of Excellence - Artificial Intelligence.

Dr. Ritu Rani
Research Associate
Centre of Excellence-AI, IGDTUW

Prof. Arun Sharma
Dean Academic Affairs, IGDTUW
Coordinator - Centre of Excellence-AI, IGDTUW