**Data Science Last Project**

### ****Introduction****

* **Objective**:Objective of this project to address class imbalance in classification tasks using various CI (Class Imbalance) techniques to make data balanced.
* **Datasets Used**
  + **Credit Card dataset**
  + **Brest Cancer Dataset**
  + **Customer Churn Dataset**
* **CI Techniques Explored**
  + Resampling Method (Smote)
  + Algorithmic Method like class weight
  + Data Level (Feature Engineering )

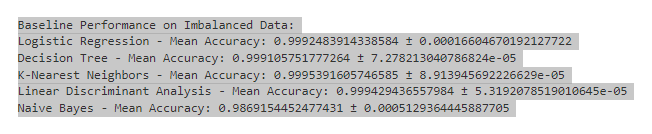
### ****Data Preparation and Initial Analysis****

* **Data Cleaning and EDA**:
  + handling missing values
  + Remove duplicate entries.
  + Change type
  + Remove unnecessary feature
  + The exploratory data analysis (EDA), focusing on the target class distribution to whether data is balanced or not.
* **Feature Selection and Cross(baseline Performance)**
  + **Wrapper Method Recursive feature elimination using cross validation and this is I use to see accuracy on imbalanced data then I apply techniques of class imbalance and compare their performance with baseline**

### ****Credit Card Dataset****

### Resampling Method using Smote on ****Credit Card data****

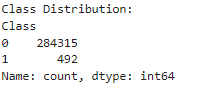
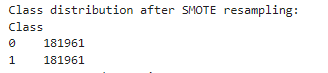
### **Baseline Performance:**



### **Performance After Class Imbalance:**

1. Logistic Regression: **accuracy 0.97**
2. Decision Tree:  **accuracy 1.00**
3. K-Nearest Neighbors: **accuracy 1.00**
4. Linear Discriminant: **accuracy 0.99**
5. Naive Bayes: **accuracy 0.98**

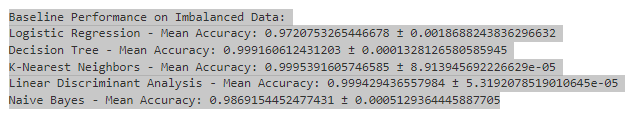
### **Distribution Before and After Class Imbalance:**

### 

### Algorithmic Method using Class Weight and Ensemble method on ****Credit Card**** data

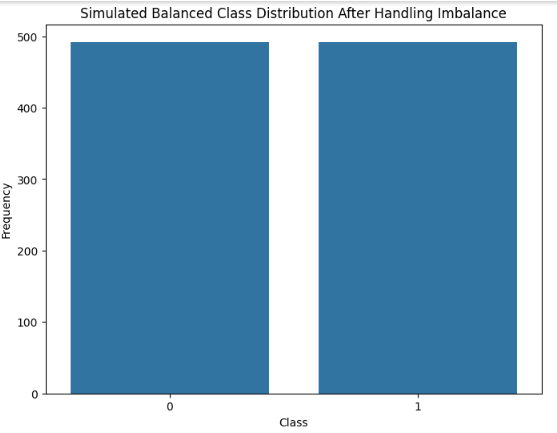
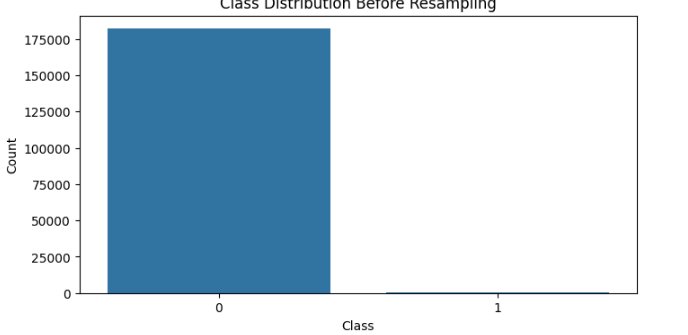
**Baseline Performance:**



### **Performance After Class Imbalance:**

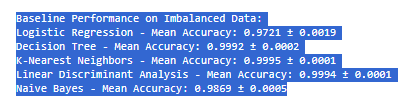
1. Logistic Regression: **accuracy 0.99**
2. Decision Tree:  **accuracy 1.00**
3. K-Nearest Neighbors: **accuracy 1.00**
4. Linear Discriminant: **accuracy 1.00**
5. Naive Bayes: **accuracy 0.99**

### **Distribution Before and After Class Imbalance:**



### One Class SVM on ****Credit Card****

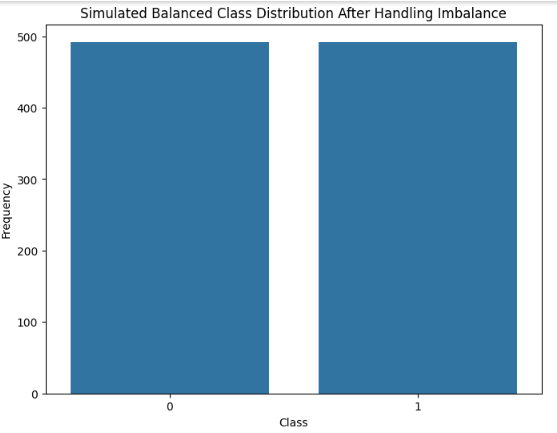
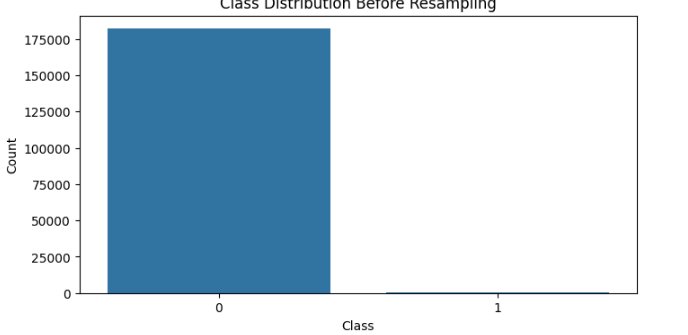
**Baseline Performance:**



**Performance After Class Imbalance:**

1. Logistic Regression: **accuracy 0.97**
2. Decision Tree:  **accuracy 1.00**
3. K-Nearest Neighbors: **accuracy 1.00**
4. Linear Discriminant: **accuracy 1.00**
5. Naive Bayes: **accuracy 0.99**

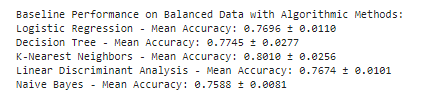
### **Distribution Before and After Class Imbalance:**



### ****Customer Churn Dataset****

### Resampling Method using Smote on Customer Churn ****data****

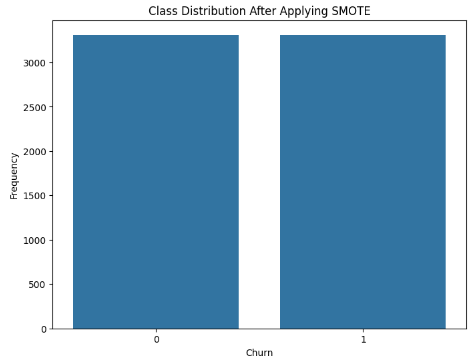
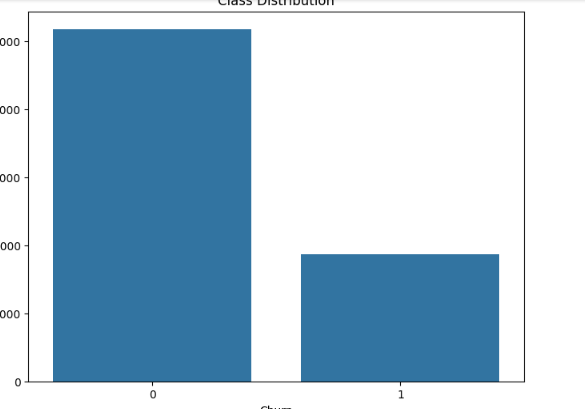
**Baseline Performance:**



**Performance After Class Imbalance:**

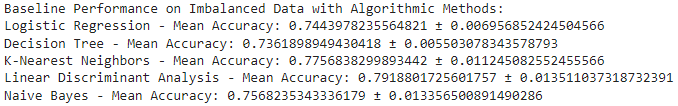
1. Logistic Regression: **accuracy 0.74**
2. Decision Tree:  **accuracy 0.72**
3. K-Nearest Neighbors: **accuracy 0.73**
4. Linear Discriminant: **accuracy 0.73**
5. Naive Bayes: **accuracy 0.73**

**Distribution Before and After Class Imbalance:**



### Algorithmic Method using Class Weight on Customer Churn ****data****

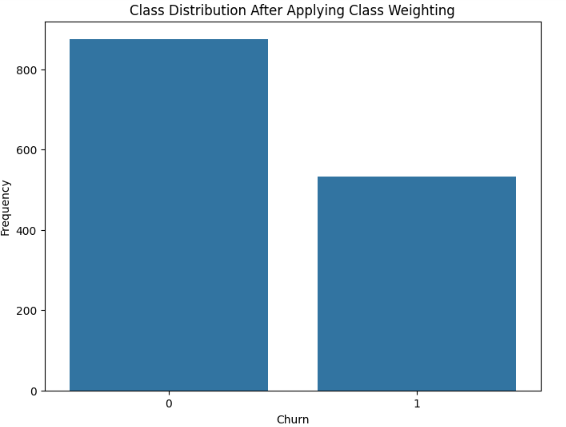
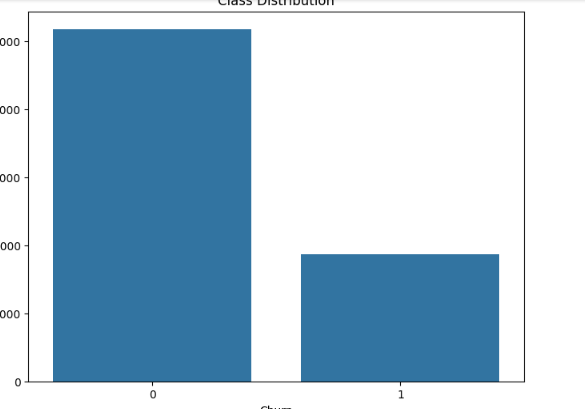
**Baseline Performance:**



**Performance After Class Imbalance:**

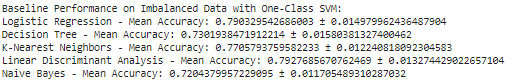
1. Logistic Regression: **accuracy 0.73**
2. Decision Tree:  **accuracy 0.73**
3. K-Nearest Neighbors: **accuracy 0.77**
4. Linear Discriminant: **accuracy 0.79**
5. Naive Bayes: **accuracy 0.75**

**Distribution Before and After Class Imbalance:**



### One Class SVM on ****Credit Card****

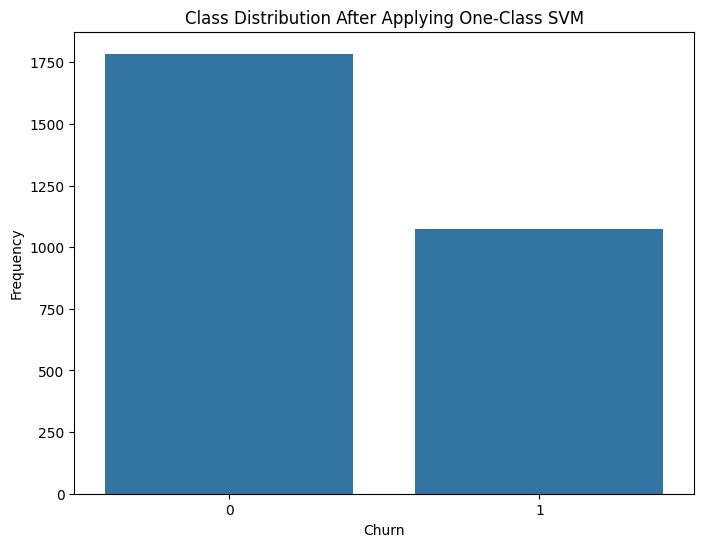
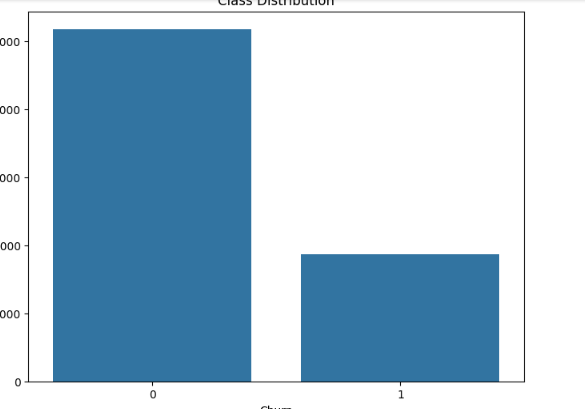
**Baseline Performance:**



**Performance After Class Imbalance:**

1. Logistic Regression: **accuracy 0.73**
2. Decision Tree:  **accuracy 0.73**
3. K-Nearest Neighbors: **accuracy 0.77**
4. Linear Discriminant: **accuracy 0.79**
5. Naive Bayes: **accuracy 0.75**

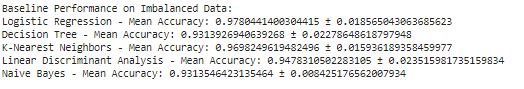
**Distribution Before and After Class Imbalance:**



### ****Breast Cancer Dataset****

### Resampling Method using Smote on ****Breast Cancer**** ****data****

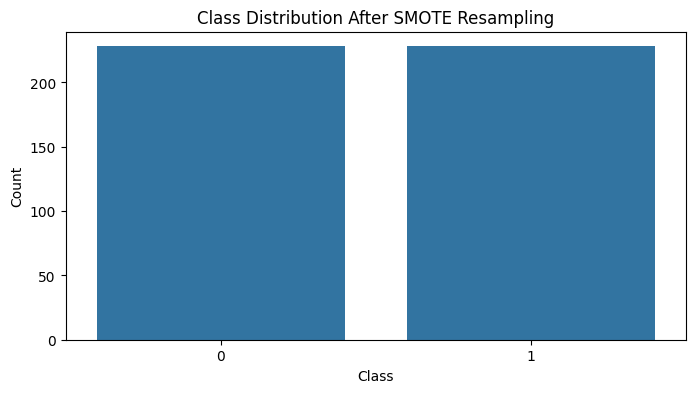
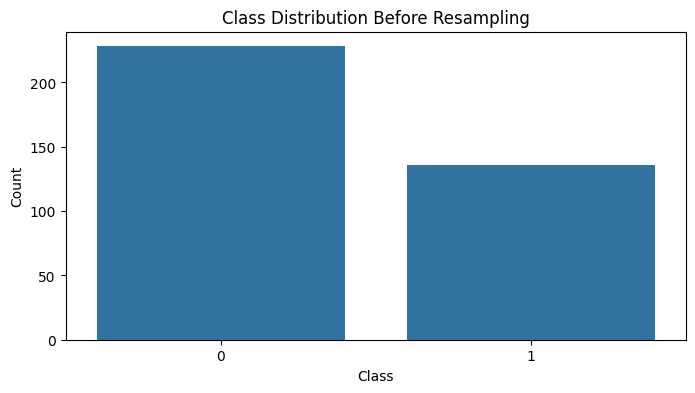
**Baseline Performance:**



**Performance After Class Imbalance:**

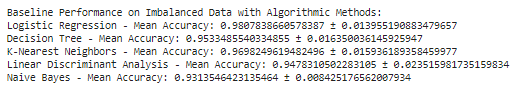
1. Logistic Regression: **accuracy 0.97**
2. Decision Tree:  **accuracy 0.92**
3. K-Nearest Neighbors: **accuracy 0.96**
4. Linear Discriminant: **accuracy 0.97**
5. Naive Bayes: **accuracy 0.94**

**Distribution Before and After Class Imbalance:**



### Algorithmic Method using Class Weight on ****Breast Cancer data****

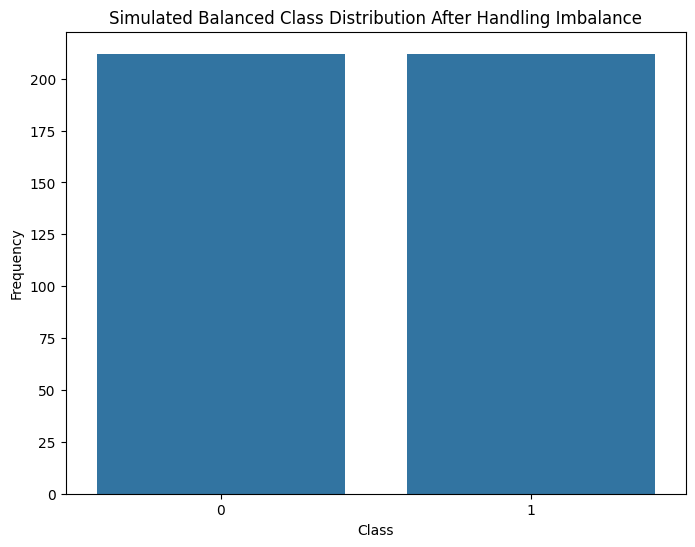
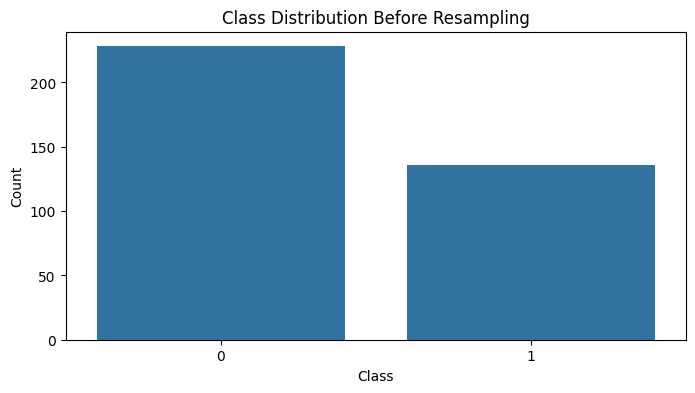
**Baseline Performance:**



**Performance After Class Imbalance:**

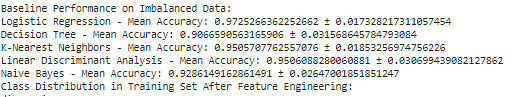
1. Logistic Regression: **accuracy 0.94**
2. Decision Tree:  **accuracy 0.97**
3. K-Nearest Neighbors: **accuracy 0.96**
4. Linear Discriminant: **accuracy 0.95**
5. Naive Bayes: **accuracy 0.93**

**Distribution Before and After Class Imbalance:**



### Data Level (feature Engineering) on ****Breast Cancer Data****

**Baseline Performance:**



**Performance After Class Imbalance:**

1. Logistic Regression: **accuracy 0.96**
2. Decision Tree:  **accuracy 0.91**
3. K-Nearest Neighbors: **accuracy 0.88**
4. Linear Discriminant: **accuracy 0.69**
5. Naive Bayes: **accuracy 0.76**

**Distribution Before and After Class Imbalance:**

