<https://arxiv.org/pdf/1903.06740.pdf>

1. This paper uses random forest, SVM, Gradient Boosting Classifier and k-nearest neighbor to analyze flight arrival delay to find the best performing classifier for predicting whether a flight will be delayed 15 minutes or not. The result from this paper is that Gradient Boosting Classifier had the best performance of 79.7%. This paper suggests that Gradient Boosting Classifier would be an effective model to use on our flight dataset.   
   Link:<https://www.researchgate.net/publication/327389509_Flight_Arrival_Delay_Prediction_Using_Gradient_Boosting_Classifier>  
   Chakrabarty, Navoneel, et al. ”Flight Arrival Delay Prediction Using Gradient Boosting Classifier.” Emerging Technologies in Data Mining and Information Security. Springer, Singapore, 2019. 651-659
2. This paper uses multiple linear regression to model arriving flights and predict flight delays. From their model, they got an accuracy of 80%, an improvement from Naive Bayes and C4.5 models. This paper suggests that multiple linear regression should also be considered for the model. Link: <https://iopscience.iop.org/article/10.1088/1755-1315/81/1/012198>  
     
   Yi Ding ”Predicting flight delay based on multiple linear regression”, IOP Conference Series: Earth and Environmental Science.
3. This paper is an example of using big data tools (in this case, MapReduce) to implement a predictor of flight arrival times using weather data. Although we are not using MapReduce, this paper serves as an example for data analysis and mining on a cloud platform to predict flight arrivals. Link: <https://pdfs.semanticscholar.org/c743/57537069272d19b9cf8f4b2b2cbb5c71e73b.pdf>  
     
   Belcastro, L. & Marozzo, Fabrizio & Talia, Domenico & Trunfio, Paolo. (2016). Using Scalable Data Mining for Predicting Flight Delays. ACM Transactions on Intelligent Systems and Technology. 8. 10.1145/2888402.

5) [Machine Learning with PySpark and MLlib — Solving a Binary Classification Problem](https://towardsdatascience.com/machine-learning-with-pyspark-and-mllib-solving-a-binary-classification-problem-96396065d2aa)

**Machine Learning Models**

This will be a review of existing machine learning approaches on flight data. One paper by Chakrabarty et al [4] uses random forest, SVM, Gradient Boosting Classifier and k-nearest neighbor to analyze flight arrival delay to find the best performing classifier for predicting whether a flight will be delayed 15 minutes or not. The result from this paper is that Gradient Boosting Classifier had the best performance of 79.7%. This paper suggests that Gradient Boosting Classifier would be an effective model to use on our flight dataset. Another paper by Ding [5] uses multiple linear regression to model arriving flights and predict flight delays. From their model, they got an accuracy of 80%, an improvement from Naive Bayes and C4.5 models. This paper suggests that multiple linear regression should also be considered for the model. This paper by Belcastro et al [6] is an example of using big data tools (in this case, MapReduce) to implement a predictor of flight arrival times using weather data. Although we are not using MapReduce, this paper serves as an example for data analysis and mining on a cloud platform to predict flight arrivals.