# AIRA-ML: Auto Insurance Risk Assessment-Machine Learning Model using Resampling Methods

Ahmed Shawky Elbhrawy[1], Mohamed A. Belal[2], Mohamed Sameh Hassanein[3]
Business Information System Department-Faculty of Commerce and Business Administration, Helwan University
Cairo, Egypt[1]
Professor, Computer Science Department-Faculty of Computers and Artificial Intelligence, Helwan University
Cairo, Egypt[2]
Integrated Thebes Institutes for Computing & Management Science, Cairo, Egypt[3]

***Abstract*—Predicting underwriting risk has become a major challenge due to the imbalanced datasets in the field. A real-world imbalanced dataset is used in this work with 12 variables in 30144 cases, where most of the cases were classified as "accepting the insurance request", while a small percentage classified as "refusing insurance". This work developed 55 machine learning (ML) models to predict whether or not to renew policies. The models were developed using the original dataset and four data-level approaches resampling techniques: random oversampling, SMOTE, random undersampling, and hybrid methods with 11 ML algorithms to address the issue of imbalanced data (11 ML× (4 resampling techniques + unbalanced datasets) = 55 ML models). Seven classifier efficiency measures were used to evaluate these 55 models that were developed using 11 ML algorithms: logistic regression (LR), random forest (RF), artificial neural network (ANN), multilayer perceptron (MLP), support vector machine (SVM), naive Bayes (NB), decision tree (DT), XGBoost, k-nearest neighbors (KNN), stochastic gradient boosting (SGB), and AdaBoost. The seven classifier efficiency measures namely are accuracy, sensitivity, specificity, AUC, precision, F1-measure, and kappa. CRISP-DM methodology is utilisied to ensure that studies are conducted in a rigorous and systematic manner. Additionally, RapidMiner software was used to apply the algorithms and analyze the data, which highlighted the potential of ML to improve the accuracy of risk assessment in insurance underwriting. The results showed that all ML classifiers became more effective when using resampling strategies; where Hybrid resampling methods improved the performance of machine learning models on imbalanced data with an accuracy of 0.9967 and kappa statistics of 0.992 for the RF classifier.**

***Keywords*—Risk assessment; machine learning; imbalanced data; rapid miner; CRISP-DM methodology***

## I. Introduction

Insurance underwriting is a critical process that assesses and selects risks. In exchange for a premium payment, an insurance company agrees to compensate the insured for financial losses under the terms of a contract between the person or organization and the insurer. Insurance is a vital risk management tool that protects individuals and businesses from unforeseen occurrences that could cause financial losses. Car insurance is one of the most important types of insurance, as it protects owners and drivers financially from a variety of dangers and uncertainties [1].

Risk assessment in the insurance sector is a critical process for evaluating the likelihood and severity of potential losses or damages for a specific policyholder. In auto insurance, risk assessment considers several factors that may increase the likelihood of an accident, such as the driver's age, driving record, vehicle type, and location. Risk assessment is essential in the world of auto insurance for accurately pricing policies and ensuring financial viability [2].

The global usage-based auto insurance market is projected to grow from $57.86 billion in 2023 to $174.33 billion by 2030, at a compound annual growth rate (CAGR) of 17.1%. This growth is being driven by the increasing adoption of usage-based insurance (UBI) programs by consumers, as well as the growing availability of telematics devices that can collect the data needed to calculate UBI premiums. The total direct written premium for private passenger auto insurance in the United States was $247.1 billion in 2020, which underscores the significant size of the car insurance industry and the importance of risk assessment in ensuring that insurance companies can cover losses and remain financially stable [3] [4].

Machine learning (ML) has been used to improve the accuracy and effectiveness of risk assessment in auto insurance. ML algorithms are trained on large amounts of data to identify patterns and trends that would be difficult to find using traditional methods. This information can then be used to make more accurate predictions about the likelihood of an accident occurring, which can help insurers to price their policies more accurately and make better underwriting decisions [5].

Imbalance learning is a long-standing challenge in machine learning. In the context of auto insurance risk assessment, the majority class would be the group that reflects the majority of risks. The minority class, on the other hand, would be the group that makes up less of the total, such as policyholders who are denied insurance renewal. This category may have very little data. As a result, the data distribution across dataset classifications is often inconsistent in real-world settings. To improve the reliability of risk assessment, it is necessary to correct erroneous data. Data imbalances can be addressed using resampling techniques [6].

RapidMiner Studio is a data science platform that provides a graphical user interface for designing and deploying ML models. It enables users to preprocess data, build ML models, and apply reshaping techniques for unbalanced data. RapidMiner supports a wide range of ML algorithms and provides tools for evaluating model performance and selecting the best model. It is an important tool for handling ML algorithms and applying reshaping techniques for unbalanced data because it provides a user-friendly interface for implementing these techniques without the need for advanced programming skills [7] [8].

The motivation for this paper is the need for accurate risk assessment in auto insurance. Risk assessment is critical for accurately pricing policies and ensuring the financial viability of insurance companies. However, traditional risk assessment methods are often inaccurate due to imbalanced datasets, which makes it difficult to predict the risk of underwriting a new policy. Machine learning (ML) techniques have been proposed to improve accuracy, but they are also susceptible to data imbalance problems.

This paper proposes a new approach to address the challenge of imbalanced data in auto insurance datasets by using resampling techniques to create a more balanced dataset. This could lead to more accurate risk predictions and better pricing decisions for insurance companies. Additionally, the proposed approach could help reduce the risk of losses and automate the risk assessment process, thereby improving underwriting efficiency.

The paper is divided into five sections. Section I introduces the topic, followed by Section II, which reviews the related work. Section III discusses the methodology, which separately describes the phases of the proposed model. Section IV, Results and Discussion presents the final steps of the methodology and its results. Section V presents the conclusion, and discusses the future research directions.

## II. Related Work

In this section, the current work will review previous research efforts in risk assessment, claim prediction, and the use of machine learning algorithms and resampling methods. In the study of [9], the authors investigated the use of data mining tools and methods to develop models for analyzing risk levels for the Ethiopian Insurance Corporation (EIC). The study found that a decision tree model achieved an accuracy rate of 0.75 in classifying 3100 policies, while a neural network model achieved an accuracy rate of 58. And in [10] , they investigated the use of telematics data to predict accident claims. They compared the effectiveness of XGBoost and LR methods. LR was found to be a suitable model for this task because it is interpretable and has good predictive performance with accuracy rate of 0.8397. XGBoost requires more effort to interpret and requires several model-tuning strategies to match the predictive performance of logistic regression. And in [11] , the authors aimed to classify participants in the insurance renewal process to help companies reduce the claim ratio by being more selective in approving them. The proposed method involved classifying insurance participants' data using 3,803 datasets with four attributes and five algorithms to find significant features when generating the model. The study found that the decision tree (DT) algorithm was the most accurate, with an accuracy rate of 0.9540. The DT algorithm also showed that the most significant feature in defining prospective company assessment was the average age. And in [12], the authors used bootstrapping for resampling to evaluate two classifiers, RF and SVM, using four metrics: Accuracy, Precision, Recall, and F-measure. The experimental results showed that the two classifiers scored an overall accuracy of 0.9836 and 0.9817, respectively.

And in [13], the authors investigated the use of machine learning techniques by auto insurance companies to analyse large amounts of insurance-related data and forecast claim incidence. They applied a variety of machine learning techniques, including LR, XGBoost, RF, DT, NB, and K-NN. They also applied the random over-sampling technique to address the problem of unbalanced data. The results of the study showed that the RF model outperformed all other approaches with an accuracy of 0.8677. And in [14], they developed 32 machine learning models using various data-level approaches to address this challenge. The study found that the AdaBoost classifier with oversampling and the hybrid method had the most accurate predictions. The study concluded that the AdaBoost classifier, using oversampling or the hybrid process, can generate more accurate models for analyzing imbalanced data in the insurance industry than other models.

## III. Methodology and Proposed Model

CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is adopted to develop the proposed model called AIRA-ML (Auto Insurance Risk Assessment-machine learning) shown in Fig. 1 for predicting risk assessment in car insurance. This methodology is used to ensure that the model is developed correctly and that it meets the needs of the business [15]. AIRA-ML consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment [15]that will be explain in this sections and the following two sections in details:

### *A. Business Understanding*

This phase focuses on understanding business needs, which is then used to create an accurate predictive model using machine learning techniques. This phase consists of the following two sub phases:

*1) Determine business objective*: Classify customers using historical data such as insured data, vehicle data, and claims data for a deeper understanding of the data due to its different sources as shown in Table I. The Remark column in Table I provides additional insights such as relationships between independent variables were determined by preliminary examination and expert opinions to identify the dependent variable.

*2) Determine machine learning goals*: ML techniques in auto insurance can provide valuable information but face challenges like risk assessment using unbalanced data. Thus, this work aims to illustrate the effects of imbalanced data and select the most effective resampling technique.

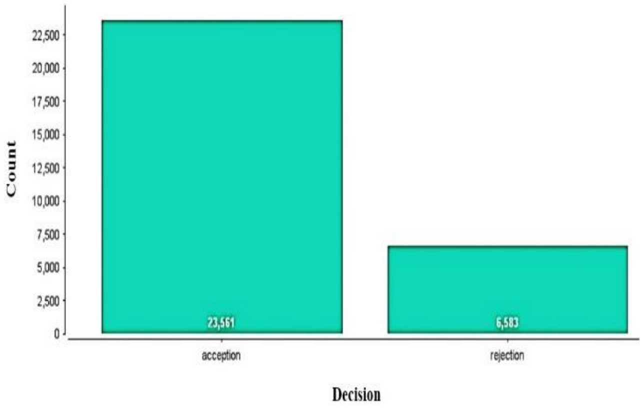

Fig. 1. Imbalanced data in the dataset.

TABLE I. SUMMARY OF THE ATTRIBUTES USED WITH THEIR DESCRIPTION

| No. | Factors Group | Features Name | Description | Remark |
|---|---|---|---|---|
| 1 | Insured data | Age | Age of Policyholder | Independent |
| 2 | | Location | Address | |
| 3 | | Job | Class of business | |
| 4 | | Hstatus | Health status | |
| 5 | | Qualification | Educational Qualification | |
| 6 | Vehicle data | Make | Make of vehicle | |
| 7 | | Model | Model of vehicle | |
| 8 | | Body | Body type of vehicle | |
| 9 | | Cc | Horsepower or CC | |
| 10 | | Vage | Age of vehicle | |
| 11 | | Use | Sub class of business (based on the purpose of use of vehicle) | |
| 12 | | Availparts | Availability of spare parts | |
| 13 | | Mileage | Mileage | |
| 14 | Claim data | Premium | Premium (averaged) | |
| 15 | | Tclaimc | Total claim cost | |
| 16 | | Insurance Val | Insurance Value | |
| 17 | | Category | The level of probability of risk | |
| 18 | Decision | Decision | The decision to accept or refuse a policy | Dependent |

## B. Data Understanding

This phase focuses attention on finding, gathering, and analyzing the data sets that are used in AIRA-ML model phases. This phase is composed of four main tasks:

*1) Collect initial data*: a real-life data from an Egyptian car insurance company's policy and claims database were used, as well as manual formats for collecting vehicle and owner information during underwriting and claim requests.

*2) Describe the data*: The features in the dataset are 18, as shown in Table I, and 30144 records for insurance policy renewal that construct the used dataset, where insurance policy for 23561 client were renewed and 6583 were refused as shown in Fig. 2. This figure shows an imbalanced dataset problem that this work will address.

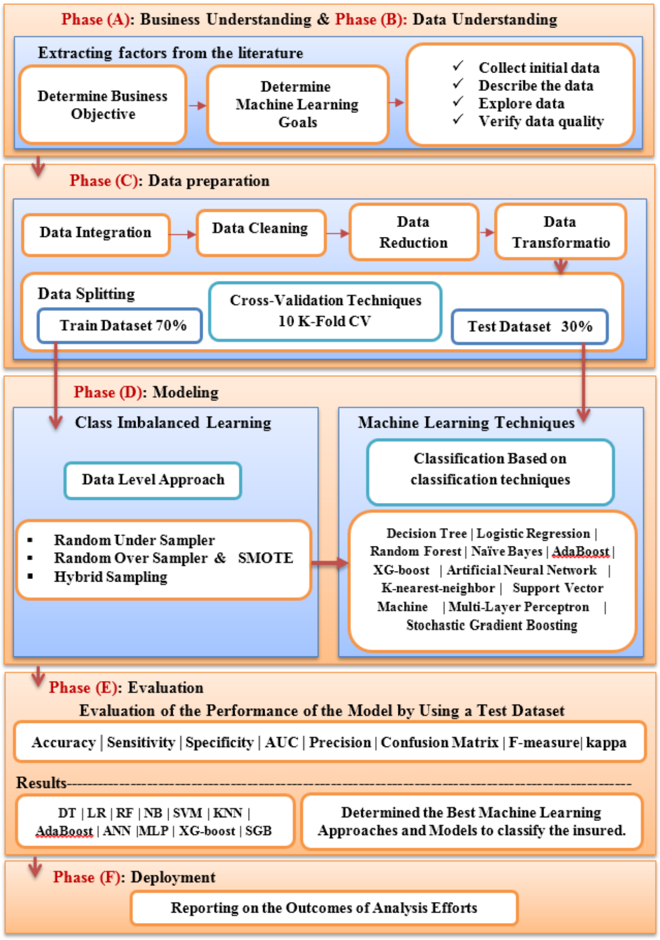
Phase (A): Business Understanding & Phase (B): Data Understanding
Extracting factors from the literature
Determine Business Objective
Determine Machine Learning Goals
✓ Collect initial data
✓ Describe the data
✓ Explore data
✓ Verify data quality
Phase (C): Data preparation
Data Integration
Data Cleaning
Data Reduction
Data Transformatio
Data Splitting
Train Dataset 70%
Cross-Validation Techniques
10 K-Fold CV
Test Dataset 30%
Phase (D): Modeling
Class Imbalanced Learning
Data Level Approach
▪ Random Under Sampler
▪ Random Over Sampler & SMOTE
▪ Hybrid Sampling
Machine Learning Techniques
Classification Based on classification techniques
Decision Tree | Logistic Regression | Random Forest | Naïve Bayes | AdaBoost | XG-boost | Artificial Neural Network | K-nearest-neighbor | Support Vector Machine | Multi-Layer Perceptron | Stochastic Gradient Boosting
Phase (E): Evaluation
Evaluation of the Performance of the Model by Using a Test Dataset
Accuracy | Sensitivity | Specificity | AUC | Precision | Confusion Matrix | F-measure| kappa
Results
DT | LR | RF | NB | SVM | KNN | AdaBoost | ANN |MLP | XG-boost | SGB
Determined the Best Machine Learning Approaches and Models to classify the insured.
Phase (F): Deployment
Reporting on the Outcomes of Analysis Efforts

Fig. 2. AIRA-ML model.

*3) Explore data*: It is possible to infer remark column to explains that relationships between the independent variables and determine the dependent variable through preliminary examination and expert opinions. The insurance policies were classified into one of three possible categories risk (low, medium, or high) based on an annual assessment made by the insurance company. 17 features were used to classify the policies, including "accept the insurance request" and "refusal of insurance" as illustrated in Table I.

*4) Verify data quality*: This process plays a crucial role in enhancing the quality and integrity of the data, through checking out data completeness and correctness to enable more effective data-driven decision-making and insights extraction for the AIRA-ML model.

## C. Data Preparation

Data preparation involves cleaning, transformation, feature engineering, integration, reduction,and splitting the organizing raw data for ML algorithm, which form training of classification model [8] [16]. These processes were conducted through a data science platform offering tools called RapidMiner that will be also used throughout all phases of AIRA-ML model.

*1) Data cleaning*: Data cleaning is a critical process that was performed on identifying and correcting or removing errors, inconsistencies, and inaccuracies within a dataset. By eliminating duplicate entries, handling missing values, correcting formatting issues, and dealing with outliers. In order to ensure that the dataset is accurate, reliable, and ready for further analysis.

*2) Data transformation*: The data transformation is performed on the feature that captures whether a car is used privately or commercially, which determines a specific feature. To facilitate analysis, the "Nominal to Numerical" operator is employed to convert categorical feature into numerical values. For example, in Table I, "private" is denoted as (1) and "commercial" as (0) for car usage. This conversion is applied to all categorical features in the analysis and in integral sub-process step within the AIRA-ML model, as depicted in Fig. 3. This ensures that the knowledge obtained from this transformation can be effectively utilized across the categorical features.

*3) Data integration*: This step "Join" combines data from different sources, as in the Factors Group in Table I, collecting vehicle and owner information during underwriting and claim requests. As shown in Fig. 3.

*4) Data reduction*: The "Select Attributes" selects only the most relevant features for analyzing relationships between the independent variables and determining the dependent variable through preliminary examination and expert opinions, as shown in Table I, and this can be determined through the RapidMiner as shown in Fig. 3.

*5) Data splitting*: The "Split Data" divides datasets into training and testing sets in the AIRA-ML model, splitting the data into 30% for testing and 70% for training [14], improving model performance and accuracy by ensuring data format and relevant features. This comprehensive tool helps organize raw data for machine learning training model as illustrated in Fig. 3[8].

## *D. Modeling*

*1) Imbalanced data and a data-level approach:* Learning from imbalanced data is a challenging problem in machine learning, as real-life datasets often have imbalanced class distributions where a minority class has fewer samples than the majority class. As illustrated in Fig 2. This leads to biased results in standard machine learning algorithms. Minority classes may have more critical information and higher value, making it crucial to distinguish them. To overcome the bias, various techniques have been proposed in the field of imbalanced learning [16] [17].

Unbalanced data has a big impact on classification algorithm performance [8]. So this paper discusses resampling techniques like Random OverSampler, Random UnderSampler, and SMOTE to address data imbalance and improve machine learning algorithms' performance. The authors compare these techniques and suggest improvements in classification algorithm performance.

*a) Data level methods*: The data-level approach involves oversampling and undersampling techniques to maintain balance between classes before classification [18]. And this is what it was applied for in the AIRA-ML model, as shown in Fig 1, and the implementation is illustrated in Fig. 3.

- Under-Sampling Methods

Under-sampling, also known as random under-sampling, is a technique to address unbalanced data by removing cases from the majority class of the training dataset [18].

- Over-sampling methods

Random Over-Sampling is a bootstrap-based technique for binary classification in imbalanced classes. It creates synthetic samples using conditional density estimation, working with continuous and categorical data. The technique maintains constant sample diversity without creating new samples [19].

SMOTE (Synthetic Minority Oversampling Technique) is a method of oversampling that creates fresh minority samples by mixing two minorities with one of their K nearest neighbours [6].

- Hybrid methods

Hybrid methods are a mixture of over-sampling and under-sampling methods at the data level. Hybrid sampling combines oversampling and undersampling to increase minority class numbers while decreasing majority class numbers.

*2) Implementing the data-level methods*

*a) RUS (Random Under Sampling)*: RUS was implemented by simply choosing a random sample from the acceptance class ("majority class") that corresponds to the number of samples from the rejection class ("minority class"). Random undersampling of the majority class was accomplished as illustrated in Fig. 3.

*b) ROS (Random Over Sampling)*: To implement ROS, a procedure that produces an equal number of replicants of the minority class as samples of the majority class was developed. The bootstrapping operator in step resampling Techniques, as shown in Fig. 3. By doing several resamples of the original dataset using random oversampling of the rejection class with replacement, this operator generates a bootstrapped sample from the original dataset.

*c) The ROS/RUS (Random Over Sampling / Random Under Sampling)*: The aforementioned approach is modified in the ROS/RUS, where a sample size of the acceptance class and rejection class is selected in the step with resampling techniques as illustrated in Fig. 3.

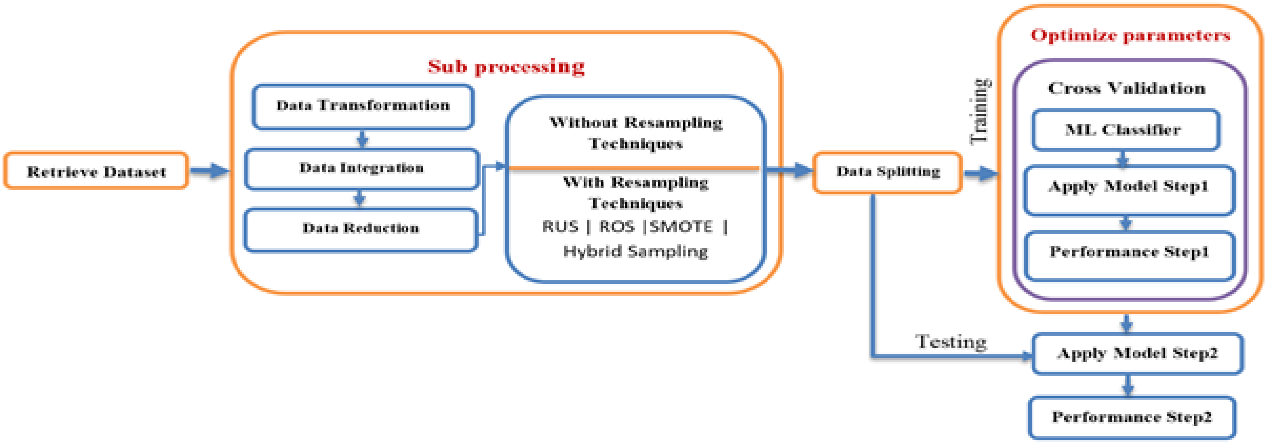

Fig. 3. The Process performed in RapidMiner with AIRA-Ml Model.

*d) SMOTE (Synthetic Minority Oversampling Technique)*: The SMOTE technique was used on the dataset until the ratio of minority samples to majority samples is equal as illustrated in Fig. 3.

*3) Modeling techniques:* In the AIRA-ML model, 11 machine learning classifiers are used, as previously stated in Fig. 1. The AIRA-ML model incorporates 10-fold cross-validation to ensure fair comparisons by selecting machine learning parameters for each model. RapidMiner parameter optimization feature is employed to determine the ideal values for the chosen parameters, resulting in optimal outcomes across various machine learning models. The 10-fold cross-validation method is utilized by dividing the data into two groups: approximately 30% for test data and 70% for training data. Models are developed using the training data and evaluated using the test data.

## IV. Results and Discussion

### A. Evaluation

Evaluation techniques calculate the effectiveness of classifiers in selecting the best applied model. Accuracy alone may not solve classification problems due to bias in majority class results, especially in imbalanced data [13] [16]. Consequently, Confusion Matrix evaluation criteria are used for measuring accuracy, sensitivity, specificity, precision, recall, AUC stands for "Area Under the Receiver Operating Characteristic Curve". It is a metric used to evaluate the performance of a binary classifier, and F-Measure, are utilized, beside Kappa Statistics to accurate more precise insurance policy renewal acceptance/ rejection.

*1) Confusion matrix*: A confusion Matrix is employed in binary classification problems to ensure accurate predictions of class outputs and facilitate the comparison between predicted and actual classes [16]. The matrix, as shown in Table II, displays the proportions of correctly classified samples (TP) and incorrectly classified samples (FP/FN). In other words, TP signifies renewal acceptance, while TN indicates rejection.

TABLE II. Confusion Matrix

| | ***Predicted Positive*** | ***Predicted Negative*** |
|---|---|---|
| *Actual Positive* | True positive (TP) | False negative(FN) |
| *Actual Positive* | False positive (FP) | True negative (TN) |

Accuracy is the fraction of predictions rate for insurance state are correct which is calculated in “(1)”, Sensitivity is the true positive ratio of positively classified cases that are actually positive for rejected insurance policies. Which is calculated using “(2)” Specificity is true negative rate of negatively classified cases that are actually negative for accepted insurance policies which is calculated using “(3)” [13] [20].

$$\text{Accuracy}= (TP+TN)/ (TP+FP+TN+FN) \qquad (1)$$

$$\text{Sensitivity}=TP/(TP+FN) \qquad (2)$$

$$\text{Specificity}=TN/(FP+TN) \qquad (3)$$

The metric “(4)” is the percentage of the relevant outcomes that assess the reliability of the classification and determine the appropriateness of acceptance or rejection decision. A recall “(5)” is a measurement of how many positive instances were correctly identified as positive, especially for imbalanced datasets [13] [20].

$$\text{Precision} =TP/(TP+FP) \qquad (4)$$

$$\text{Recall} =TP/ (TP + FN) \qquad (5)$$

The F-measure also known as the F1 score “(6)” is a measure of a model's performance that takes into account the averaging of precision and recall.

$$\text{F-measure}= (2^* \text{Precision}^* \text{Recall}) / (\text{Precision}+\text{Recall}) \qquad (6)$$

*2) Kappa statistics*: Kappa statistics play a valuable role in assessing prediction success for both class acceptance and rejection, considering factors beyond accuracy. This is particularly important when dealing with datasets that exhibit significant imbalance class. Kappa “(7)” takes into account the agreement between model predictions and actual labels, providing a measure of agreement that goes beyond accuracy alone [13].

$$K=(Pr(a)-Pr(e))/(1-Pr(e)) \qquad (7)$$

Where:

Pr(a) represents the observed agreement between the raters, which is the proportion of cases where the raters agree.

Pr(e) represents the expected agreement between the raters by chance.

Results of 11 machine learning classifiers on unbalanced data without any resampling models applied are shown in Fig. 4. The result highlights the highest accuracy rate in the classifiers with DT achieving an accuracy of 0.9015, because of machine learning techniques often ignore the minority class (rejection class) and allocate most cases to the majority class (acceptance class). Moreover, a direct proportion relationship can be seen between accuracy and specificity, whenever accuracy is high, the specificity of 0.949 is high as well due to the model consistently predicting the majority class, but reduced sensitivity to 0.6333. Precision is low at 0.8768 because the model frequently predicts the minority class incorrectly. The F-measure is also low at 0.1482 due to the combination of low precision and recall of 0.6333. Among the classifiers, kappa is low at 0.3267 because the model is not very good at predicting the minority class, and AUC is misleading because it is high and the model is not very good at predicting the minority class. On the other hand, the MPL classifier had the lowest performance measures (accuracy = 0.6888, sensitivity = 0.0825, specificity = 0.92121, AUC = 0.5021, precision = 0.7294, F1-measure = 0.14823, and kappa = 0.0871).

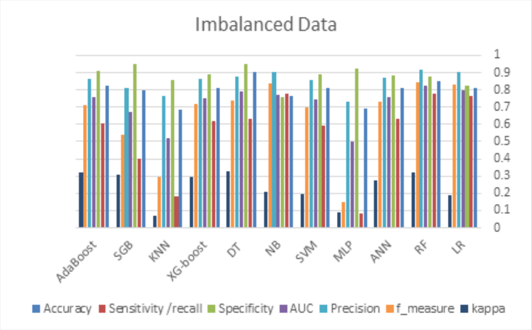

Fig. 4. Classifiers metrics for imbalanced data.

Fig. 5 shows the results of machine learning algorithms on imbalanced data that has been resampled using random oversampling. The accuracy results are not significantly improved, with the SVM classifier achieving an accuracy of 0.8736. This is understandable given that most models predict with poorer accuracy on balanced data, as they consider all classes at the same time. Accuracy is a simple metric to understand, but it overlooks several important factors that must be considered when evaluating a classifier's output. Therefore, we used additional metrics. Additionally, it is noteworthy that the sensitivity for all models with imbalanced data is lower than the sensitivity for balanced data created by random oversampling. This is because random oversampling does not address the class imbalance problem in a principled way. It simply increases the number of minority class samples lead to overfitting. The specificity for the SVM classifier is 0.8601, with a high F-measure of 0.9366. This is due to the combination of high precision 0.9662 and recall 0.9087. The kappa is low 0.4514 because the model is not very good at predicting the minority class. The AUC is also not very good 0. 8841. The KNN classifier has the lowest performance measures. The accuracy is 0.5563, the sensitivity is 0.4594, the specificity is 0.5934, the AUC is 0.5261, the precision is 0.7444, the F-measure is 0.5681, and the kappa is 0. 112.

Overall, the results show that random oversampling is not an effective way to address the class imbalance problem in machine learning. Other principled approaches, such as SMOTE or hybrid method, are applied to improve the performance of imbalanced data.

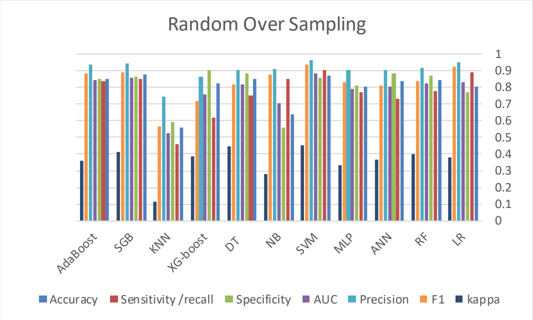

Fig. 5. Classifiers metrics using the ROS method.

The results of machine learning algorithms on imbalanced data that has been resampled using random undersampling are shown in Fig. 6. The high precision for many classifiers, such as random forest (RF) 0.9823 and Artificial Neural Network ANN 0.9505, is because the model frequently predicts the minority class correctly. The F-measure is also high for RF 0.9669 due to the combination of high precision and recall 0.9521, the kappa is high with comparison with Fig. 4 and Fig. 5. Moreover, such as ANN 0.4514, the model is very good at predicting the minority class. The AUC is 0.8391, and accuracy is 0.8713, specificity is 0.799. On the other hand, the KNN classifier has the lowest performance measures. With accuracy of 0.5001, sensitivity of 0.6913, specificity of 0.4268, AUC of 0.5591, precision is 0.7833, F1-measure of 0.7344, and kappa of 0.254.

Additionally, Random undersampling is a technique that reduces the number of majority class samples in a dataset. This can help to improve the performance of machine learning models on imbalanced data, as it reduces the bias towards the majority class. However, the results show that random undersampling can be an effective way to address the class imbalance problem in machine learning.

The best performance with SMOTE is DT classifier as shown in Fig. 7 With an accuracy of 0.8575, sensitivity of 0.9521, specificity of 0.8212, AUC of 0.9521, precision of 0.8871, F1-measure of 0.9184, and a kappa of 0.483, while the lowest performance at SMOTE is the KNN classifier with an accuracy of 0.5844, sensitivity of 0.4159, specificity of 0.6490, AUC of 0.4159, precision of 0.5321, F1-measure of 0.4668, and kappa of 0. 225. The result of the SMOTE resampling method has shown the improvement for sensitivity and specificity of the model. For example, the ANN classifier of the SMOTE results has significantly improved with a sensitivity of 0.9087 and a specificity of 0.8212; in comparison with the ANN classifier without SMOTE, which has a sensitivity of 0.6333, and a specificity of 0.8823 as shown in Fig. 4. The SMOTE resampling method is also improving the performance of other classifiers, such as the DT classifier and the KNN classifier. Overall, the results show that the SMOTE resampling method is an effective way to improve the performance of machine learning models on imbalanced data, which leads to a more accurate and reliable model.

The hybrid resampling methods are more effective than random oversampling or undersampling alone. This is because they are able to create a more balanced dataset without overfitting the models.

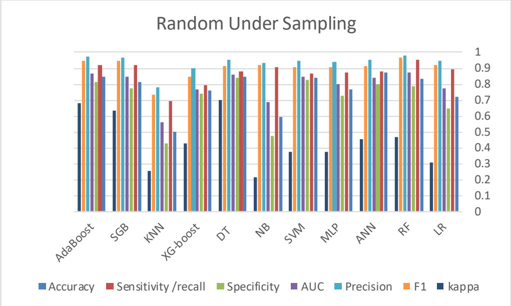

Fig. 6. Classifiers metrics using the RUS method.

The results are shown in Fig. 8. The random forest (RF) classifier with hybrid resampling has the best performance, with an accuracy of 0.9967, an AUC of 0.994, a precision of 0.9977, an F-measure of 0.9975, and a kappa of 0.992. This means that the RF classifier is very accurate in predicting whether accept or reject the renewal of the policies. It also has the smallest gap between sensitivity of 0.9977 and specificity of 0.9959, which is an important performance indicator while the MLP classifier has the lowest performance, with an accuracy of 0.5242, a sensitivity of 0.5029, a specificity of 0.5323, an AUC of 0.5171, a precision of 0.7388, an F-measure of 0.5984, and a kappa of 0.0497.

Table III compare the purposed model (AIRA-ML model) with earlier studies that used other model and resampling technique. The data was preprocessed in both its balanced and unbalanced states to improve the accuracy of training result and the effectiveness of machine learning algorithms.

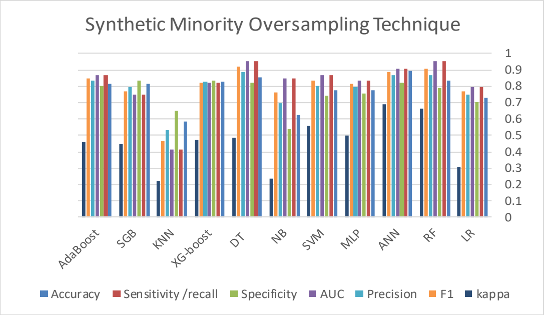

Fig. 7. Classifiers metrics using the SMOTE method.

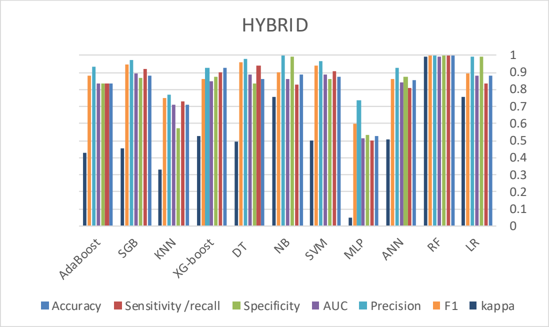

Fig. 8. Classifiers metrics using the hybrid method.

TABLE III. COMPARATIVE RESULTS FOR NEW APPROACH PERFORMANCE AGAINST RELATED WORKS

| **Related work** | **Resampling Methods** | **Machine Learning Techniques** | **The best model** | **Performance Measures** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | ***Accuracy*** | ***Sensitivity / Recall*** | ***Specificity*** | ***AUC*** | ***Precision*** | ***F-measure*** | ***Kappa*** |
| [9] | Random Sampling, SMOTE | KNN,<br>DT | DT | 0.75 | 0.749 | × | × | × | × | × |
| [10] | × | LR,<br>XGBoost | XGBoost | 0.8397 | 0.0790 | 0.9022 | × | × | × | × |
| [11] | × | NB, SVM, LR, NN, DT | DT | 0.9540 | × | × | × | × | × | × |
| [12] | Bootstrapping | SVM , RF | RF | 0.9836 | 0.9471 | × | × | 0.9515 | 0.9490 | × |
| [13] | ROS | LR,C50,J48,XGBoost, DT, NB, K-NN,RF | RF | 0.8677 | 0.9717 | 0.71 | 0.840 | 0.9429 | 0.8101 | 0.7117 |
| [14] | ROS, RUS, Hybrid, SMOTE | C5.0, C4.5, CART, Bagged CART, RF, XGBoost, SGB, AdaBoost | AdaBoost | 0.9940 | 0.9294 | 0.9982 | × | × | × | × |
| AIRA-ML Model | ROS, RUS, SMOTE, Hybrid | LR, ANN, M LP, SVM, NB, DT, XG-boost, KNN, SGB, AdaBoost,RF | RF | 0.9967 | 0.9977 | 0.9959 | 0.8701 | 0.9977 | 0.9975 | 0.992 |
| ×- Not used; | | | | | | | | | | |

Finally, the hybrid approach performs is much better than the results of the earlier research. This approach, which is very accurate with predicting the decision of policies renewal at the same time as an essential performance indicator, that has the smallest gap between sensitivity and specificity. AUC, F1-score, and Kappa statistics are used as other Performance Measures to ensure that the model is effective as a trustworthy instrument for risk assessment and insurance policies activities in the insurance industry.

### B. Deployment

The deployment of the model is beyond the scope of this work and is the responsibility of the insurance company. The purpose of the model is to increase knowledge of the data, and the knowledge gained will need to be organized and presented in a way that the customer can use it. The research in this paper was conducted primarily for academic purposes, but the results can be used by the financial sector to apply machine learning technology to improve their business practices and gain a competitive edge.

However, the research has identified a task that needs more consideration in future work. Proper handling and concern for information are strongly recommended in data mining research.

## V. Conclusion and Future Work

The insurance industry faces a significant challenge in predicting risk assessment in insurance policies. Thus, this paper proposes an accurate predictive model using machine learning (ML) and resampling techniques to assist insurance companies in making better pricing decisions. The results demonstrate that ML can be used to create an accurate predictive model for auto insurance risk, which can improve insurance acceptance and pricing decisions. Additionally, the results demonstrate that ML can be effective in addressing data imbalance problems in the auto insurance sector. The hybrid resampling technique outperformed all other resampling techniques, achieving an accuracy of 99.6% for the random forest (RF) classifier. This suggests that the hybrid resampling method is a promising approach for dealing with class imbalance problems in ML.

Further research is required to compare the efficiency measures using various datasets from various fields to prove the prediction efficiency of a random forest classifier with resampling methods to solve the imbalanced data problem. And future work may be done in the following directions: Using hybrid resampling techniques to improve comparison and performance with machine learning classifiers, apply this methodology to other sectors of insurance or any other sector with the same problem, which has an imbalance of data.

## References

[1] Et. al., Nadia Yas. 2021. “Implications of Compulsory Car Accident Insurance Comparative Study.” Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12 (2): 2410–20. https://doi.org/10.17762/turcomat.v12i2.2052.

[2] Radic, M., P. Herrmann, P. Haberland, and Carla R. Riese. 2022. “Development of a Business Model Resilience Framework for Managers and Strategic Decision-Makers.” Schmalenbach Journal of Business Research 74 (4). https://doi.org/10.1007/s41471-022-00135-x.

[3] F. B. Insights, "Automotive Usage Based Insurance Market Size | Growth [2028]," Jun 2023. [Online]. Available: https://www.fortunebusinessinsights.com/automotive-usage-based-insurance-market-104103. [Accessed 1 8 2023].

[4] Drakulevski, Ljubomir, and Tamara Kaftandzieva. 2021. “Risk Assessment Providing Solid Grounds For Strategic Management In The Insurance Industry.” European Scientific Journal ESJ 17 (15): 38–56. https://doi.org/10.19044/esj.2021.v17n15p38.

[5] Rawat, Seema, Aakankshu Rawat, Deepak Kumar, and A. Sai Sabitha. 2021. “Application of Machine Learning and Data Visualization Techniques for Decision Support in the Insurance Sector.” International Journal of Information Management Data Insights 1 (2): 1–15. https://doi.org/10.1016/j.jjimei.2021.100012.

[6] T Wongvorachan, Tarid, Surina He, and Okan Bulut. 2023. “A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining.” Information (Switzerland) 14 (1): 1–15. https://doi.org/10.3390/info14010054.

[7] Andry, Johanes Fernandes, Henny Hartono, Honni, Aziza Chakir, and Rafael. 2022. “Data Set Analysis Using Rapid Miner to Predict Cost Insurance Forecast with Data Mining Methods.” Journal of Hunan University Natural Sciences 49 (6): 167–75. https://doi.org/10.55463/issn.1674-2974.49.6.17.

[8] Madyatmadja, Evaristus Didik, Samuel Imanuel Jordan, and Johanes Fernandes Andry. 2021. “Big Data Analysis Using Rapidminer Studio to Predict Suicide Rate in Several Countries.” ICIC Express Letters, Part B: Applications 12 (8): 757–64. 10.24507/icicelb.12.08.757.

[9] Wuyu, Sisay, and Patrick Cerna. 2018. “Risk Assessment Predictive Modelling in Ethiopian Insurance Industry Using Data Mining.” Software Engineering 6 (4): 121–27. https://doi.org/10.11648/j.se.20180604.13.

[10] Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. 2019. “Predicting Motor Insurance Claims Using Telematics Data—XGboost versus Logistic Regression.” Risks 7 (2). https://doi.org/10.3390/risks7020070.

[11] Utomo, Deddy, Noperida Damanik, and Indra Budi. 2021. “Classification on Participants Renewal Process in Insurance Company: Case Study PT XYZ.” In 2021 9th International Conference on Information and Communication Technology (ICoICT), 576–81. IEEE. https://doi.org/10.1109/ICoICT52021.2021.9527479.

[12] Alamir, Endalew, Teklu Urgessa, Ashebir Hunegnaw, and Tiruveedula Gopikrishna. 2021. “Motor Insurance Claim Status Prediction Using Machine Learning Techniques.” International Journal of Advanced Computer Science and Applications 12 (3): 457–63. https://doi.org/10.14569/IJACSA.2021.0120354.

[13] Hanafy, Mohamed, and Ruixing Ming. 2021. “Machine Learning Approaches for Auto Insurance Big Data.” Risks 9 (2): 1–23. https://doi.org/10.3390/risks9020042.

[14] Hanafy, Mohamed, and Ruixing Ming. 2021. “Improving Imbalanced Data Classification in Auto Insurance by the Data Level Approaches.” International Journal of Advanced Computer Science and Applications 12 (6): 493–99. https://doi.org/10.14569/IJACSA.2021.0120656.

[15] Martinez-Plumed, Fernando, Lidia Contreras-Ochando, Cesar Ferri, Jose Hernandez-Orallo, Meelis Kull, Nicolas Lachiche, Maria Jose Ramirez-Quintana, and Peter Flach. 2021. “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories.” IEEE Transactions on Knowledge and Data Engineering 33 (8): 3048–61. https://doi.org/10.1109/TKDE.2019.2962680.

[16] Baran, Sebastian, and Przemysław Rola. 2022. “Prediction of Motor Insurance Claims Occurrence as an Imbalanced Machine Learning Problem.” ArXiv abs/2204.0: 1–12. https://doi.org/10.48550/arXiv.2204.06109.

[17] Mohammed, Roweida, Jumanah Rawashdeh, and Malak Abdullah. 2020. “Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results.” 2020 11th International Conference on Information and Communication Systems, ICICS 2020, no. May: 243–48. https://doi.org/10.1109/ICICS49469.2020.239556.

[18] Dasari, Siva Krishna, Abbas Cheddad, Jonatan Palmquist, and Lars Lundberg. 2022. “Clustering-Based Adaptive Data Augmentation for Class-Imbalance in Machine Learning (CADA): Additive Manufacturing

Use Case.” Neural Computingand Applications 6. https://doi.org/10.1007/s00521-022-07347-6.

[19] Le, Tuong, Minh Thanh Vo, Bay Vo, Mi Young Lee, and Sung Wook Baik. 2019. “A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction.” Complexity 2019. https://doi.org/10.1155/2019/8460934.

[20] Uddin, Moin, Mohd Faizan Ansari, Mohd Adil, Ripon K. Chakrabortty, and Michael J. Ryan. 2023. “Modeling Vehicle Insurance Adoption by Automobile Owners: A Hybrid Random Forest Classifier Approach.” Processes 11 (2): 1–16. https://doi.org/10.3390/pr11020629.