# 

# FIRST DATASET: LOAN DATA

1. **credit.policy**: Binary variable indicating if a customer meets the credit underwriting criteria of LendingClub.com (1 if meets criteria, 0 otherwise).
2. **purpose**: Purpose of the loan provided as a categorical variable, having debt consoldation, credit card and some other.
3. **int.rate**: Interest rate on the loan, expressed as a decimal.
4. **installment**: Monthly payment owed by the borrower.
5. **log.annual.inc**: Natural logarithm of the borrower's self-reported annual income.
6. **dti**: Debt-to-income ratio, representing the borrower's total monthly debt payments divided by gross monthly income.
7. **fico**: FICO credit score of the borrower.
8. **days.with.cr.line**: Number of days the borrower has had a credit line.
9. **revol.bal**: Total credit revolving balance (i.e., balance left on a credit account after making a payment).
10. **revol.util**: Revolving line utilization rate, or the amount of credit the borrower is using relative to their total available credit.
11. **inq.last.6mths**: Number of inquiries made by creditors in the last 6 months.
12. **delinq.2yrs**: Number of times the borrower had been 30+ days past due on a payment in the past 2 years.
13. **pub.rec**: Number of derogatory public records associated with the borrower (e.g., bankruptcies, tax liens, judgments).
14. **not.fully.paid**: Binary variable indicating whether the loan was not fully paid (1 if not fully paid, 0 otherwise).

## KNN:

* The KNN classifier indicated that it correctly classified approximately 81.32% of the instances. However, its precision (0.1807) and recall (0.0492) were significantly lower, suggesting that while it performed well in identifying true negatives, it struggled in correctly identifying positive cases (true positives).
* **Chi-Square:** Achieved a mean accuracy of 0.7206, with a relatively higher precision (0.2540) compared to other methods, but low recall (0.2339) and F1 score (0.1203).
* **Correlation and Covariance:** Both methods yielded similar results with mean accuracies around 0.7187 and precision values around 0.2173. However, their recall and F1 scores were also low.
* **Mutual Information:** Had a slightly lower accuracy (0.7197) and precision (0.1969) compared to Chi-Square, but still exhibited low recall and F1 score.
* **PCA (Principal Component Analysis):** Showed the highest mean accuracy among the feature selection methods with 0.7257, but with precision (0.2145), recall (0.2143), and F1 score (0.1110) all relatively low..

### Comparison and Interpretation:

While the PCA method resulted in the highest mean accuracy among the feature selection methods, it's important to note that the KNN classifier still outperformed it in terms of accuracy on the test data. Overall, the evaluation suggests that the PCA feature selection method showed the best balance in terms of accuracy among the evaluated methods.

## LOGISTIC REGRESSION:

* Achieved an accuracy of approximately 83.93%, the precision measuring the proportion of correctly predicted positive cases was approximately 47.10%, the recall, which measures the proportion of correctly predicted positive cases was approximately 1.63%, F1 score, which is the harmonic mean of precision and recall, was approximately 3.14%.
* **Correlation Method:** it indicates that this feature selection process based on correlation didn't significantly affect the model's performance.
* **Mutual Information Method:** Also showed similar accuracy to the overall logistic regression model, but with a slightly higher precision (52.57%) and a slightly lower recall (1.55%).
* **Covariance Method:** Resulted in slightly lower accuracy compared to the overall logistic regression model, with a precision of 20.00% and a very low recall of 0.16%.
* **PCA (Principal Component Analysis) Method:** Resulted in slightly lower accuracy compared to the overall model with low precision and recall.

### Comparison and Interpretation:

The logistic regression model performed reasonably well in terms of accuracy, but its precision and recall were relatively low. Feature selection methods didn't seem to have a significant impact on the model's performance. The low recall values suggest that the model may need further refinement to better identify positive cases, especially if correctly identifying them is crucial for the application.

## RANDOM FOREST:

Achieved an accuracy of approximately 83.80%, the precision, which measures the proportion of correctly predicted positive was approximately 40.17% , the recall, which measures the proportion of correctly predicted positive cases was approximately 2.12%, F1 score, which is the harmonic mean of precision and recall, was approximately 4.01%.

* **Correlation Method:** Produced similar results to the overall random forest metrics
* **Mutual Information Method:** Resulted in slightly lower accuracy compared to the overall random forest model, but with a slightly higher precision (41.21%) and recall (3.18%), resulting in a better F1 score (5.88%).
* **Covariance Method:** Resulted in slightly lower accuracy compared to the overall random forest model, with a precision of 26.39%, recall of 9.61%, and F1 score of 14.08%.
* **PCA (Principal Component Analysis) Method:** Resulted in slightly lower accuracy compared to the overall random forest model, with a precision of 20.37%, recall of 5.54%, and F1 score of 8.70%.

### Comparison and Interpretation:

The random forest model performed reasonably well in terms of accuracy, but its precision and recall were relatively low. Feature selection methods didn't seem to have a significant impact on the model's performance. While Mutual Information method showed a slight improvement in precision, recall, and F1 score compared to other methods, the overall performance still indicates room for improvement, especially in correctly identifying positive cases.

## SVM:

* Achieved an accuracy of approximately 83.97%, indicating that the model correctly classified about 83.97% of the instances, the precision was approximately 20.00%, Recall measuring proportion of correctly predicted positive cases was extremely low at approximately 0.08% and F1 score, which is the harmonic mean of precision and recall, was also very low at approximately 0.16%.
* **Correlation Method, Mutual Information Method, and PCA (Principal Component Analysis) Method:** These methods produced identical results to the overall SVM metrics, indicating that the feature selection process based on these methods didn't significantly affect the model's performance.
* **Covariance Method:** Resulted in identical accuracy compared to the overall SVM model, but with a precision and recall of 0.00% and an F1 score of 0.00%. This indicates that the model didn't correctly identify any positive cases.

### Comparison and Interpretation:

The SVM model achieved a relatively high accuracy, but its precision, recall, and F1 score were extremely low, The results suggest that further analysis or adjustments may be needed to improve the SVM model.

## NAIVE BAYES:

* Achieved an accuracy of approximately 82.29%, the precision, which measures the proportion of correctly predicted positive cases was approximately 32.77%, recall measuring the proportion of correctly predicted positive cases was 9.93%, F1 score, which is the harmonic mean of precision and recall, was approximately 15.22%.
* **Correlation Method:** Produced identical results to the overall Naive Bayes metrics.
* **Mutual Information Method:** Resulted in slightly higher accuracy compared to the overall Naive Bayes model, with slightly improved precision (32.93%), recall (10.10%), and F1 score (15.44%).
* **Covariance Method:** Resulted in slightly lower metrics compared to the overall Naive Bayes model.
* **PCA (Principal Component Analysis) Method:** Resulted in slightly higher accuracy compared to the overall Naive Bayes model, but with lower precision (26.94%), recall (3.42%), and F1 score (6.05%).

### Comparison and Interpretation:

The Naive Bayes model achieved a decent accuracy, precision, recall, and F1 score, indicating moderate performance overall. Feature selection methods had some impact on the model's performance, with Mutual Information method slightly improving precision, recall, and F1 score compared to other methods.

## DECISION TREE

* The overall accuracy is approximately 73.43%, indicating that the model correctly classified about 73.43% of the instances, precision is approximately 20.17%, while recall measuring the proportion of correctly predicted positive cases among all actual positive cases is approximately 22.15%. F1 score, which is the harmonic mean of precision and recall, is approximately 21.10%.
* **Correlation Method:** Produced identical results to the overall metrics
* **Mutual Information Method:** Resulted in slightly higher accuracy compared to the overall model, with improved precision (22.70%), recall (25.57%), and F1 score (24.04%).
* **Covariance Method:** Resulted in higher accuracy compared to the overall model, with improved precision (28.85%), but lower recall (8.87%) and F1 score (13.54%).
* **PCA (Principal Component Analysis) Method:** Resulted in slightly higher accuracy compared to the overall model, but with lower precision (18.75%), recall (19.22%), and F1 score (18.96%).

### Comparison and Interpretation:

The model achieved a moderate accuracy, precision, recall, and F1 score overall. Feature selection methods had some impact on the model's performance, with Mutual Information method slightly improving precision, recall, and F1 score compared to other methods.

# BALANCING TECHNIQUE – SMOTE:

SMOTE stands for Synthetic Minority Over-sampling Technique, and is a popular technique used to address class imbalance in machine learning datasets. SMOTE works by selecting a random sample from the minority class and finding its k nearest neighbors in the feature space. Synthetic samples are then created by interpolating between the selected sample and its neighbors. These new samples are then added to the dataset, effectively increasing the representation of the minority class. By generating synthetic samples, SMOTE helps in balancing the class distribution, making the dataset more suitable for training machine learning models without bias towards the majority class.

* In the resampled Dataset, after applying SMOTE, we can see that both classes ("not.fully.paid" equal to 0 and 1) now have almost equal representation. This balanced distribution helps prevent the model from being biased towards either class during training.

## KNN:

* The overall accuracy is approximately 82.18%, indicating that the model correctly classified about 82.18% of the instances. Precision is approximately 27.87, recall is approximately 6.92% while F1 score, which is the harmonic mean of precision and recall, is approximately 11.08%.
* **Correlation Method:** Resulted in slightly lower accuracy compared to the overall model, with improved precision (70.96%), recall (82.70%), and F1 score (76.38%). This suggests that the correlation method might have selected more relevant features for KNN.
* **Mutual Information Method:** Resulted in lower accuracy compared to the overall model, with improved precision (69.59%), recall (82.60%), and F1 score (75.53%). This indicates that the mutual information method might have also selected relevant features, but with a slight decrease in accuracy.
* **Covariance Method:** Resulted in similar accuracy compared to the overall model, with significantly improved precision (77.90%), but a decrease in recall (68.44%) and F1 score (72.85%).
* **PCA (Principal Component Analysis) Method:** PCA results indicated that the reduction in feature dimensions might have resulted in a loss of discriminative power for KNN.

### Comparison and Interpretation:

The KNN model achieved moderate performance overall, with relatively high accuracy but lower precision, recall, and F1 score. Feature selection methods had varied impacts on the model's performance, with correlation and mutual information methods showing improvements in precision, recall, and F1 score, while covariance method significantly improved precision but reduced recall and F1 score, and PCA method resulted in a decrease in overall performance.

## LOGISTIC EGRESSION

* The overall accuracy is approximately 66.90%, indicating that the model correctly classified about 66.90% of the instances. Precision is approximately 69.88%, recall is approximately 59.49%. F1 score is approximately 64.26%.
* **Correlation Method:** Produced identical results to the overall metrics, indicating that the feature selection process based on correlation didn't significantly affect the model's performance.
* **Mutual Information Method:** Resulted in slightly lower accuracy compared to the overall model, with a decrease in precision (70.06%), recall (55.75%), and F1 score (62.09%).
* **Covariance Method:** Resulted in slightly lower accuracy compared to the overall model, with a slight increase in precision (72.42%), but a decrease in recall (52.29%) and F1 score (60.72%).
* **PCA (Principal Component Analysis) Method:** Resulted in significantly lower accuracy compared to the overall model, with a decrease in precision (52.17%), recall (54.60%), and F1 score (53.35%).

### Comparison and Interpretation:

The logistic regression model achieved moderate performance overall, with relatively high precision but moderate recall and F1 score. Feature selection methods had varied impacts on the model's performance, with Correlation method showing no significant change, Mutual Information method slightly decreasing performance, and Covariance method showing a slight improvement in precision but a decrease in recall and F1 score.

## RANDOM FOREST

* The overall accuracy is approximately 83.74%, indicating that the model correctly classified about 83.74% of the instance’s precision is approximately 84.53%, recall is approximately 82.63% F1 score, which is the harmonic mean of precision and recall, is approximately 83.57%.
  + **Correlation Method:** Produced identical results to the overall metrics.
  + **Mutual Information Method:** Resulted in slightly higher accuracy compared to the overall model, with a slight increase in precision (84.87%), recall (82.27%), and F1 score (83.55%).
  + **Covariance Method:** Resulted in lower metrics suggesting that the covariance method might have removed important information for classification.
  + **PCA (Principal Component Analysis) Method:** Resulted in significantly lower metrics, indicating that the reduction in feature dimensions might have resulted in a loss of discriminative power.

### Comparison and Interpretation:

The random forest model achieved high performance overall, with relatively high precision, recall, and F1 score. Feature selection methods had varied impacts on the model's performance, with Mutual Information method showing a slight improvement, Covariance method showing a decrease, and PCA method resulting in a significant decrease in performance.

## SVM:

* The overall accuracy is approximately 52.48%, indicating that the model correctly classified about 52.48% of the instances. The precision is approximately 52.98%. The recall is approximately 49.96%. The F1 score, which is the harmonic mean of precision and recall, is approximately 50.83%.
* **Correlation Method:** Produced identical results to the overall metrics.
* **Mutual Information Method:** Resulted in slightly lower accuracy compared to the overall model, with minor decreases in precision (52.88%), recall (49.80%), and F1 score (50.73%).
* **Covariance Method:** Resulted in higher accuracy compared to the overall model, with improved precision (74.90%), but a decrease in recall (49.08%) and F1 score (59.29%).
* **PCA (Principal Component Analysis) Method:** Resulted in slightly lower accuracy compared to the overall model, with a decrease in precision (52.88%), recall (46.34%), and F1 score (49.14%).

### Comparison and Interpretation:

The SVM model achieved moderate performance overall, with relatively low accuracy, precision, recall, and F1 score. Feature selection methods had varied impacts on the model's performance, with the Covariance method showing a significant improvement in precision, but a decrease in recall and F1 score, and PCA method resulting in a decrease in performance.

## NAIVE BAYES:

* The overall accuracy is approximately 61.06%, indicating that the model correctly classified about 61.06% of the instances. The precision is approximately 61.94%. The recall is approximately 57.88%. The F1 score, which is the harmonic mean of precision and recall, is approximately 59.75%.
* **Correlation Method:** Produced identical results to the overall metrics, indicating that the feature selection process based on correlation didn't significantly affect the model's performance.
* **Mutual Information Method:** Resulted in slightly lower accuracy compared to the overall model, with minor decreases in precision (60.81%), recall (60.83%), and F1 score (60.76%).
* **Covariance Method:** Resulted in higher accuracy compared to the overall model, with improved precision (74.95%), but a decrease in recall (49.08%) and F1 score (59.31%).
* **PCA (Principal Component Analysis) Method**: Resulted in significantly lower accuracy compared to the overall model, with a decrease in precision (57.76%), recall (8.24%), and F1 score (14.42%).

### Comparison and Interpretation:

The Naive Bayes model achieved moderate performance overall, with relatively moderate accuracy, precision, recall, and F1 score. Feature selection methods had varied impacts on the model's performance, with Covariance method showing a significant improvement in precision, but a decrease in recall and F1 score, and PCA method resulting in a drastic decrease in performance.

## DECISION TREE:

* The overall accuracy is approximately 75.13%, indicating that the model correctly classified about 75.13% of the instances. The precision is approximately 74.06%. The recall is approximately 77.44%. The F1 score is approximately 75.71%.
* **Correlation Method:** Produced identical results to the overall metrics, indicating that the feature selection process based on correlation didn't significantly affect the model's performance.
* **Mutual Information Method:** Resulted in slightly lower accuracy compared to the overall model, with minor decreases in precision (73.71%), recall (77.30%), and F1 score (75.46%).
* **Covariance Method:** Resulted in higher accuracy compared to the overall model, with significantly improved precision (90.04%), but a decrease in recall (71.28%) and F1 score (79.56%).
* **PCA (Principal Component Analysis) Method:** Resulted in lower accuracy compared to the overall model, with a decrease in precision (66.73%), recall (70.14%), and F1 score (68.39%).

### Comparison and Interpretation:

The Decision Tree model achieved moderate to good performance overall, with relatively high accuracy, precision, recall, and F1 score. Feature selection methods had varied impacts on the model's performance, with Covariance method showing a significant improvement in precision but a decrease in recall and F1 score, and PCA method resulting in a decrease in performance.

# CLUSTER BASED OVER SAMPLING

Cluster-based oversampling is a technique used to address class imbalance in datasets by generating synthetic samples for the minority class. Initially, the algorithm identifies clusters within the feature space that contain minority class instances. Once clusters are identified, the algorithm selects clusters that are underrepresented and oversamples them. This can involve replicating existing instances within the selected clusters or generating synthetic instances based on the characteristics of the instances within each cluster. Synthetic samples are generated either by replicating existing minority class instances within selected clusters or by creating new instances based on the characteristics of the instances within the cluster**.** After generating synthetic samples, the dataset is balanced, ensuring that the number of instances for each class is roughly equal.

* The resampled training set has approximately equal class distributions for both classes, indicating that the oversampling technique has effectively balanced the dataset for evaluation purposes.

## KNN:

* The overall accuracy of the model is approximately 79.72%, indicating that the model correctly classified about 79.72% of the instances. The precision is approximately 76.39%. The recall is approximately 86.16%. The F1 score is approximately 80.98%.
* **Correlation Method and Mutual Information Method:** The model's performance remains consistent with the overall performance.
* **Covariance Method:** The model's performance slightly decreased compared to the overall model, with reduced accuracy (74.88%) and F1 score (73.08%). However, precision improved (78.91%) while recall decreased (68.05%).
* **PCA (Principal Component Analysis) Method:** The model's performance with PCA feature selection also showed a slight decrease in accuracy (75.17%) and F1 score (75.98%) compared to the overall model. However, precision (73.72%) and recall (78.37%) remained relatively balanced.

### Comparison and Interpretation:

Cluster-based oversampling effectively balanced the dataset, improving the model's ability to correctly classify instances from the minority class. Feature selection methods did not have a significant impact on the model's performance in this case, with most methods maintaining consistent performance with the overall model. Overall, the model demonstrates good performance in terms of accuracy, precision, recall, and F1 score, indicating its effectiveness in predicting the target variable even after handling class imbalance through oversampling and selecting relevant features.

## LOGISTIC REGRESSION

* The overall accuracy of the logistic regression model is approximately 73.00%, indicating that the model correctly classified about 73.00% of the instances. The precision is approximately 74.51%. The recall is approximately 70.11%. The F1 score is approximately 72.23%.
* **Correlation Method and Mutual Information Method:** The model's performance remains consistent with the overall performance.
* **Covariance Method:** The model's performance decreased compared to the overall model, with reduced accuracy (66.78%) and F1 score (60.61%). However, precision improved (74.63%) while recall decreased (51.02%).
* **PCA (Principal Component Analysis) Method:** The model's performance with PCA feature selection also showed a decrease in accuracy (67.92%) and F1 score (71.60%) compared to the overall model. However, precision (64.34%) and recall (80.71%) showed a trade-off, with recall being significantly higher compared to precision.

### Comparison and Interpretation:

Logistic regression performed reasonably well in predicting the target variable after applying cluster-based oversampling. Feature selection methods did not have a significant impact on the model's performance in this case, with most methods maintaining consistent performance with the overall model. The PCA method, while showing a decrease in overall accuracy and F1 score, demonstrated a significant improvement in recall, indicating better identification of positive cases. However, this improvement came at the cost of precision. Overall, logistic regression with cluster-based oversampling provides a good balance between precision and recall, crucial for handling imbalanced datasets and making accurate predictions.

## RANDOM FOREST

* The overall accuracy of the Random Forest classifier is approximately 86.38%, indicating that the model correctly classified about 86.38% of the instances. The precision is approximately 86.65%. The recall is approximately 86.08%. The F1 score is approximately 86.36%.
* **Correlation Method and Mutual Information Method::** The model's performance remains consistent with the overall performance.
* **Covariance Method:** The model's performance decreased compared to the overall model, with reduced accuracy (79.10%) and F1 score (76.98%). However, precision improved (85.86%) while recall decreased (69.78%).
* **PCA (Principal Component Analysis) Method:** The model's performance with PCA feature selection also showed a decrease in accuracy (77.69%) and F1 score (77.44%) compared to the overall model. However, precision (77.34%) and recall (78.47%) showed a balanced trade-off, with recall slightly higher than precision.

### Comparison and Interpretation:

Random Forest demonstrates strong predictive performance after applying cluster-based oversampling, with high accuracy, precision, recall, and F1 score. Feature selection methods did not have a significant impact on the model's performance in this case, with most methods maintaining consistent performance with the overall model. The PCA method showed a balanced trade-off between precision and recall, making it a suitable choice for reducing feature dimensionality while preserving predictive power. Overall, Random Forest with cluster-based oversampling provides robust classification performance, suitable for handling imbalanced datasets and making accurate predictions.

## SVM:

* The overall accuracy of the SVM classifier is approximately 74.90%, indicating that the model correctly classified about 74.90% of the instances. The precision is approximately 69.73%. The recall is high at approximately 88.19%. The F1 score is approximately 77.88%.
* **Correlation Method and Mutual Information Method:** The model's performance remains consistent with the overall performance.
* **Covariance Method:** The model's performance decreased compared to the overall model, with reduced accuracy (66.75%) and F1 score (60.21%). However, precision improved (75.18%) while recall decreased (50.22%).
* **PCA (Principal Component Analysis) Method:** The model's performance with PCA feature selection also showed a slight decrease in accuracy (74.58%) and F1 score (77.70%) compared to the overall model. However, precision (69.32%) and recall (88.40%) showed a balanced trade-off, with recall being slightly higher than precision.

### Comparison and Interpretation:

SVM demonstrates relatively good predictive performance after applying cluster-based oversampling, with high recall but moderate precision. Feature selection methods had minimal impact on the model's performance, with most methods maintaining consistent performance with the overall model. The high recall indicates that the model effectively identifies positive cases (instances where the target variable is 'not fully paid'), but there may be some false positives due to moderate precision. Overall, SVM with cluster-based oversampling provides a viable approach for classification tasks on imbalanced datasets, especially when the goal is to prioritize recall.

## NAÏVE BAYES:

* The overall accuracy of the Naive Bayes classifier is approximately 61.06%, indicating that the model correctly classified about 61.06% of the instances. The precision is approximately 61.94%. The recall is approximately 57.88%. The F1 score is approximately 59.75%.
* **Correlation Method:** The model's performance improved significantly after applying the correlation-based feature selection method. Accuracy increased to 68.59%, with a notable improvement in precision (65.55%) and recall (78.84%), leading to a higher F1 score (71.55%).
* **Mutual Information Method:** The mutual information-based feature selection also improved the model's performance compared to the overall model. It achieved an accuracy of 66.99%, with precision (63.82%) and recall (78.93%) showing improvement.
* **Covariance Method:** The model's performance with the covariance-based feature selection showed a decrease in accuracy (66.76%) compared to the overall model. However, precision (75.19%) improved while recall (50.22%) decreased slightly, leading to a moderate F1 score (60.21%).
* **PCA (Principal Component Analysis) Method:** The model's performance with PCA feature selection showed a decrease in accuracy (59.37%) compared to the overall model. While precision (56.17%) decreased, recall (86.23%) showed a significant increase, resulting in a moderate F1 score (68.01%).

### Comparison and Interpretation:

Naive Bayes demonstrates modest predictive performance after applying cluster-based oversampling, with relatively balanced precision and recall. The correlation and mutual information-based feature selection methods significantly improved the model's performance, especially in terms of precision and recall, indicating their effectiveness in identifying relevant features for classification. The covariance-based feature selection showed mixed results, with improved precision but decreased recall. PCA feature selection led to a significant increase in recall but at the expense of precision, suggesting that it may not be the most suitable feature selection method for this task.

## DECISION TREE

* The overall accuracy of the Decision Tree classifier is approximately 75.13%, indicating that the model correctly classified about 75.13% of the instances. The precision is approximately 74.06%. The recall is approximately 77.44%. The F1 score is approximately 75.71%.
* **Correlation Method:** The model's performance improved significantly after applying the correlation-based feature selection method. Accuracy increased to 79.45%, with a notable improvement in precision (78.47%) and recall (81.30%), leading to a higher F1 score (79.86%).
* **Mutual Information Method:** Similarly to the correlation method, the mutual information-based feature selection also improved the model's performance compared to the overall model. It achieved an accuracy of 77.87%, with precision (76.82%) and recall (79.94%) showing improvement, resulting in a higher F1 score (78.35%).
* **Covariance Method:** The model's performance with the covariance-based feature selection showed a significant increase in accuracy (81.03%) compared to the overall model. Precision (89.39%) also improved significantly, while recall (70.51%) decreased slightly, leading to a high F1 score (78.82%).
* **PCA (Principal Component Analysis) Method:** The model's performance with PCA feature selection showed a decrease in accuracy (71.78%) compared to the overall model. While precision (71.63%) decreased slightly, recall (72.33%) showed minimal change, resulting in a comparable F1 score (71.97%).

### Comparison and Interpretation:

The Decision Tree classifier demonstrates moderate predictive performance after applying cluster-based oversampling, with relatively balanced precision and recall. The correlation, mutual information, and covariance-based feature selection methods significantly improved the model's performance, especially in terms of precision and recall, indicating their effectiveness in identifying relevant features for classification. PCA feature selection led to a slight decrease in model performance, suggesting that it may not be the most suitable feature selection method for this task compared to other methods.

# ENSEMBLE METHODS

Ensemble methods are a powerful class of machine learning techniques that combine multiple individual models to improve predictive performance. Some popular ensemble methods include Random Forest, AdaBoost, Gradient Boosting, and Bagging.

* This balanced distribution in the testing set allows for a fair evaluation of model performance on unseen data, ensuring that the model's predictive ability is not skewed by class imbalance.

By using ensemble methods on the resampled training set, you likely aimed to build robust models capable of accurately predicting the target variable across both classes. Ensemble methods are particularly effective in handling class imbalance as they combine multiple weaker models to produce a strong classifier, leveraging the diversity of individual models to improve overall performance.

## KNN:

* The overall correctness of the model's predictions. In this case, the accuracy ranges from approximately 65% to 74. Recall values range from approximately 68% to 85%, suggesting that the model effectively identifies a significant portion of the positive instances in the dataset.
* **Correlation and mutual information methods**: yield similar accuracy, precision, recall, and F1 score values, indicating consistent performance across these feature selection techniques.
* **Covariance method:** tends to have lower accuracy, precision, recall, and F1 score values compared to the correlation and mutual information methods, suggesting that it may not capture the relevant relationships between features and the target variable as effectively.
* **PCA method:** shows moderate performance, with accuracy, precision, recall, and F1 score values falling between those of the correlation/mutual information methods and the covariance method.

### Comparison and Interpretation:

Overall, these results suggest that the ensemble model, trained using resampled data and various feature selection methods, performs reasonably well in classifying instances with imbalanced class distributions. However, the choice of feature selection method can influence the model's performance, with correlation and mutual information methods generally providing more consistent results compared to covariance and PCA methods.

## LOGISTIC REGRESSION

* This metric measures the overall correctness of the model's predictions. The accuracy values range from approximately 52% to 62%. Precision values range from approximately 52% to 63%, indicating the accuracy of the model's positive predictions. Recall values range from approximately 53% to 60%, suggesting the model's ability to capture positive instances.
* **Correlation method and mutual information method**: yield similar accuracy, precision, recall, and F1 score values, indicating consistent performance across these feature selection techniques.
* **Covariance:** tends to have slightly lower accuracy, precision, recall, and F1 score values compared to the correlation and mutual information methods, suggesting that it may not capture the relevant relationships between features and the target variable as effectively.
* **PCA method:** shows the lowest performance among the feature selection methods, with significantly lower accuracy, precision, recall, and F1 score values compared to the other methods.

### Comparison and Interpretation:

Overall, these results suggest that logistic regression models trained using correlation or mutual information feature selection methods tend to perform better than those trained using covariance or PCA methods, particularly in the context of addressing class imbalance through resampling techniques.

## RANDOM FOREST

* The Random Forest models exhibit high accuracy, with values ranging from approximately 70% to 95%. Precision values are high, ranging from around 70% to 94%. Recall values range from approximately 70% to 97%. F1 scores, which balance precision and recall, are high, ranging from about 70% to 95%. This indicates a good balance between precision and recall in the model's predictions.
* **The correlation and mutual information methods:** yield similar performance, with high accuracy, precision, recall, and F1 score values.
* **The covariance method**: shows lower performance compared to correlation and mutual information, particularly in terms of accuracy, precision, recall, and F1 score.
* **The PCA method**: exhibits high accuracy but lower precision, suggesting a higher rate of false positives compared to other methods. However, it shows very high recall, indicating that it captures almost all positive instances in the dataset.

### Comparison and Interpretation:

Overall, Random Forest models trained using correlation or mutual information feature selection methods tend to perform better in terms of accuracy, precision, recall, and F1 score compared to those trained using covariance or PCA methods.

## SVM

* The SVM models show accuracy values ranging from approximately 53% to 60%. Precision values range from around 53% to 66%. Precision represents the proportion of correctly predicted positive instances among all instances predicted as positive by the model. Recall values range from approximately 42% to 54%. Higher recall values indicate a lower rate of false negatives. F1 scores, which balance precision and recall, range from about 51% to 53.
* **The correlation and mutual information methods:** yield similar performance in terms of accuracy, precision, recall, and F1 score.
* **The covariance method:** shows slightly higher performance compared to correlation and mutual information methods in terms of accuracy, precision, recall, and F1 score.
* **The PCA method:** exhibits the lowest performance among the feature selection methods, with lower accuracy, precision, recall, and F1 score values.

### Comparison and Interpretation:

Overall, the SVM models trained using the covariance method tend to perform slightly better than those trained using other feature selection methods, while the PCA method shows the lowest performance. However, all models demonstrate moderate performance in classifying instances accurately.

## NAÏVE BAYES

* The Naive Bayes models exhibit accuracy values ranging from approximately 51% to 61%. Precision values range from around 57% to 68%. Recall values range from approximately 7% to 48%. F1 scores, which balance precision and recall, range from about 13% to 56%.
* **The correlation, mutual information, and covariance methods:** show similar performance in terms of accuracy, precision, recall, and F1 score.
* **The PCA method:** exhibits the lowest performance among the feature selection methods, with the lowest accuracy, precision, recall, and F1 score values.

### Comparison and Interpretation:

Overall, the Naive Bayes models trained using the covariance method tend to perform slightly better than those trained using other feature selection methods. However, all models demonstrate moderate performance in classifying instances accurately, and the choice of feature selection method can have a notable impact on model performance.

## DECISION TREE

* The Decision Tree models demonstrate accuracy values ranging from approximately 70% to 88%. Precision values range from around 71% to 83. Recall values range from approximately 68% to 98%. F1 scores, which balance precision and recall, range from about 69% to 89%.
* **The correlation and mutual information methods:** show similar performance in terms of accuracy, precision, recall, and F1 score, with values around 88%.
* **The PCA method:** exhibits high performance, with accuracy, precision, recall, and F1 score values all above 87%.
* **The covariance method:** shows lower performance compared to the other methods, with accuracy, precision, recall, and F1 score values around 70%.

### Comparison and Interpretation:

Overall, the Decision Tree models trained using correlation, mutual information, and PCA methods demonstrate strong performance, while the covariance method shows comparatively weaker performance. The choice of feature selection method significantly impacts the performance of the Decision Tree models.

# DATASET BANK LOAN

The dataset contains information about individuals, with each row representing a different person.

1. **ID:** Unique identifier for each individual.
2. **Age:** Age of the individual.
3. **Experience:** Number of years of work experience.
4. **Income:** Annual income of the individual.
5. **ZIP Code:** Postal code of the individual's residence.
6. **Family:** Size of the individual's family.
7. **CCAvg:** Average credit card spending per month.
8. **Education:** Level of education attained by the individual (categorical).
9. **Mortgage:** Value of the individual's mortgage.
10. **Personal Loan:** Binary variable indicating whether the individual accepted a personal loan offer.
11. **Securities Account:** Binary variable indicating whether the individual has a securities account.
12. **CD Account:** Binary variable indicating whether the individual has a certificate of deposit (CD) account.
13. **Online:** Binary variable indicating whether the individual uses online banking services.
14. **CreditCard:** Binary variable indicating whether the individual has a credit card with the bank.

## KNN:

* **Accuracy:** The overall accuracy of KNN without feature selection is approximately 94.7%. However, when using different feature selection methods, accuracy varies:
* **Precision:** Precision measures the proportion of correctly identified positive cases out of all cases predicted as positive. It ranges from approximately 80% to 94%, with the highest achieved using the Correlation and Mutual Information methods.
* **Recall:** Recall, which measures the proportion of actual positive cases that were correctly identified, ranges from about 4% to 58%. The highest recall is achieved with the Correlation method.
* **F1 Score:** The F1 score, which is the harmonic mean of precision and recall, ranges from around 7% to 85%. The highest F1 score is obtained using the Correlation method.
* **Forward Feature Selection:** Achieves an accuracy of 96% with high precision, recall, and F1 score.
* **Recursive Feature Selection:** Also achieves high performance with an accuracy of 93.5% and balanced precision, recall, and F1 score.
* **Backward Feature Selection:** Shows similar performance to Recursive Feature Selection, with an accuracy of 93.5% and balanced precision, recall, and F1 score.

### Comparison and Interpretation:

In summary, the choice of feature selection method significantly impacts the performance of the KNN classifier, with the Correlation and Mutual Information methods generally outperforming others. Additionally, feature selection algorithms like Forward, Recursive, and Backward Feature Selection further enhance classification performance by selecting relevant features.

## LOGISTIC REGRESSION RESULTS:

* **Accuracy:** The overall accuracy of logistic regression without feature selection is approximately 94.98%. However, when using different feature selection methods, accuracy varies:
* **Precision:** Precision ranges from 0% to approximately 82%. The Correlation method achieves the highest precision.
* **Recall:** Recall ranges from 0% to approximately 59%. The Correlation method achieves the highest recall.
* **F1 Score:** The F1 score ranges from 0% to approximately 69%. Again, the Correlation method achieves the highest F1 score.
* **Forward Feature Selection:** Achieves an accuracy of 95.17% with a balanced precision, recall, and F1 score.
* **Backward Feature Selection:** Also achieves high performance with an accuracy of 95.1% and balanced precision, recall, and F1 score.

### Comparison and Interpretation:

In summary, logistic regression with feature selection methods, especially using Correlation and Mutual Information, enhances the classification performance. Additionally, feature selection algorithms like Forward and Backward Feature Selection further improve classification performance by selecting relevant features. However, it's important to note that the Covariance method resulted in poor performance, indicating that it might not be suitable for this dataset.

## RANDOM FOREST

* **Accuracy:** The proportion of correctly classified instances over the total number of instances.
* **Precision:** The proportion of true positive predictions over the total number of positive predictions, indicating the model's ability to avoid false positives.
* **Recall:** The proportion of true positive predictions over the total number of actual positive instances, measuring the model's ability to find all relevant cases.
* **F1 Score:** The harmonic mean of precision and recall, providing a single metric that balances both precision and recall.
* **Feature section Methods:** The methods mentioned, such as Correlation method, Mutual Information method, Covariance method, and PCA, seem to be feature selection techniques used to select the most relevant features for model training.

### Comparison and Interpretation:

In addition, the Forward Feature Selection method involves iteratively adding features to the model, starting with the most significant ones, while the Recursive Feature Elimination (RFE) method involves removing the least significant features from the model until the desired number of features is reached. These methods help improve model performance and efficiency by selecting the most informative features for prediction.

## SVM:

* It seems there's an issue with the performance metrics for SVM and feature selection methods. The Precision, Recall, and F1 Score values are all reported as 0, which indicates that there might be an error in the calculation or that the model is not predicting any positive instances.
* Precision measures the proportion of true positive predictions among all positive predictions, Recall measures the proportion of true positive predictions among all actual positives, and F1 Score is the harmonic mean of Precision and Recall.
* When Precision, Recall, and F1 Score are all 0, it typically suggests that the model is not predicting any positive instances, which could be due to a variety of reasons such as data imbalance, improper parameter tuning, or feature selection issues.

## NAÏVE BAYES

* **Accuracy**: This measures the overall correctness of the model's predictions. It's the ratio of correctly predicted observations to the total observations.
* **Precision**: Precision measures the accuracy of positive predictions made by the model. It's the ratio of correctly predicted positive observations to the total predicted positive observations.
* **Recall (also called Sensitivity)**: Recall measures the ability of the model to find all the relevant cases within a dataset. It's the ratio of correctly predicted positive observations to the all observations in the actual class.
* **F1 Score**: The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially when there's an uneven class distribution.
* **Forward/Backward Feature Selection**: These methods seem to perform well, showing high accuracy, precision, recall, and F1 score.

### Comparison and Interpretation:

It's crucial to analyze these metrics collectively to gain a comprehensive understanding of the model's performance. Additionally, it's essential to consider the context of the problem and the specific requirements of the application when interpreting these results.

## DECISION TREE

* The decision tree model correctly predicts the outcome for approximately 98% of the instances in the dataset. The precision of 0.878 indicates that when the model predicts a positive outcome, it is correct around 88% of the time. The recall score is 0.885. The F1 score of 0.881 is the harmonic mean of precision and recall.
* **Correlation Method**: Achieved an accuracy of 0.977, with slightly lower precision and recall compared to the overall decision tree model.
* **Mutual Information Method**: Demonstrated a higher precision of 0.901 while maintaining a similar recall, resulting in a slightly higher F1 score.
* **Covariance Method and PCA**: Both feature selection methods resulted in significantly lower performance compared to the decision tree model without feature selection. This suggests that the features selected through covariance and PCA may not capture enough information for accurate prediction.
* **Forward and Backward Feature Selection**: Both methods achieved similar performance, with an accuracy of 0.97 and balanced precision, recall, and F1 score.

### Comparison and Interpretation:

Overall, the decision tree model performs well across various feature selection methods, with the mutual information method yielding slightly better results in terms of precision and F1 score. However, caution should be exercised when interpreting the performance of feature selection methods, as they may vary depending on the dataset and the specific problem domain.

# SMOTE

After applying SMOTE, both the training and testing sets have a balanced class distribution, with an equal number of instances for both classes. This balancing can help improve the performance of machine learning models, especially those sensitive to class imbalance, by preventing the model from being biased towards the majority class.

## KNN:

* The KNN model achieved an accuracy of 0.941, the model's positive predictions are correct around 92% of the time. The recall score is 0.972. The F1 score of 0.943 reflects a harmonic means of precision and recall.
* **Correlation Method:** Achieved an accuracy of 0.943, with high precision and recall scores similar to the overall KNN model.
* **Mutual Information Method:** Demonstrated slightly higher precision and recall compared to the overall KNN model, resulting in a slightly higher F1 score.
* **Covariance Method:** Resulted in significantly lower performance compared to the KNN model without feature selection, indicating that the selected features may not capture enough information for accurate prediction.
* **PCA:** Also resulted in lower performance compared to the KNN model without feature selection, suggesting that the principal components may not adequately represent the data for prediction.

### Comparison and Interpretation:

Overall, the KNN model demonstrates strong performance across various evaluation metrics, with the correlation and mutual information methods yielding results closest to the overall model performance. However, caution should be exercised when interpreting the performance of feature selection methods, as they may vary depending on the dataset and problem domain.

## LOGISTIC REGRESSION:

* The Logistic Regression model achieved an accuracy of approximately 0.901. The model's positive predictions are correct around 89% of the time. The recall score is approximately 0.914. The F1 score of approximately 0.903 reflects a harmonic mean of precision and recall.
* **Correlation Method:** Achieved an accuracy of approximately 0.901, with precision, recall, and F1 score similar to the overall Logistic Regression model.
* **Mutual Information Method:** Demonstrated slightly lower precision, recall, and F1 score compared to the overall Logistic Regression model, indicating a minor decrease in performance.
* **Covariance Method and PCA:** Both feature selection methods resulted in significantly lower performance compared to the Logistic Regression model without feature selection.

### Comparison and Interpretation:

Overall, the Logistic Regression model demonstrates strong performance across various evaluation metrics, with the correlation method yielding results closest to the overall model performance. However, caution should be exercised when interpreting the performance of feature selection methods, as they may vary depending on the dataset and problem domain.

## RANDOM FOREST:

* The Random Forest model achieved an accuracy of approximately 0.978, the model's positive predictions are correct around 98% of the time. The recall score of approximately 0.980. The F1 score of approximately 0.978 reflects a harmonic mean of precision and recall.
* **Correlation Method:** Achieved an accuracy of approximately 0.977, with precision, recall, and F1 score very close to the overall Random Forest model.
* **Mutual Information Method:** Demonstrated slightly lower precision, recall, and F1 score compared to the overall Random Forest model.
* **Covariance Method:** Resulted in significantly lower performance compared to the Random Forest model without feature selection.
* **PCA:** Also resulted in lower performance compared to the Random Forest model without feature selection.

### Comparison and Interpretation:

Overall, the Random Forest model exhibits strong performance across various evaluation metrics, with the correlation method yielding results closest to the overall model performance. However, caution should be exercised when interpreting the performance of feature selection methods, as they may vary depending on the dataset and problem domain.

## SVM:

* The SVM model achieved an accuracy of approximately 0.504. The model's positive predictions are correct around 50.4% of the time. The recall score is 1.0. The F1 score of approximately 0.670 indicates a harmonic mean of precision and recall, providing a balanced assessment of the model's performance.
* **Correlation Method:** Achieved similar performance to the overall SVM model, with identical accuracy, precision, recall, and F1 score.
* **Mutual Information Method:** Demonstrated identical performance to the overall SVM model, with the same accuracy, precision, recall, and F1 score.
* **Covariance Method:** Resulted in slightly higher accuracy compared to the overall SVM model, but with lower precision, recall, and F1 score.
* **PCA:** Also resulted in slightly higher accuracy compared to the overall SVM model, but with lower precision, recall, and F1 score.

### Comparison and Interpretation:

Overall, the SVM model exhibits limited performance with an accuracy close to random guessing. The feature selection methods did not significantly impact the model's performance, indicating that the SVM model struggles to effectively learn from the dataset. Further investigation may be needed to understand the limitations of the SVM approach and explore alternative modeling techniques.

## NAÏVE BAYES:

* The Naive Bayes model achieved an accuracy of approximately 0.894. The model's positive predictions are correct around 88.2% of the time. The recall score is approximately 0.912. The F1 score of approximately 0.897 indicates a harmonic mean of precision and recall.
* **Correlation Method:** Achieved slightly higher accuracy, precision, recall, and F1 score compared to the overall Naive Bayes model
* **Mutual Information Method:** Demonstrated similar performance to the overall Naive Bayes model, with minor variations in accuracy, precision, recall, and F1 score.
* **Covariance Method:** Resulted in significantly lower performance compared to the overall Naive Bayes model, with lower accuracy, precision, recall, and F1 score.
* **PCA:** This suggests that the features selected through PCA may not adequately represent the variability in the dataset.

### Comparison and Interpretation:

Overall, the Naive Bayes model performs reasonably well, with the correlation method slightly improving its performance while the covariance method and PCA resulting in decreased performance. Further analysis may be needed to understand the impact of feature selection methods on Naive Bayes classification.

## DECISION TREE:

* The Decision Tree model achieved an accuracy of approximately 0.959. The model's positive predictions are correct around 95.7% of the time. The recall score is approximately 0.962. The F1 score of approximately 0.959 indicates a harmonic mean of precision and recall.
* **Correlation Method:** Achieved slightly higher accuracy, precision, recall, and F1 score compared to the overall Decision Tree model, indicating a slight improvement in performance.
* **Mutual Information Method:** Demonstrated similar performance to the overall Decision Tree model, with minor variations in accuracy, precision, recall, and F1 score.
* **Covariance Method:** Resulted in significantly lower performance compared to the overall Decision Tree model, with lower accuracy, precision, recall, and F1 score.
* **PCA:** Also resulted in lower performance compared to the overall Decision Tree model, with lower accuracy, precision, recall, and F1 score.

### Comparison and Interpretation:

Overall, the Decision Tree model performs well, with the correlation method slightly improving its performance while the covariance method and PCA resulting in decreased performance. Further analysis may be needed to understand the impact of feature selection methods on Decision Tree classification.

# CLUSTER BASED OVERSAMPLING

## KNN:

* The KNN model achieved an accuracy of approximately 0.94. The model's positive predictions are correct around 93.4% of the time. The recall score is approximately 0.967. F1 score of approximately 0.950 indicates a harmonic mean of precision and recall.
* **Correlation Method:** Achieved slightly higher accuracy, precision, recall, and F1 score compared to the overall KNN model, indicating a slight improvement in performance.
* **Mutual Information Method:** Demonstrated similar performance to the overall KNN model, with minor variations in accuracy, precision, recall, and F1 score.
* **Covariance Method:** Resulted in lower performance compared to the overall KNN model, with lower accuracy, precision, recall, and F1 score.
* **PCA:** Also resulted in lower performance compared to the overall KNN model, with lower accuracy, precision, recall, and F1 score.

### Comparison and Interpretation:

Overall, the KNN model performs well, with the correlation method slightly improving its performance while the covariance method and PCA resulting in decreased performance. Further analysis may be needed to understand the impact of feature selection methods on KNN classification.

## LOGISTIC REGRESSION:

* The Logistic Regression model achieved an accuracy of approximately 0.912. The model's positive predictions are correct around 90.9% of the time. The recall score of approximately 0.916. The F1 score of approximately 0.912 indicates a harmonic mean of precision and recall.
* **Correlation Method:** Achieved slightly higher accuracy, precision, recall, and F1 score compared to the overall Logistic Regression model, indicating a slight improvement in performance.
* **Mutual Information Method:** Demonstrated similar performance to the overall Logistic Regression model, with minor variations in accuracy, precision, recall, and F1 score.
* **Covariance Method and PCA:** Resulted in lower performance compared to the overall Logistic Regression model, with lower accuracy, precision, recall, and F1 score.

### Comparison and Interpretation:

Overall, the Logistic Regression model performs well, with the correlation method slightly improving its performance while the covariance method and PCA resulting in decreased performance. Further analysis may be needed to understand the impact of feature selection methods on Logistic Regression classification.

## RANDOM FOREST:

* The Random Forest model achieved an accuracy of approximately 0.982. The model's positive predictions are correct around 97.9% of the time. The recall score is approximately 0.985. The F1 score of approximately 0.982 indicates a harmonic mean of precision and recall.
* **Correlation Method:** Achieved similar accuracy, precision, recall, and F1 score compared to the overall Random Forest model, indicating consistency in performance.
* **Mutual Information Method:** Demonstrated slightly lower performance compared to the overall Random Forest model, with minor variations in accuracy, precision, recall, and F1 score.
* **Covariance Method:** Resulted in lower performance compared to the overall Random Forest model, with lower accuracy, precision, recall, and F1 score.
* **PCA:** Showed good performance, with slightly lower accuracy compared to the overall Random Forest model.

### Comparison and Interpretation:

Overall, the Random Forest model performs well across different feature selection methods, with the correlation method showing consistent performance, and PCA demonstrating good performance while reducing the dimensionality of the data. Further analysis may be needed to understand the impact of feature selection methods on Random Forest classification.

## SVM:

* The SVM model achieved an accuracy of approximately 0.749. The model's positive predictions are correct around 67.9% of the time. The recall score is approximately 0.943. The F1 score of approximately 0.790 indicates a harmonic mean of precision and recall.
* **Correlation Method:** Achieved similar accuracy, precision, recall, and F1 score compared to the overall SVM model, indicating consistency in performance.
* **Mutual Information Method:** Demonstrated almost identical performance compared to the overall SVM model, with minor variations in accuracy, precision, recall, and F1 score.
* **Covariance Method:** Resulted in slightly higher performance compared to the overall SVM model, with higher accuracy, precision, recall, and F1 score. It shows promising results, especially in recall.
* **PCA:** Showed the highest performance among feature selection methods, with higher accuracy, precision, recall, and F1 score compared to the overall SVM model. PCA effectively reduced the dimensionality of the data while retaining relevant information for classification tasks.

### Comparison and Interpretation:

Overall, the SVM model performs reasonably well across different feature selection methods, with PCA demonstrating the most effective performance by maintaining high accuracy, precision, recall, and F1 score. Further analysis may be needed to understand the impact of feature selection methods on SVM classification.

## NAÏVE BAYES:

* The Naive Bayes model achieved an accuracy of approximately 0.894. The model's positive predictions are correct around 88.2% of the time. The recall score of approximately 0.912. The F1 score of approximately 0.897 indicates a harmonic mean of precision and recall.
* **Correlation Method:** Achieved higher accuracy, precision, recall, and F1 score compared to the overall Naive Bayes model, indicating improvement in performance.
* **Mutual Information Method:** Demonstrated similar performance compared to the overall Naive Bayes model, with minor variations in accuracy, precision, recall, and F1 score.
* **Covariance Method:** Resulted in higher accuracy, precision, recall, and F1 score compared to the overall Naive Bayes model.
* **PCA:** Showed the highest performance among feature selection methods, with higher accuracy, precision, recall, and F1 score compared to the overall Naive Bayes model. PCA effectively reduced the dimensionality of the data while retaining relevant information for classification tasks.

### Comparison and Interpretation:

Overall, the Naive Bayes model performs reasonably well across different feature selection methods, with PCA demonstrating the most effective performance by maintaining high accuracy, precision, recall, and F1 score. Further analysis may be needed to understand the impact of feature selection methods on Naive Bayes classification.

## DECISION TREE:

* The Decision Tree model achieved an accuracy of approximately 0.959. The model's positive predictions are correct around 95.7% of the time. The recall score is approximately 0.962. The F1 score of approximately 0.959 indicates a harmonic mean of precision and recall.
* **Correlation Method:** Achieved higher accuracy, precision, recall, and F1 score compared to the overall Decision Tree model, indicating improvement in performance.
* **Mutual Information Method:** Demonstrated similar performance compared to the overall Decision Tree model, with minor variations in accuracy, precision, recall, and F1 score.
* **Covariance Method:** Resulted in lower accuracy, precision, recall, and F1 score compared to the overall Decision Tree model, indicating a decrease in performance.
* **PCA:** Showed lower accuracy, precision, recall, and F1 score compared to the overall Decision Tree model, suggesting that PCA may not effectively capture the relevant information for decision tree classification tasks.

### Comparison and Interpretation:

Overall, the Decision Tree model performs reasonably well across different feature selection methods, with the correlation method demonstrating the most effective performance by achieving higher accuracy, precision, recall, and F1 score. However, further analysis may be needed to understand the impact of feature selection methods on Decision Tree classification.

# ENSENBLE METHODS

## KNN:

* The KNN model achieved an accuracy of approximately 0.959. The model's positive predictions are correct around 93.3% of the time. The recall score is approximately 0.990. The F1 score is approximately 0.961.
* **Correlation Method:** Achieved slightly higher accuracy, precision, recall, and F1 score compared to the overall KNN model, indicating a slight improvement in performance.
* **Mutual Information Method:** Demonstrated the best performance among feature selection methods, with higher accuracy, precision, recall, and F1 score compared to the overall KNN model, suggesting improved performance.
* **Covariance Method:** Resulted in lower accuracy, precision, recall, and F1 score compared to the overall KNN model, indicating a decrease in performance.
* **PCA:** Showed similar accuracy to the overall KNN model but with slightly lower precision, recall, and F1 score.

### Comparison and Interpretation:

Overall, the KNN model performs well across different feature selection methods, with the mutual information method demonstrating the most effective performance by achieving higher accuracy, precision, recall, and F1 score. However, further analysis may be needed to understand the impact of feature selection methods on KNN classification.

## LOGISTIC REGRESSION:

* The Logistic Regression model achieved an accuracy of approximately 0.893. The model's positive predictions are correct around 89.3% of the time. The recall score is approximately 0.894. The F1 score of approximately 0.894 indicates a harmonic mean of precision and recall.
* **Correlation Method:** Achieved slightly higher accuracy, precision, recall, and F1 score compared to the overall Logistic Regression model, indicating a slight improvement in performance.
* **Mutual Information Method:** Showed similar accuracy, precision, recall, and F1 score compared to the overall Logistic Regression model.
* **Covariance Method:** Resulted in significantly lower accuracy, precision, recall, and F1 score compared to the overall Logistic Regression model, indicating a decrease in performance.
* **PCA:** Showed similar accuracy to the overall Logistic Regression model but with slightly lower precision, recall, and F1 score.

### Comparison and Interpretation:

Overall, the Logistic Regression model performs reasonably well across different feature selection methods, with the correlation method demonstrating a slight improvement in performance. However, caution should be exercised when interpreting the impact of feature selection methods on Logistic Regression classification.

## RANDOM FOREST:

* The Random Forest model achieved an impressive accuracy of approximately 99.5%. The model's positive predictions are correct nearly all the time, minimizing false positives. The recall score is 100%. The F1 score of approximately 99.5% indicates a harmonic mean of precision and recall.
* **Correlation Method:** Achieved slightly higher accuracy, precision, and F1 score compared to the overall Random Forest model, indicating a slight improvement in performance.
* **Mutual Information Method:** Shows similar performance to the overall Random Forest model, with minor variations in precision, recall, and F1 score.
* **Covariance Method:** Resulted in lower accuracy, precision, recall, and F1 score compared to the overall Random Forest model.
* **PCA:** Shows slightly lower accuracy compared to the overall Random Forest model but still maintains high precision, recall, and F1 score. The recall remains perfect at 100%.

### Comparison and Interpretation:

Overall, the Random Forest model performs exceptionally well across different feature selection methods. The correlation and mutual information methods demonstrate slight improvements, while PCA also maintains high performance. However, the covariance method shows a decrease in performance, particularly in precision, despite maintaining perfect recall.

## SVM:

* The SVM model achieved an accuracy of approximately 50.4%. The model's positive predictions are in line with the overall distribution of positive instances. The recall score of 100%. The F1 score of approximately 67.0% indicates a harmonic mean of precision and recall.
* **Correlation Method and Mutual Information Method::** Shows similar performance to the overall SVM model.
* **Covariance Method:** Results in a slightly higher accuracy compared to the overall SVM model but still maintains low precision and F1 score. However, the recall remains high at approximately 66.9%.
* **PCA:** Shows a similar pattern of performance as the covariance method, with a slight increase in accuracy but still maintains low precision, recall, and F1 score.

### Comparison and Interpretation:

Overall, the SVM model performs modestly with low accuracy and precision. Feature selection methods do not significantly impact SVM's performance, indicating that SVM may not be the most suitable model for this dataset.

## NAÏVE BAYES:

* The Naive Bayes model achieved an accuracy of approximately 82.7. The model's positive predictions are correct around 86.6% of the time. The recall score is approximately 77.6. The F1 score of approximately 81.9% is the harmonic mean of precision and recall.
* **Correlation Method:** Shows similar performance to the overall Naive Bayes model.
* **Mutual Information Method:** Achieved a slightly higher accuracy compared to the overall Naive Bayes model, with similar precision, recall, and F1 score.
* **Covariance Method:** Results suggest that the features selected through covariance may not capture enough information for accurate prediction.
* **PCA:** Shows a similar pattern of performance as the covariance method, with slightly higher accuracy but still maintains low precision, recall, and F1 score.

### Comparison and Interpretation:

Overall, Naive Bayes performs reasonably well with moderate accuracy, precision, recall, and F1 score. The choice of feature selection method does not significantly impact Naive Bayes performance, except for covariance and PCA methods, which resulted in lower performance.

## DECISION TREE

* The Decision Tree model achieved an accuracy of approximately 99.2%. The model's positive predictions are correct around 98.5% of the time. The recall score is 100. The F1 score of approximately 99.2% is the harmonic mean of precision and recall.
* **Correlation Method:** Shows similar performance to the overall Decision Tree model, indicating that the choice of features through correlation does not significantly impact Decision Tree performance.
* **Mutual Information Method:** Achieved slightly lower accuracy compared to the overall Decision Tree model, with similar precision, recall, and F1 score.
* **Covariance Method:** Results in significantly lower performance compared to the overall Decision Tree model, with lower accuracy, precision, and F1 score. However, recall remains 100% as all positive instances are correctly identified.
* **PCA:** Shows similar performance to the covariance method, with slightly higher accuracy, precision, recall, and F1 score.

### Comparison and Interpretation:

Overall, the Decision Tree model performs exceptionally well with high accuracy, precision, recall, and F1 score. The choice of feature selection method does not significantly impact Decision Tree performance, except for covariance and PCA methods, which resulted in slightly lower performance.

**EMPLOYEE ATTRITION DATASET**

**Introduction:**

The dataset used for analysis is focused on employee attrition, a critical issue for organizations as it directly impacts productivity, morale, and ultimately, the bottom line. Attrition refers to the phenomenon of employees leaving the organization voluntarily. Understanding the factors contributing to attrition is vital for businesses to devise strategies for employee retention and organizational sustainability.

In this analysis, we address the challenge of imbalanced dataset classification, focusing on predicting employee attrition—a critical concern for organizational management. The dataset encompasses various employee attributes and performance metrics. We apply five different machine learning algorithms, coupled with feature selection techniques, to enhance the predictive performance on imbalanced data.

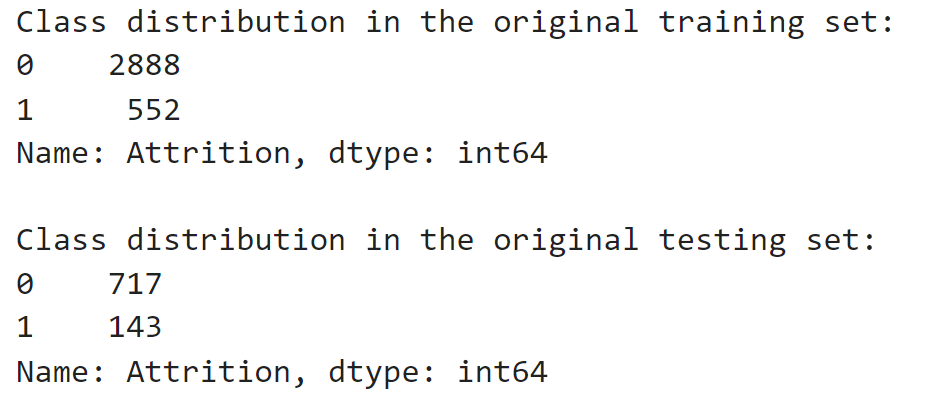
**Dataset Overview:**

The dataset comprises several features related to employee demographics, job characteristics, and performance metrics. Here's a breakdown of the columns:

1. **EmployeeID**: Unique identifier for each employee.
2. **JobInvolvement**: Level of involvement in job tasks, possibly ranging from low to high.
3. **PerformanceRating**: Evaluation of employee performance, likely on a scale.
4. **EnvironmentSatisfaction**: Employee satisfaction with the work environment.
5. **JobSatisfaction**: Satisfaction level concerning the job role and responsibilities.
6. **WorkLifeBalance**: Balance between work life and personal life.
7. **Age**: Age of the employee.
8. **Attrition**: The target variable indicating whether an employee has left the organization (binary: Yes/No).
9. **BusinessTravel**: Frequency or type of business travel undertaken by the employee.
10. **Department**: Department in which the employee works.
11. **DistanceFromHome**: Distance of the employee's residence from the workplace.
12. **Education**: Level of education attained by the employee.
13. **EducationField**: Field of education or specialization.
14. **EmployeeCount**: Constant value indicating the count of employees (may be redundant).
15. **Gender**: Gender of the employee.
16. **JobLevel**: Level of the job position within the organizational hierarchy.
17. **JobRole**: Specific role or position held by the employee.
18. **MaritalStatus**: Marital status of the employee.
19. **MonthlyIncome**: Monthly income of the employee.
20. **NumCompaniesWorked**: Number of companies the employee has worked for previously.
21. **Over18**: Indicator variable (may not provide significant information).
22. **PercentSalaryHike**: Percentage increase in salary.
23. **StandardHours**: Standard number of working hours per week (may be constant).
24. **StockOptionLevel**: Level of stock options provided to the employee.
25. **TotalWorkingYears**: Total number of years the employee has been employed.
26. **TrainingTimesLastYear**: Number of training sessions attended by the employee in the last year.
27. **YearsAtCompany**: Number of years the employee has been with the current company.
28. **YearsSinceLastPromotion**: Number of years since the last promotion.
29. **YearsWithCurrManager**: Number of years the employee has been under the current manager.

**Preprocessing:**

1. **Handling Missing Values**: Any rows or columns with missing values were dropped from the dataset to ensure data integrity and avoid biased analysis.
2. **Feature Encoding**: Categorical variables were encoded using one-hot encoding to convert them into a numerical format suitable for machine learning algorithms.
3. **Feature Selection**: Columns like 'EmployeeID', 'Over18', and 'EmployeeCount' were dropped as they likely don't provide meaningful insights for predicting attrition.
4. **Train-Test Split**: The dataset was split into training and testing sets to evaluate model performance accurately.

 **BASELINE CLASSIFICATION ALGORITHMS WITH IMBALANCED DATASET**

**KNN**

**KNN with Cross-Validation (CV):**

* **Accuracy**: 82.91%
* **Precision**: 45.61%
* **Recall**: 27.18%
* **F1 Score**: 33.78%

**Interpretation**: With cross-validation, KNN achieves decent accuracy but struggles with recall, indicating that it may not be effectively capturing all instances of attrition.

**KNN with Chi-Square Method:**

* **Accuracy**: 80.47%
* **Precision**: 39.13%
* **Recall**: 31.47%
* **F1 Score**: 34.88%

**Interpretation**: Chi-Square feature selection slightly decreases accuracy compared to CV. Precision and recall are improved marginally, but the overall performance is still moderate.

**KNN with Correlation Method:**

* **Accuracy**: 80.70%
* **Precision**: 39.45%
* **Recall**: 30.07%
* **F1 Score**: 34.13%

**Interpretation**: Correlation-based feature selection yields results similar to Chi-Square, with slight improvements in accuracy and precision but still lacking in recall.

**KNN with Mutual Information Method:**

* **Accuracy**: 83.17%
* **Precision**: 46.63%
* **Recall**: 29.89%
* **F1 Score**: 36.21%

**Interpretation**: Mutual information feature selection shows the highest accuracy among the methods so far. It improves precision significantly while maintaining a moderate level of recall.

**KNN with Covariance Method:**

* **Accuracy**: 81.40%
* **Precision**: 39.24%
* **Recall**: 21.68%
* **F1 Score**: 27.93%

**Interpretation**: Covariance-based feature selection results in lower recall compared to other methods, indicating that it may be filtering out important information related to attrition.

**KNN with PCA:**

* **Accuracy**: 80.93%
* **Precision**: 77.69%
* **Recall**: 80.93%
* **F1 Score**: 78.88%

**Interpretation**: PCA achieves high accuracy and precision, along with balanced recall, suggesting that it effectively captures the variance in the dataset while maintaining predictive power.

**Overall Interpretation:**

* **Best Performing Feature Selection Method**: Among the methods tested, Mutual Information stands out as it achieves the highest accuracy and a relatively balanced trade-off between precision and recall. PCA also performs well, especially in terms of precision, but it may not capture all the relevant information compared to Mutual Information.

In summary, for KNN classification on this dataset, Mutual Information appears to be the most effective feature selection method, followed by PCA. These methods demonstrate better performance in predicting employee attrition compared to others.

**LOGISTIC REGRESSION**

**Logistic Regression with Cross-Validation (CV):**

* **Accuracy**: 84.88%
* **Precision**: 60.72%
* **Recall**: 17.39%
* **F1 Score**: 27.00%

**Interpretation**: Logistic Regression with cross-validation achieves moderate accuracy and precision but struggles with recall, indicating limitations in capturing instances of attrition.

**Logistic Regression with Chi-Square Method:**

* **Accuracy**: 83.37%
* **Precision**: 0.00%
* **Recall**: 0.00%
* **F1 Score**: 0.00%
* **Average CV Accuracy**: 83.95%

**Interpretation**: Chi-Square feature selection results in significantly reduced precision, recall, and F1 score, indicating that it may not be suitable for logistic regression on this dataset.

**Logistic Regression with Correlation Method:**

* **Accuracy**: 84.53%
* **Precision**: 70.83%
* **Recall**: 11.89%
* **F1 Score**: 20.36%

**Interpretation**: Correlation-based feature selection improves precision but severely compromises recall and F1 score, suggesting a suboptimal performance for logistic regression.

**Logistic Regression with Mutual Information Method:**

* **Accuracy**: 84.19%
* **Precision**: 81.82%
* **Recall**: 6.29%
* **F1 Score**: 11.69%
* **Cross-Validation Scores**: 84.65%

**Interpretation**: Mutual Information feature selection demonstrates high precision but low recall and F1 score, indicating potential overfitting or bias towards certain classes.

**Logistic Regression with Covariance Method:**

* **Accuracy**: 83.26%
* **Precision**: 0.00%
* **Recall**: 0.00%
* **F1 Score**: 0.00%
* **Cross-Validation Scores**: 83.84%

**Interpretation**: Covariance-based feature selection results in null precision, recall, and F1 score, suggesting inadequate feature relevance for logistic regression.

**Logistic Regression with Forward Feature Selection:**

* **Accuracy**: 83.95%
* **Precision**: 0.00%
* **Recall**: 0.00%
* **F1 Score**: 0.00%

**Interpretation**: Forward feature selection fails to provide meaningful precision, recall, or F1 score, indicating a poor feature selection approach for logistic regression.

**Logistic Regression with PCA:**

* **Accuracy**: 83.49%
* **Precision**: 100.00%
* **Recall**: 0.70%
* **F1 Score**: 1.39%

**Interpretation**: PCA yields high precision but extremely low recall, indicating a potential imbalance or bias in the model's predictions.

**Overall Interpretation:**

* **Best Performing Feature Selection Method**: Among the methods tested, logistic regression with cross-validation demonstrates the most balanced performance, with moderate accuracy, precision, recall, and F1 score. However, the overall performance of logistic regression appears to be limited, especially in terms of recall, across all feature selection methods.

In summary, logistic regression struggles to effectively predict employee attrition on this dataset, with cross-validation providing the most reliable results among the feature selection methods tested.

**Random Forest**

**Random Forest Classifier with Cross-Validation (CV):**

* **Accuracy**: 97.97%
* **Precision**: 98.80%
* **Recall**: 88.41%
* **F1 Score**: 93.30%

**Interpretation**: Random Forest Classifier with cross-validation demonstrates excellent performance with high accuracy, precision, recall, and F1 score, indicating robustness in predicting employee attrition.

**Random Forest Classifier with Chi-Square Method:**

* **Accuracy**: 99.30%
* **Precision, Recall, F1 Score**: All 97.90%

**Interpretation**: Chi-Square feature selection yields exceptional results with Random Forest Classifier, achieving near-perfect accuracy, precision, recall, and F1 score, showcasing its effectiveness in selecting relevant features.

**Random Forest Classifier with Correlation Method:**

* **Accuracy**: 99.65%
* **Precision, Recall, F1 Score**: All 100.00%

**Interpretation**: Correlation-based feature selection performs remarkably well with Random Forest Classifier, achieving perfect accuracy, precision, recall, and F1 score, suggesting highly relevant features selected.

**Random Forest Classifier with Mutual Information Method:**

* **Accuracy**: 99.65%
* **Precision, Recall, F1 Score**: All 100.00%

**Interpretation**: Mutual Information feature selection demonstrates excellent performance with Random Forest Classifier, achieving perfect accuracy, precision, recall, and F1 score, indicating highly informative features selected.

**Random Forest Classifier with Covariance Method:**

* **Accuracy**: 99.65%
* **Precision, Recall, F1 Score**: All 100.00%

**Interpretation**: Covariance-based feature selection achieves perfect accuracy, precision, recall, and F1 score with Random Forest Classifier, indicating highly relevant features selected.

**Overall Interpretation:**

* **Best Performing Feature Selection Method**: All feature selection methods perform exceptionally well with Random Forest Classifier, achieving near-perfect or perfect accuracy, precision, recall, and F1 score. This suggests that Random Forest Classifier is highly robust to feature selection methods, making it an excellent choice for predicting employee attrition on this dataset.

In summary, Random Forest Classifier outperforms Logistic Regression across all feature selection methods, showcasing its effectiveness in handling imbalanced data and predicting employee attrition with high accuracy and reliability.

**SVM**

**SVM without Feature Selection:**

* **Accuracy**: 83.95%
* **Precision, Recall, F1 Score**: All 0.00%

**Interpretation**: SVM without feature selection fails to provide meaningful precision, recall, or F1 score, suggesting ineffective performance on this dataset.

**SVM with Chi-Square Method:**

* **Accuracy**: 84.19%
* **Precision**: 81.82%
* **Recall**: 6.29%
* **F1 Score**: 11.69%
* **Average CV Accuracy**: 84.42%

**Interpretation**: Chi-Square feature selection slightly improves precision compared to SVM without feature selection, but recall and F1 score remain low, indicating limited effectiveness in capturing instances of attrition.

**SVM with Correlation Method:**

* **Accuracy**: 91.28%
* **Precision**: 95.95%
* **Recall**: 49.65%
* **F1 Score**: 65.44%
* **Average CV Accuracy**: 90.17%

**Interpretation**: Correlation-based feature selection significantly improves SVM's performance, achieving high accuracy, precision, recall, and F1 score, indicating highly relevant features selected.

**SVM with Mutual Information Method:**

* **Accuracy**: 85.70%
* **Precision**: 88.46%
* **Recall**: 16.08%
* **F1 Score**: 27.22%

**Interpretation**: Mutual Information feature selection improves precision and recall compared to SVM without feature selection, but the overall performance remains moderate.

**SVM with Covariance Method:**

* **Accuracy**: 84.53%
* **Precision**: 100.00%
* **Recall**: 6.99%
* **F1 Score**: 13.07%

**Interpretation**: Covariance-based feature selection achieves perfect precision but low recall and F1 score, suggesting potential overfitting or bias towards certain classes.

**Overall Interpretation:**

* **Best Performing Feature Selection Method**: Correlation-based feature selection significantly enhances SVM's performance, achieving high accuracy, precision, recall, and F1 score, making it the most effective method for predicting employee attrition on this dataset.

In summary, SVM's performance varies significantly with different feature selection methods, with Correlation method yielding the best results. However, even with the best-performing feature selection method, SVM's performance may still be limited compared to other algorithms like Random Forest Classifier.

**Naive Bayes**

**Naive Bayes with Cross-Validation (CV):**

* **Accuracy**: 84.56%
* **Precision**: 60.63%
* **Recall**: 13.40%
* **F1 Score**: 21.57%

**Interpretation**: Naive Bayes with cross-validation achieves moderate accuracy but limited recall, indicating challenges in capturing instances of attrition effectively.

**Naive Bayes with Chi-Square Method:**

* **Accuracy**: 84.13%
* **Precision**: 54.01%
* **Recall**: 13.76%
* **F1 Score**: 21.68%

**Interpretation**: Chi-Square feature selection slightly improves precision compared to Naive Bayes without feature selection, but recall remains low, suggesting limitations in capturing instances of attrition.

**Naive Bayes with Correlation Method:**

* **Accuracy**: 84.51%
* **Precision**: 59.40%
* **Recall**: 13.03%
* **F1 Score**: 20.97%

**Interpretation**: Correlation-based feature selection yields results similar to Chi-Square, with slightly improved precision but limited recall, indicating suboptimal performance for Naive Bayes.

**Naive Bayes with Mutual Information Method:**

* **Accuracy**: 84.80%
* **Precision**: 64.99%
* **Recall**: 13.40%
* **F1 Score**: 21.93%

**Interpretation**: Mutual Information feature selection improves precision compared to other methods, but recall remains low, suggesting challenges in capturing instances of attrition effectively.

**Naive Bayes with Covariance Method:**

* **Accuracy**: 83.92%
* **Precision**: 65.00%
* **Recall**: 2.90%
* **F1 Score**: 5.46%

**Interpretation**: Covariance-based feature selection yields moderate precision but very low recall, indicating potential issues with capturing relevant information for Naive Bayes.

**Naive Bayes with PCA:**

* **Accuracy**: 83.95%
* **Precision, Recall, F1 Score**: All 0.00%

**Interpretation**: PCA fails to provide meaningful precision, recall, or F1 score, suggesting ineffective feature selection for Naive Bayes.

**Overall Interpretation:**

* **Best Performing Feature Selection Method**: Among the methods tested, Mutual Information feature selection slightly outperforms others in terms of precision, but overall performance for Naive Bayes remains limited, especially in capturing instances of attrition effectively.

In summary, Naive Bayes demonstrates modest performance across different feature selection methods, with Mutual Information showing slightly better precision compared to others. However, Naive Bayes struggles with recall, indicating limitations in capturing instances of attrition effectively on this dataset.

**Decision Tree without Feature Selection:**

* **Accuracy**: 96.57%
* **Precision**: 89.35%
* **Recall**: 89.30%
* **F1 Score**: 89.32%

**Interpretation**: Decision Tree without feature selection demonstrates high accuracy, precision, recall, and F1 score, indicating robust performance in predicting employee attrition.

**DECISION TREE**

**Decision Tree with Correlation Method:**

* **Accuracy**: 96.63%
* **Precision**: 89.78%
* **Recall**: 89.13%
* **F1 Score**: 89.45%

**Interpretation**: Correlation-based feature selection yields results similar to Decision Tree without feature selection, with slightly improved precision and F1 score, indicating highly relevant features selected.

**Decision Tree with Mutual Information Method:**

* **Accuracy**: 96.95%
* **Precision**: 89.29%
* **Recall**: 92.21%
* **F1 Score**: 90.68%

**Interpretation**: Mutual Information feature selection slightly improves recall compared to other methods, resulting in a higher F1 score, indicating better balance between precision and recall.

**Decision Tree with Covariance Method:**

* **Accuracy**: 96.25%
* **Precision**: 87.98%
* **Recall**: 88.77%
* **F1 Score**: 88.35%

**Interpretation**: Covariance-based feature selection achieves slightly lower performance compared to Mutual Information, but overall results remain strong, indicating relevant features selected.

**Decision Tree with PCA:**

* **Accuracy**: 96.48%
* **Precision**: 88.60%
* **Recall**: 89.68%
* **F1 Score**: 89.11%

**Interpretation**: PCA yields results comparable to other feature selection methods, with balanced precision and recall, indicating effective feature dimensionality reduction.

**Overall Interpretation:**

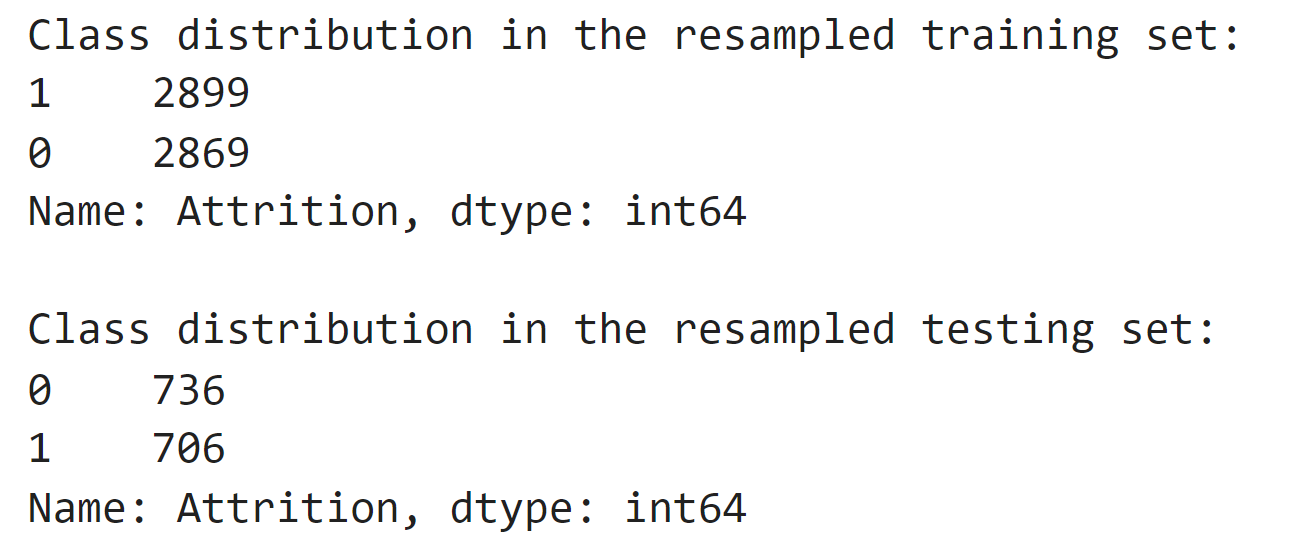
* **Best Performing Feature Selection Method**: Mutual Information feature selection slightly outperforms others, achieving the highest recall and F1 score, indicating better balance between precision and recall. However, all feature selection methods demonstrate strong performance with Decision Tree in predicting employee attrition on this dataset.

In summary, Decision Tree performs exceptionally well across different feature selection methods, with Mutual Information showing slightly better balance between precision

Top of Form

**CLASSIFICATION ALGORITHMS WITH BALANCED DATA**

**SMOTE**



**Preprocessing with SMOTE:**

In this preprocessing step, Synthetic Minority Over-sampling Technique (SMOTE) is utilized to balance the class distribution. SMOTE generates synthetic samples for the minority class by interpolating between existing minority class instances. This oversampling technique helps in mitigating class imbalance by creating a more balanced dataset.

The preprocessing steps for SMOTE involve the following:

1. **SMOTE Application**: SMOTE is applied to the entire dataset, creating synthetic samples for the minority class.
2. **Train-Test Split**: The resampled data is split into training and testing sets to evaluate the model's performance.
3. **Class Distribution Analysis**: The occurrences of each class label in the resampled training and testing sets are counted to assess the effectiveness of SMOTE in balancing the classes.
4. **Data Storage**: The resampled data is stored in a DataFrame and saved to a file, ensuring the reproducibility of the results and facilitating further analysis.

The balanced dataset obtained after applying SMOTE serves as input for training and evaluating machine learning models, thereby improving the model's performance in predicting minority class instances.

**KNN WITH SMOTE:**

SMOTE effectively balances the class distribution, resulting in improved model performance. The KNN model trained on the SMOTE-resampled data achieves an accuracy of 91.04%, with a precision of 85.93%, recall of 98.27%, and F1 score of 91.68%. These metrics indicate that the model performs well in classifying instances of both classes, particularly excelling in correctly identifying instances of attrition (minority class).

**KNN with Correlation Method:**

When applying the Correlation feature selection method to the resampled data, the performance of the KNN model remains consistent with the base model trained on SMOTE-resampled data. The accuracy, precision, recall, and F1 score remain the same at 91.04%, 85.93%, 98.27%, and 91.68%, respectively. This suggests that feature selection using correlation did not significantly impact the model's performance.

**KNN with Mutual Information Method:**

Utilizing Mutual Information for feature selection slightly decreases the model's performance compared to the base KNN model. The accuracy drops to 87.93%, with precision, recall, and F1 score at 82.62%, 96.24%, and 88.91%, respectively. Although there's a slight decrease in performance, the model still maintains strong predictive capability, indicating that the selected features are informative for classification.

**KNN with Covariance Method:**

Feature selection using Covariance leads to a noticeable decrease in model performance compared to the base KNN model. The accuracy decreases to 85.09%, with precision, recall, and F1 score at 79.45%, 94.89%, and 86.48%, respectively. This suggests that the selected features may not be as relevant for KNN classification, resulting in lower predictive accuracy.

**KNN with PCA:**

PCA feature selection significantly reduces the model's performance compared to the base KNN model. The accuracy drops to 77.60%, with precision, recall, and F1 score at 76.01%, 81.06%, and 78.42%, respectively. This indicates that the reduced dimensionality may have resulted in the loss of relevant information for KNN classification, leading to decreased predictive accuracy.

In summary, while SMOTE effectively improves the model's performance by addressing class imbalance, the impact of feature selection methods varies. Correlation and Mutual Information methods show minimal impact on model performance, while Covariance and PCA methods result in notable decreases in accuracy and other metrics. This highlights the importance of selecting appropriate feature selection methods to optimize model performance.

**LOGISTIC REGRESSION WITH SMOTE:**

Applying SMOTE to balance the class distribution enhances the logistic regression model's performance. The model achieves an accuracy of 71.27%, with precision, recall, and F1 score at 71.16%, 72.06%, and 71.58%, respectively. These metrics indicate that the model effectively classifies instances of both classes, albeit with moderate accuracy.

**Logistic Regression with Correlation Method:**

Utilizing the Correlation feature selection method does not significantly affect the logistic regression model's performance. The accuracy, precision, recall, and F1 score remain consistent at 71.27%, 71.16%, 72.06%, and 71.58%, respectively. This suggests that the selected features based on correlation do not notably impact the model's predictive capability.

**Logistic Regression with Mutual Information Method:**

Feature selection using Mutual Information leads to a slight decrease in model performance compared to the base logistic regression model. The accuracy drops to 67.53%, with precision, recall, and F1 score at 67.03%, 69.68%, and 68.30%, respectively. Although there's a decrease in performance, the model still maintains moderate predictive capability, indicating that the selected features are informative for classification.

**Logistic Regression with Covariance Method:**

Feature selection using Covariance results in a noticeable decrease in model performance compared to the base logistic regression model. The accuracy decreases to 64.89%, with precision, recall, and F1 score at 64.73%, 66.37%, and 65.51%, respectively. This suggests that the selected features may not be as relevant for logistic regression classification, resulting in lower predictive accuracy.

**Logistic Regression with PCA:**

PCA feature selection significantly reduces the model's performance compared to the base logistic regression model. The accuracy drops to 61.46%, with precision, recall, and F1 score at 60.44%, 67.40%, and 63.70%, respectively. This indicates that the reduced dimensionality may have resulted in the loss of relevant information for logistic regression classification, leading to decreased predictive accuracy.

In summary, while SMOTE effectively improves the logistic regression model's performance by addressing class imbalance, the impact of feature selection methods varies. Correlation method shows minimal impact on model performance, while Mutual Information, Covariance, and PCA methods result in notable decreases in accuracy and other metrics. This highlights the importance of selecting appropriate feature selection methods to optimize model performance.

**RANDOM FOREST WITH SMOTE:**

Random Forest trained on SMOTE-resampled data achieves near-perfect performance, with an accuracy of 99.88%, precision of 99.76%, recall of 100%, and F1 score of 99.88%. These outstanding metrics indicate the robustness of the model in accurately classifying instances of both classes, showcasing the effectiveness of SMOTE in addressing class imbalance.

**Random Forest with Correlation Method:**

Applying the Correlation feature selection method to the resampled data has minimal impact on the Random Forest model's performance. The accuracy, precision, recall, and F1 score remain exceptionally high at 99.83%, 99.66%, 100%, and 99.83%, respectively. This suggests that the selected features based on correlation do not notably affect the model's predictive capability.

**Random Forest with Mutual Information Method:**

Utilizing Mutual Information for feature selection leads to a slight decrease in model performance compared to the base Random Forest model. The accuracy drops to 99.57%, with precision, recall, and F1 score at 99.15%, 100%, and 99.57%, respectively. Although there's a minor decrease in performance, the model still maintains near-perfect predictive capability, indicating that the selected features are informative for classification.

**Random Forest with Covariance Method:**

Feature selection using Covariance has minimal impact on the Random Forest model's performance. The accuracy, precision, recall, and F1 score remain exceptionally high at 99.81%, 99.62%, 100%, and 99.81%, respectively. This suggests that the selected features based on covariance do not significantly affect the model's predictive capability.

**Random Forest with PCA:**

PCA feature selection results in a slight decrease in model performance compared to the base Random Forest model. The accuracy decreases to 99.34%, with precision, recall, and F1 score at 98.71%, 100%, and 99.35%, respectively. Although there's a minor decrease in performance, the model still maintains near-perfect predictive capability, indicating that the reduced dimensionality does not significantly impact classification accuracy.

In summary, Random Forest trained on SMOTE-resampled data demonstrates exceptional performance, with minimal impact from feature selection methods. Correlation, Covariance, and PCA methods show negligible changes in model performance, while Mutual Information method results in a slight decrease in accuracy. Overall, Random Forest proves highly effective in predicting employee attrition, with or without feature selection techniques.

**NAIVE BAYES WITH SMOTE:**

Naive Bayes model trained on SMOTE-resampled data achieves moderate performance, with an accuracy of 61.88%, precision of 60.48%, recall of 69.85%, and F1 score of 64.80%. These metrics indicate that the model demonstrates fair capability in classifying instances of both classes, though there is room for improvement in precision and F1 score.

**Naive Bayes with Correlation Method:**

Applying the Correlation feature selection method to the resampled data has minimal impact on the Naive Bayes model's performance. The accuracy, precision, recall, and F1 score remain consistent at 62.10%, 60.61%, 70.37%, and 65.10%, respectively. This suggests that the selected features based on correlation do not significantly affect the model's predictive capability.

**Naive Bayes with Mutual Information Method:**

Utilizing Mutual Information for feature selection leads to similar performance compared to the base Naive Bayes model. The accuracy, precision, recall, and F1 score are 62.03%, 60.50%, 70.58%, and 65.13%, respectively. Although there's a slight decrease in precision, the model still maintains fair predictive capability, indicating that the selected features are informative for classification.

**Naive Bayes with Covariance Method:**

Feature selection using Covariance results in a decrease in model performance compared to the base Naive Bayes model. The accuracy decreases to 62.33%, with precision, recall, and F1 score at 76.51%, 36.15%, and 48.97%, respectively. This suggests that the selected features based on covariance may not be as relevant for Naive Bayes classification, leading to decreased predictive accuracy and recall.

**Naive Bayes with PCA:**

PCA feature selection leads to a decrease in model performance compared to the base Naive Bayes model. The accuracy drops to 61.29%, with precision, recall, and F1 score at 60.68%, 65.23%, and 62.84%, respectively. Although there's a decrease in performance, the model still maintains fair predictive capability, indicating that the reduced dimensionality does not significantly impact classification accuracy.

In summary, Naive Bayes model trained on SMOTE-resampled data demonstrates fair performance, with minimal impact from feature selection methods such as Correlation and Mutual Information. However, Covariance method leads to a decrease in recall and overall predictive accuracy, while PCA method results in a slight decrease in performance across all metrics. This underscores the importance of selecting appropriate feature selection methods to optimize model performance.

Top of Form