Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

Forecasting failure-prone air pressure systems (FFAPS) in vehicles using machine learning

Mohamed Safiyur Rahman & V. Sumathy

To cite this article: Mohamed Safiyur Rahman & V. Sumathy (2024) Forecasting failure-prone air pressure systems (FFAPS) in vehicles using machine learning, Automatika, 65:1, 1-13, DOI: 10.1080/00051144.2023.2269514

To link to this article: https://doi.org/10.1080/00051144.2023.2269514

Published online: 21 Nov 2023.

AUTOMATIKA
2024, VOL. 65, NO. 1, 1–13
https://doi.org/10.1080/00051144.2023.2269514

Taylor & Francis
Taylor & Francis Group

Forecasting failure-prone air pressure systems (FFAPS) in vehicles using machine learning

Mohamed Safiyur Rahman[a,b] and V. Sumathy[c]

[a]Government College of Engineering, Dharmapuri, Tamilnadu, India; [b]AAM India Manufacturing Corporation Pvt Ltd., Chennai, Tamilnadu, India; [c]Electronics and Communication Engineering, Government College of Engineering, Settikarai, Dharmapuri, Tamilnadu, India

ABSTRACT

Vehicles become an inevitable factor in everyone's life. Sometimes it becomes a threat to human lives and society. For any real-time-based applications, everyone should focus on predicting failure-prone components. A vehicle's air pressure system (APS) is one of its most important parts. If any system failure happens against APS it leads to core-financial losses, which in turn sometimes leads to loss of human lives. Prediction of APS negligence in a real-time application requires a deep diagnosis and diligent solution. In this study, we developed a machine learning model to predict system failure against APS. A real-time dataset that includes the 170 features and the presence of high-class imbalance data and missing values has been taken and experimentally validated with existing linear and nonlinear classifiers. The performance metrics results show that the Random Forest classifier exceeds other algorithms for training and testing data with an accuracy and F1 score of 99.5 and 99.5 percent respectively.

ARTICLE HISTORY
Received 22 August 2023
Accepted 5 October 2023

KEYWORDS
Air pressure system (APS); machine learning; classifier algorithms; forecasting failure prone air pressure systems (FFAPS)

1. Introduction

Air Pressure System (APS) takes part an indispensable role in braking system components (gauging brakes), Engine system components (shifting gears), body control modules (adjusting seats), and controlling suspensions. Brakes, gears, and suspension systems start to malfunction if any APS faults occur, resulting in system misbehaving and leading to unexpected or unpleasant incidents. Finally, it impacts the sudden breakdown of a vehicle and sometimes leads to the loss of human lives. Always ensure the activeness and performance of APS to deliver the expected percentage of compressed air to the above systems on time.

System level functionality of APS works by taking the incoming compressed air as input, delivering, and distributing the clean and dry air to different vehicle system component circuits [1]. At the component level, APS can be subdivided into three major parts: air dryers, control units, and circuit protection valves. The air drier removes the dampness from the input air. The circuit protection valves control various pneumatic circuits such as auxiliary control circuits, and braking system component circuits in the vehicle by enabling and disabling valves with predetermined pressures at different levels based on vehicle component requirements. The control (temperature and pressure sensors) units add more values by sense and decide to trigger the circuits based on the temperature and pressure inputs. Predicting APS failure before it happens is the primary and key area in this research and to detect whether the complete system failure occurred because of APS failure.

In recent years, advancements in technology have provided more innovative solutions to any real-time-based problems. Using Machine learning algorithms, we can categorize the problems [2] in real-time. For any real-time, application prognosis, approaching solutions via machine learning is getting attention and becoming popular nowadays. Many researchers recently faced difficulties in this classification problem due to the presence of highly imbalanced class distribution and a huge volume of missing values. Addressing and developing the machine learning algorithm for the highly imbalanced distribution of classes and huge volume of missing values in a large dataset needs exhaustive and iterative solutions. In this paper new machine learning methodology is identified at each iterative level to improve all the performance metrics needed for classification problems.

This paper describes a prognosis representation that can predict whether a vehicle faces forthcoming system failure against specific component(s) of the Air Pressure System (APS).

The remaining sections of this paper are categorized as follows. Section 2 represents related works on APS negligence forecasting from literature. The methodology is depicted in Section 3. Experiments, Performance Metrics, and Results are discussed in Section 4.

CONTACT Mohamed Safiyur Rahman mohamedsafiyur@gmail.com, mohamed.safiyur@outlook.com

Conclusion and Future works are summarized in Section 5.

2. Related works

Machine learning algorithms and data analysis methods are frequently applied in the prognosis of vehicle transportation systems [3–5]. The key focus of this section is to elaborate on the method and techniques followed against forecasting APS failure in vehicles. In addition, some of the works related to an imbalanced dataset and missing values are also presented.

Costa and Nascimento [6] handled the problem of class imbalance by using weighted data classifiers. Similarly, the weights defined to the classes in the classifiers for Logistic Regression (LR) and SVM are more specific and reciprocal in proportion to samples present in a class. In some classifiers such as Random Forest and K-Nearest Neighbours (KNN), authors migrated the thresholds with respect to part of samples in each class for predicting a sample. To handle the Missing values in features, the Author deploys soft impute, which is one of the Expectation Maximization (EM) based imputation techniques.

The number of multiple existing linear and nonlinear data classifiers was analysed by Gondek *et al.* [7] against APS failure for system faults. The authors used feature engineering methods to extract target features from the data. In their analysis, they used a feature selection approach along with the existing engineered features by giving preference to feature ranking. The author tried to bring some modifications with the help of feature ranking and replacing all missing values with medians uniformly to provide a cost-effective solution. As a result, the author achieved 0.6 as an average cost where the cost of 10 units for wrong prediction and the cost of 500 units for missing failures.

Fatlawi *et al.* [8] presented a feature reduction classification model with a high number of features by K-means clustering and bagging. It is mentioned that every feature performance feature is recorded by several metrics followed by weak and strong clusters categorization. Weak clusters are getting filtered as they are irrelevant. Strong cluster values are taken into consideration. For training classification models, Author used bagging-based decision trees finally. Similarly, in [9] the decision tree is formed using a feature reduction approach, and the weighted Gini index value of the feature is used to select the nodes.

It is widely known that there are two techniques being followed for imbalanced data sets. One is at the data level by means of oversampling (duplication) of minority class labels and undersampling (random) of majority class labels. The above-mentioned techniques were used in (cf. [10]). Unfortunately, it ended with noises in the dataset (information loss or unwanted information). The second one is employed at the identification of the best classifier to meet the problem requirements. Similarly to the first one, the Enhanced technique Synthetic Minority Oversampling Technique (SMOTE) [11] is also frequently used for imbalanced data sets by means of synthetically creating the data for minority class labels between class samples and their neighbours. Borderline-SMOTE [10] is one such modified technique discussed in the paper where the minority class labels are getting over-sampled by choosing the samples close to the boundary.

To deal with minority class [12] concerns, an additional noise filter was getting introduced with an undersampling technique to reduce noise. Ertekin *et al.* [13] presented a new technique to handle imbalanced data where online SVM adds samples to the training set one at a time incrementally based on an active learning strategy. The author described a new pattern to add the samples near the boundary to the training set where random sampling of those 59 data points is being performed. Similarly, at one stage to stop the training, the author proposed an early stopping criterion as the new criterion that once potential vectors have been chosen, training can end.

Nguyen *et al.* [14] justified that for classification problems, oversampling can be done for minority classes by considering the samples close to the borderline instead of all samples to handle the imbalanced data. This could be feasible where overlap between the feature classes is minimal which is not suitable for a high imbalance data set. Oh *et al.* [15] proposed the subset selection as an incremental technique that randomly selects the subset out of the complete data and adds the necessary data which can be identified by improvement in the classifier to gain information iteratively to the existing training set.

The RB-Boost classifier collection was introduced by Dez-Pastor [16]. It was planned to incorporate AdaBoost with a random sample selection for an AdaBoost instance in a training set. SMOTE is being used for the minority class and random under-sampling is used for the majority class. The inclination ratio between the class samples was the authors' goal. To handle the imbalance data for the binary classification problem, Shao *et al.* [17] described the Weighted LaGrange Twin Support Vector Machine (WLTSVM) where growth-based under-sampling is being discussed for the majority class. In addition to that, Weighted bias was used to enhance Minority sample performance.

Rafsunjani *et al.* [18] have taken five different classifiers and five different imputations been taken for comparison and analysis. As a result, the author proposed Multiple Imputation by Chained Equation (MICE) techniques that provided effective results to deal with missing values. Similarly, to handle the high imbalance data, random under-sampling was the productive performance technique. The author used precision, recall along with accuracy as performance metrics to get better performance results.

In their modified random forest method, Jose and Gopakumar [19] suggested training the incorrectly classified data individually before integrating it with the original random forest classifier as a training strategy. To achieve better performance outcomes for missing value imputation, the author chose to employ the KNN approach with a K value of 33. The proposed technique was applied to the dataset of Scania trucks, and results were obtained. Precision, F-measure, and Matthews Correlation Coefficient (MCC) were employed to assess performance. On the dataset, accuracy of 0:46 and F-measure of 0:62 was obtained.

Akarte and Hemachandra [20] introduced gradient-boosting trees for forecasting APS failure. The authors assigned the weights depending on the proportion of Minority and majority classes. The weights assignment is more on the minority class samples in the training set. In addition to that, authors removed the feature columns with over 70 per cent missing values, by using the median remaining missing values were filled. The authors employ the optimization of the determined parameters using cross-validation results. Apart from misclassification cost, authors measured other performance metrics as well.

From the literature review, it is identified that many researchers considered highly imbalanced data in the public APS dataset as the most important and challenging one (real-time) and provided solutions as well. Techniques handled by researchers and results obtained were nominal. After the literature survey, it is understood that continuous improvement is required at all stages to provide better performance results. The Methodology of forecasting system failure against APS is depicted (Table 1).

The current research work contributions are made as follows:

(1) Dataset Categorization: Implementation of new sequential steps of data exploitation in machine-learning prediction/classification method based on categorizing the dataset in multiple aspects to explore the information present in the data without modifying the data.
(2) Data Exploration: Instantiate and iterative solution is proposed to maneuver highly imbalanced data and missing values.
(3) F-Measure ($F\beta$): F score (Performance metrics) is discussed in detail for the classification problem with high imbalance data and missing values.

3. Methodology

3.1. Dataset description

In this research analysis, The real-time dataset was collected from https://www.kaggle.com/datasets/uciml/aps-failure-at-scania-trucks-data-set [21]. Dataset contains information from heavy Scania trucks against APS component failure. Based on data analysis, it is identified that the samples or instances present in the training dataset are categorized as positive or negative. The positive occurrences point to system Failure because of Air pressure system (APS) component failure, while negative instances indicate component failures in the system not related to the APS.

Among 60,000 samples collected in the APS dataset provided, 98.33 per cent (59,000) samples correspond to negative instances and only 1.67 per cent (1000) samples correspond to positive instances. It clearly shows that the provided dataset from Scania trucks has imbalanced data samples (More negative instances (59,000) and fewer positive instances (1000)) as shown in Figure 1. Every feature present in the dataset has missing values (Values that are not required at time instances when some of the functional applications executed are not stored for some features). Hence, we categorize the given dataset belongs to a classification problem with an imbalanced dataset and missing values.

The proposed solution is well designed in such a way that data present in the dataset is efficiently categorized and handled effectively at each processing level of machine learning as depicted in the flow chart in Figure 2. It is widely known that there are two techniques being followed for imbalanced data sets. One at the data level by means of oversampling (duplication) of the minimum no of samples present (minority class) labels and under-sampling (random) of the maximum no of samples (majority class) labels. The above-said techniques were used in (cf. [10]). Unfortunately, it ended with noises in the dataset (information loss or unwanted information). The second one is the identification of the best classifier to meet the problem requirements. Figure 3 shows the model of FFAPS.

For the given dataset, 171 attributes are available for the prediction of APS failure. An attribute named class "Class" is being taken to predict positive instances or negative instances. The remaining 170 attributes available in the reference forum (For more details please refer to (https://ida2016.blogs.dsv.su.se/?pageid = 1387)) do not have the specific meaning of the attribute, all the attributes are named with notations only. The classes to be predicted are binary values of ("negative", "positive"). The proposed method begins processing the data in the dataset right from the pre-processing stage to handle the imbalance of data concerns and missing values to get the real samples for prediction. Initially, during the pre-processing stage, we categorize the data into different aspects and identify the missing values in every feature column and how it is distributed throughout the dataset. In this approach, as we have a greater number of samples, it is better to filter a feature have more than 70 per cent of missing values as depicted in Figure 4.

Table 1. Summary of the related work.

Name of the author's	Year	Algorithm used	Prediction/Inference
Costa and Nascimento	2016	LR, SVM, RF and KNN	Expectation maximization (EM) technique used for Missing values.
Gondek *et al.*	2016	Multiple linear and non-linear classifiers	Cost reduction Feature Ranking
Fatlawi *et al.*	2018	Decision Tree classifier	Feature reduction by K means clustering and bagging weighted Gini index to select the nodes.
H. Han *et al.*	2005	Oversampling techniques and undersampling techniques	Noises get introduced in the dataset. Oversampling of minority class samples happens only if it is close to the chosen boundary.
N. V. Chawla *et al.*	2002	Synthetic minority oversampling technique (SMOTE)	Synthetically creates data between class samples and neighbours.
S. Ertekin *et al*.	2007	SVM	Add the training set incrementally based on an active learning strategy. Early stopping criteria are introduced once potential vectors have been chosen.
H. M. Nguyen *et al.*	2009	Borderline- Oversampling Technique	The borderline method can be chosen only if the feature class gap is minimal. Not feasible for an imbalanced dataset.
S. Oh *et al.*	2011	Sampling Technique	Selecting Subset randomly and adding necessary data iteratively to gain information.
J. F. Dez-Pastor *et al.*	2015	Random balance (RB) Boost Classifier	Incorporate Ada Boost Technique with a random sample Training set. SMOTE is being applied to the Minority class. Random undersampling is also being used for the Majority class. Aimed to achieve and maintain the inclination ratio between the class samples.
Y.-H. Shao, *et al*	2014	Weighted LaGrange Twin Support Vector Machine (WLTSVM)	Growth-based under-sampling technique being handled for majority class samples. Weighted bias was introduced for Minority class samples.
.	2019	Five different classifiers and different imputations	Proves Multiple Imputation by Chained Equation (MICE) as an effective technique for Missing values. Random undersampling for majority class samples. Precision and recall are used as performance metrics along with accuracy.
C. Jose *et al*	2017	Random forest classifier KNN	Train the incorrectly classified data before integrating. $K = 33$ as the best result for Missing value imputation. Precision, F-measure, and Matthews Correlation Coefficient (MCC) were used and achieved 0.46 as accuracy and 0.62 as F measure.
M. M. Akarte *et al.*	2018	Different classifiers	Used Scania Trucks dataset. Assigned weights based on positive and negative classes. Removed 70 per cent of missing values data from the dataset. The remaining were filled by the median. Cross-validation was used for optimization.

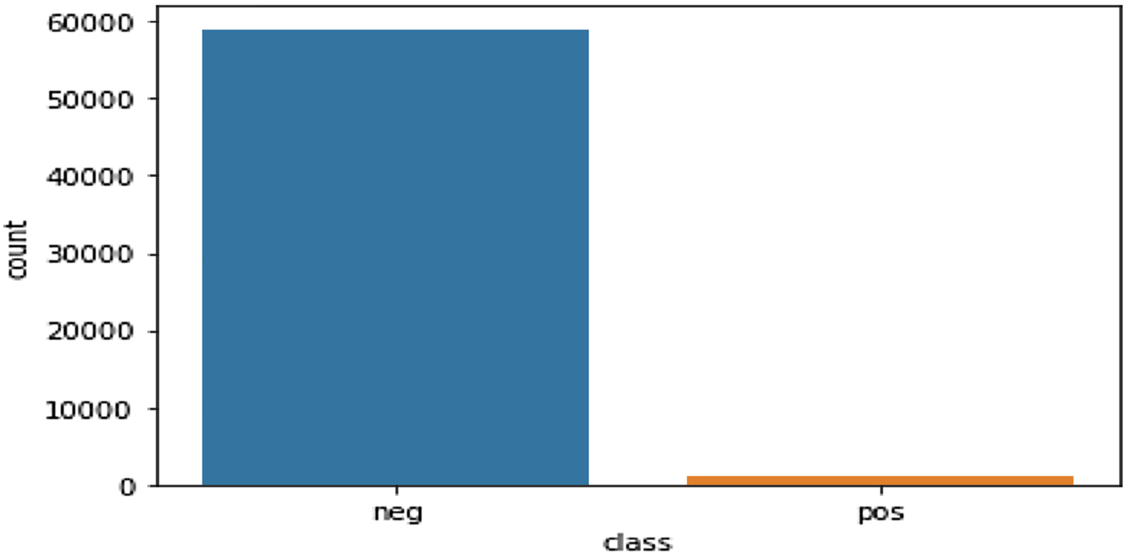
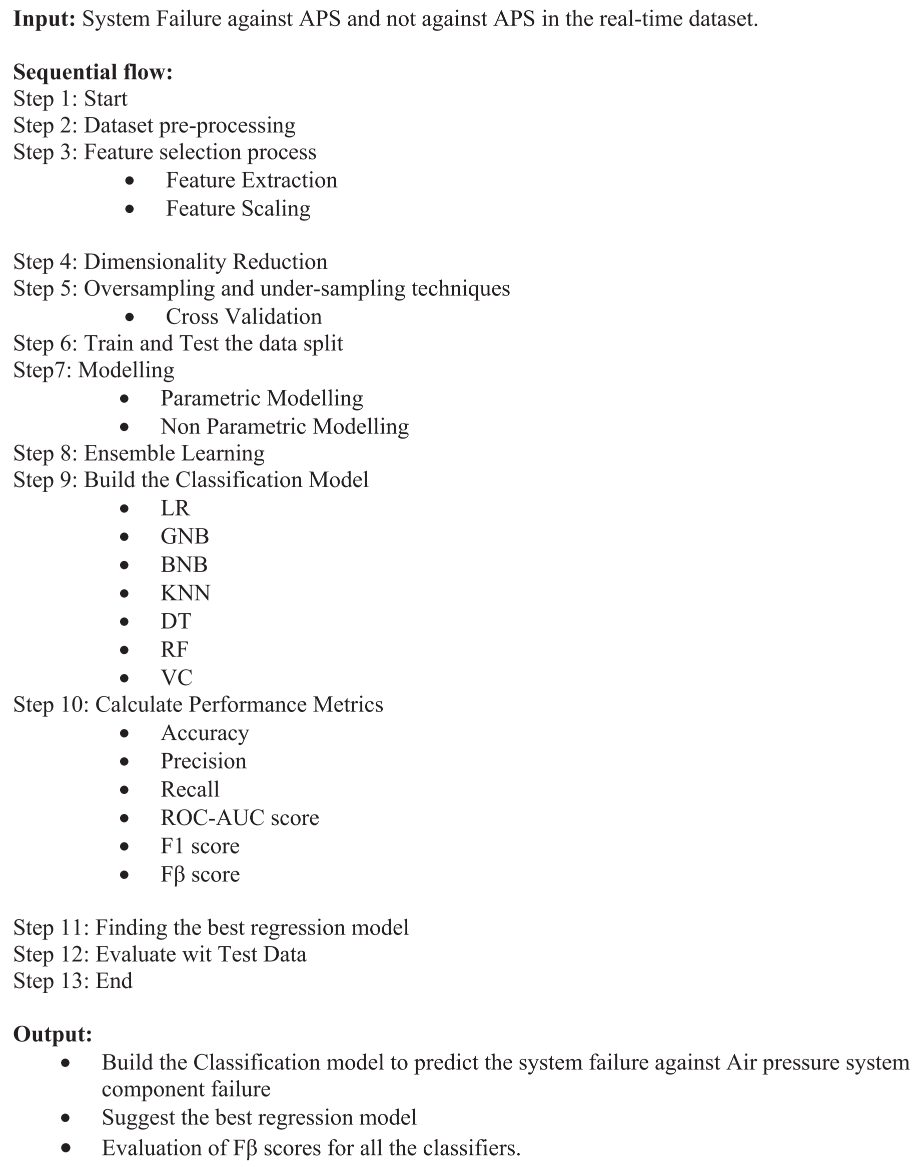


Figure 1. Positive and negative data samples from Scania trucks dataset.

Similarly, once the dataset is filtered, feature scaling is an important pre-processing step to ensure that all the features are on the same scale and have similar statistical properties, such as mean and variance. Overall, it improves the model by reducing the computational complexity and improving interpretability.

Cross-validation is one of the resampling methods that ensure the accuracy of a predictive model by using different combinations of the data in a particular dataset to train and test a model on different iterations until each and every instance of the data gets a chance in the validation set. k-Fold Cross-validation is one such



```
Input: System Failure against APS and not against APS in the real-time dataset.

Sequential flow:
Step 1: Start
Step 2: Dataset pre-processing
Step 3: Feature selection process
        • Feature Extraction
        • Feature Scaling

Step 4: Dimensionality Reduction
Step 5: Oversampling and under-sampling techniques
        • Cross Validation
Step 6: Train and Test the data split
Step7: Modelling
        • Parametric Modelling
        • Non Parametric Modelling
Step 8: Ensemble Learning
Step 9: Build the Classification Model
        • LR
        • GNB
        • BNB
        • KNN
        • DT
        • RF
        • VC
Step 10: Calculate Performance Metrics
        • Accuracy
        • Precision
        • Recall
        • ROC-AUC score
        • F1 score
        • Fβ score

Step 11: Finding the best regression model
Step 12: Evaluate wit Test Data
Step 13: End

Output:
    • Build the Classification model to predict the system failure against Air pressure system
      component failure
    • Suggest the best regression model
    • Evaluation of Fβ scores for all the classifiers.
```

Figure 2. Algorithm for FFAPS model.

technique that uses k−1 folds(sets) out of k folds for training the algorithm and the remaining set to be used for validation [22–24]. During the K folds generation, ordered sampling confirms the same portion of observations across each classification in each of the sets. One can use normalization or standardization to scale the data. In this approach, we can put it under cross-validation to validate the predictive model if any unknown data comes for validation or training. Once the dataset is scaled, the Random Forest algorithm is used for variable selection and recommended as the powerful algorithm for large dimensional data to identify the important dataset among thousands of samples using its built-in variable selection mechanism.

The dataset has experimented with both parametric modelling and non-parametric modelling for training and test split to identify the best classifier and compare the performance metrics as well. Algorithms that use assumptions to simplify the function to a known form are called parametric machine learning algorithms. These assumptions are typically based on the form of the distribution of the data and the functional form of the model. Hence, model assumptions ensure that they are appropriate for the data. Algorithms that are free to learn any functional form from the training data without making any assumptions to form the mapping function are called nonparametric machine learning algorithms. This can be useful in situations where the data do not meet the assumptions of parametric models, or when the functional form of the relationship is not known beforehand. Nonparametric modelling is the best suited for the huge number of variables and no need for distributional assumptions.

Random Forest (RF) is a non-parametric modelling method, which does not make any assumptions about the intrinsic data distribution. This makes it a good choice for variable selection, especially when the dataset has many samples and is not sure which ones are important.

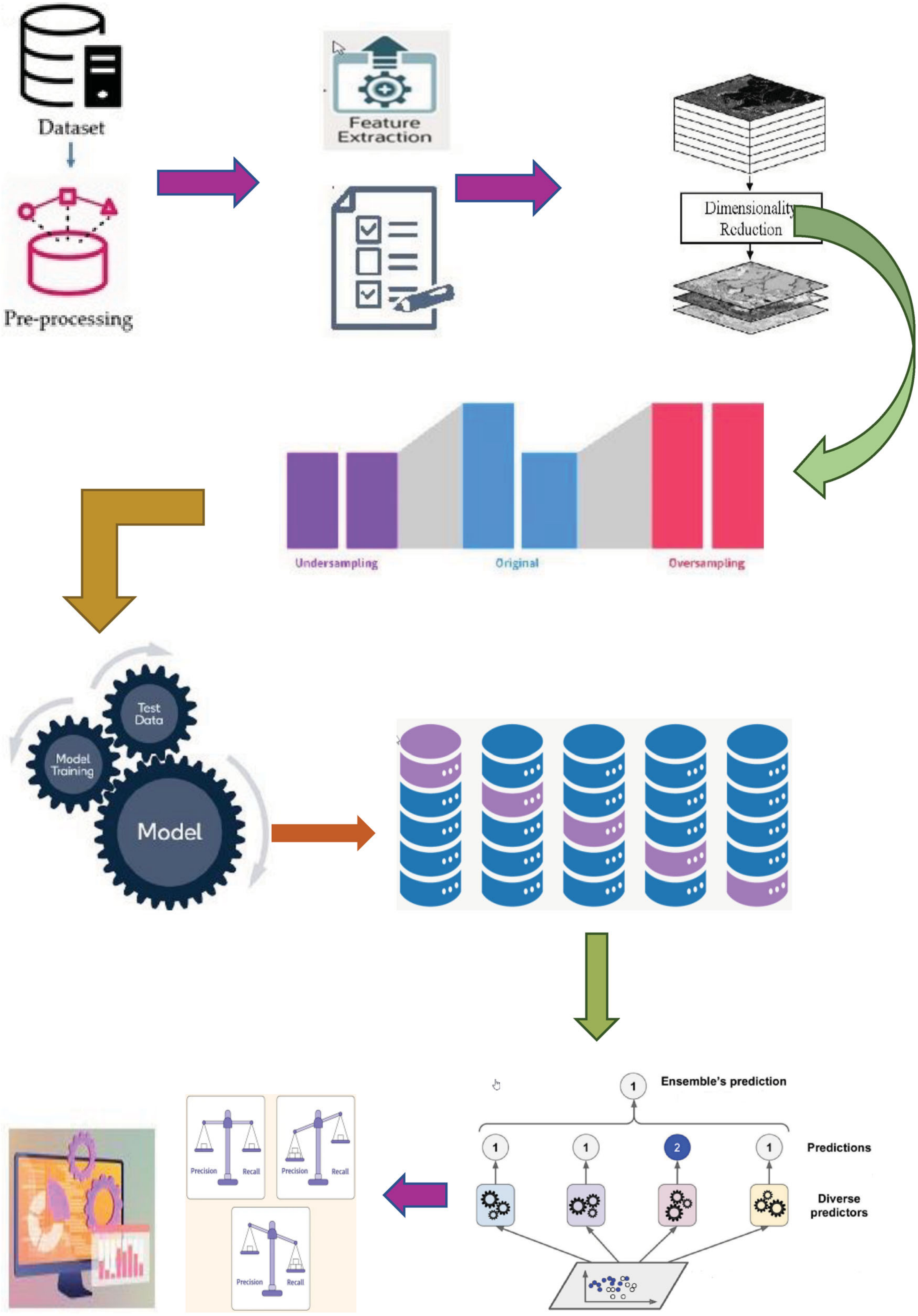


Figure 3. FFAPS model.

The dataset is imbalanced data where only 1000 samples belong to the positive class. Oversampling and under-sampling are two techniques that can be used to address imbalanced datasets, where one class is significantly more or less represented than the other. Under-sampling entails lowering the number of samples from the majority class whereas oversampling entails raising the number of samples from the minority class in the training dataset. Both techniques can be useful in situations where you have a highly imbalanced dataset and, in a need, to equalize the class distribution to enhance the performance of the model. However, it's important to be careful when using these techniques, as they can also introduce bias into the model if not

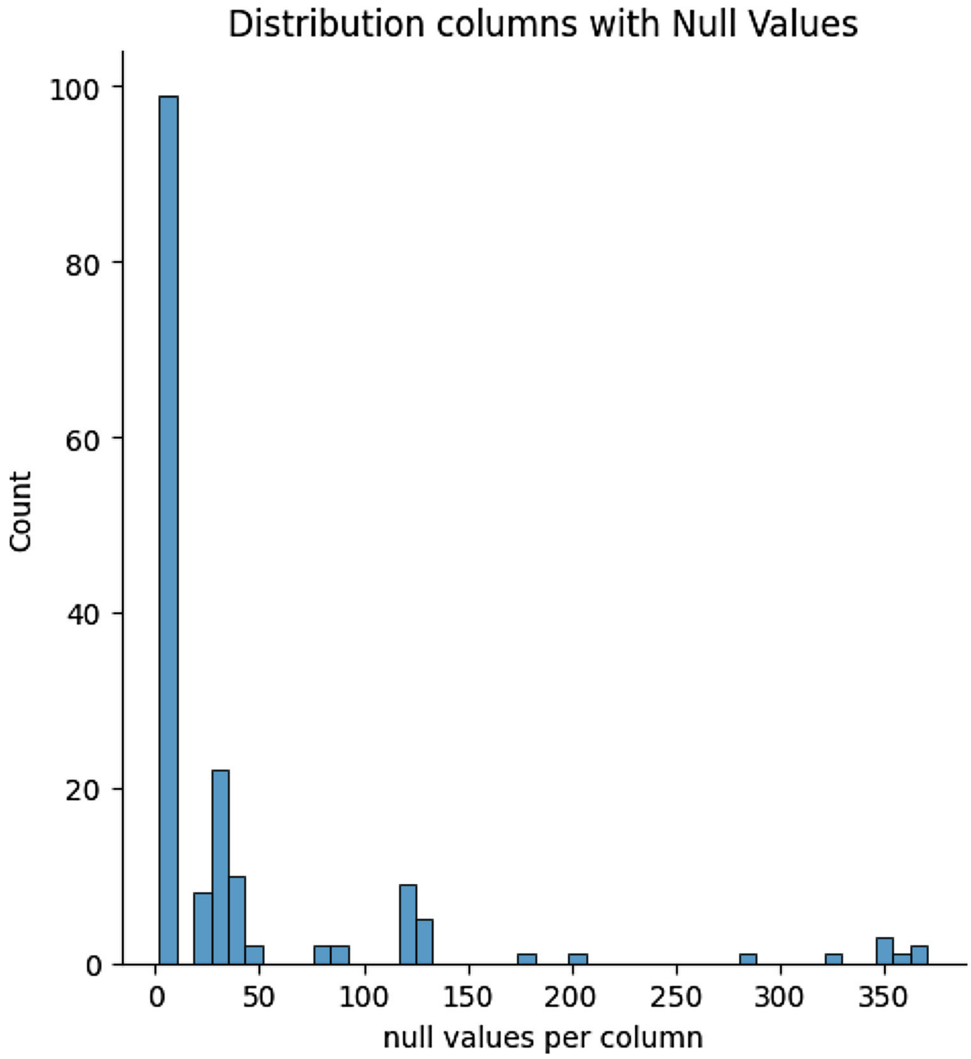


Figure 4. Distribution columns with missing values.

used correctly. Oversampling techniques that simply replicate the minority class can create a model that is overly sensitive to the minority class and may not generalize well to new data. Similarly, undersampling techniques that simply remove instances of the majority class can create a model that is overly sensitive to the majority class and may not generalize well to new data. Hence, it's generally a good idea to try both oversampling and under-sampling techniques and compare their performance on your dataset to see which one works best.

To make more accurate and robust predictions, we need to train the multiple models and combine their individual predictions. Ensemble learning is a type of machine learning that does the same. With the help of ensemble learning, the performance metrics like Accuracy, Recall, Precision, Fβ score, and AUC score were tremendously improved. Experimental results are depicted as shown in Table 2.

Table 2. Performance metrics.

Evaluation metrics	Equivalent equation
Accuracy	$(TP + TN)/(TP + TN + FP + FN)$
Precision	$TP/(TP + FP)$
Recall	$TP/(TP + FN)$
F-Beta score	$((1+\beta^2)*$ *(Precision* $*$ *Recall))/(Precision* $+$ *Recall)*

Table 3. Confusion matrix.

Confusion matrix	Actual value	
Predicted value	Positive	Negative
Positive	*True positive(TP)*	*False Positive(FP)*
Negative	*False Negative(FN)*	*True negative(TN)*

3.2. Mathematical model for Fβ

It is clearly known that the dataset taken for consideration is a highly imbalanced dataset where 59,000 samples related to class "negative" and only 1000 samples belong to class "positive". For a binary classification with an imbalanced dataset, Performance metrics like precision, Recall, and Fβ score would be the right metrics to judge whether the predicted value is right or not. Let us see the basic acronyms one by one to understand how they impact the proposed solution.

Understanding the basic performance metrics such as precision and recall is very much needed before proceeding with F-measure (F1 and Fβ) calculation.

Table 4. The performance metrics of comparison algorithms.

Model	Accuracy score	Precision score	Recall score	ROC-AUC score	F1_score	Fβ score
			Training data set (60,000 samples)			
Logistic Regression	0.818051	0.748708	0.957458	0.818051	0.840312	0.906887
GaussianNB	0.923602	0.967718	0.876441	0.923602	0.91982	0.893292
BernoulliNB	0.856737	0.859879	0.852373	0.856737	0.856109	0.853864
KNeighbors Classifier	0.98661	0.974859	0.998983	0.98661	0.986774	0.994063
DecisionTree Classifier	0.990551	0.987453	0.993729	0.990551	0.990581	0.992467
Random Forest Classifier	**0.995085**	**0.992829**	**0.997373**	**0.995085**	**0.995096**	**0.996461**
VotingClassifier	0.988898	0.981713	0.996356	0.988898	0.98898	0.993393
			Test data set(16,000 samples)			
Random Forest Classifier	0.990313	0.786458	0.85333	0.900043	0.795784	0.83961

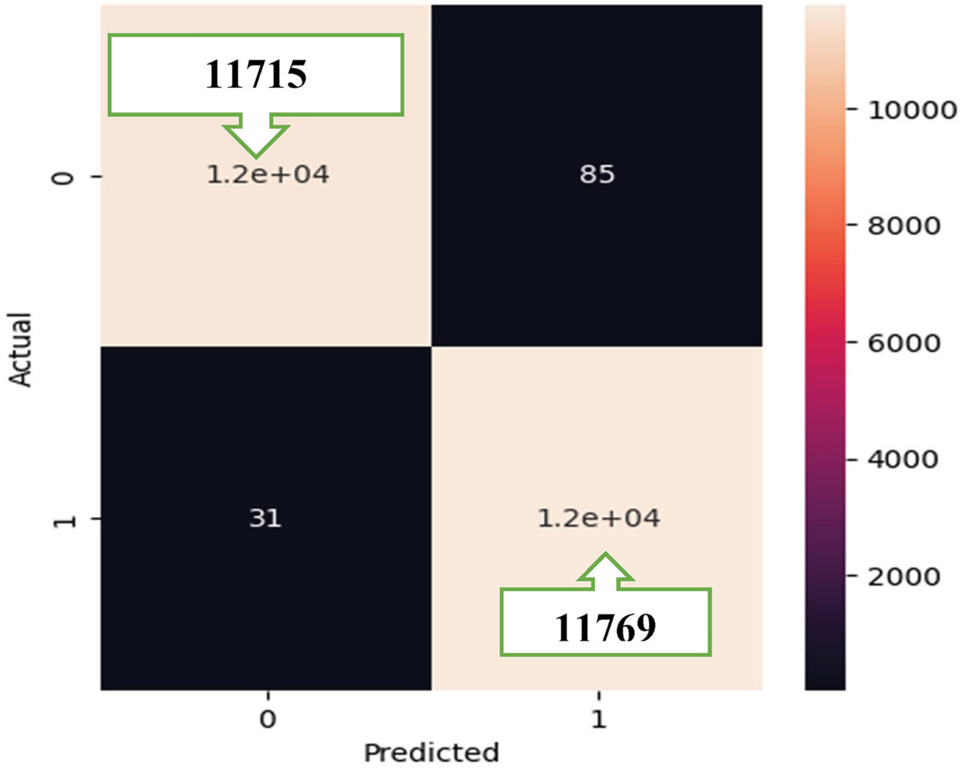


Figure 5. Confusion matrix.

The recall represents among the total actual positives, how many positives were predicted correctly. Recall can be also called a True Positive rate (TPR) or sensitivity. Similarly, Precision represents the total positive results predicted; how many were positive. Precision can be also called a "Positive prediction value". As sensitivity and positive prediction value play vital roles in highly imbalanced datasets, we need to evaluate which beta value suits the application for better performance results as shown in Table 4.

F measure (F1 and Fβ) can be explained as the weighted harmonic mean of precision and recall. Harmonic mean generally encourages closely related values for precision and recall. With more deviation between the precision and recall scores, the harmonic mean would be less which leads to better performance results. The Fβ score is needed for the evaluation of performance when there is the necessity of prioritizing one measure over the other.

3.2.1. Confusion matrix

Table 3 confirms the prediction against actual values based on the performance metrics measurement. After evaluation, the confusion matrix for the considered dataset for the RF is as below. Accuracy is well known for balanced datasets Here our classification problem involves many imbalanced data samples, hence there would be a possibility of Type 1 and Type 2 errors. To ensure the prediction, we should consider the recall and precision metrics as well (Figure 5).

Most of the researchers used F1 score for the performance results for the APS dataset which means researchers treated both Precision and recall equally. But in real-time samples with the high imbalanced dataset (APS), we should analyse the possibility of reduction of the false positives and false negatives errors. In the training dataset provided, out of 60,000 samples collected, 59,000 samples were recorded as negative, and only 1000 samples as positive. It clearly

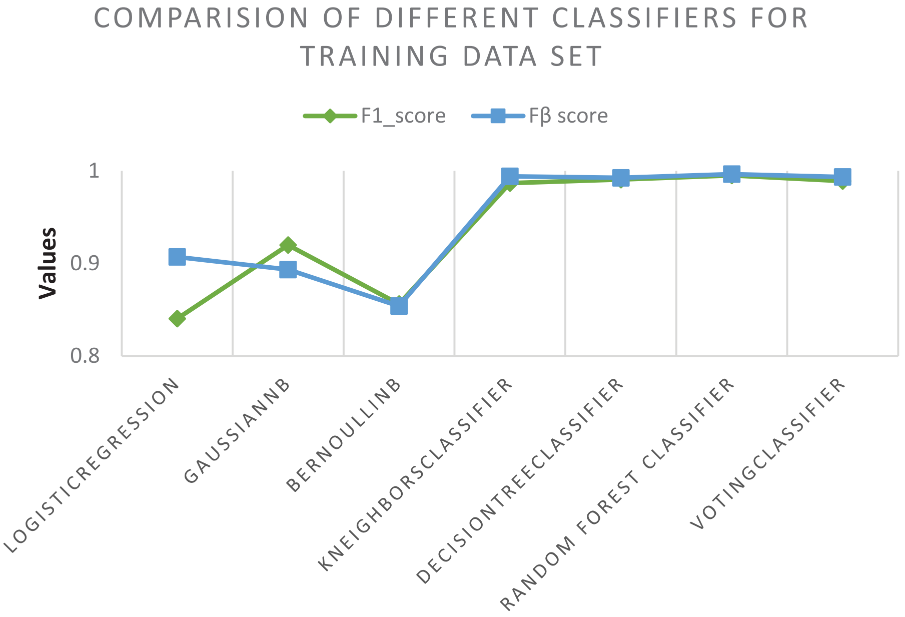


Figure 6. Classifiers comparisons against F1 and Fβ score.

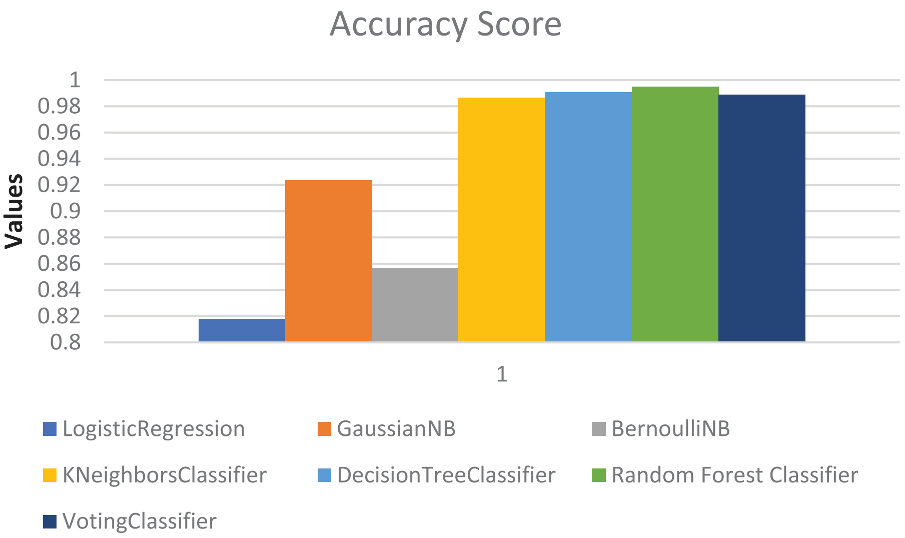


Figure 7. Classifiers comparisons against accuracy score.

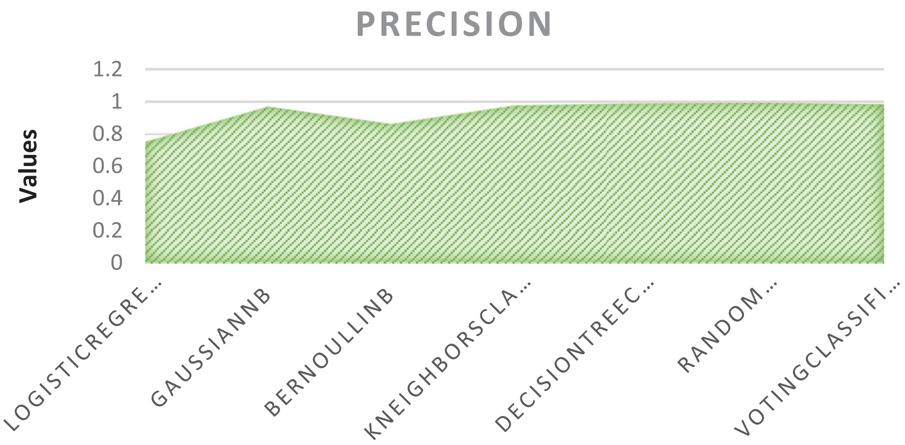


Figure 8. Classifiers comparisons against precision score.

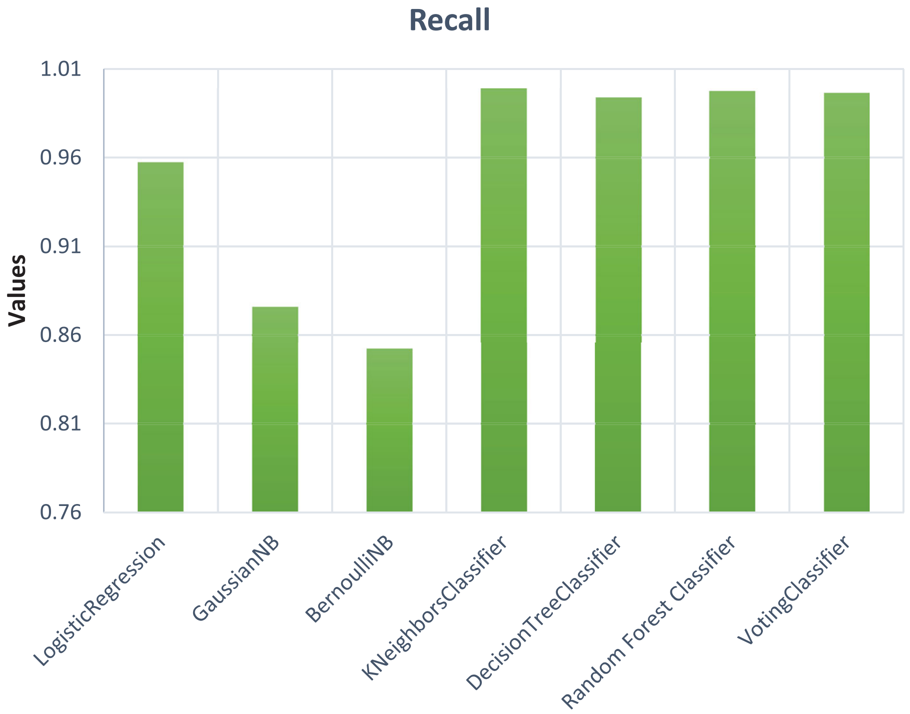


Figure 9. Classifiers comparisons against recall score.

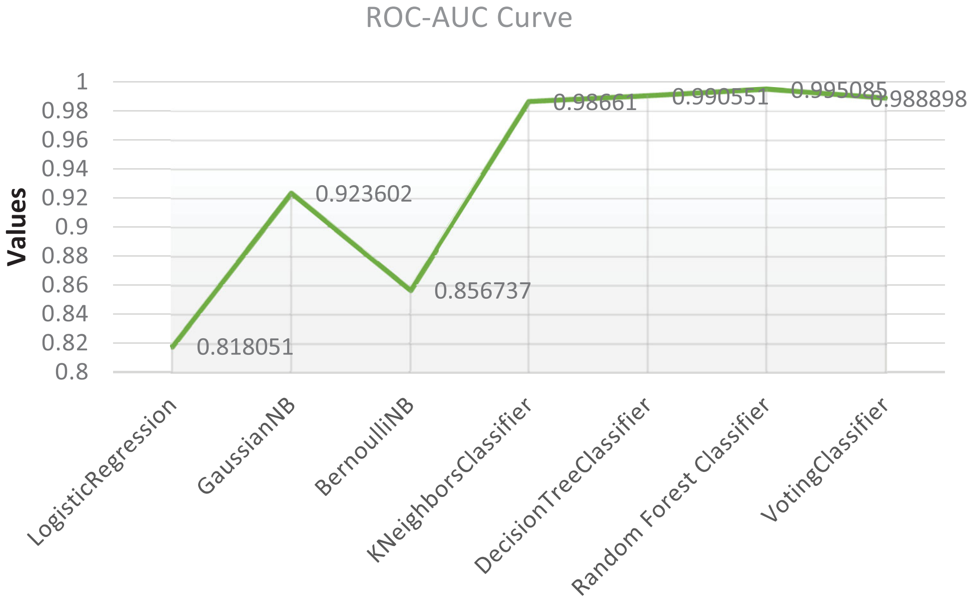


Figure 10. Classifiers comparisons against ROC-AUC score.

says that only 1000 samples recorded APS as a root cause for the vehicle failure. As no of samples which says the true positive rate is less when compared to true negative rates. It is sensitive that the prediction model not only predicts true positives correctly but also the model should focus mainly on reducing the false negatives. Hence, Prediction model output should yield better Recall (performance metrics) values. Similarly, as mentioned above, the β value also should be 2, as we are focussing mainly on Recall, then only the provided dataset on the proposed prediction model will yield the best prediction results.

4. Experiments, results and comparison

In this section, we compare the linear and nonlinear algorithms such as Logistic Regression (LR), Gaussian NB (GNB), Bernoulli NB (BNB), k-Neighbours Classifier (k-NN), Decision Tree Classifier (DT), Random Forest Classifier (RF), Voting Classifier (VC) against the performance metrics such as Accuracy, Precision, Recall, ROC_AUC Score, F1_Score, Fβ score as shown in Table 4. The comparison is being performed to understand which classifier suits best binary classification problems with highly imbalanced datasets with our proposed solution.

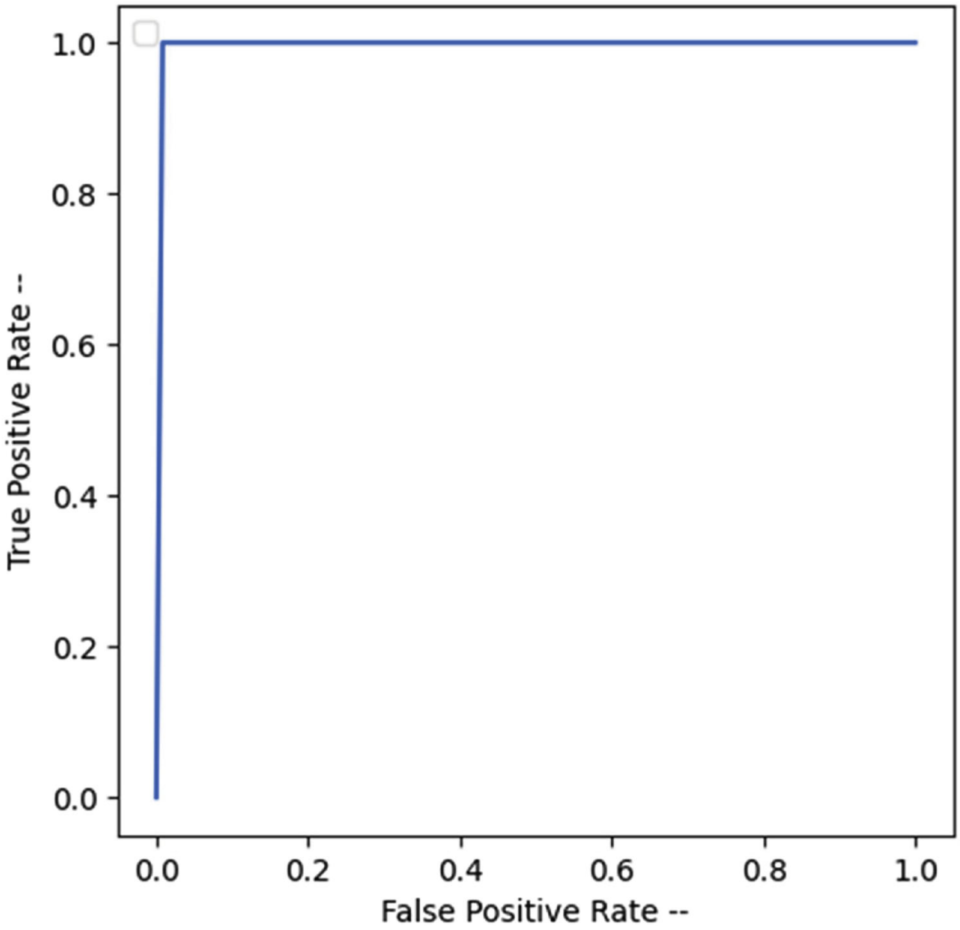


Figure 11. ROC-AUC score for random forest classifier.

4.1. F1 and Fβ score comparison

The results clearly show a great improvement in the Fβ score level against the F1 score of the existing linear and nonlinear classifiers which have good Recall metrics. The classifiers Gaussian Naive Bayes, and Bernoulli Naive Bayes precision score are better than the Recall score, that's why the F1 score shows better than the Fβ score. As explained before, recall metrics are much more important than other performance metrics in case of a highly imbalanced set.

We can see a good improvement in the testing dataset (Unknown data given to prediction model for validation) provided, Hence we would recommend the value of β as 2 for the highly imbalanced large volume dataset. Comparison metrics are shown in Figure 6.

4.2. Accuracy score

Figure 7 confirms that the accuracy score of the Random forest classifier reached 99.5 per cent against other comparative classifiers. Overall Accuracy has improved for all classifiers as per the algorithm of the FFAPS Model. As the dataset is an imbalanced data set, we can't predict the best model only with Accuracy metrics, hence below performance metrics are also performed to identify the best classifier.

4.3. Precision

Prediction performance metrics are actually needed when there is a situation highlighted that, Out of positively predicted values against a high volume of samples, how many were actually positive. Precision metrics strengthen the model and give better performance results in case of an imbalanced dataset. In Figure 8 shows that there is a considerable variation among classifiers to predict the values, Random forest classifiers predict the Positive prediction as 99.28 per cent when compared to other classifiers.

4.4. Recall

Recall, that the other performance metrics are usually used to reduce the false negatives in the imbalanced data set, it adds more values to identify the best classifier for proper prediction. As more the recall value, the better the prediction results. In Figure 9 as depicted, the Random forest classifier and KNN neighbour classifier have a good recall value of more than 99.5 per cent when compared to other classifiers.

4.5. ROC-AUC curve

ROC-AUC curve is one of the performance metrics to see the visual illustration of prediction for the balanced

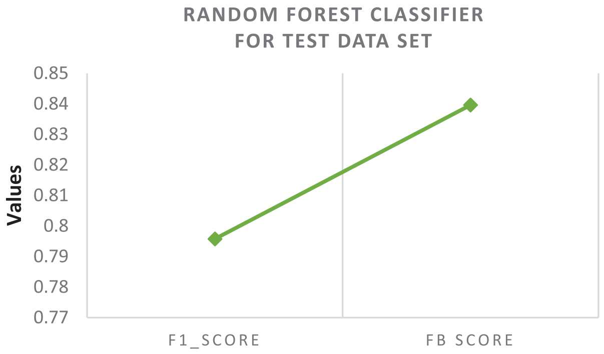


Figure 12. F1 and Fβ score comparison for the test dataset.

data, but if the data is highly imbalanced, we can take into consideration but not conclude the best model only with the help of the ROC-AUC curve. Because the False positive rate may not be drastically reduced when the actual total negatives are high. Even with the reasons mentioned, the Random forest classifier provides the best ROC-AUC performance at 99.50 when compared to other classifiers as shown in Figures 10 and 11.

Overall, it is found that the performance of the Random Forest classifier with the proposed Fβ score is better than the other classifier algorithms for an imbalanced dataset. Random forest classifier-based results for the test data set are shown in Figure 12.

5. Conclusions and future work

This paper proposes the handling of a highly imbalanced large volume of datasets and missing values right from the pre-processing stage itself. Mapping of data samples with respect to numerical, continuous, and discrete variables is employed at the pre-processing stage to enhance the performance. In this paper, we explored the imputation of the categorical variable with mode and impute the remaining all left skewed variable with the median for missing values after filtration. Furthermore, we have explored the dimensionality reduction, cross-validation of the processed data for feature enhancements, oversampling, and undersampling techniques also carried out of the sampled data to stabilize the highly imbalanced data. This paper proposes the Fβ score value as 2 for the highly imbalanced data as the major performance metrics for the comparison of algorithms along with other popular and commonly used metrics in the literature, namely F1-score, Recall, Precision, and Accuracy for the same imbalanced datasets. The Tables and the Figures presented in the experimental section validate that the performance metrics and results (proposed Fβ-score) of the Random Forest classifier are the best comparable with other algorithms.

As a continuous improvement, Data exploitation sequential steps with the proposed technique will be further enhanced to explore, and predict the real-time problems and provide better performance results in all principle components of automated vehicles.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Peruffo E. Improving predictive maintenance classifiers of industrial sensor data using entropy. A case study for a master's thesis in Computer Science at National College Ireland in Dublin, Ireland; 2018.

[2] Ranasinghe GD, Parlikad AK. Generating real-valued failure data for prognostics under the constraints of restricted data availability. Proceedings IEEE Internati onal Conference Prognostics Health Manage (ICPHM); Jun. 2019. p. 1–8.

[3] Wang F, Xu T, Tang T, et al. Bilevel feature extraction-based text mining for defect diagnosis of railway systems. IEEE Trans Intell Transp Syst. Jan. 2017;18(1): 49–58. doi:10.1109/TITS.2016.2521866

[4] Yuan D, Lu Z, Zhang J, et al. Integrative design of an emergency resource predicting-scheduling-repairing technique for rail track failures. IEEE Access. 2019;7: 155686–155700. doi:10.1109/ACCESS.2019.2949289

[5] Veres M, Moussa M. Deep learning for intelligent transportation systems: a survey of emerging trends. IEEE Trans Intell Trans Syst. Aug. 2020;21(8):3152–3168. doi:10.1109/TITS.2019.2929020

[6] Costa CF, Nascimento MA. Ida 2016 industrial challenge: using machine learning for failure prediction. Proceedings International Symposium on Intelligent Data Analysis, Cham, Switzerland: Springer; 2016. p. 381–386.

[7] Gondek C, Hafner D, Sampson OR. Prediction of failures in the air pressure system of Scania trucks using a random forest and feature engineering. Proceedings International Symposium on Intelligent Data Analysis, Cham, Switzerland: Springer; 2016. p. 398–402.

[8] Fatlawi HK, Alharan A, Ali NS. An effective hybrid model for trustworthy classification of high-dimensional data combining K-means clustering and bagging ensemble classifier. J Theor Appl Inf Technol. 2018;96(24):8379–8398.

[9] Liu H, Zhou M, Liu Q. An embedded feature selection method for imbalanced data classification. IEEE/CAA J Automatica Sinica. 2019;6(3):703–715. doi:10.1109/JAS.2019.1911447

[10] Han H, Wang W-Y, Mao B-H. Borderline-SMOTE: a new oversampling method in unbalanced data sets learning. Proceedings International Conference Intelligent, Berlin, Germany: Springer; 2005. p. 878–887.

[11] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. June 2002;16:321–357. doi:10.1613/jair.953

[12] Kang Q, Chen X, Li S, et al. A noise-altered undersampling approach for imbalanced classification. IEEE Trans Cybern. 47, 2016. doi:10.1109/TCYB.2016.2606104

[13] Ertekin S, Huang J, Bottou L, et al. Learning on the border: active learning in imbalanced data classification. Proceedings 16th ACM Conference on Information Knowledge Management (CIKM); 2007. p. 127–136.

[14] Nguyen HM, Cooper EW, Kamei K. Borderline oversampling for imbalanced data classification. Proceedings

5th International Workshop on Computer Intelligence Applications, no. 1. Hiroshima, Japan: IEEE SMC Hiroshima Chapter; 2009, p. 24–29.

[15] Oh S, Lee MS, Zhang B-T. Ensemble learning with active example selection for imbalanced biomedical data classification. IEEE/ACM Trans. Comput Biol Bioinf. Mar. 2011;8(2):316–325. doi:10.1109/TCBB.2010.96

[16] Dez-Pastor JF, Rodrguez JJ, Garca-Osorio C, et al. Random balance: ensembles of variable priors classifiers for unbalanced data. Knowl-Based Syst. Sep. 2015;85:96–111. doi:10.1016/j.knosys.2015.04.022

[17] Shao Y-H, Chen W-J, Zhang J-J, et al. An effective weighted Lagrangian twin support vector machine for unbalanced data classification. Pattern Recognit Sep. 2014;47(9):3158–3167. doi:10.1016/j.patcog.2014.03.008

[18] Rafsunjani S, Safa RS, Imran AA, et al. An empirical comparison of missing value imputation techniques on APS failure prediction. Int J Indust Appl Comput Sci. Feb. 2019;11(2):21–29.

[19] Jose C, Gopakumar G. An improved random forest algorithm for classification in an imbalanced dataset. Proc URSI Asia-Pacific Radio Sci Conf (AP-RASC). Mar. 2019;12:1–4. p. 4263–4274, Dec. 2017.

[20] Akarte MM, Hemachandra N. Predictive maintenance of air pressure systems using boosting trees: a machine learning approach. Proceedings of the 51st Annual Conference of the ORSI International Conference, Kolkata, India: ORSI; 2018.

[21] Air pressure system failures in Scania Trucks | Kaggle. María Ren. [cited 2023 Jan 14]. Available from: https://www.kaggle.com/datasets/uciml/aps-failure-at-scania-trucks-data-set.

[22] James G, Witten D, Hastie T, et al. Resampling methods. In An Introduction to statistical learning: with Applications in Python. Cham: Springer; 2023. p. 201–228.

[23] Muideen AA, Lee CKM, Chan J, et al. Broad embedded logistic regression classifier for prediction of air pressure systems failure. Mathematics. 2023;11(4):1014. doi:10.3390/math11041014

[24] Pardeshi SS, Patange AD, Jegadeeshwaran R, et al. Tyre pressure supervision of two wheeler using machine learning. Struct Durab Health Monitor. 2022;16(3):271. doi:10.32604/sdhm.2022.010622