JUSIKOM PRIMA (Jurnal Sistem Informasi dan Ilmu Komputer Prima)
Vol. 6 No. 1, Agustus 2022 E-ISSN : 2580-2879

COMPARISON OF CLASSIFICATION ALGORITHM IN CLASSIFYING AIRLINE PASSENGER SATISFACTION

[1]Jacky Suwanto, [2]Daniel Ryan Hamonangan Sitompul, [3]Stiven Hamonangan Sinurat,
[4]Andreas Situmorang, [5]Ruben, [6]Dennis Jusuf Ziegel, [7]*Evta Indra
Prodi Sistem Informasi, Fakultas Teknologi dan Ilmu Komputer, Universitas Prima Indonesia
Jl. Sampul No.3, Sei Putih Barat, Medan Petisah
E-mail : *evtaindra@unprimdn.ac.id

ABSTRACT- In order to revive the airline industry, which is being hit by the current recession, it is essential to restore passenger confidence in airlines by improving the services provided by airlines. With the influence of technology in all industrial fields, airlines can now use Machine Learning to find the essential points that can make passengers feel satisfied with airline services and classify passenger satisfaction. This study presents the making of Machine Learning models starting from Data Acquisition, Data Cleaning, Exploratory Data Analysis, Preprocessing, and Model Building. It is concluded that Random Forest is the best algorithm used in this case study, with an F1 accuracy score of 89.4, ROC-AUC score of 0.90, and a shorter modeling period than other algorithms used in this study.

Kata kunci : Machine Learning, Random Forest, AdaBoost, XGBoost, Classification.

1. INTRODUCTION

The airline industry experienced a setback during the pandemic. Based on data from the International Civil Aviation Organization (ICAO), in 2020, the aviation industry suffered a loss of $372 billion in 2020 with a decrease in the number of passengers by -60% [1]. In order to revive the airline industry, which is being hit by a recession, it is essential to restore passenger confidence by improving the services provided by airlines. With the influence of technology in all industrial fields, airlines can now use Machine Learning to find the essential points that can make passengers feel satisfied with airline services. The airline can also classify the rating given by the passenger to find out whether the passenger is satisfied or not with the service that has been provided.

Many previous studies have carried out the classification of airline passenger comfort, for example, research entitled "Comparison of Feature Selection Optimization in Nave Bayes for Airline Passenger Satisfaction Classification" [2] and "Predicting Airline Passenger Satisfaction With Classification Algorithms" [3]. Research [2] performs classification using the K-Nearest Neighbors (KNN) algorithm, Logistic Regression, Gaussian Naïve Bayes, Decision Tree, and Random Forest. The results of each algorithm will be compared based on the highest accuracy value. Research [3] carried out the classification using the Naïve Bayes algorithm, which is configured by default, with Particle Swarm Optimization (PSO) and with Genetic Algorithm (GA).

Based on the problems mentioned above, to help airlines know the essential points to provide passengers with the best service and also be able to classify passenger satisfaction, the authors suggest making an analysis using the Exploratory Data Analysis method and making Machine Learning models that can carry out the classification process. The parameters used in making the model include the services provided by the airline to passengers. In this study, several classification algorithms will be presented, such as Random Forest (RF), Adaptive Boosting (AdaBoost), and Extreme Gradient Boosting (XGBoost). The highest accuracy value and the shortest creation time will be concluded to be the best algorithm that can be used in this case study.

2. METHODOLOGY

This research was conducted at the Data Analyst Laboratory of Prima Indonesia University. The notebook used in making the model uses the help of Google Collaboratory. The workflow of this research is presented in Figure 1. The flow of the research was carried out by retrieving data from the web page of the dataset provider, namely Kaggle (TJ Klein, 2020); after the data is obtained, data cleaning will be carried out, which includes checking whether there is data that is not balanced (Imbalance) and checking whether there is empty data (Null Value); then the Exploratory Data Analysis process is carried out where the search for essential points that can be visualized from the dataset will be carried out; after the understanding of the data is complete, the Pre-processing stage will be carried out where there is a Label Encoding and Outlier Removal process; Then the last one will be making a model with a predetermined configuration.

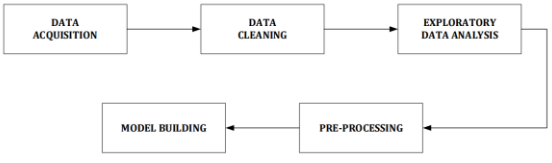


Figure 1. Research Methodology

id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service
70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	...	5	4	3	4	4	5
5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	...	1	1	5	3	1	4
110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	...	5	4	3	4	4	4
24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	...	2	2	5	3	1	4
119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	...	3	3	4	4	3	3

imns

Figure 2. Dataset Detail

2.1 Data Acquisition

The data acquisition stage aims to download data from trusted sources before the data is stored, processed, pre-processed, and used for other purposes. This process begins with retrieving relevant information, changing the data as needed, and calling the dataset into the notebook [4], [5]. In this study, the dataset used came from Kaggle, namely "Airline Passenger Satisfaction" (TJ Klein, 2020). This dataset contains passenger survey data from an airline. The survey covers passenger numbers to passenger satisfaction. This dataset contains 25 columns and 103904 rows of data. Details of the dataset can be seen in Figure 2.

2.2 Data Cleaning

The data cleaning stage prepares data for analysis by removing irrelevant or inappropriate data. The data in question has a negative impact on the model or algorithm to be made. Data cleaning is not only to dispose of data but can also be interpreted as a step to improve data [7], [8]. In this study, data cleaning is done by checking the null value. Details of this stage can be seen in Table 1.

Table 1. Data Cleaning Process

Column	**After Data Cleaning**
Arrival_Delay_in_Minutes	Filled with mean value
Gender	Filled with categorical data (eg. 0/1)
Customer_Type	
Type_of_Travel	
Class	

2.3 Exploratory Data Analysis (EDA)

The EDA stage is an essential process for conducting an initial investigation of the data used to find patterns and anomalies and test hypotheses with the help of statistical visualization [9], [10]. In this study, EDA was conducted to visualize data such as the number of satisfied or dissatisfied passengers based on the distance traveled and the type of trip. An example of the EDA stages can be seen in Figure 3.

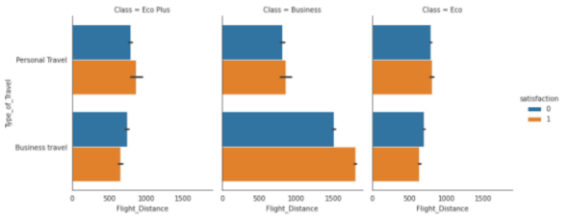


Figure 3. Exploratory Data Analysis

2.4 Preprocessing

The pre-processing stage is transforming or encoding data so that the data can be parsed easily by machine learning. The main task of this stage is to create an accurate and predictable model [11], [12]. In this study, pre-processing was carried out to encode the data and remove outliers contained in the data. Details of the data encoding process and the outlier removal process can be seen in Figure 4.

id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance
70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460
5047	Male	disloyal Customer	25	Business travel	Business	235
110028	Female	Loyal Customer	26	Business travel	Business	1142
24026	Female	Loyal Customer	25	Business travel	Business	562
119299	Male	Loyal Customer	61	Business travel	Business	214

Gender	Customer_Type	Age	Type_of_Travel	Class
1	0	13	1	2
1	1	25	0	0
0	0	26	0	0
0	0	25	0	0
1	0	61	0	0

```
# Mengecek jumlah baris data dan kolom dataset
print("Dataset memiliki", train.shape[0], "data")
print("Dataset memiliki", train.shape[1], "kolom")
```

```
Dataset memiliki 103904 data
Dataset memiliki 25 kolom
```

```
# Mengecek jumlah baris data dan kolom dataset
print("Dataset memiliki", train.shape[0], "data")
print("Dataset memiliki", train.shape[1], "kolom")
```

```
Dataset memiliki 61197 data
Dataset memiliki 23 kolom
```

Figure 4. Preprocessing

2.5 Model Building

2.5.1 Random Forest (RF)

The random forest algorithm is an ensemble learning method used to carry out the process of classification, regression, and other things that are made using many decision trees when doing model training. For the classification process, the result of a random forest is the class chosen by most trees. [13], [14]. In this study, the random forest algorithm was created with the configuration max_depth=16,min_samples_leaf=1,min_samples_split=2, n_estimators=100 and random_state=12345. The process of making the random forest algorithm can be seen in Figure 5.

```
from sklearn.ensemble import RandomForestClassifier

params_rf = {'max_depth': 16,
             'min_samples_leaf': 1,
             'min_samples_split': 2,
             'n_estimators': 100,
             'random_state': 12345}
```

Figure 5. Random Forest Detail

2.5.2 Adaptive Boosting (AdaBoost)

The adaptive boosting algorithm is an ensemble learning method that creates a model by assigning a balanced weight value to each data point, then giving more weight to the incorrectly classified points [15], [16]. In this study, the AdaBoost algorithm was created with the configuration of n_estimators 500 and random_state 12345. The process of making the random forest algorithm can be seen in Figure 6.

```
from sklearn.ensemble import AdaBoostClassifier as adab
params_adab ={'n_estimators': 500,
              'random_state': 12345}
```

Figure 6. AdaBoost Detail

2.5.3 Extreme Gradient Boosting (XGBoost)

The xgboost algorithm is a scalable and easily distributed Gradient-Boosted Decision Tree (GDBR) based machine learning library. This algorithm presents a parallel tree and is a machine learning library that is most often used for the process of regression and classification [17], [18]. In this study, the XGBoost algorithm was created with the configuration of n_estimators 500 and max_depth 16. The process of making the random forest algorithm can be seen in Figure 7.

```
import xgboost as xgb
params_xgb ={'n_estimators': 500,
             'max_depth': 16}
```

Figure 7. XGBoost Detail

3. RESULT AND DISCUSSION

3.1 Exploratory Data Analysis

This study's results of the Exploratory Data Analysis stage are divided into three points. The first point is the comparison of the number of satisfied or dissatisfied passengers by gender; the second point is the comparison of the number of satisfied or dissatisfied passengers based on baggage handling services and terminal location (gate location), and the third point is the comparison of the number of satisfied or dissatisfied passengers based on aircraft media entertainment services (Inflight Entertainment) and in-flight wi-fi services (Inflight wi-fi).

In the first point, it can be seen in Figure 8 that the comparison of the number of satisfied or dissatisfied passengers is evenly distributed by gender, and it can be concluded from this comparison that there are more dissatisfied passengers than satisfied passengers.

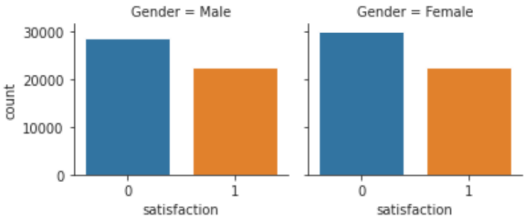


Figure 8. Comparison of Gender

On the second point, it can be seen in Figure 9 that for business class, more dissatisfied passengers are carried out if Baggage Handling is carried out imperfectly (rating <=4). For eco plus and eco classes, when the location of the terminal is not good/far (<=2), the passenger is not satisfied, even if Baggage Handling is usually carried out (range 2-4).

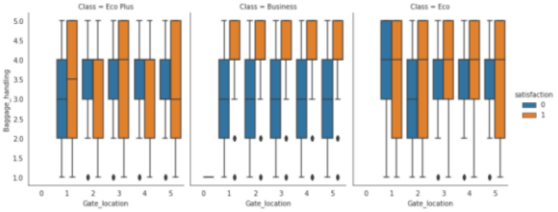


Figure 9. Comparison Based on Baggage Handling

On the third point, it can be seen in Figure 10 that Eco plus passengers are more satisfied with flights without wi-fi service (rating 0) and the usual media entertainment (rating 2-4). For business class

passengers, only the best entertainment media (rating 5) can satisfy them. For Eco passengers, good media entertainment (rating 3-5) and the best wi-fi service (rating 5) can make them satisfied.

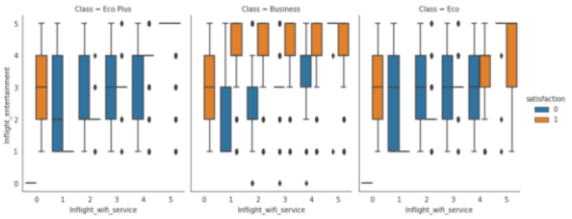


Figure 10. Comparison Based on Inflight Media and Wi-Fi

3.2 Important Features

In this study, the essential features of the dataset were obtained using Permutation Importance. From the method used, it is concluded that there are five most important features: the type of travel (Type of Travel), aircraft wi-fi service (Inflight WiFi Service), online boarding service, and seat comfort. Details of the essential features can be seen in Figure 11.

Weight	Feature
0.2723 ± 0.0039	Type_of_Travel
0.1278 ± 0.0026	Inflight_wifi_service
0.0435 ± 0.0011	Online_boarding
0.0424 ± 0.0013	Seat_comfort
0.0355 ± 0.0009	Checkin_service
0.0294 ± 0.0014	Inflight_service
0.0289 ± 0.0008	Baggage_handling
0.0246 ± 0.0006	Cleanliness
0.0177 ± 0.0007	On-board_service
0.0172 ± 0.0007	Class
0.0097 ± 0.0006	Leg_room_service
0.0095 ± 0.0004	Flight_Distance
0.0087 ± 0.0004	Inflight_entertainment
0.0082 ± 0.0003	Age
0.0062 ± 0.0007	Arrival_Delay_in_Minutes
0.0058 ± 0.0003	Ease_of_Online_booking
0.0044 ± 0.0004	Gate_location
0.0033 ± 0.0002	Departure/Arrival_time_convenient
0.0027 ± 0.0002	Departure_Delay_in_Minutes
0.0021 ± 0.0003	Food_and_drink
… 2 more …	

Figure 11. Important Features

3.3 Model Result and Comparison

3.3.1 Model Result

This study will use three algorithm models to carry out the classification process: Random Forest, AdaBoost, and XGBoost. The results of model fitting will be divided into three points.

The first point is the Random Forest algorithm, with an accuracy value of 0.894 (89.4%) and the ROC-AUC value of 0.9003 (90%). The precision of a true negative is 12375 data, a false positive is 2197 data, a false negative is 553 data, and a true positive is 10850 data. The random forest can also get 90% precision in this case. Details of the results can be seen in Figure 12.

```
Accuracy = 0.8941330458885125
ROC Area under Curve = 0.9003728693084586
Time taken = 5.19550895690918
              precision    recall  f1-score   support

           0    0.95723   0.84924   0.90001     14573
           1    0.83161   0.95150   0.88753     11403

    accuracy                        0.89413     25976
   macro avg    0.89442   0.90037   0.89377     25976
weighted avg    0.90208   0.89413   0.89453     25976
```

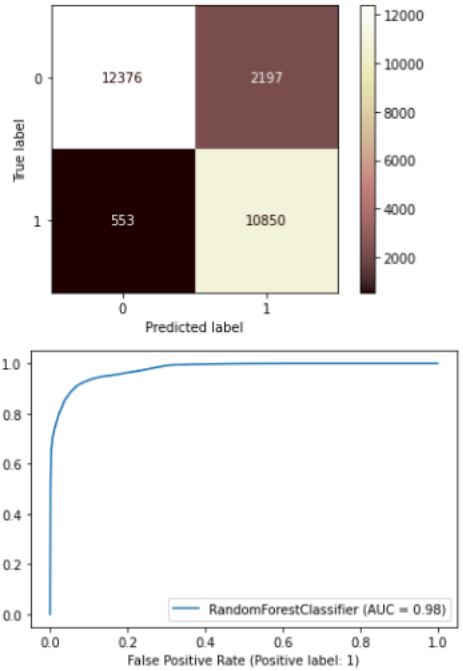


Figure 12. Resultt of Random Forest

The second point is the AdaBoost algorithm, with an accuracy value of 0.895 (89.5%) and ROC-AUC value of 0.899 (89.9%). The precision of a true negative is 12620 data, a false positive is 1953 data, a false negative is 761 data, and a true positive is 10642 data. AdaBoost also scores 90% precision in this case. Details of the results can be seen in Figure 13.

```
Accuracy = 0.8955189405605174
ROC Area under Curve = 0.8996241085930146
Time taken = 24.31750512123108
              precision    recall  f1-score   support

           0    0.94313   0.86599   0.90291     14573
           1    0.84494   0.93326   0.88691     11403

    accuracy                        0.89552     25976
   macro avg    0.89403   0.89962   0.89491     25976
weighted avg    0.90002   0.89552   0.89589     25976
```

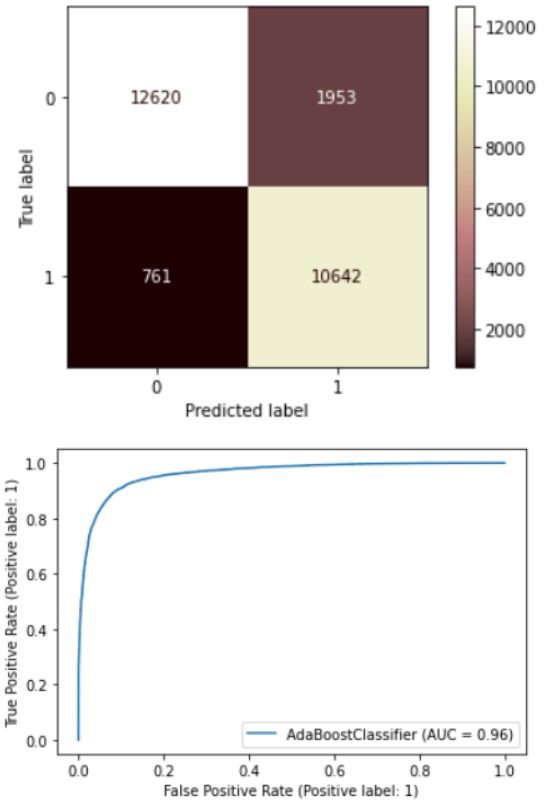


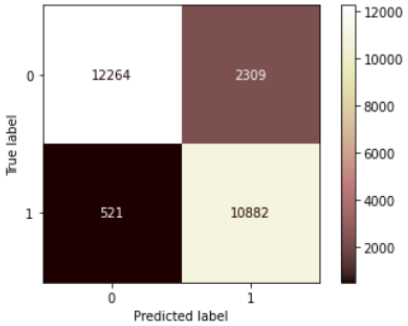
Figure 13. Result of AdaBoost

The third point is the XGBoost algorithm, with an accuracy value of 0.891 (89.1%) and ROC-AUC value of 0.897 (89.7%). The precision of the actual negative is 12264 data, the false positive is 2309 data, the false negative is 521 data, and the true positive is 10882 data. XGBoost also scores 89% precision in this case. Details of the results can be seen in Figure 14.

```
Accuracy = 0.8910532799507237
ROC Area under Curve = 0.8979332859895301
Time taken = 81.9677631855011
              precision    recall  f1-score   support

           0    0.95925   0.84156   0.89656     14573
           1    0.82496   0.95431   0.88493     11403

    accuracy                        0.89105     25976
   macro avg    0.89210   0.89793   0.89074     25976
weighted avg    0.90030   0.89105   0.89145     25976
```



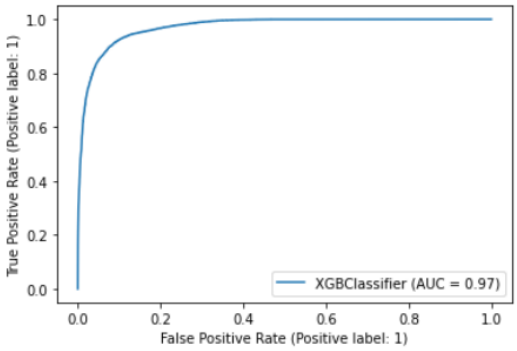


Figure 14. Result of XGBoost

3.3.2 Model Comparison

After the modeling is complete and the model fitting results have been obtained, the model will be compared based on the ROC-AUC value and the duration of manufacture. In this study, it can be concluded that the Random Forest and AdaBoost algorithms have the same ROC-AUC value of 90%, while the XGBoost algorithm has a value of 89.7%. The algorithm with the fastest creation time is Random Forest, and the longest is XGBoost. The comparison visualization of the model can be seen in Figure 15. The bar chart represents the manufacturing time, and the line represents the ROC-AUC value.

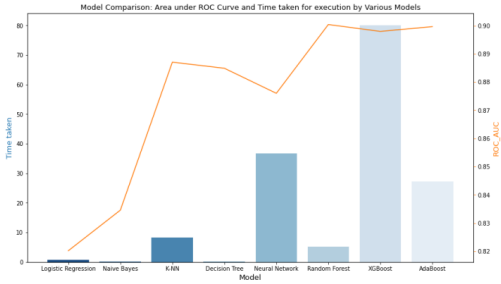


Figure 15. Algorithm Comparison

4. CONCLUSION

With the influence of technology in all industrial fields, airlines can now use Machine Learning to find the essential points that can make passengers feel satisfied with airline services. The airline can also classify the rating given by the passenger to find out whether the passenger is satisfied or not with the service that has been provided. It can be concluded that the best model in this case study is Random Forest with a ROC-AUC value of 90% and the model generation time is the fastest compared to other algorithms that have been made in this study.

BIBLIOGRAPHY

[1] [1] ICAO and ICAO, “Effects of Novel Coronavirus (COVID-19) on Civil Aviation: Economic Impact Analysis Economic Development-Air Transport Bureau,” 2022.

[2] K. Hulliyah, "Predicting Airline Passenger Satisfaction with Classification Algorithms," IJIIS: International Journal of Informatics and Information Systems, vol. 4, no. 1, pp. 82–94, Mar. 2021, doi: 10.47738/ijiis.v4i1.80.

[3] Yoga Religia and A. Amali, "Perbandingan Optimasi Feature Selection pada Naïve Bayes untuk Klasifikasi Kepuasan Airline Passenger," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 5, no. 3, pp. 527–533, Jun. 2021, doi: 10.29207/resti.v5i3.3086.

[4] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, and Md. S. H. Sunny, "Application of Big Data and Machine Learning in Smart Grid, and Associated Security Concerns: A Review," IEEE Access, vol. 7, pp. 13960– 13988, 2019, doi: 10.1109/ACCESS.2019.2894819.

[5] T. Wei, K. Liu, J. Shi, X. Liu, and L. Liu, "Research on Data Acquisition of Communication Equipment Based On Machine Learning," in 2019 International Conference on Computer Network, Electronic and Automation (ICCNEA), Sep. 2019, pp. 106–112. doi: 10.1109/ICCNEA.2019.00030.

[6] TJ KLEIN, "Airline Passenger Satisfaction," 2020. https://www.kaggle.com/datasets/teejmahal2 0/airline-passenger-satisfaction (accessed May 21, 2022).

[7] K. H. Tae, Y. Roh, Y. H. Oh, H. Kim, and S. E. Whang, "Data Cleaning for Accurate, Fair, and Robust Models," in Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning - DEEM'19, 2019, pp. 1–4. doi: 10.1145/3329486.3329493.

[8] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," Electronic Commerce Research and Applications, vol. 31, pp. 24–39, Sep. 2018, doi: 10.1016/j.elerap.2018.08.002.

[9] N. Verbeeck, R. M. Caprioli, and R. van de Plas, "Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry," Mass Spectrometry Reviews, vol. 39, no. 3, pp. 245–291, May 2020, doi: 10.1002/mas.21602.

[10] M. Abukmeil, S. Ferrari, A. Genovese, V. Piuri, and F. Scotti, "A Survey of Unsupervised Generative Models for Exploratory Data Analysis and Representation Learning," ACM Computing Surveys, vol. 54, no. 5, pp. 1–40, Jun. 2022, doi: 10.1145/3450963.

[11] P. Sharma, P. Hans, and S. C. Gupta, "Classification Of Plant Leaf Diseases Using Machine Learning And Image Preprocessing Techniques," in 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Jan. 2020, pp. 480–484. doi: 10.1109/Confluence47617.2020.905788.

[12] V. Gómez-Escalonilla, P. Martínez-Santos, and M. Martín-Loeches, "Preprocessing approaches in machine-learning-based groundwater potential mapping: an application to the Koulikoro and Bamako regions, Mali," Hydrology and Earth System Sciences, vol. 26, no. 2, pp. 221–243, Jan. 2022, doi: 10.5194/hess-26-221-2022.

[13] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 6308–6325, 2020, doi: 10.1109/JSTARS.2020.3026724.

[14] J. S. Cobb and M. A. Seale, "Examining the effect of social distancing on the compound growth rate of COVID-19 at the county level (United States) using statistical analyses and a random forest machine learning model," Public Health, vol. 185, pp. 27–29, Aug. 2020, doi: 10.1016/j.puhe.2020.04.016.

[15] K. W. Walker and Z. Jiang, "Application of adaptive boosting (AdaBoost) in demanddriven acquisition (DDA) prediction: A machine-learning approach," The Journal of Academic Librarianship, vol. 45, no. 3, pp. 203–212, May 2019, doi: 10.1016/j.acalib.2019.02.013.

[16] S. Chen, B. Shen, X. Wang, and S.-J. Yoo, "A Strong Machine Learning Classifier and Decision Stumps Based Hybrid AdaBoost Classification Algorithm for Cognitive Radios," Sensors, vol. 19, no. 23, p. 5077, Nov. 2019, doi: 10.3390/s19235077.

[17] H. Jiang, Z. He, G. Ye, and H. Zhang, "Network Intrusion Detection Based on PSOXgboost Model," IEEE Access, vol. 8, pp. 58392–58401, 2020, doi: 10.1109/ACCESS.2020.2982418.

[18] C. Wang, C. Deng, and S. Wang, "ImbalanceXGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost," Pattern Recognition Letters, vol. 136, pp. 190–197, Aug. 2020, doi: 10.1016/j.patrec.2020.05.035.