A Study over Supervised and Unsupervised learning in predicting the impact of bird strike in Aviation Industry.

MSc Research Project
Msc Data Analytics

Srivathsav Venugopal
Student ID: 20130660

School of Computing
National College of Ireland

Supervisor: Bharathi Chakravarthi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Srivathsav Venugopal
Student ID:	20130660
Programme:	Msc Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Bharathi Chakravarthi
Submission Due Date:	31/01/2021
Project Title:	A Study over Supervised and Unsupervised learning in predicting the impact of bird strike in Aviation Industry.
Word Count:	8407
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Srivathsav Venugopal
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Study over Supervised and Unsupervised learning in predicting the impact of bird strike in Aviation Industry.

Srivathsav Venugopal
x20130660

Abstract

Various reasons urge people to plan journeys all across the world. Air travel is by far the most popular mode of transportation is because of its numerous advantages. Airlines collisions may occur for various reasons despite their popularity as a form of transportation. One hazard that has constantly concerned the aviation business is bird hits. Large open areas like airport runways and other locations are occupied by birds, resulting in an accident or other emergency actions by the airlines. Predicting damage to aircraft caused by bird hits by using machine learning models and comparing the results of the two subsets of Machine learning, such as Supervised and Unsupervised learning, will be the aim of the research in the end.

1 Introduction

There are several terms for bird strikes in the aviation industry, including "bird strike," "bird impact," and "bird ingestion." The phrase refers to a bird and any other animals that may contact the plane as explained in Bradbeer et al. (2017). In the aviation business, bird strikes are a severe issue since they may lead to structural damage to aircraft or even accidents or aborted take-offs. Nowadays, it is common practice in the aviation business to deploy a variety of ideas to keep planes above the clouds safe. The damage to the plane's body caused by a bird strike is seen in the image below. Injury or an accident necessitates a repair bill, which is always incurred. Only civil aircraft repairs are anticipated to cost $1 billion per year. The below picture shows how far an bird strike can create an impact to the aircraft. Figure 1 Depending on the altitude, the effect of a bird attack might vary significantly. Technology evolves as the globe grows more technologically savvy, and newer technologies are applied in numerous industries. When it comes to developing and upgrading their sector, the aviation industry has never been lower than it is now. There has been a rise in the number of bird strikes on planes, but their effect is still enormous since they may cause catastrophic damage to the aircraft or even result in aborted take-offs and landings that may disrupt many passengers' travel plans. Predicting bird strikes has been applied for various reasons, including determining the bird strike's height or comparing the impacts recorded when taking off with the hits reported after landing. With this forecast, the airline industry has a general concept to advise that the Air Traffic Control [ATC] be more acquainted with bird strike events to alert the aircraft based on this. This has resulted in 219 deaths, and it is important to

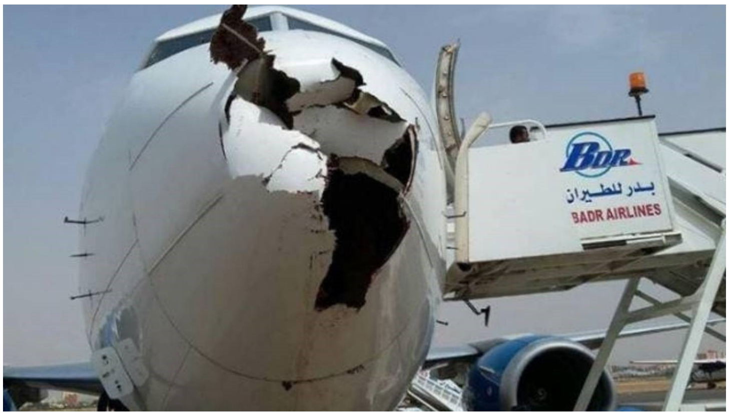


Figure 1: Impact Of Bird Strike

know what kind of birds are engaged in an attack and how much damage they do to a plane's body. A bird hit and the resulting influence on the plane's trajectory as evaluated in Robinson et al. (2021) .
In various fields, machine learning is being used as advanced technology. As it is a subset of artificial intelligence, it can be more accurate in predicting the outcomes of experiments. Prediction is always used in various areas to seek a specific problem, which is no exception. Two categories of machine learning methodologies namely supervised learning and unsupervised learning, have been selected for this study from among the many available. The data will be fed to them, and their performance will be assessed in this research based on their accuracy in predicting the actual result of the experiment. Because they both work on different approaches, they will be used in this research to examine how far they can go and which sorts of techniques seem to be the most effective based on the results. In the case of a bird strike, does it significantly influence the body of flights when utilizing various approaches of machine learning and developing multiple machine learning models in forecasting the future conclusion of the flight by carrying out an analysis? As it is also customary for birds, they are on the lookout for a large open region to nest in, and by default, they prefer airport runways. Through more data analysis, it is possible to determine the next step in avoiding animal strikes. As a result, investing in the repair of body parts can be avoided as analyzed in Robinson et al. (2021), and the cost of investing in wildlife strikes can be avoided. Suppose a client is dissatisfied with the service given and, as a result, does not return to the same airline or promote the same to others which will affect the aviation business. In that case, the airline industry will be one of the most important businesses in the world.For the reasons stated above, it is required to anticipate the primary phase most adversely impacted by bird attacks, as well as the various phases adversely affected and the pace of environmental damage. To ascertain how far the hit has spread and whether or whether aircraft or passengers have been harmed, this investigation's main purpose is to identify A bird's reaction to an approaching plane differs by species. A number of algorithms were constructed utilising data from Kaggle to make the prediction. Here, we use supervised and unsupervised learning to exploit bird attack aircraft structural damage as implemented in AISSAOUI et al. (2019). Based on their results, we can recommend the best machine learning approach for the aviation industry.

2 Research Question

How effective the supervised and unsupervised machine learning approaches are in predicting the effect of bird strikes in the aviation industry?

3 Related Work

3.1 Introduction

This chapter from the report mainly discusses the related works by examining both advantages and disadvantages of the implemented technologies. In addition, from the previous research, the strength and limitations of the implementation of the approach are to be discussed.

3.2 Bird Strikes and their impacts

In this study Altringer et al. (2021), the author has analysed the repair cost due to bird strikes. Aviation has more bird strikes. Machine learning is used to estimate the cost of bird strike repairs at US airports. A strike impacts fuel costs, rescheduling, and crew accommodations. The data was analysed using random forest and ANNs. A better model was chosen based on this. The implementation's regression tree separates data into categories based on each observation. The categorization performance of random forest and artificial neural networks has been explored. The random forest is an overfitting ensemble method. The regression trees are built using a bootstrapped training set. Then came the ANN model. Four. The ANN utilised the same approach. To improve neural network performance, some hyperparameters were adjusted using grid search. LR, RF, and NN predicted the training sample cost better than the real aircraft repair cost. A bird strike model predicted sample and repair costs correctly. On average, random forests outperform neural networks. According to previous studies, machine learning algorithms outperform other prediction methods.
In this research (Mehta et al.; 2021), machine learning techniques are utilized to anticipate the severity of an aviation collision. As previously said, plane travel is a popular choice. Air collisions occur due to engine failure, Air Traffic control (ATC) misdirection, and other factors. Using Support vector machines, Random forests, Gradient Boosting, k nearest neighbours classifier, Logistic regression, and artificial neural networks, the author has considered providing a solution for air crashes. The data has been read programmatically, and any non-supportive characteristics have been removed. The data is then filtered. The data is then further studied by picking additional connected characteristics, and the aforesaid models are developed and tested. The model's accuracy is assessed. The ensemble model seems to be the most accurate at 91.66 percent. Then a 91.51 percent accurate artificial neural network. Many others have shown their efficacy by giving above 90% accuracy. The accuracy and recall scores for each algorithm suggest that the models are balanced and have made correct predictions in this investigation. The author also proposes predicting the severity of aviation crashes. The conclusion of machine learning was flawless and demonstrated that machine learning might be more accurate in aviation prediction.
According to the author of this research (Dİkbayir and Bülbül; 2018), animal attacks cause structural damage to aeroplanes. Predicting damage may help avert accidents.

Using data from bird strikes, supervised classification systems were used to predict the result. Structural damage estimate is a vehicle collision evaluation method. The remaining life of the components may be forecast, and future probable mishaps can be averted. To prevent engine failures or accidents in the aviation business, scientists broke into two groups: structure damage rate and structure status monitoring. The effect of bird strikes will be studied and concentrated in this research utilising machine learning approaches. The algorithms used in this investigation were SVM, Gaussian Naive Bayes, and Decision Tree. The data is appropriately handled and analysed, and a flawless data frame with supporting pieces is generated. Based on the performance, Gaussian Naive Bayes is the top method with 79.31 percent accuracy, followed by Decision Tree with 74.13 percent. Third, the least accurate SVM.

In this research Baranzini and Zanin (2015)e,the Risk-Intelligence-Patterns-Theory will be used to forecast risk variables in the aviation sector. Pilot misconceptions occur when aircraft crash or are mislead by ATC. The risk patterns are employed in commercial aviation and their performance is compared to a case study. It is a risk intelligence hypothesis utilised in numerous projects particularly in the aviation business. It's a feedback measure. In this work, the author created a risk pattern recognition method that uses descriptive and predictive analysis to verify the data. In descriptive analysis, a single event, such as bird strikes, may be examined and determined to be reliant on another component. These implementations are used to discover operational trends in severe bird attacks. Forecasting utilising the airline's name, when a bird attack is logged with the airline's status. The author studied patterns to see whether Data Science can spot them.

In this study Valletta et al. (2017), Machine learning algorithms will be used to anticipate animal behaviour to test their accuracy. Supervised and Unsupervised Machine Learning techniques are used in the Ml framework. Different models were constructed to better anticipate the result. Support Vector Machines (SVM), Decision trees, Gradient Boosting Trees (GBT), and Gaussian Mixture Models (GMM). This is how the model was made. Before that, the data was fed into machine learning algorithms and its accuracy was evaluated. The research finds that instead of using statistical analysis in prediction, machine learning algorithms may be utilised to overcome any statistical flaws. In the evaluation, all models worked well and predicted the outcome better. The author concludes that Ml algorithms are important in interpreting complicated information and will help anticipate animal behaviour.

In this study Robinson et al. (2021), the data from the incidents between animals and aircraft at Oliver Tambo Airport in South Africa were used to evaluate the strikes. Airports are one of the large locations where birds readily attract and settle, resulting in wildlife collisions with aeroplanes. The author proposes many methods to forecast strikes. Bird attacks are more prevalent in South Africa. The research found that the laughing dove was the most involved species at 95.77 percent, followed by the blacksmith lapwing at 95.52 percent, the hadeda ibis at 91.79 percent, and the Egyptian goose at 86.07 percent. The top ten birds engaged in strikes have been tallied from the research. Egyptian goose (2.1 kg), Spur-winged goose (3.5 – 5.1 kg) are the most deadly bird species that have been reported with strikes. The birds indicated above were level 1, which is more harmful and risky to the aeroplane since they would disrupt the flight. The author finds that birds in level 1 hazard stage are more risky for the aviation sector and may have a major effect as a consequence of strike occurrences in the aviation industry.

3.3 Machine learning approach

This research Viswanath and Suresh Babu (2009) will analyse and assess the DBSCAN algorithm's performance on huge datasets. Density-based clustering may locate arbitrary-shaped clusters with noisy outliers. The author claims that the DBSCAN algorithm can overcome temporal complexity and work more efficiently than DSCAN. The author uses hybrid clustering to get density-based arbitrary form clusters. Rough DBSCAN is known to generate the prototype naming leaders. The suggested solution is based on crude set theory and is being improved to store each leader's followers. The study's rough-dbscan is stated to be easier and take less time per execution than DBSCAN. With the DBSCAN and rough-DBSCAN, the rough running time is linear, but DBSCAN seems to be quadratic. Because the threshold needed to determine the leaders is low as the dataset size increases, the result based on both models' performance is comparable, the rough-DBSCAN and DBSCAN have an identical output when they are so near. The author finds that the rough-DBSCAN is more scalable to produce density-based clusters whenever the dataset is huge.

The Birch algorithm Lorbeer et al. (2018) and its performance will be explored in this paper. Clustering methods have two major flaws: scaling when the dataset is large and requires a parameterization strategy to overcome. The author created a BIRCH automated threshold estimation technique in this study. This method computes the BIRCH algorithm parameters from the data, resulting in a suitable clustering without the need for the final global clustering step. Performance is only achievable if the data meets specific criteria. Or the algorithm will alert you before showing the findings. This technique has two benefits. The BIRCH method will get more sophisticated and quicker if the number of clusters is not known. For really big datasets, the author suggests using the MBD-BIRCH model. To improve accuracy in deep trees, the author has developed a new algorithm called BIRCH that extends the accuracy of the original BIRCH algorithm. Based on the assessment, the author recommends utilising the tree-BIRCH model since it is quicker and better than the other models.

In this research Deng et al. (2015), the OPTICS clustering technique is better with huge data. The author proposes Tra – POPTICS, a novel big data method that modifies the clustering algorithm point data (POPTICS). Tra –POPTICS is a novel technique that uses dispersed data to increase scalability. To propose a speedier answer to the huge data clustering algorithm challenge. The research used GPGPU, where the GPGPGU – clustering technique paralysed the Tra- OPTICS threads to test the data processing performance. The experimental findings show that the Tra- OPTICS algorithm works better than T-OPTICS and is more scalable than T-OPTICS. The G-tra-OPTICS clustering performance is comparable to the T-POPTICS. Again, comparing tra-POPTICS with threads provides over 30 speedups of average clustering approaches of data. Based on the data, the author believes that the above-mentioned algorithms have better scalability and computational performance when utilised for Big data services. These techniques are designed to handle large datasets more efficiently and precisely.

In this study Wang (2012), the usage of the AdaBoost algorithm has been examined. The study focuses on feature selection and the SVM model. This article introduces the AdaBoost Classifier algorithm. The original adaptive boosting technique is being studied for facial emotion detection. Pattern classification uses the SVM model, which has proven excellent results over a wide range of performance. The drawback is that it takes time. The AdaBoost algorithm will establish or maintain the weight across the training set. It

produces a group of bad learners by collecting weights over training data and updating them after each weak learning cycle. The algorithm's principles were known in this study. AdaBoost is one of the real-time algorithms utilised across numerous applications, not alone in feature selection. The models appear to perform well in feature extraction, with AdaBoost being one of the top functioning algorithms.
This research Singh et al. (2016) will examine the Ridge Classifier, a classification-based system, using Twitter data. The research will assess how accurate the prediction can be when combined with the various algorithms employed in the study. The author has built two classifiers to read the study's efficacy. With non-linear approaches, the author finds that ridge regression works effectively. Non-linear classifiers may yield even better results. However, it is difficult to forecast the parameters used for collecting data, and over-fitting is a serious concern. This reduces over-fitting and leads to better outcomes. Ridge regression prevents over-fitting by keeping the weights low. The feature selection should be simple with just one parameter. Moreover, avoiding over-fitting would have made model selection more challenging. Each feature is independent, and the best combination of traits is chosen. The analysis would be to assess the performance. So, instead of selecting and choosing features, regularisation may be more effective and deliver higher performance.

In this study Lieber et al. (2013), to forecast manufacturing quality, the author employed supervised and unsupervised machine learning. This study presents a model that divides the analysis into phases. Feature selection, pre-processing, and EDA approaches are used to analyse the data. After one-hot encoding, pick models to forecast quality outcomes and translate each variable into the right format. Next, supervised and unsupervised learning are combined, and predictions are formed. The quality forecasting methods applied by data mining were successful. The result is more work. It was shown that the k-Means algorithm and decision tree were the best algorithms for predicting the most results, with the decision tree having a 90% process and accuracy. The KNN model predicts a small association between modes and quality and has a 97 percent accuracy rate. These supporting features were chosen for the analysis, and the predictions were accurate. The author finds that supervised and unsupervised learning performed well at each stage. The author concluded that KNN and decision tree were the best performing algorithms. Machine learning is a popular technique used to predict the outcome of data processing. That is, machine learning improves both supervised and unsupervised prediction accuracy.

4 Data Cleaning and EDA

4.1 Methodology

Bird strikes are likely to become more common in aviation. Some strikes may impact the sector, causing cancellations of flights or vacation arrangements. The aforesaid issue is projected to be a major concern in the industry. Machine learning methods will be utilised to foresee and overcome this difficulty, predicting how far the birds assault would affect the aircraft's flying body. The investigation will be carried out in Python. The implementation flow Figure 2 that follows explains the study's implementation strategy.

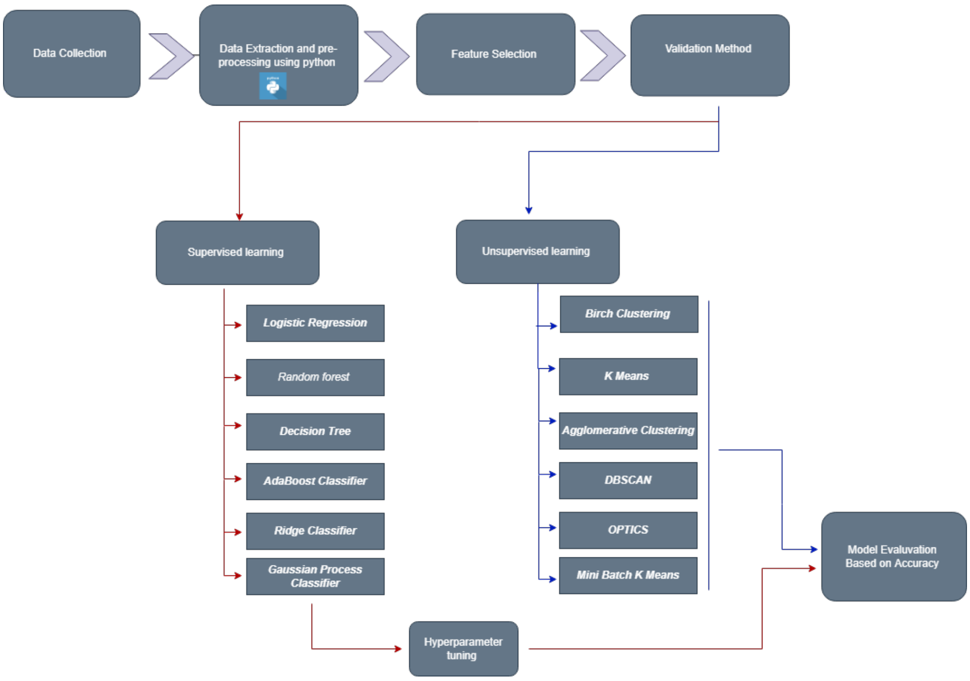


Figure 2: Strike Impact Prediction Methodology

Data Extraction: There was a first phase in the data extraction process. The dataset of real-time bird strike data from various airports across the world was acquired from the open repository Kaggle. The data was sent into Google Colab's python environment to be programmatically examined.

Data Preprocessing: After the data was delivered into the environment, the data was initially analyzed to understand the dataset's characteristics. Unwanted columns were removed from datasets as part of the data cleaning process. Those with a higher percentage of null values were taken into account and updated. No missed values would have resulted in a negative analysis due to the absence of further data. The data was first cleansed, and then data frames were generated using that information.

Feature engineering and techniques implemented: One of the metrics to be used in the continuing data collecting analysis was Exploratory Data Analysis (EDA). Insights may be gained from the EDA, which assists in confirming its usage and the value of its characteristics, and in selecting the target for further analysis. Visualizations assist derive feature relevance and show how important each dataset characteristic is for analysis and prediction. Convert categorical data to integers for machine learning models. The filtered category data was encoded into integers using a method called One Hot Encoding. Post-cleansing data is separated into test and train, with the latter being fed into the model. A test set is required to evaluate the models' performance after they have been trained on a train set. The technique is then followed by model creation and evaluation.

Machine learning algorithms: Both supervised and unsupervised machine learning algorithms will be utilized to forecast the data since machine learning was selected as the technique of technology. For the purpose of calculating each model's accuracy rating, the f1 score, precision, and recall are all taken into account.

Visualization and selection of best predictive model: The developed machine learning algorithms are shown using a heatmap to indicate their accuracy and assessed using a model based on supervised and unsupervised learning. In this research, both supervised and unsupervised learning models will be compared to see which one is the most effective. A visual representation of all 12 models was created with varying accuracy to determine which model performed the best.

4.2 Exploratory Data Analysis

The dataset being explored using different meaningful insights to carry out and have knowledge about the dataset has been achieved using EDA. This research aims to find if there is any effective damage to the airline body due to a bird strike. Whenever a bird strike has been recorded, the bird hit creates a massive impact on the aircraft's body. When the features from the data-carrying out the information were visualized, the data was unbalanced where the damage occurred rate was low compared with the non – damage rate.which has been shown below in Figure 3

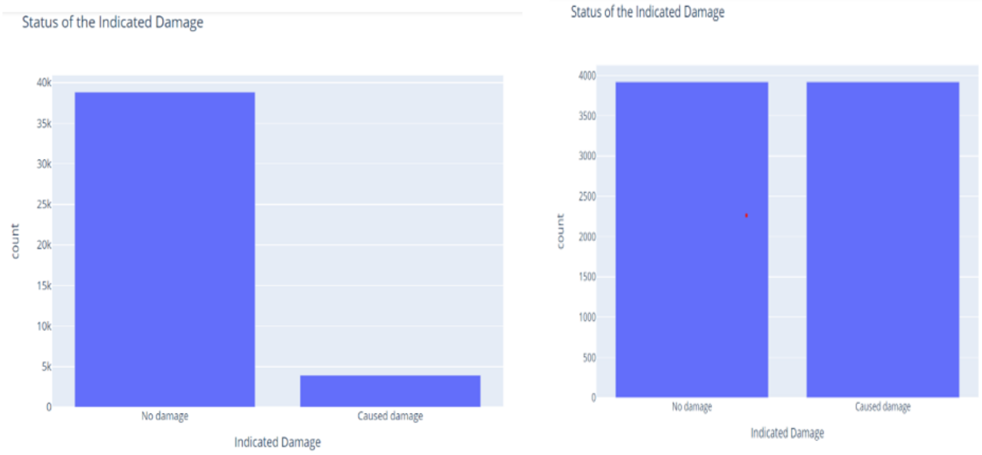


Figure 3: Handling Unbalanced Data

According to the analysis, the damage rate seems to be relatively low. In cases when the dataset includes more data that has resulted in no strikes, utilizing this data will result in a one-sided assessment since the models will be more trained to the no damage, and the analysis will be focused on the no harm, resulting in findings that are not suitable. The data were then balanced, their counts equalised, and models built using this data. After using sample techniques, the data must be balanced with the precise count and the dataset balanced. Second, the plots depicted the wildlife species that were reported as a consequence of the hits. The bird with the highest number of recorded strikes may be used for the analysis. The resulting graphic will help anticipate how far the species' influence will spread into the aircraft's body. On the other hand, a visual depiction of the birds that have been struck the most may be included. The resulting graphic will help anticipate how far the species' influence will spread into the aircraft's body. The python library has helped depict the species that are deemed more dangerous and capable of producing substantial effects, as well as the birds that have been involved in the most strikes. In cases like these, prior information may be used to predict the future and its impact on the aviation business..

The visualizations used to depict the information presented above are displayed in the next section. The two kinds of plots that will be used are bar plots and line plots, with the data being derived from the data utilizing the EDA analysis carried out programmatically. Based on the information provided below, it is evident that gulls and mourning doves are the two species that have been reported with more than 1000 strikes per year. The two types of plots that will be utilized are bar plots and line plots, with the information produced from the data using the EDA analysis that will be carried out programmatically to extract the data from the data.

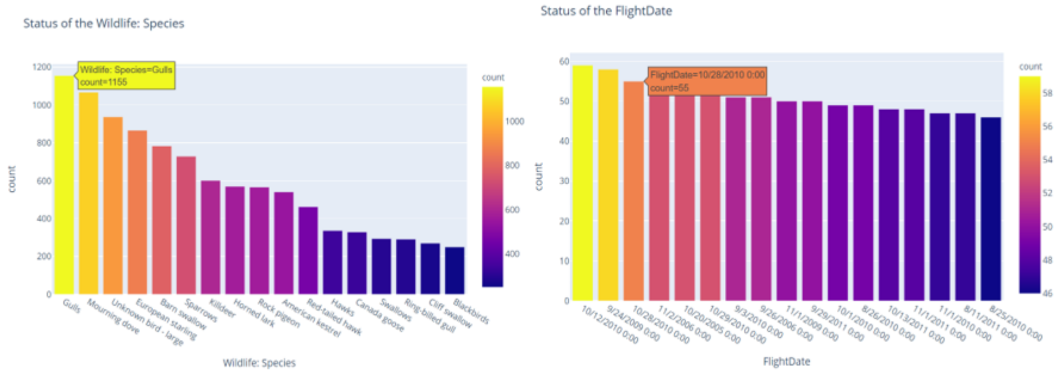


Figure 4: Bird hit Based on Species and flight date

Visualizations from Figure 4 depict the bird struck according to the species name and flying date. To see how far the birds have spread their wings, we will use this data to visually represent the damage they have caused. This data will be based on the species size, and the damage rate will be calculated based on the size of the species. This data will be used in order to see how far the birds have spread their wings. It is possible to go much deeper in the visualizing process.

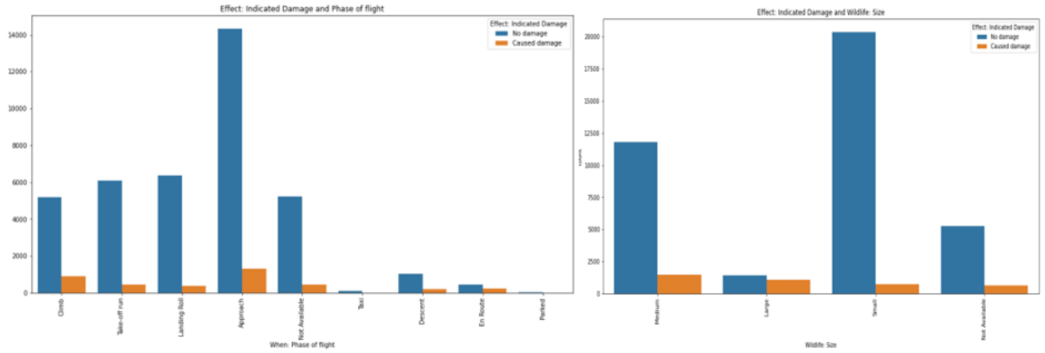


Figure 5: The phase of flight and species based on size being recorded with strike

It can be observed in the above visualization Figure 5 that the approaching and rising flights from the runway have recorded a more significant number of strikes than the other flights. As a result, it is clear that bird strikes are recorded when the flight approaches the runways and when climbing and that flights with a bird strike have been recorded once after the take-offs. If any preventive measures are to be implemented, these data visualizations would be more beneficial for the implementation of preventive measures.

The following visualization is carried out Figure 6 is the creation of insights based on the bird strikes that have happened at the various airports around the country. The airports with a high number of bird strikes are being identified and shown graphically. It is evident from the data that Denver International Airport and Memphis International Airport are the two airports that have been reported with more than 1000 attacks each, according to the data. Following then, the Sacramento International Airport has had approximately 1000 strike counts recorded in its history. Because these insights provide much more meaningful information, the primary purpose of these visualizations is to show that the recorded strikes which are expected to be more prevalent in those airports can be seen and used for further analysis and that the effect of bird strikes on aircraft in the visualized airports can be more informative for the airline industry to gain information about the bird hits in their location.

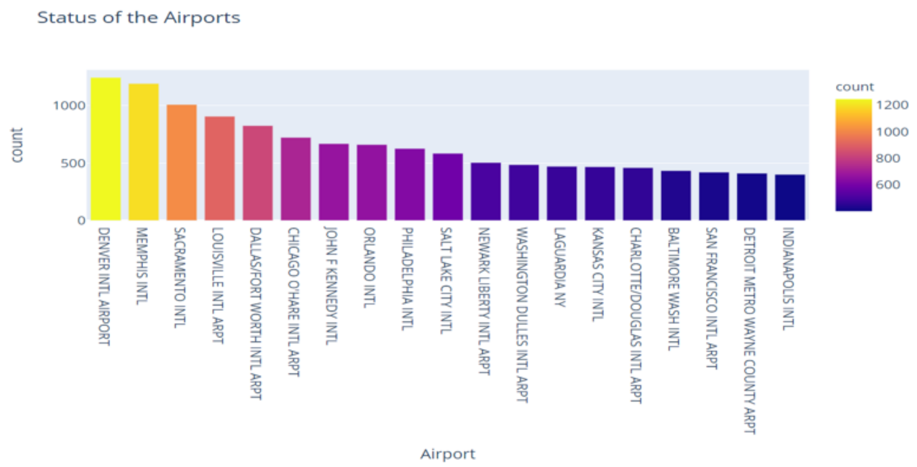


Figure 6: Status of Airports recorded with Bird strikes

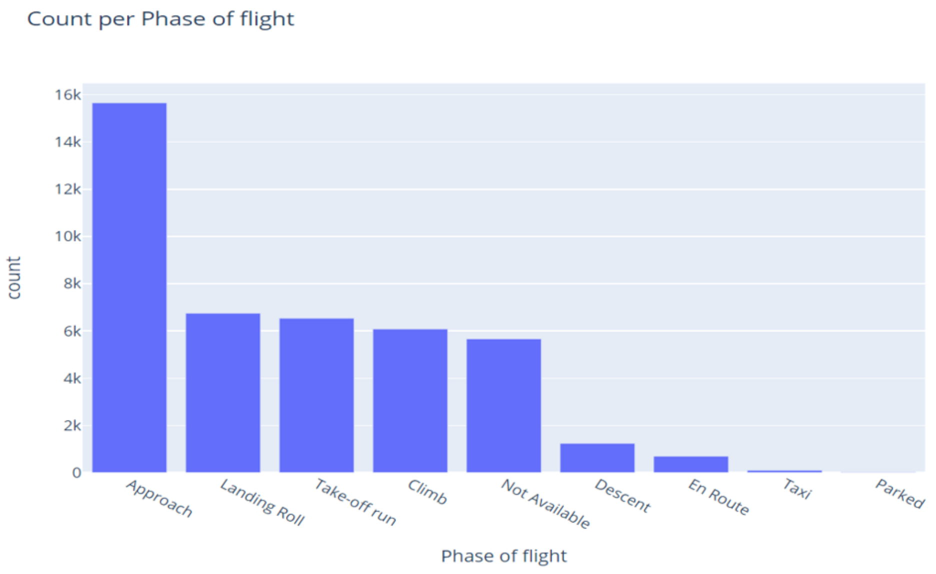


Figure 7: The phase of flight encountered with bird hit

According to the given insight Figure 7, the data indicates that the phase of flight during which a greater number of strikes has been recorded has a larger number of strikes. According to the visualization, the incoming planes had a higher number of bird attacks recorded than the departing flights, with more than 14,000 bird strikes reported on the departing flights. Landing aircraft have been seen to make more than 6,000 landings, which is the second highest total ever recorded. The approach and landing flights are the two phases that will be documented as having the highest number of strikes, according to the data. The take-off and rising stages of flights are the other phases of flights that are documented with strikes, and they are all closer to the strike count than the landing phases of flights. When the flights are in the moving phase, the visualization serves as evidence that bird hits are being recorded on the flights, which is the case. The strike has happened at an altitude before landing and some during the landing roll on flights nearing their destination airports. The birds on the runways may be to blame for the strike. As with the take-off flights, the rising flights that are recorded with strikes are quite similar to the flights that are recorded with strikes when flying into the airport during the approach phase.

4.3 Feature selection and data cleaning

Once the data has been extracted and entered into a programming environment, it is further analyzed in the pre-processing stage, where it is checked for any null values and cleaned to make the data more accurate. The missing values and null values are all treated, and the format of data cleaning that has been used, as there were null values found in the data to be dropped, there were some criteria based processes used to determine which null values should be dropped. In cases where a column contained more than 60% null values, the column was dropped, and features that were critical to the analysis were removed. The data with null values were replaced with other terms, and the data was further cleaned before being assembled into a data frame containing the appropriate information.

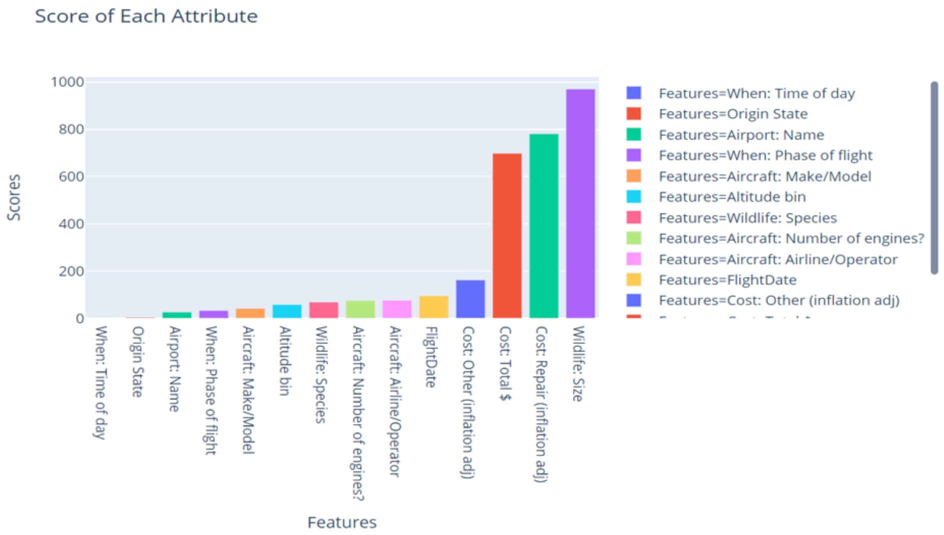


Figure 8: Scores based on the importance of each feature

While the above-mentioned visualizationFigure 8, accurately depicts the relevance of each characteristic for the study, the most critical features are filtered based on the scores and employed in the subsequent analysis; each feature is crucial. The primary characteristics that must be investigated in further depth are the size of the animals, the cost of repair, and the overall amount of money that will be spent or invested by the airlines whenever a bird hit is reported. From these insights, it is possible to identify the most significant features, and during the preprocessing step, the features that have entirely empty or null values may be changed and utilized for the implementation stage. As a result, the significance of the feature has been determined, and the most significant aspects have been picked for the feature.

4.4 Conclusion

To begin with, the data being extracted has gone through multiple phases to generate a comprehensive data set without any null or missing values that may not be acceptable for the study. That included cleaning and preprocessing the data. To better understand the information offered by the visualisations, the preprocessed data were used in the EDA section of the research. The data was skewed on the charts. However, by using this data, the data was balanced. The data from the visualisations guided the analysis.

5 Implementation, Evaluation of Machine learning models

5.1 Introduction

As previously said, machine learning is a commonly used technology for data prediction. This study employed a similar approach. The main purpose of this study was to see how well supervised and unsupervised learning systems can predict bird attacks in aviation. Six supervised and six unsupervised machine learning models were built to anticipate the impact of a strike. This section will discuss model implementation and evaluation. We've covered all of the steps in this section..

5.2 Technology and work environment

This study's workstation will include an Intel Core i5 CPU, 10th generation processors, 64-bit Windows 11, 512GB SSD storage, and 16GB RAM. To continue the inquiry, Google's collaborative Python environment was used. The study's resources will be discussed. Excel is used to see the raw data and learn more about the rows and columns. Due to the large datasets and built-in libraries used in the research, Google collaborate is the best solution. The development of supervised and unsupervised models in Colab notebook is more challenging, but it is easier to utilise a large dataset. A technical environment was used for this study.

5.3 Evaluation using metrics

In order to evaluate the accuracy of the machine learning algorithms being implemented, the f1 score, precision, and recall have been used to assess the accuracy as stated from the study Bradley et al. (2006)of the algorithms in question. The four components stated

above will be examined in further detail below.
Accuracy: One of the intuitive performances that should be taken into account as the real terms of prediction is accuracy. Even if a model has better accuracy, it doesn't mean that it's the most outstanding model or that its result prediction is flawless. A high degree of precision in a model might lead to false positives and false negatives. If we look at these indicators in connection to each other, it's evident they aren't the only ones we need to consider to forecast the proper conclusion. The other three metrics, namely precision, f1, and recall, have been incorporated in this study.
Precision: Precision is one of the good measures that need to be considered. Where precision has been based on the formula that

$$Precision = (Truepositive)/(TotalpredictedPositive))$$

From the formula, the True positive + false positive gets converted into the total predicted positive as in which the precision measure has the ability to show how precise the model is based on the precision score. From this, the real positive counts for the accuracy can be further considered based on the working precision as one of the good metrics to be implemented.
Recall: Recall metrics works as same as the precision metrics. Recall calculates the number of positive labels that the model has captured, which are to be the true positive. So when this has been implemented, the recall metric can be used to know the best model when the false negative has been linked with high-cost attributes.

$$Recall = (TruePositive)/(TotalActualPositive)$$

F1 Score: F1 score is to be a function of precision and recall, where the formula for recall is to be the,

F1=2* (Precision*Recall)/(Precision+Recall)}

From the above formula, the working of the precision has been formulated, where an F1 score is needed when there is a balance between the precision and recall metrics. In addition, the f1 score is to be implemented if there is an uneven distribution among the data where there are to be more actual negative in those situations, f1 score can be implemented.

5.4 Implementing Supervised, Unsupervised Machine learning algorithms

This research will use both supervised and unsupervised machine learning methods. Data validation follows data cleansing and preparation. The data was separated into train and test sets using Python's scikit module, and then the models were developed and validated.

5.5 Unsupervised algorithms

5.5.1 Birch Clustering:

Due to a shortage of resources, unsupervised learning is inefficient with large datasets. So the model was Python. As a result, typical clustering algorithms are slow and provide poor data. BIRCH clustering improves prediction accuracy. In this approach, all unsupervised models employ PCA to minimise data while maintaining information. Based on their performance, accuracy should be at 54.33%, which is below expectations.

5.5.2 k means clustering:

A cluster is a collection of data points that have been clustered together because of their commonalities. The K means clustering algorithm is one of the simplest algorithms. Starting with a random selection of data points, K signifies the algorithm begins its execution by updating the locations of the data points depending on the procedure. K-means is the simplest yet best model to predict quicker and more accurately than other models. Based on the evaluation using the measures implemented, the model has an accuracy of 57.39

5.5.3 Agglomerative clustering:

Hierarchical clustering, a typical kind of hierarchical clustering, was used to arrange the data according to their commonalities. Each item is treated as a singleton by the algorithm until it is fed into the clustering process, where it is combined into a single large object and shown in a tree-like flow called a dendogram. One of the motivations for using agglomerative clustering in this investigation was the difference in how well it worked. Their performance-based evaluation was excellent. They had an accuracy rate of around 54%.

5.5.4 DBSCAN:

While all clustering algorithms use the same methods, their purpose differs. The DBSCAN approach relies on clusters and noise, as detailed here. The DBSCAN approach was created because other methods like K-means and hierarchical clustering suffer from noise and outliers. When there is a lot of noise, DBSCAN, or density-based spatial clustering of applications with noise, may anticipate the outcome. This method's accuracy was found to be 50.12%.

5.5.5 OPTICS:

Method OPTICS is a well-known ordering point algorithm for figuring out clustering. Where the density-based spatial data has been deployed, the density that a cluster must accept is represented by each individual point's unique distance from the centre of the cluster. In comparison to other clustering models employed, the accuracy of the OPTICS method was expected to be just 51.41 percent, which was lower than the other models.

5.5.6 Mini Batch K – means:

As a result of the algorithm's ability to store random samples of data in allocated memory, it was used in this analysis of data. Random samples from the dataset are generated and used to update the clusters each time an iteration is started, and this process is continued until the execution completes. A convex combination of the variables is used to update the cluster in the micro-batch modifications. This may be done by using a learning rate that lowers dependent on the number of iterations, which is the iteration of numbers in clustering. The fresh data impact diminishes as the number of rounds increases. Mini Batch K-means algorithm is expected to have the greatest accuracy next to the K-means algorithm, with an accuracy of 57.27 percent.

5.6 Supervised algorithms

5.6.1 Logistic regression:

An improved accuracy rate was obtained by modifying the dataset's parameters. Using the average tuning rate of that specific model, the accuracy rate can be calculated. The data will be separated into distinct percentages using a cross-validation approach. They were making a forecast based on segmented percentages of the data. For each parameter, a random value is assigned to each parameter in the code, and the sequence of values obtained as a prediction rate, as a result, is used to determine the model's accuracy depending on parameter tuning. This model has a predicted accuracy rate of 72.57 percent.

5.6.2 Random Forest:

The model was built in Python using the data supplied into it. Our approach was based on what we learned from the logistic regression model. Thus we used Random Forest is used as a classification method for a reason. It is possible to develop more accurate uncorrelated trees by using a combination of bagging and feature randomization during the construction of the individual trees. Once the data has been fed, the best accuracy for the parameters that have been selected based on parameter tuning will be passed along. The algorithm was supposed to have a 72.57 percent accuracy rate.

5.6.3 Decision tree:

It is possible to forecast the goal based on the training model's decision rules once data has been verified into train and test data. While the model's technique is supposed to be unique, this model was used in this study. On top of that, when the model had been developed in Python, it was given test and training data from which parameter adjustment was made to improve accuracy. The algorithm has a 70.08 percent success rate.

5.6.4 AdaBoost Classifier:

As a supervised machine learning approach's boosting method. As an ensemble technique, this methodology has been utilized before. As a result, weights are continually recalculated in the model, even called the "meta estimator." In supervised learning, the "boosting implementation" technique is used to decrease bias and variation. The Adaboost algorithm is based on the same boosting idea, but with a few tweaks. Once the Ababoost classifier is in place, the data is fed into model one and then validated using the model one classifier. The model's accuracy rate was computed after implementing parameter adjustment. The model's predicted accuracy is set at 75.38 percent.

5.6.5 Ridge Classifier:

Assumptions about subspaces are more prevalent in this model, which indicates that the samples of class are dependent on linear subspaces, as shown by the Ridge regression. When it comes to algorithmic accuracy, the ridge classifier comes out on top due to its better performance and better ability to forecast outcomes in advance. Using data-driven parameter tweaking, this study's algorithm predicted an accuracy rate of 71.42 percent, based on the other metrics used to measure accuracy.

5.6.6 Gaussian Process Classifier:

It is one of the supervised machine learning algorithms known as the generalized concepts of the Gaussian probability distribution, the Gaussian process classifier (GPC). In order to test the model's accuracy, the model was fed with data and applied with parameter tweaking to see how accurate the model was based on the results that had been produced. According to the results, the model has an accuracy rate of 63.52 percent.

5.7 Conclusion

This study's main purpose is to assess how well supervised and unsupervised models predict the impacts of bird assaults on structures. There were 12 machine learning algorithms in all, six supervised and six unsupervised. The performance and functioning techniques of the models were chosen. We reviewed each model's accuracy rate and model itself in the previous section. The following part will go through how the research is assessed and the methodology employed.

6 Discussion and evaluation of the methodology

6.1 Discussion

The results, scope of work, and obstacles encountered during execution are to be compared here. It assesses the accuracy of supervised and unsupervised methods. The research separated the method into two parts: supervised and unsupervised. The models were fed data and then evaluated for efficiency. For this research, supervised and unsupervised risk assessment methods were compared for estimating the implications of bird strike hazards. They were preprocessed to form a comprehensive dataset once the data was read. Some qualities have more null values. Having more null values brought the data to analysis. After gathering the data, the EDA was performed, revealing significant insights. The most important features and ideals were visualised to make them more clear. The traits were picked from it. We made further machine learning predictions using the datasets. Given that unsupervised learning takes longer than clustering, the PCA was used to investigate a smaller set of data at a time.

This study's purpose was to assess how well each machine learning algorithm's subset predicted the expected output. The models needed the labels to be in integer format with ” ” than symbols replaced with ’less than’ symbols. As a result, the number of non-damaged hits exceeded the number of damaged hits. Predicting the aim is the basis. Based on this data, damaged and non-damaged values were equalised by sampling. Unsupervised learning would be used initially, followed by a deeper dive into the data. As a result, PCA was utilised, using just 2 features instead of 8, and 95 percent of the data returned in 2 feature format. It was simple to employ PCA-based clustering. In addition, the efficacy of the supervised machine learning algorithms was assessed, and parameter tuning was done to the models in order to increase the accuracy of the models based on the parameters that were tested. The tuned models took longer to run since tweaking each parameter took time. The research indicates that bird crashes caused the damage detected during flight. Strikes on wildlife have long been observed as aircraft approach airports, according to the EDA's species strike plot. Supervised machine learning predicts better. Unsupervised learning requires less time. Problems included data processing and estimating aircraft effect on Google Colaboratory Notebook (16vCPU, 104 GB RAM).

6.2 Evaluation Of implemented models

The research utilised 12 models, six supervised and six unstructured. Machine learning was used to assess whether bird impacts had endangered the plane's integrity. Other measures, such as the f1 score, precision, and recall, were utilised to assess each model's accuracy, and the results were used to decide which analytics approach was employed in the final analysis. The results of supervised and unsupervised model assessments were plotted to measure the models' effectiveness. It was also tested using Python libraries. The best model was chosen based on its accuracy rate, which was determined to be the model that predicted the damage rate the most accurately. It has been shown in the previous picture Figure 9, that the model performance may be displayed. It is intended to use these plots to demonstrate the ideal model derived from the unsupervised learning model and to evaluate the model performance. According to the results of this evaluation, it is obvious that tiny batch K-means and Kmeans performed well in terms of accuracy and effectiveness in predicting the effects of bird attacks on aircrafts. Mini batch K-means and K-means were shown to be the most effective algorithms for unsupervised learning when compared to the other algorithms that were being tested. When compared to other unsupervised models, their accuracy in prediction and effectiveness in prediction were much higher than other unsupervised models. According to how accurate they were, the following heatmap was created based on their results.

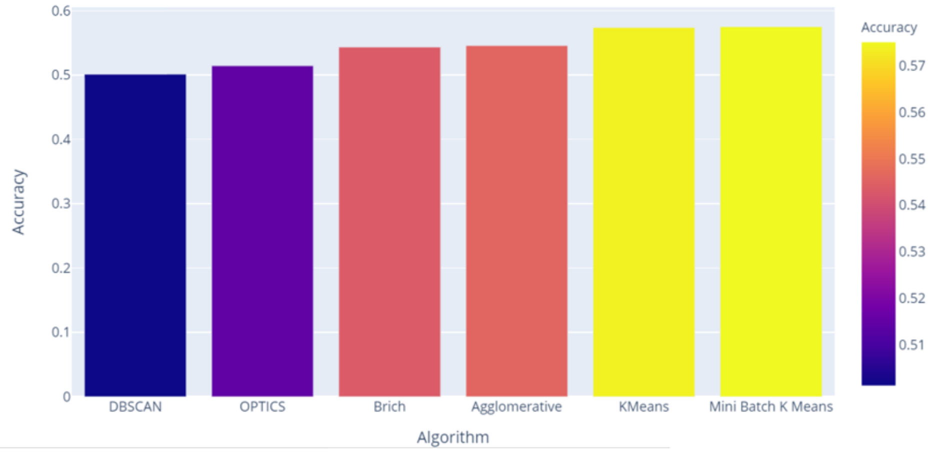
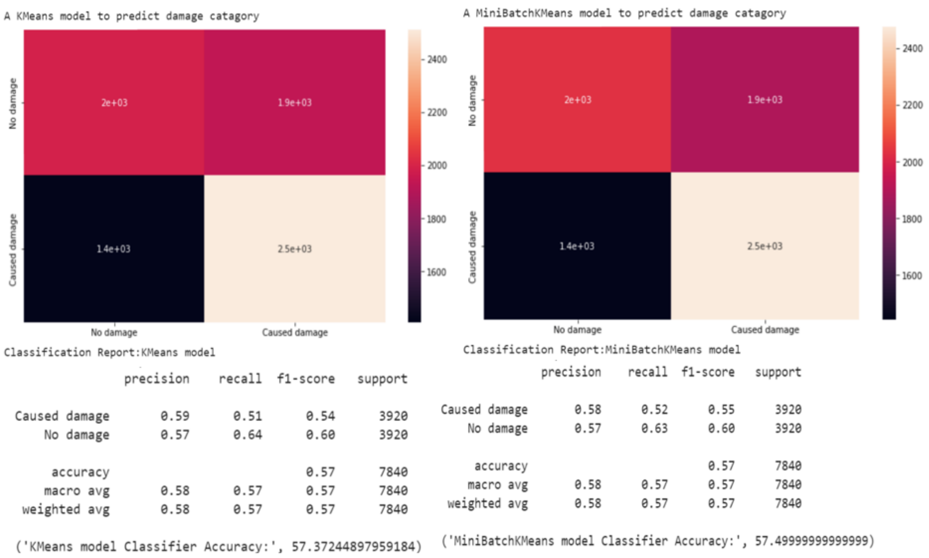


Figure 9: Accuracy comparison of Unsupervised algorithm



```
Classification Report:KMeans model
               precision    recall  f1-score   support

Caused damage       0.59      0.51      0.54      3920
    No damage       0.57      0.64      0.60      3920

     accuracy                           0.57      7840
    macro avg       0.58      0.57      0.57      7840
 weighted avg       0.58      0.57      0.57      7840

('KMeans model Classifier Accuracy:', 57.37244897959184)
```

```
Classification Report:MiniBatchKMeans model
               precision    recall  f1-score   support

Caused damage       0.58      0.52      0.55      3920
    No damage       0.57      0.63      0.60      3920

     accuracy                           0.57      7840
    macro avg       0.58      0.57      0.57      7840
 weighted avg       0.58      0.57      0.57      7840

('MiniBatchKMeans model Classifier Accuracy:', 57.49999999999999)
```

Figure 10: Heatmap based on performance for both K-means and Mini batch K-means

From the Figure 10 heat map being generated for both the k-mean and mini batch k – means model have been shown and based on the evaluation metrics, the accuracy rate of both the models have been generated where the proper accuracy has been derived based on the other metrics being implemented as to predict the damage rate of models where both well performed models from unsupervised learning had accuracy rate of 57% with slight variations in them.

The accuracy prediction rate of the supervised learning algorithm vs the unsupervised learning algorithm can tell how far the data has been forecasted. The supervised learning approach was used with parameters adjusted to see how well the negative effect was expected. Even the best-chosen algorithms have not delivered outcomes that are equivalent to or closer to the accuracy rates of supervised learning. The optimal method should be compared to other models used for the study. Using supervised learning, each model's parameter has been tuned to increase model accuracy. The accuracy visualisation given below is for your convenience.. Comparing the best method to other models that have been created for the analysis, it is essential to determine which model is the best. Each model has been applied with parameter tweaking with the assistance of supervised learning in order to increase the accuracy rate of the models. We've given a visual representation of accuracy in the following image for your convenience. According to the accuracy metrics and evaluation, the supervised machine learning approach has been more accurate in terms of prediction. Their results in predicting the damage rate due to bird hits have been more accurate. In addition, the supervised machine learning algorithm has been more perfect in terms of prediction. Their performance has been good compared to the unsupervised machine learning approach, which has been less accurate. The below visualization that has been performed Figure 14 the insights being carried compares the performance of supervised machine learning algorithms.

Each of the numerous measures that were being applied had an accuracy rate that was expected based on the data analysis. The models that were further tested for accuracy had model predictions that were more than 70% accurate, with the highest performing algorithms being AdaBoost Classifier and Random forest, which had higher accuracy when compared to other models. In order to demonstrate the model's performance, an

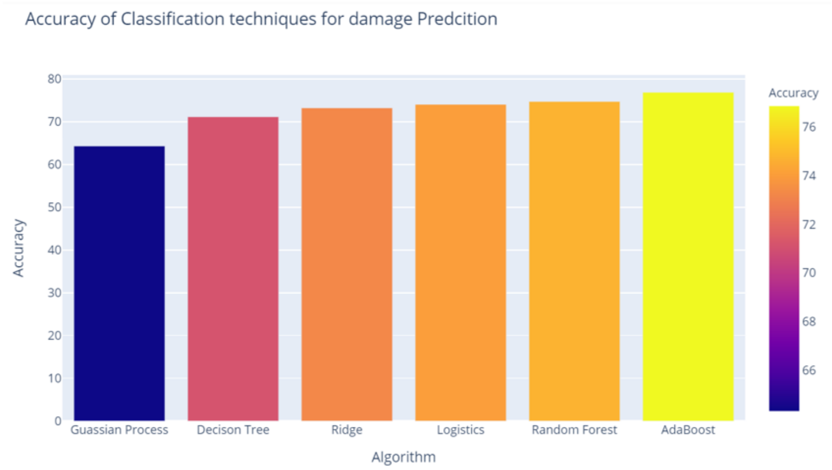
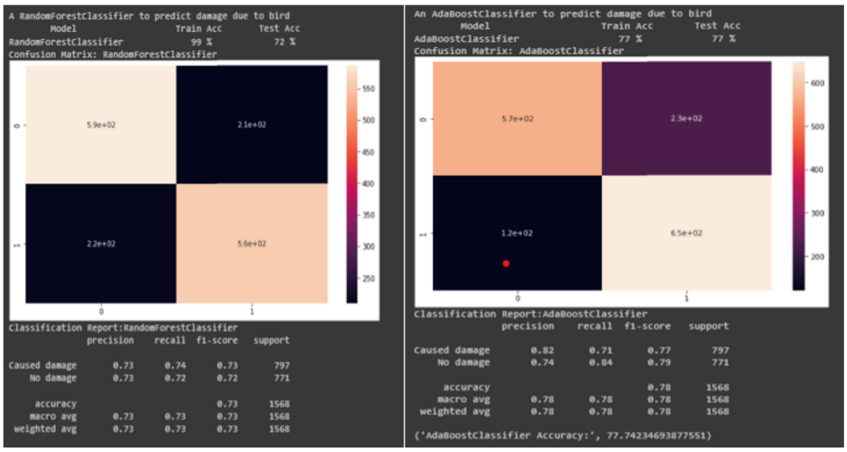


Figure 11: Accuracy comparison of supervised algorithm

accuracy-based assessment has been provided in the sections below.

When compared to the other 11 models, the model has very excellent accuracy since this algorithm has been more effective in anticipating the impact of the strikes as well as in determining how much harm has been caused by the data strike. It was found that the AdaBoost classifier was one of the most efficient algorithms when it came to predicting the target. On the basis of the other measures that have been added, it is obvious from



```
A RandomForestClassifier to predict damage due to bird
        Model                      Train Acc      Test Acc
RandomForestClassifier               99 %           72 %
Confusion Matrix: RandomForestClassifier
```

```
Classification Report:RandomForestClassifier
              precision    recall  f1-score   support

Caused damage       0.73      0.74      0.73       797
    No damage       0.73      0.72      0.72       771

     accuracy                           0.73      1568
    macro avg       0.73      0.73      0.73      1568
 weighted avg       0.73      0.73      0.73      1568
```

```
An AdaBoostClassifier to predict damage due to bird
        Model                      Train Acc      Test Acc
AdaBoostClassifier                   77 %           77 %
Confusion Matrix: AdaBoostClassifier
```

```
Classification Report:AdaBoostClassifier
              precision    recall  f1-score   support

Caused damage       0.82      0.71      0.77       797
    No damage       0.74      0.84      0.79       771

     accuracy                           0.78      1568
    macro avg       0.78      0.78      0.78      1568
 weighted avg       0.78      0.78      0.78      1568

('AdaBoostClassifier Accuracy:', 77.74234693877551)
```

Figure 12: AdaBoost and Random Forest performance

Figure 12 that the accuracy rate is being further examined and improved. The Random forest model was determined to be the second most accurate model after the Random forest model. The results of the random forest method are displayed in the table below Figure 12.

The model is presently being built, and parameters are being tuned to see how well the algorithm predicts the data. The model's accuracy has been assessed based on the data's multiple metrics. The logistic regression model outperforms other models in predicting data outcomes. Overall, the model's performance was acceptable compared to other models. The data for the model being applied has been examined using several machine learning techniques. Based on the study, the data has been fine-tuned and made more accurate and comprehensive to continue the investigation. The data was only put into the models when the validation phase had been completed successfully. The data was input into each model with the verified data that had been separated into train and test segments. As seen in the plotFigure 13,

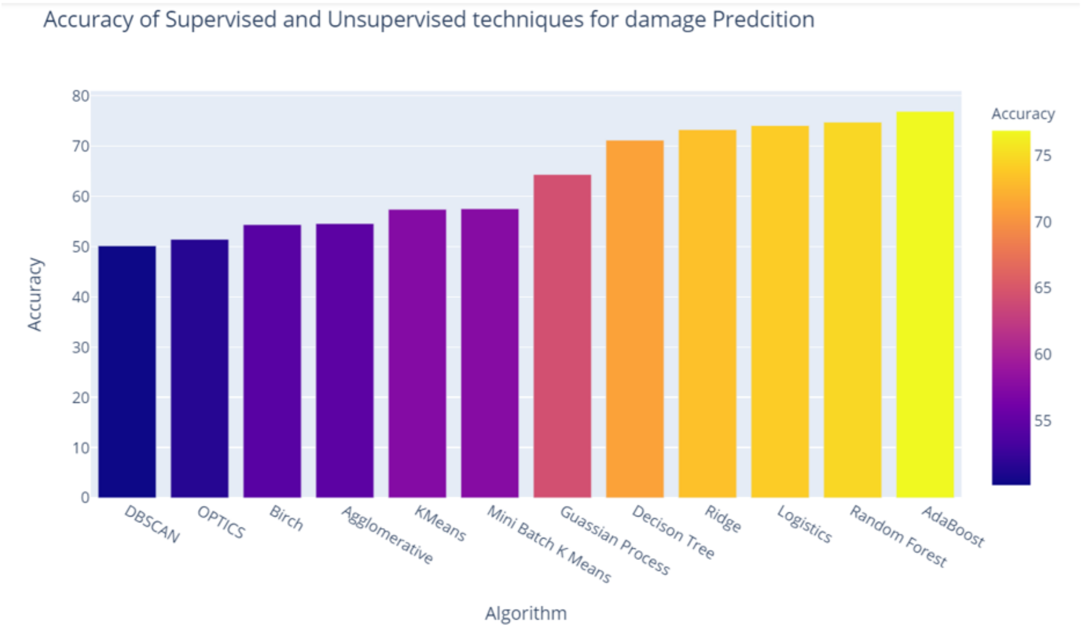


Figure 13: Overall Model Comparison

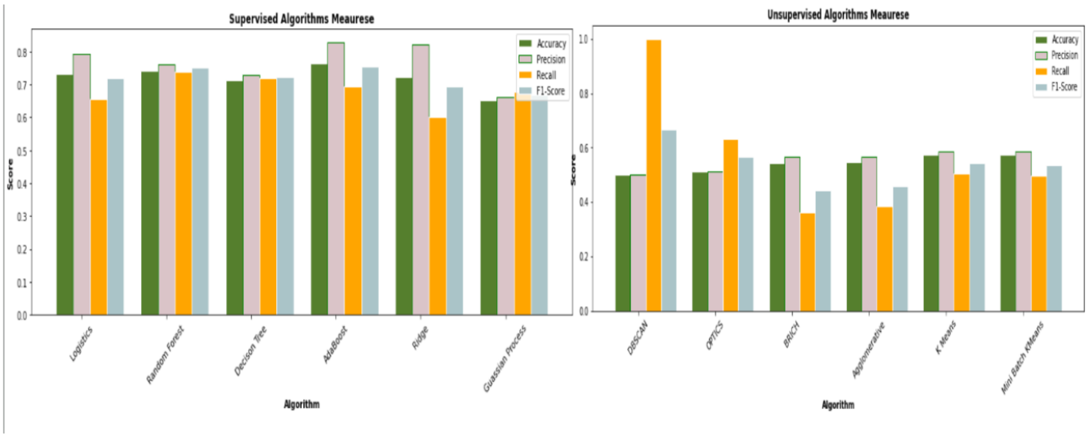


Figure 14: Model performance based on implemented metrics

Based on the data the analysis has taken place and there were 12 machine learning models has been implemented.Each single model has been evaluated using the accuracy, the main reason to use accuracy as it is more crucial to make business decisions and based on this it helps to make better business judgements. as shown in figure 14 Figure 14., the metrics being implemented has been compared to check the performance of the models. It is apparent that the supervised machine learning algorithms have done a good job with this particular study. A supervised machine learning algorithm is one in which the prediction rate of all of the algorithms is much greater than the prediction rate of the unsupervised machine learning algorithm. Using the data that was fed into the supervised machine learning technique, the data was further modified by performing parameter tuning in order to increase the accuracy rate of the model and hence improve its performance. As a result, each model has become more predictive and effective in predicting the target, as evidenced by the fact that the majority of supervised machine learning algorithms had an accuracy rate of more than 70 percent and, in general, more than 60 percent, whereas the majority of unsupervised machine learning algorithms had an accuracy rate of more than 50%.

6.3 Conclusion

The machine learning algorithms that were implemented for this research were based on both the supervised and unsupervised machine learning approaches, as demonstrated by the results of the algorithm evaluation, which revealed that the prediction rate of supervised machine learning algorithms was higher when compared with unsupervised machine learning algorithms and that the accuracy rate of the models was evaluated using a variety of metrics that had been implemented in the research.

7 Future Work and Conclusion

The outcomes were totally predicted by the analysis. The study has accurately predicted the damage caused by bird hits, as well as the extent of the injuries. Strikes against aircraft approaching airports have been greater than other phases of flight. The statistics show an increase in bird attacks when aircraft are flying at different altitudes. Using more accurate data and testing alternate supervised and unsupervised learning methods may improve the accuracy of the prediction. The whole data may be used for this study since the data must be more precise and thorough to bypass the sampling techniques. More animal data should be included in the research to see more strikes. The research could also employ additional flight and strike data to better forecast the animals being reported. This research aimed to compare supervised and unsupervised algorithms, with data input programmatically and outcomes tweaked depending on accuracy. The supervised vs unsupervised technique was fine-tuned using programmatic data input and accuracy-based outcomes. The data was cleansed at every step of preparation, and the data was implemented in 12 different ways with care and skill. The data was displayed whenever any useful parameters would give important insights were plotted, and the information was communicated between each visualisation. The data was then split into two groups: the training set and the test set, and the information was examined.

To examine the usefulness of supervised and unsupervised machine learning algorithms in forecasting bird strike damage in the aviation sector. The findings demonstrated that supervised machine learning was better at forecasting bird hit damage in the aviation sector. Supervised learning was more accurate than unsupervised learning because it was better at predicting. Surveillance machine learning worked effectively in this study, and the bird impacts did affect the plane's structural integrity. The different phases of flight documented with bird assaults have had a range of implications on the observed aircraft.

8 Acknowledgement

I would like to appreciate and express my heartfelt gratitude to Bharathi Chakravarthi, who helped me through the whole proposal report process. As a fresher and student studying overseas, I benefited from the in-depth information offered by my lecturer about doing research. I want to express my gratitude to my father, Late Mr P.Venugopal, whose hard work and blessings have made me who I am today, and to my mother, S.Rajarajeswari, who raised me all by herself as a single parent. I will be eternally grateful to all of the people mentioned above who have assisted me in pursuing my international study goals. Finally, I would like to express my gratitude to God for guiding me through all of my troubles. I have seen your leadership daily, and I will continue to put my faith in you in the future.

References

AISSAOUI, O. E., EL MADANI, Y. E. A., OUGHDIR, L. and ALLIOUI, Y. E. (2019). Combining supervised and unsupervised machine learning algorithms to predict the learners' learning styles, *Procedia Computer Science* **148**: 87–96. THE SECOND INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING IN DATA SCIENCES, ICDS2018.

Altringer, L., Navin, J., Begier, M. J., Shwiff, S. A. and Anderson, A. (2021). Estimating wildlife strike costs at us airports: A machine learning approach, *Transportation Research Part D: Transport and Environment* **97**: 102907.

Baranzini, D. and Zanin, M. (2015). Baranzini, d., and zanin, m. (2015). risk prediction risk intelligence in aviation – the next generation of aviation risk concepts from prospero fp7 project. esrel 2015 - 25th european safety and reliability conference.

Bradbeer, D. R., Rosenquist, C., Christensen, T. K. and Fox, A. D. (2017). Crowded skies: Conflicts between expanding goose populations and aviation safety, *Ambio* **46**(2): 290–300.

Bradley, A., Duin, R., Paclik, P. and Landgrebe, T. (2006). Precision-recall operating characteristic (p-roc) curves in imprecise environments, *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 4, pp. 123–127.

Deng, Z., Hu, Y., Zhu, M., Huang, X. and Du, B. (2015). A scalable and fast optics for clustering trajectory big data, *Cluster Computing* **18**(2): 549–562.

DİkbayIr, H. S. and Bülbül, H. (2018). Estimating the effect of structural damage on the flight by using machine learning, *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1333–1337.

Lieber, D., Stolpe, M., Konrad, B., Deuse, J. and Morik, K. (2013). Quality prediction in interlinked manufacturing processes based on supervised unsupervised machine learning, *Procedia CIRP* **7**: 193–198. Forty Sixth CIRP Conference on Manufacturing Systems 2013.

Lorbeer, B., Kosareva, A., Deva, B., Softić, D., Ruppel, P. and Küpper, A. (2018). Variations on the clustering algorithm birch, *Big Data Research* **11**: 44–53. Selected papers from the 2nd INNS Conference on Big Data: Big Data Neural Networks.

Mehta, J., Vatsaraj, V., Shah, J. and Godbole, A. (2021). Airplane crash severity prediction using machine learning, *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–6.

Robinson, L., Mckay, T. and Mearns, K. (2021). Oliver tambo international airport, south africa: Land-use conflicts between airports and wildlife habitats, *Frontiers in Ecology and Evolution* **9**: 1–9.

Singh, A., Prakash, B. S. and Chandrasekaran, K. (2016). A comparison of linear discriminant analysis and ridge classifier on twitter data, *2016 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 133–138.

Valletta, J. J., Torney, C., Kings, M., Thornton, A. and Madden, J. (2017). Applications of machine learning in animal behaviour studies, *Animal Behaviour* **124**: 203–220.

Viswanath, P. and Suresh Babu, V. (2009). Rough-dbscan: A fast hybrid density based clustering method for large data sets, *Pattern Recognition Letters* **30**(16): 1477–1488.

Wang, R. (2012). Adaboost for feature selection, classification and its relation with svm, a review, *Physics Procedia* **25**: 800–807. International Conference on Solid State Devices and Materials Science, April 1-2, 2012, Macao.