Flight Delay Prediction for Mitigation of Airport Commercial Revenue Losses Using Machine Learning on Imbalanced Dataset

Rae Arun Sugara, Diana Purwitasari
Technology Management Department
Sepuluh Nopember Institute of Technology
Surabaya, Indonesia
sugara.206032@mhs.its.ac.id, diana@if.its.ac.id

2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM) | 978-1-6654-7650-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/CENIM56801.2022.10037369

***Abstract*—Flight delay is one of the factors that affect the decline in customer satisfaction and airport revenue. In addition to influencing customer perceptions of airport services, flight delay also has an impact on decreasing airport revenue and operation. This study models a flight delay prediction, and the process is carried out using Decision Tree, Random Forest, Gradient Boosted Tree, and XGBoost Tree algorithms. This study has also used and merged the weather characteristic data as secondary data to the airport operational flight data. To anticipate the imbalanced class, several sampling techniques were applied. Synthetic Minority Over Sampling Technique (SMOTE), Random Over-Sampling (ROS), Random Under-Sampling (RUS), and combining ROS with RUS are being used. The result of processing the analysis is in the form of a model to predict the category of flight delay. The model has been evaluated by using the Confusion Matrix and Area Under ROC Curve (AUC) value. The result of this study shows the Random Forest classifier with the combination of ROS + RUS technique and data split ratio of 90:10 gave the highest accuracy, error rate, and AUC value as shown as 82.58%, 17.42%, and 81.1% respectively on data testing. The result of the flight delay prediction model is expected to be a strategic recommendation for determining airport policies in the future. By implementing the best strategy related to the airport operation, it could carry out commercial planning in order to optimize airport commercial revenue.**

Keywords*— *flight delay, classifier, confusion matrix, Area Under Curve (AUC), sampling technique, commercial revenue.

I. Introduction

Flight delays will not only increase economic losses, but can also weaken the performance of air transportation operational systems, adversely affecting passengers, airlines, and airport planning [1, 2]. Any discrepancy between the scheduled departure or arrival of the aircraft and the actual time can be recognized as a delay [1]. Flight delay is unavoidable and contributes significantly to the advantages and disadvantages of airports, airlines, and passengers [3, 4]. Flights in the United States were delayed by 31.1% by 15 minutes in 2013, flights in Europe were delayed by 36% by 5 minutes, and flights in Brazil were delayed by approximately 16.3% by 30 minutes [2]. The data shows how important the flight delay topic is and how it affects [1]. The annual cost in China caused by flight delays is estimated at more than USD 7.4 billion [5]. As for the economy in the United States, flight delays in 2007 have cost implications that are estimated at USD 32.9 billion [6]. The impact of flight delays can practically affect the passenger experience at the airport and form a bad judgment for the airport. Research [7] reveals the fact that passengers who experience delays, especially service failures, usually react immediately with negative emotional dominance.

The non-aeronautical business takes advantage of the potential of the airport as a place visited by the public to earn income in addition to the aeronautical business. As a result, terminal design has taken this need into account [8], with check-in and departure areas being the most important elements [9]. Flight delays resulted in the mapping of passengers on the flow of passengers from check-in to the departure waiting room not being in accordance with the commercial space plan. Aviation regulators have set flight service standards, including on-time performance and consequences. However, flight delays are unavoidable. As a result, the airport operator has to mitigate the commercial losses caused by the flight delay. High-accuracy prediction models are widely recognized as effective tools to reduce flight delays, reduce economic costs, and passenger dissatisfaction [10]. Making flight status prediction can be a solution in terms of mitigating the airport's commercial losses. This can be a reference for managing the flow of passengers after carrying out the check-in process, directing passengers through the right commercial area, and placing passengers at the boarding gate that has commercial space with the right passenger target.

Research [11] by combining flight and weather data using Random Forest (RF) algorithm resulted in predictive modeling with accuracy, precision, and recall values, respectively of 96.48%, 94.39%, and 90.26%. Predictive modeling research conducted [12] also combines flight data with weather using the RF algorithm and produces an accuracy of 85.8% and a recall of 86.9%. Research [13] has made different classifications and regression models by testing Logistic Regression (LR), Single Classification Trees, Bagging, Boosting, Linear Regression (LR), Neural Networks (NN), and Random Forests (RF) algorithms. Then the RF algorithm was chosen because of its superior performance. The performance of four machine learning models in [14] which are K-Nearest Neighbors (K-NN) [15], Support Vector Machines (SVM) [16], Naïve Bayes (NB), and Random Forests (RF) has been compared for predicting flight delays. The result showed RF as the best classifier that gave an accuracy of 78.02% among the four classifiers. The air route information (e.g., traffic flow and size of each route) in this study was not considered, which restricts the model's ability to reach higher accuracy.

However, several reasons are restricting the existing methods from improving the accuracy of flight delay prediction. The reasons are summarized, such as the variety of causes that affect the flight delays, the complexity and relevancy between causes, and also the insufficient availability of flight data [17].

In this study, we aim to model the prediction of flight delay. Since research on how bad weather affects airline delays is essential for the effectiveness of flight operations [18], we combined the airport's flight operation data and weather data related to the airport environment. The method for processing data and analyzing the model results by using the machine learning algorithm. Machine learning techniques are increasingly being used to solve real-world problems. Machine learning plays an important role in the aviation industry, e.g., in predicting flight delays. The various machine learning techniques for classification can be used to classify the flight status. In this study, we apply Decision Tree (DT), Random Forest (RF), Gradient Boosted Tree (GBT), and XGBoost Tree (XGBoost) algorithms. We also used several sampling techniques in terms of handling the imbalanced dataset. In addition, the results of this flight delay prediction model study can be considered as recommendations to determine whether commercial planning policies at the airport terminal operation should be resumed or evaluated. Thus, the results of an accurate flight delay prediction model are expected to produce effective strategic and operational policies in terms of optimizing airport commercial revenues.

II. Methodology

A. Data Collection

This study was conducted by combining flight data owned by the airport with weather data. Airport data is selected using airport data that has a majority of indirect routes or requires a hub (indirect flight). The concentration of traffic at several hub airports on one side is increasingly recognized as a major cause of congestion and delay, as well as the complexity of airline and airport operations, particularly for connecting passengers [19, 20]. Fig. 1 shows the flow process of this study.

The flight operation data has been collected from Pattimura International Airport (Pattimura Airport), which also considers the percentage of the status of flight punctuality with the data shown daily from January 2020 to December 2021. The total number of samples is 19,910, which will be used for further training and testing processes. The secondary data to be combined with the primary one is the weather data that has been collected by the Meteorological, Climatological, and Geophysical Agency (BMKG). We have combined the flight and weather data by using the actual date as a primary key. Furthermore, the dataset features were devided by class "no delay" (class = 0) and "delay" (class = 1). The variable description used in the models is shown in Table I.

TABLE I. Description of Variable

Variable	*Description*
Independent Variable	
STDinMinutes	Scheduled departure time in minute dimension
PAX_TOTAL	The number of total passengers on a flight
BAG1	The amount of hand-carried cabin load in a flight
CARGO	The total amount of cargo carried by an aircraft during a flight
CARGO_TOTAL	The amount of total aircraft cargo load, mail, and hand-carried cabin load in a flight
RW	The runway number of aircraft that take-off or land
APRONum	Parking lot number of the aircraft parking bay
PAX_CAP	Aircraft capacity for passengers
DEST_NUM	Numeric flight destination or origin
OPERATOR_NUM	Aircraft operators in numeric
TYPEAC_NUM	Numeric aircraft type
DayofWeek	Contact day of the week
DayofMonth	Contact day of the month
DayofYear	Contact day of the year
MONTH	The contact month
Tn	The environment's minimum temperature
Tx	The environment's average temperature
RH_avg	Relative humidity on average
RR	Rainfall
ss	The duration of the sun
ff_x	Maximum wind speed
ddd_x	Wind direction at maximum wind speed
ff_avg	Average wind speed
ddd_carNum	The majority of wind directions
Dependent Variable	
Flight status	Delay or no delay

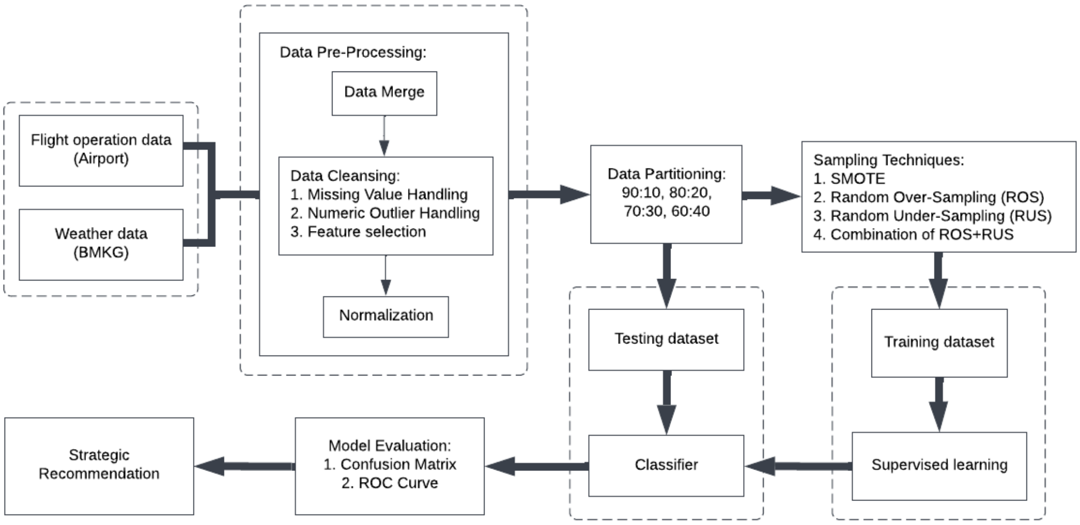

Fig. 1. Methodology flow to predict the status of flight delays

TABLE II. CORRELATION COEFFICIENT OF DATA FEATURE

	DayofWeek	DayofMonth	DayofYear	MONTH	STDinMinutes	Tn	Tx	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_carNum	PAX_TOTAL	BAG1	CARGO	CARGO_TOTAL	RW	APRONum	PAX_CAP	DEST_NUM	OPERATOR_NUM	TYPEAC_NUM
DayofWeek	1.0	0.023	0.010	0.005	-0.001	0.009	-0.030	0.071	0.023	0.012	0.010	-0.041	-0.019	0.012	-0.005	0.017	-0.005	0.020	0.001	0.018	-0.009	-0.000	0.042	-0.137
DayofMonth	0.023	1.0	0.118	0.058	-0.014	-0.046	-0.018	-0.004	-0.009	0.038	0.031	-0.009	0.023	0.023	0.005	-0.012	-0.018	0.015	0.001	0.030	0.012	0.019	0.011	0.009
DayofYear	0.010	0.118	1.0	0.530	-0.029	-0.199	-0.284	0.288	0.127	-0.161	-0.066	-0.039	-0.014	0.145	-0.002	0.158	0.081	0.062	-0.084	0.117	-0.116	-0.136	0.036	-0.069
MONTH	0.005	0.058	0.530	1.0	-0.024	-0.274	-0.266	0.339	0.071	-0.197	-0.138	0.041	-0.161	0.094	-0.013	0.117	0.039	0.121	-0.077	0.098	0.234	-0.133	-0.038	-0.066
STDinMinutes	-0.001	-0.014	-0.029	-0.024	1.0	0.016	0.049	-0.054	-0.031	0.030	0.011	0.011	-0.006	0.000	0.081	-0.037	0.069	-0.059	0.000	-0.009	0.086	0.052	-0.085	0.161
Tn	0.009	-0.046	-0.199	-0.274	0.016	1.0	0.205	-0.174	-0.177	0.187	0.098	0.049	0.088	-0.068	0.030	-0.033	-0.028	-0.031	0.007	-0.016	-0.349	0.057	0.005	0.013
Tx	-0.030	-0.018	-0.284	-0.266	0.049	0.205	1.0	-0.750	-0.435	0.465	0.070	0.195	0.023	-0.226	0.055	-0.085	-0.105	-0.105	0.029	-0.003	0.032	0.158	-0.048	0.036
RH_avg	0.071	-0.004	0.288	0.339	-0.054	-0.174	-0.750	1.0	0.35	-0.404	-0.117	-0.146	-0.137	0.116	-0.062	0.078	0.095	0.048	-0.026	0.019	0.005	-0.147	0.032	-0.030
RR	0.023	-0.009	0.127	0.071	-0.031	-0.177	-0.435	0.350	1.0	-0.369	0.011	-0.108	0.014	0.068	-0.053	0.027	0.054	-0.000	-0.014	0.009	-0.037	-0.085	0.029	-0.020
ss	0.012	0.038	-0.161	-0.197	0.030	0.187	0.465	-0.404	-0.369	1.0	-0.013	0.097	-0.038	-0.023	0.029	-0.050	-0.066	-0.022	0.024	-0.017	-0.018	0.093	-0.016	0.017
ff_x	0.010	0.031	-0.066	-0.138	0.011	0.098	0.070	-0.117	0.011	-0.013	1.0	-0.068	0.427	0.119	0.016	-0.023	-0.014	-0.018	0.037	0.001	0.006	0.030	0.001	0.000
ddd_x	-0.041	-0.009	-0.039	0.041	0.011	0.049	0.195	-0.146	-0.108	0.097	-0.068	1.0	-0.010	-0.088	0.036	0.008	-0.038	0.017	0.000	0.021	0.074	0.050	-0.035	0.008
ff_avg	-0.019	0.023	-0.014	-0.161	-0.006	0.088	0.023	-0.137	0.014	-0.038	0.427	-0.010	1.0	0.206	-0.003	-0.012	0.011	-0.028	0.018	-0.025	-0.057	0.002	0.016	0.017
ddd_carNum	0.012	0.023	0.145	0.094	0.000	-0.068	-0.226	0.116	0.068	-0.023	0.119	-0.088	0.206	1.0	-0.007	0.046	0.034	0.016	-0.003	-0.003	-0.033	-0.059	0.005	-0.010
PAX_TOTAL	-0.005	0.005	-0.002	-0.013	0.081	0.030	0.055	-0.062	-0.053	0.029	0.016	0.036	-0.003	-0.007	1.0	0.340	-0.179	0.453	0.153	-0.051	0.330	0.312	-0.319	0.415
BAG1	0.017	-0.012	0.158	0.117	-0.037	-0.033	-0.085	0.078	0.027	-0.050	-0.023	0.008	-0.012	0.046	0.340	1.0	0.148	0.606	0.100	-0.112	0.140	-0.294	-0.418	0.245
CARGO	-0.005	-0.018	0.081	0.039	0.069	-0.028	-0.105	0.095	0.054	-0.066	-0.014	-0.038	0.011	0.034	-0.179	0.148	1.0	0.054	-0.124	-0.005	-0.078	-0.294	-0.132	0.019
CARGO_TOTAL	0.020	0.015	0.062	0.121	-0.059	-0.031	-0.105	0.048	-0.000	-0.022	-0.018	0.017	-0.028	0.016	0.453	0.606	0.054	1.0	0.169	-0.100	0.340	-0.043	-0.420	0.410
RW	0.001	0.001	-0.084	-0.077	0.000	0.007	0.029	-0.026	-0.014	0.024	0.037	0.000	0.018	-0.003	0.153	0.100	-0.124	0.169	1.0	-0.009	0.168	0.102	-0.193	0.210
APRONum	0.018	0.030	0.117	0.098	-0.009	-0.016	-0.003	0.019	0.009	-0.017	0.001	0.021	-0.025	-0.003	-0.051	-0.112	-0.005	-0.100	-0.009	1.0	0.020	0.096	0.170	-0.121
PAX_CAP	-0.009	0.012	-0.116	0.234	0.086	-0.349	0.032	0.005	-0.037	-0.018	0.006	0.074	-0.057	-0.033	0.330	0.140	-0.078	0.340	0.168	0.020	1.0	0.174	-0.378	0.295
DEST_NUM	-0.000	0.019	-0.136	-0.133	0.052	0.057	0.158	-0.147	-0.085	0.093	0.030	0.050	0.002	-0.059	0.312	-0.294	-0.294	-0.043	0.102	0.096	0.174	1.0	0.076	0.207
OPERATOR_NUM	0.042	0.011	0.036	-0.038	-0.085	0.005	-0.048	0.032	0.029	-0.016	0.001	-0.035	0.016	0.005	-0.319	-0.418	-0.132	-0.420	-0.193	0.170	-0.378	0.076	1.0	-0.515
TYPEAC_NUM	-0.137	0.009	-0.069	-0.066	0.161	0.013	0.036	-0.030	-0.020	0.017	0.000	0.008	0.017	-0.010	0.415	0.245	0.019	0.410	0.210	-0.121	0.295	0.207	-0.515	1.0

B. *Data Pre-processing*

Airport operation flight data and associated weather data are extracted. Referring to the Regulation of the Minister of Transportation of the Republic of Indonesia, Conditions for flight are those having more than or equal to 30 minutes from the scheduled time. We split the time stamps into four parts [17, 11] to enrich the content of the dataset: day of week, day of month, day of year, and month.

By pre-processing, all categorical variable data are converted to numerical. Machine learning algorithms earn a better performance by using numerical variables [18]. The first step in pre-processing data is by checking for wrong data, missing values, and outliers. Then, as the second step, we handled them and normalization (min, max) is also implemented, which increases the convergence speed and the model's prediction accuracy [17]. In this study, the dataset is randomly partitioned into four scenario ratios for model training and model testing, respectively: 90:10, 80:20, 70:30, and 60:40.

1) Feature Selection: Feature selection is one of the essential contents of machine learning, which is intended to eliminate redundant features, increase the accuracy of the model, and also reduce the processing time [21]. Variable selection is carried out if there is missing variable data and the variable is entered into the next modeling to obtain a less accurate model. If there is more than 50% missing variable data, then the variable will be eliminated. If less than 50% of the data is lost, then data imputation will be carried out. Data imputation is a way of dealing with missing data by providing possible values based on available data information. The mean equation is used to handle missing value data as shown in (1).

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad (1)$$

In addition, the selected variables are determined based on the correlation coefficient calculation between the candidate features and selected features. Regarding to Table II, the correlation demonstrates the features used to predict a delay label. We have a used 0.75 correlation coefficient as a cutoff. If the correlation coefficient between features is greater than 0.75, then the candidate feature is erased [22] because a strong correlation may possibly leads the developing models to be overfitting [23]. Thus, to avoid overfitting, we have reduced the initial feature set consisting of twenty-five features to a final feature set consisting of twenty-four features. The result shows there is no strong correlation coefficient that has a coefficient greater than 0.75 between features. Low correlation is also shown by features that have minus values, represented by the light blue color in Table II. Since the raw data of our model contains some extreme values, long tails, and outliers, we handled them by removing the data [24]. As [25] set the primary and weather data, which have a large difference between the min and max limits, we also applied the normalization (min, max) range to improve the performance of the estimation methods [26] and reduce the processing time [21]. The formula is as follows:

$$x' = \frac{X - X_{mean}}{X_{max} - X_{min}}, \qquad (2)$$

Where X_{mean} is the mean value, X_{max} is the maximum value, and X_{min} is the minimum value.

2) Sampling Technique: The statistical data distribution we have for the flight delay class or category is imbalanced. From 19,909 data, it contains 21.34% of flight data for class as delayed and 78.66% of flight data for no delay. An imbalance class may cause the classifier to learn concepts related to the majority class and dominate the minority class [24]. Simple random over-sampling [13] and random under-sampling techniques [12] are recommended to balance the flight delay dataset to make it suitable for classification. In this study, we apply several sampling techniques, which are Synthetic Minority Over Sampling Technique (SMOTE), Random Over-Sampling (ROS), Random Under-Sampling (RUS), and combining ROS and RUS. RUS only applies to the majority class, which works by randomly balancing the majority class to bring it equal to the minority class [24]. In the RUS technique, the chance of losing potentially useful data is increased because it works by eliminating examples of the majority class [27]. Opposite to RUS, ROS balances the dataset by randomly adding examples of the minority class [24].

C. Prediction Model

We proposed several machine learning methods by using Decision Tree (DT), Random Forest (RF), Gradient Boosted Tree (GBT), and XGBoost Tree (XGBoost) algorithms.

1) Decision Tree: DT is a classification methodology arranged in a tree structure where each node shows a feature, each link shows a decision, and each leaf shows the outcome (the value of categorical or continuous) [23, 28]. Decision trees partition the input dataset at a node for randomly chosen features. A local model storing the distribution over class labels is defined at each leaf node. The model for the input variable *x* can be written in the following form:

$$f(\mathrm{x}) = \mathbb{E}[y|x] = \textstyle\sum_{m=1}^{M} w_m \phi(\mathrm{x}; \mathrm{v}_m) \tag{3}$$

where w_m is the distribution over class labels in the m^{th} region and v_m encodes the choice of variable to split and the threshold value on the path from the root to the m^{th} leaf [18].

2) Random Forest: The RF algorithm is a collection of individual decision trees [4]. RF has a wide range of applications because of its better stability and generalization [29]. This algorithm works by creating a group of decision-making structures in training and generating classes, which are the class mode or average predictions of each decision tree [2].

Since RF is an ensemble of individual decision trees [4], it builds a large collection of de-correlated trees which noisy but unbiased and averages them to reduce the variance. RF obtains a class vote from each tree, then classifies a sample using the majority vote [18]. Let $\hat{C}_b(x)$ be the class prediction of the b^{th} tree, then the class obtained from RF, $\hat{C}_{\mathrm{rf}}^{B}(x)$, is

$$\hat{C}_{\mathrm{rf}}^{B}(x) = \text{majority vote } \{\hat{C}_b(x)\}_1^B \tag{4}$$

3) Gradient Boosted Tree: Gradient-descent based formulation of boosting methods and the corresponding models are termed as gradient boosting machines (GBT) [30]. Boosting works by sequentially applying weak learners to repeatedly re-weighted versions of the training data [31]. GBT constructs base learners iteratively by re-weighting observations that were misclassified. Furthermore, GBT determines the weights by operating on the negative partial derivatives of the loss function at each training observation.

4) XGBoost Tree: XGBoost algorithm adopts a residual learning method to improve the model like the traditional boosting tree. The distinction is that the weak classifier's split node selection process does not always follow the least squares loss principle. Similarly, to predict a sample's score based on its feature, it will land on the relevant leaf node on each tree, and each leaf node corresponds to a score. [32]. The expression shown in (5), $w_{q(x)}$ is the score of leaf node q, and f(x) is the expression of one of the regression trees.

$$\hat{y} = \phi(x_i) = \sum_{\mathrm{k}=1}^{\mathrm{M}} \mathrm{f_k}\,(x_i) \tag{5}$$

$$\text{Where } \mathrm{F} = \{\mathrm{f(x)} = \mathrm{w_{q(x)}}\}(\mathrm{q}: \mathrm{R^m} \rightarrow \mathrm{T}, \mathrm{w} \epsilon R^{T}_{//\mathrm{n}})$$

D. Model Evaluation

1) Confusion matrix: Confusion matrix is a popular method for assessing classification quality. It contains information about instances in the current class and the predicted class. Specifically, each row of the confusion matrix represents an instance in the current class, while each column represents an instance in the predicted class [12]. We used a common performance measure from the confusion matrix, which is the accuracy and the error rate (6, 7).

$$\text{accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)} \tag{6}$$

$$\text{error rate} = \frac{FP+FN}{(TP+TN+FP+FN)} \tag{7}$$

2) Receiver Operating Characteristic (ROC) Curve: The Receiver Operating Characteristic (ROC) curve is the plot for a binary classifier's performance [18]. Technically, the ROC curve, also known as the ROC graph, is a two-dimensional graph where the rate of TP is plotted on the Y axis and the rate of FP is plotted on the X axis. In this way, the ROC graph depicts the relative trade-off between true positives and false positives [33]. The accuracy of ROC classification is worked out by calculating the area under the ROC curve. The area under the ROC curve is an area that shows the level of accuracy of the empirical model and is calculated by a calculation method called Area Under the ROC Curve (AUC). The AUC is a rectangular area whose value is always between 0 and 1. A higher AUC value implies better classification performance, making it as the maximization goal.

E. SHAP (SHapley Additive exPlanations)

We also present a unified framework which is SHAP (SHapley Additive exPlanations) for interpreting predictions. It is aimed to specify the impact of a feature on the classifier's output, [34] of a feature i, which is denoted as φ_i, for each classified flight, as follows [34]:

$$\varphi_i = \sum_{S \subseteq F\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

where F is the set of all features considered for the classification algorithm, $S \subseteq F$ is a subset of features obtained from the set F except feature i, and $f(S)$ is the expected classification output given by the set S of features. [34] note that, when a SHAP value of a given feature in log odds is close to zero, it does not contribute to deciding in classify a flight status as delay or no delay.

TABLE III. THE MODELING RESULT ON DATA TESTING

Classifier + Sampling Technique	**90:10**			**80:20**			**70:30**			**60:40**		
	Accuracy	***Error Rate***	***AUC Value***	***Accuracy***	***Error Rate***	***AUC Value***	***Accuracy***	***Error Rate***	***AUC Value***	***Accuracy***	***Error Rate***	***AUC Value***
DT	77.85%	22.15%	69.9%	77.56%	22.44%	66.5%	78.19%	21.81%	67.0%	78.52%	21.48%	66.6%
DT with SMOTE	69.34%	30.66%	67.4%	70.18%	29.82%	67.9%	70.07%	29.93%	67.0%	66.98%	33.02%	64.7%
DT with ROS	69.75%	30.25%	71.1%	69.80%	30.20%	68.5%	68.03%	31.97%	68.9%	65.36%	34.64%	65.9%
DT with RUS	64.06%	35.94%	70.4%	61.56%	38.44%	66.3%	62.99%	37.01%	67.5%	61.21%	38.79%	68.3%
DT with ROS + RUS	64.27%	35.73%	67.1%	65.96%	34.04%	66.5%	64.53%	35.47%	67.0%	62.13%	37.87%	64.0%
RF	81.28%	18.72%	81.1%	80.65%	19.35%	77.5%	80.57%	19.43%	77.7%	80.78%	19.22%	76.4%
RF with SMOTE	81.89%	18.11%	78.9%	78.48%	21.52%	74.2%	78.60%	21.40%	74.2%	79.24%	20.76%	73.2%
RF with ROS	82.30%	17.70%	81.1%	80.54%	19.46%	77.3%	81.05%	18.95%	77.4%	80.78%	19.22%	76.5%
RF with RUS	77.43%	22.57%	78.3%	75.81%	24.19%	75.0%	75.48%	24.52%	75.4%	74.46%	25.54%	74.0%
RF with ROS + RUS	82.58%	17.42%	81.1%	79.89%	20.11%	76.3%	80.07%	19.93%	76.4%	80.23%	19.77%	75.4%
GBT	80.80%	19.20%	80.3%	80.82%	19.18%	77.5%	81.41%	18.59%	77.9%	81.22%	18.78%	77.2%
GBT with SMOTE	77.16%	22.84%	77.3%	75.33%	24.67%	74.5%	76.05%	23.95%	74.6%	76.01%	23.99%	74.5%
GBT with ROS	77.30%	22.70%	81.0%	75.26%	24.74%	76.9%	76.44%	23.56%	77.6%	76.37%	23.63%	76.6%
GBT with RUS	70.64%	29.36%	78.6%	68.74%	31.26%	75.0%	69.21%	30.79%	76.2%	67.77%	32.23%	74.5%
GBT with ROS + RUS	71.74%	28.26%	78.6%	70.97%	29.03%	75.4%	72.51%	27.49%	75.9%	71.45%	28.55%	74.6%
XGBoost	80.38%	19.62%	80.1%	80.54%	19.46%	77.1%	80.21%	19.79%	77.2%	80.57%	19.43%	76.2%
XGBoost with SMOTE	77.50%	22.50%	77.0%	77.25%	22.75%	74.3%	77.32%	22.68%	75.4%	76.97%	23.03%	73.8%
XGBoost with ROS	78.60%	21.40%	79.9%	77.76%	22.24%	77.0%	77.55%	22.45%	77.2%	77.38%	22.62%	75.9%
XGBoost with RUS	70.71%	29.29%	77.7%	69.56%	30.44%	75.6%	67.64%	32.36%	75.3%	67.22%	32.78%	74.9%
XGBoost with ROS + RUS	75.93%	24.07%	80.7%	73.68%	26.32%	75.5%	74.80%	25.20%	76.1%	74.77%	25.23%	75.8%

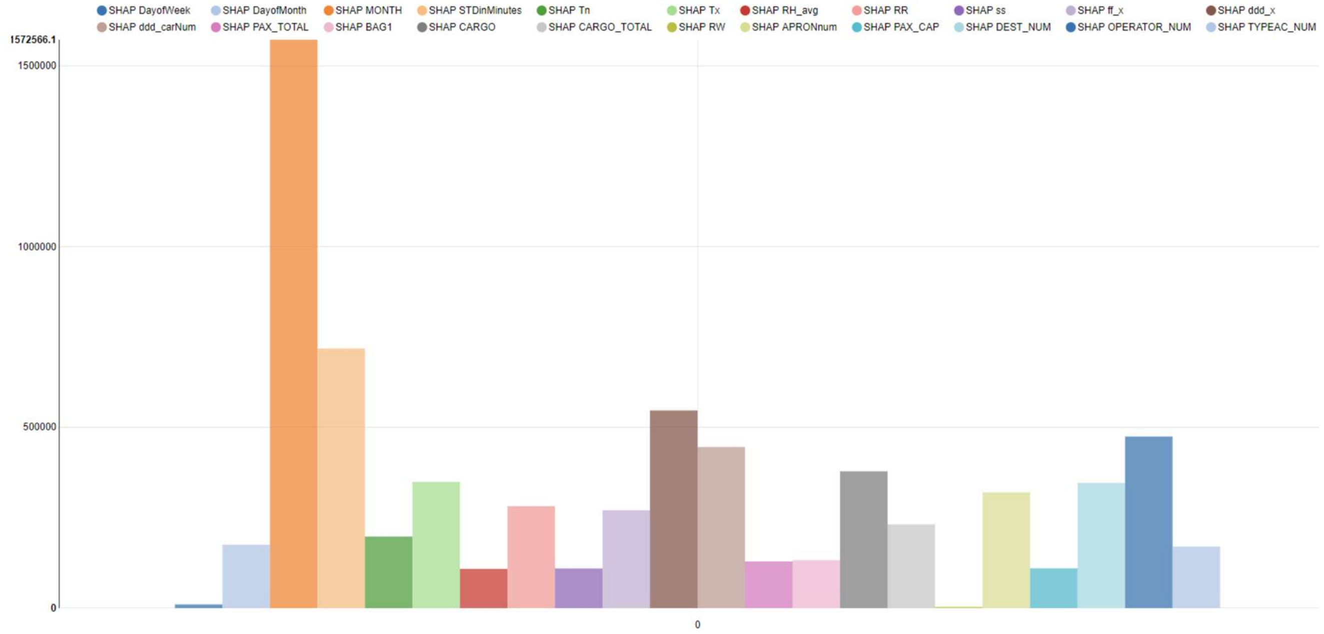

Fig. 2. SHAP value of feature given

The SHAP values display which features significantly affect how to define the aircraft delay status and how much of an impact they have. It refers to the degree to which a particular feature value supports the designation of a flight as delayed. A feature's large positive (or large negative) SHAP value for a particular flight scenario denotes that the feature significantly influences whether the flight is classed as delayed or not [35].

III. Experiment Result and Analysis

We interpret the result yielded by all classifiers with different sampling techniques and data split ratio scenarios. Table III shows model performance concerning the accuracy, error rate, and AUC value for each classifier with several data sampling techniques and data split ratios.

The RF classifier with the ROS+RUS technique and a data split ratio of 90:10 gave the highest accuracy and error rate, as shown in Table III as 82.58% and 17.42%, respectively on data testing. The AUC value earned from the ROC curve was shaped by our prediction model. The highest AUC value obtained was 81.1%, which was also consistently provided by the RF classifier using the ROS+RUS technique and a data split ratio of 90:10.

To determine the impact of a feature on the output of the RF since it shows the highest accuracy, error rate, and AUC value, we determine the Shapley additive explanations (SHAP). We only choose the top five features with the highest SHAP value, as shown in Fig. 2, that contribute the most to classifying the flight status as delayed or no delay.

TABLE IV. Top five highest SHAP value

Rank	*Variable*	*SHAP Value*
1	MONTH	1,572,566
2	STDinMinutes	717,663
3	ddd_x	546,422
4	OPERATOR_NUM	474,329
5	ddd_carNum	445,400

Table IV shows the feature's SHAP value, respectively, from the highest to the lower occupied by "MONTH" associated with time-stamped, "STDinMinutes" as scheduled departure time, "ddd_x" as a weather characteristic, "OPERATOR_NUM" as the airline operator, and "ddd_carNum" as the weather characteristic. We can interpret these five features as the largest contributors that drive the classification of flight status. We found two of the top five features that have the largest influence on the flight category status belong to weather characteristics, which are rather difficult to interpret or control, we just try to find the insight of the other three.

To validate the feature's SHAP value that drives the classification of flight status, we can see Fig. 3 shows the statistics related to "MONTH" and flight frequency in the research period of 2020-2021. It shows the month of December, which is symbolized as number 12, has the highest flight frequency in the research period. It also contained the most delayed status, which impacted the next month.

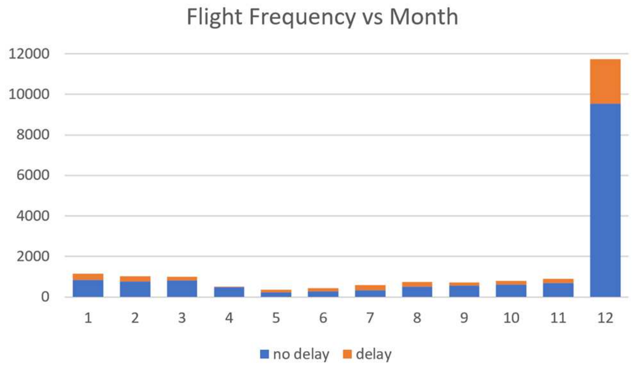

Fig. 3. Flight Frequency vs. Month of Flight Chart

Fig. 4 shows the distribution of flights in the time dimension of Pattimura Airport's operation hours. We can see that the flight schedule has an uneven distribution. There is a significant difference between hours that could potentially cause flight delays in the next few hours.

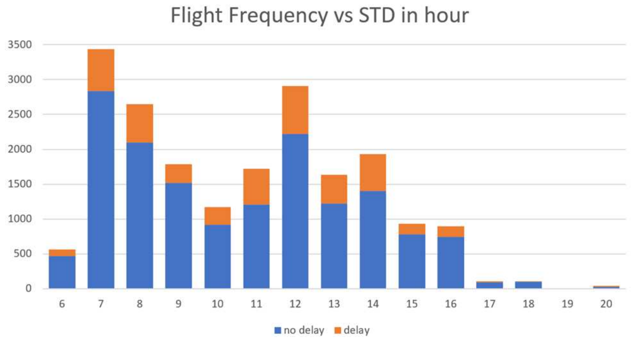

Fig. 4. Flight Frequency vs. STD in an hour chart

Since the airline operator feature is also the most important in determining flight classification status, we can see in Fig. 5 that the WON operator has the highest flight frequencies and the most flight delays during the 2020-2021 research period. We can gain insight about WON operator being the most contributor to flight delay that can impact the next flight.

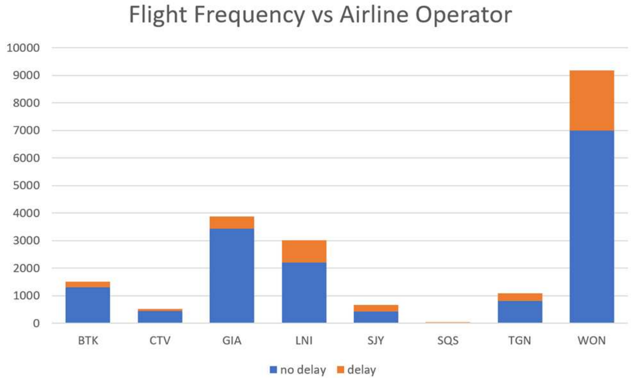

Fig. 5. Flight frequency vs. airline operator chart

IV. Conclusion and Future Works

We have developed this paper through a machine learning approach by using several classifiers in terms of predicting flight delays in the future. We also applied several sampling techniques for every classifier with split data ratio scenarios. Based on the prediction model built, we found the RF classifier with ROS+RUS technique and data split ratio of 90:10 has the best performance by giving the highest accuracy, error rate, and AUC value of 82.58%, 17.42%, and 81.1%, respectively, in predicting flight delay.

Since we found two of the top five features have the largest influence on the flight category status belonging to weather characteristics shown by SHAP value and are rather difficult to interpret and control, we can use the remaining three features as a reference to the airport strategy in order to mitigate the airport commercial revenue losses caused by flight delays. By practice, our methodology can provide insight into potential delays associated with strategic planning. The airport operator can manage the flow of passengers after the check-in process by directing passengers through the right commercial area and placing passengers at the right boarding gate, which has commercial space with the right passenger target. The airport operator can also do some acts, such as evaluating the existing flight schedules or changing or offering the flight schedule by spreading slot time with no bottleneck or critical schedule time to avoid flight delays.

In future work, enhance the dataset and features for the classification algorithm to improve the prediction accuracy. In addition, appending and configuring more sampling techniques would be good to be done.

REFERENCES

[1] K. K. H. Ng, C. K. M. Lee and F. T. S. Chan, "An Alternative Path Modelling Method for Air Traffic Flow Problem in Near Terminal Control Area," *2019 2nd International Conference on Intelligent Autonomous Systems (ICoIAS),* pp. 171-174, 2019.

[2] J. Huo, K. L. Keung, C. K. M. Lee, K. K. H. Ng and K. C. Li, "The Prediction of Flight Delay: Big Data-driven Machine Learning Approach," in *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2020.

[3] W.-B. Du, M.-Y. Zhang, Y. Zhang, X.-B. Cao and J. Zhang, "Delay causality network in air transport systems," *Transportation Research Part E: Logistics and Transportation Review,* vol. 118, pp. 466-476, 2018.

[4] L. Breiman, "Random Forests," *Machine Learning,* vol. 1, no. 45, pp. 5-32, 2001.

[5] S. Cheng, Y. Zhang, S. Hao, R. Liu, X. Luo and Q. Luo, "Study of flight departure delay and causal factor using spatial analysis," *Journal of Advanced Transportation,* vol. 2019, 2019.

[6] J. P. Pita, C. Barnhart and A. . P. Antunes, "Integrated Flight Scheduling and Fleet Assignment Under Airport Congestion," *Transportation Science,* vol. 47, no. 4, pp. 455-628, 2013.

[7] C. Diaz, A. B., M. Ruiz and F. J., "The consumer's reaction to delays in service," *International Journal of Service Industry Management,* vol. 13, no. 2, pp. 118-140, 2002.

[8] B. Edwards, The Modern Airport Terminal: New Approaches to Airport Architecture, 2005.

[9] M. Bandeira and A. Correia, "Qualitative analysis of the relationship between the profile of departing passengers and their perception of the airport terminal," *JATS,* vol. 3, no. 1, pp. 78-102, 2012.

[10] B. Thiagarajan, L. Srinivasan, A. V. Sharma, D. Sreekanthan and V. Vijayaraghavan, "A machine learning approach for prediction of on-time performance of flights," *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC),* pp. 1-6, 2017.

[11] Q. Li and R. Jing, "Generation and prediction of flight delays in air transport," *IET Intelligent Transport Systems,* vol. 15, no. 6, pp. 740-753, 2021.

[12] L. Belcastro, F. Marozzo, D. Talia and P. Trunfio, "Using scalable data mining for predicting flight delays," *ACM Transactions on Intelligent Systems and Technology,* vol. 8, no. 1, pp. 1-20, 2016.

[13] J. J. Rebollo and H. Balakrishnan, “Characterization and prediction of air traffic delays,” *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 231-241, 2014.

[14] L. Moreira, C. Dantas, L. Oliveira, J. Soares and E. Ogasawara, "On Evaluating Data Preprocessing Methods for Machine Learning Models for Flight Delays," *2018 International Joint Conference on Neural Networks (IJCNN),* pp. 1-8, 2018.

[15] S. Zhang, X. Li, M. Zong, X. Zhu and R. Wang, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 29, no. 5, pp. 1774-1785, 2018.

[16] J. Sun, Z. Wu, Z. Yin and Z. Yang, "SVM-CNN-based fusion algorithm for vehicle navigation considering a typical observations," *IEEE Signal Processing Letters,* vol. 26, no. 2, p. 212–216, 2018.

[17] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou and D. Zhao, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning," *IEEE Transactions on Vehicular Technology,* vol. 69, no. 1, pp. 140-150, 2020.

[18] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC),* pp. 1-6, 2016.

[19] P. Baumgarten, R. Malina and A. Lange, "The impact of hubbing concentration on flight delays within airline networks: An empirical analysis of the US domestic market," *Transportation Research Part E-logistics and Transportation Review,* vol. 66, pp. 103-114, 2014.

[20] A. I. Czerny, P. Forsyth and H.-M. Niemeier, Airport Slots: International Experiences and Options for Reform, 2008.

[21] J. Yi, H. Zhang, H. Liu, G. Zhong and G. Li, "Flight Delay Classification Prediction Based on Stacking Algorithm," *Journal of Advanced Transportation,* vol. 2021, 2021.

[22] N. Vandenbroucke, L. Macaire and J. G. Postaire, "Unsupervised color texture feature extraction and selection for soccer image segmentation," in *Proceedings 2000 International Conference on Image Processing*, 2000.

[23] V. A. Dev and M. R. Eden, "Formation lithology classification using scalable gradient boosted decision trees," pp. 392-404, 2019.

[24] W. A. Khan, H.-L. Ma, S.-H. Chung and X. Wen, "Hierarchical integrated machine learning model for predicting flight departure delays and duration in series," *Transportation Research Part C: Emerging Technologies,* vol. 129, 2021.

[25] R. Indralaksono, M. A. Wakhid, N. U. A, G. H. Wibowo, M. Abdillah, A. B. Rahardjo and D. Purwitasari, "Hierarchical Clustering and Deep Learning for Short-Term Load Forecasting with Influenced Factors," *J. RESTI (Rekayasa Sist. Teknol. Inf.),* vol. 6, no. 4, pp. 692-701, 2022.

[26] R. Sahoo, A. K. Pasayat, B. Bhowmick, K. Fernandes and M. K. Tiwari, "A hybrid ensemble learning-based prediction model to minimize delay in air cargo transport using bagging and stacking," *International Journal of Production Research,* vol. 60, no. 2, pp. 644-660, 2022.

[27] G. E. A. P. A. Batista, R. C. Prati and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explor. Newsl.,* vol. 6, no. 1, p. 20–29, 2004.

[28] S. D. Jadhav and H. Channe, "Efficient Recommendation System Using Decision Tree Classifier and Collaborative Filtering," 2016.

[29] M. Mursalin, Y. Zhang, Y. Chen and N. V. Chawla, "Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier," *Neurocomputing,* vol. 241, pp. 204-214, 2017.

[30] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front Neurorobot,* vol. 27, pp. 1-21, 2013.

[31] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2009.

[32] M. Lu, P. Wei, M. He and Y. Teng, "Flight Delay Prediction Using Gradient Boosting Machine Learning Classifiers," *Journal of Quantum Computing,* vol. 3, no. 1, pp. 1-12, 2021.

[33] F. Gorunescu, Data Mining: Concepts, Models and Techniques, 2011.

[34] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *arXiv*, 2017.

[35] M. Lambelho, M. Mitici, S. Pickup and A. Marsden, "Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions," *Journal of Air Transport Management,* vol. 82, no. C, 2020.