Literature Survey:

Flight delay prediction using Isolation Forest algorithm[1]:

In this paper, an analysis has been conducted to discover the main causes of flight delays and to explore a few machine learning algorithms to find the most suitable one. According to statistics from the Bureau of Transportation, more than 69% of flight delays are caused due to unexpected weather conditions. This paper mainly focuses on considering weather as the main factor for flight delay anomalies.

The flight dataset has been collected from the Bureau of Transportation(BTS) website, which has over 20 variables for each flight. The weather data has been collected from the National Oceanic and Atmospheric Administration (NOAA), with 4 or 5 determining factors(such as temperature, precipitation, etc..) consisting of values for each factor reported at different points in a day. The mean of the values at different points of the day are taken and used for integration with the Flight data.

In the initial phase of implementation, algorithms such as random forest and KNN have been used, but their accuracy was lower (up to 62%). This is because of anomalies (in our case, flight cancellation/delay is the anomaly) in the dataset. As the main goal is to capture those deviations, the next suitable method to model this dataset is the **Isolation Forest** method. This is an **unsupervised** learning algorithm that is mainly used if the dataset contains anomalies and gives high accuracy in those cases as its main purpose is to capture the anomalies.

This model gives an accuracy of 76%. Certain improvements can be done to improve the accuracy but it involves additional overhead. For example, Instead of calculating the mean of weather values at different points in time, real-time airport weather for each flight could be collected, which turns out to be a tedious task. Also, accuracy may rise a little if factors other than the weather are also considered.

Analysis of classification models on flight delay prediction[2]:

This paper explores the parameters affecting flight delays and analyses various machine learning models that can be used for the prediction of flight delays. Different studies conducted in different places show that flights are delayed due to different parameters. For example, the main parameters that affect the airline network in the US are visibility, wind, and departure time, and in Iran are fleet age and aircraft type.

Different classification models such as Logistic Regression, K-Nearest Neighbor (KNN), Gaussian Naïve Bayes, Decision Tree, Support Vector Machine (SVM), Random Forest, and Gradient Boosted Tree have been used for prediction. The main objective of this study is to predict flight delays based on labeled data. Therefore, a **supervised** learning classification algorithm was selected as the appropriate one. They calculated the values accuracy, precision, recall, and f1 score to conclude the better model among the chosen models.

The result shows that the highest values of accuracy, precision, recall, and f1-score are generated by the **Decision Tree** model (accuracy: 0.9778; precision: 0.9777; recall: 0.9778; f1-score: 0.9778). Other tree-based ensemble classifiers also show good performance. Random Forest and Gradient Boosted Tree have an accuracy of 0.9240 and 0.9334, significantly higher than the rest of the models. The other four base classifiers Logistic Regression, KNN, Gaussian Naïve Bayes, and SVM, are not tree-based and therefore do not show a good performance in classification. The KNN model has the least performance since its precision and f1-score are the lowest among the seven models.

Flight Delay Prediction Using XGBoost[3]:

In this paper, an analysis has been performed using XGBoost which is one of the most popular machine learning algorithms regardless of the type of prediction task at hand

- regression or classification. It has become the state-of-the-art machine learning algorithm to deal with structured data.

A publicly available Kaggle dataset collected from United States domestic air traffic has been used for training. The dataset consists of over 3 million samples with 19 features. XGBoost is software that can be installed on our machine and accessed from a variety of interfaces. The library is focused on computational speed and model and the performance of the model. In this implementation, Mean absolute error(MAE) has been used for delay prediction. In statistics, MAE is the average vertical distance between each point and the identity line or it is also the average horizontal distance between each point and the identity line.

Based on the analysis of the results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and regression can provide promising tools for inference in the current domain

Flight Delay Classification Prediction Using Stacking Algorithm[4]:

This paper aims to prove that the stack algorithm has advantages in airport flight delay prediction, especially for the algorithm selection problem of machine learning technology. The authors use SMOTE to preprocess an imbalanced dataset and use Boruta Algorithm for feature Selection. The author uses five supervised ML algorithms for first-level learners(KNN, Gaussian Naive Bayes, Random forest, decision Tree, and Logistic regression) and Logistic regression for second-level learners.

The data set is from Logan International Airport in Boston, Massachusetts, the United States, which contains 298914 flight datasets and 67822 delayed flights. The SMOTE algorithm is used to balance the dataset. The features are selected using the Boruta algorithm that is applied to nine features(weather is not included in feature

selection) and 4 features were marked as critical (arrival time, day of month, month, and departure time), and stack-based learning is used with k-fold cross validation.

Comparing the level one learners, the Random forest and KNN had good prediction results with accuracy exceeding 0.8 and 0.7 respectively. Whereas Gaussian Naive Bayes and Logistic regression performed poorly. The stacking algorithm also yields good results with an accuracy of over 0.8. The stacking algorithm is also stable as the result does not vary significantly even if learners are removed from the stacking algorithm. The authors conclude that stacking algorithms are good for algorithm selection and it is also stable even when learners are removed from the stacking algorithm.

References

1. Miloš Vereš, Flight delay/cancellation prediction using machine learning Adapting new ways to help stranded passengers
2. Yuemin Tang, Airline Flight Delay Prediction Using Machine Learning Models
3. K.P. Surya Teja, Vigneswara Reddy, Dr. Shaik Subhani. Flight Delay Prediction Using Machine Learning Algorithm XGBoost
4. Jia Yi, Honghai Zhang, Hao Liu, et al. Flight Delay Classification Prediction Based on Stacking Algorithm