**PREDICTING FLIGHT DELAYS WITH ERROR CALCULATION**

**A Major-Project Report submitted in partial fulfilment of the requirements for the award of the degree of,**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

Submitted by:

**GOLI SATYA DARSHAN 321910303001**

**NAGABANDI AKSHAY KUMAR 321910303004**

**ROOPESH REDDY P 321910303006**

**GUNTUPALLI HEMANTH KUMAR 321910303018**

**MEREDDY KRISH REDDY 321910303028**

**Under the esteemed guidance of**

**Mr. G SPANDAN**

Calendar

Description automatically generated with low confidence**Assistant Professor**

# Department of Computer Science & Engineering,

**GITAM SCHOOL OF TECHNOLOGY**

**(Deemed to be University)**

**Bengaluru Campus.**

**March 2023**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**GITAM SCHOOL OF TECHNOLOGY**

**GITAM**

## A picture containing calendar Description automatically generated (Deemed to be University*)*

# DECLARATION

We, hereby declare that the project report entitled **“**Predicting Flight Delays with Error Calculation Using Machine Learning**”** is an original work done in the Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University) Bengaluru submitted in partial fulfilment of the requirements for the award of the degree of B.Tech. in Computer Science and Engineering. The work has not been submitted to any other college or University for the award of any degree.

**Date:**

|  |  |  |
| --- | --- | --- |
| **Registration No(s).** | **Name(s)** | **Signature(s)** |
| 321910303001 | GOLI SATYA DARSHAN |  |
| 321910303004 | NAGABANDI AKSHAY KUMAR |  |
| 321910303006 | ROOPESH REDDY P |  |
| 321910303018 | GUNTUPALLI HEMANTH KUMAR |  |
| 321910303028 | MEREDDY KRISH REDDY |  |

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**GITAM SCHOOL OF TECHNOLOGY**

**GITAM**

## A picture containing calendar Description automatically generated(Deemed to be University)

# CERTIFICATE

## This is to certify that the project report entitled “Predicting Flight Delays With Error Calculation Using Machine Learning” is a **Bonafide** carried out by the students of B3-4th year –Roopesh reddy p (321910303006), Mereddy sai krish reddy (321910303028), Guntupalli Hemanth Kumar (321910303018), Goli Satya Darshan (321910303001), and Akshay kumar (321910303004) submitted in partial fulfilment of requirement for the award of degree of Bachelors of Technology in Computer Science and Engineering.

|  |  |
| --- | --- |
| **Project Guide** | **Head of the Department** |
| **SIGNATURE OF THE GUIDE** | **SIGNATURE OF THE HOD** |
| **Mr.G.Spandan** | **Dr Vamsidhar Yendapalli** |
| Assistant Professor | Professor |
| Computer Science and Engineering | Computer Science and Engineering |
| GITAM School of Technology | GITAM School of Technology |

# ACKNOWLEDGEMENT

Firstly, we would like to sincerely thank “**GANDHI INSTITUTE OF TECHNOLOGY**” for giving us a chance to do a mini project on the topic “**Predicting Flight Delays With Error Calculation Using Machine Learning**” and everyone who has been there and given every possible hand to make our make our project work successfully.

But some names like Mr.G.Spandan sir and Ankita Bose deserve a special thanks for investing their valuable time in guiding us and checking upon to get the best out of us (Roopesh reddy p (321910303006), Mereddy krish reddy (321910303028), Guntupalli Hemanth Kumar (321910303018), Goli Sathya Darshan (321910303001), and Akshay kumar (321910303004)).

|  |  |
| --- | --- |
| **Registration No.** | **Name** |
| 321910303001 | GOLI SATYA DARSHAN |
| 321910303004 | NAGABANDI AKSHAY KUMAR |
| 321910303006 | ROOPESH REDDY P |
| 321910303018 | GUNTUPALLI HEMANTH KUMAR |
| 321910303028 | MEREDDY KRISH REDDY |

**ABSTRACT**

Accurate flight delay prediction is fundamental to establish the more efficient airline business. An important business of airlines is to get customer satisfaction. Their prediction is crucial during the decision-making process for all players of commercial aviation. Due to bad weather, a mechanical reason, and the late arrival of the aircraft to the point of departure, flights delay and lead to customer dissatisfaction. A predictive model of on-time arrival flight is proposed with using flight data and weather data. In this paper, using machine learning models such as Decision Tree Regression, Bayesian Ridge, Random Forest Regression and Gradient Boosting Regression we predict whether the arrival of a particular flight will be delayed or not.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Title** | **Page No.** |
|  | Declaration | I |
| Acknowledgement | II |
| Abstract | III |
| Table of Contents | IV |
|  | List of Figures | V |
| 1 | Introduction | 1 |
| 2 | Literature Survey | 2 |
| 3 | Software and Hardware Specifications | 6 |
| 3.1 Specific Requirements |
| 3.1.1 Functional Requirements |
| 3.1.2 Non-Functional Requirements |
| 3.2 Hardware and Software Requirements | 7 |
| 3.2.1 Hardware Requirements |
| 3.2.2 Software Requirements |
| 4 | Problem Statement | 9 |
| 4.1 Objectives | 10 |
| 5 | Designing | 11 |
| 5.1 Proposed System |
| 5.2 System Architecture |
| 5.3 Methodology | 20 |
| 6 | Implementation | 21 |
| 6.1 Dataset |
| 6.2 Emotion detection | 22 |
| 6.3 Music Recommendation | 27 |
| 6.4 Displaying video feed and playlist |
| 7 | Experimental Results | 39 |
|  | 7.1 Outputs | 39 |
| 8 | Conclusion | 43 |
| 9 | Future Work | 44 |
| 10 | References | 45 |

**LIST OF FIGURES**

**CHAPTER 1**

**INTRODUCTION**

A flight delay is said to occur when an airline lands or takes off later than its scheduled arrival or departure time respectively. AIR traffic load has experienced rapid growth in recent years. The aviation industry around the globe incurs huge losses due to various factors, one of these factors is Airline Delay. Airline delay tends to be onerous for every entity involved i.e., airports, airlines, and passengers. Precise and meticulous prediction of Airline delay using the factors which play prodigious role will be the key to minimize the losses and increase customer satisfaction.

In the United States, the FAA believes that a flight is delayed when the scheduled and actual arrival times differs by more than 15 minutes. Since it becomes a serious problem in the United States, analysis and prediction of flight delays are being studied to reduce large costs. Notable reasons for commercially scheduled flights to delay are adverse weather conditions, air traffic congestion, late reaching aircraft to be used for the flight from previous flight, maintenance, and security issues.

In the paper, several machine learning algorithms have been employed to produce a comparative study with respect to the accuracy of each algorithm. we are using machine learning models such as Decision Tree Regression, Bayesian Ridge, Random Forest Regression and Gradient Boosting Regression we predict whether the arrival of a particular flight will be delayed or not.

**CHAPTER 2**

**LITERATURE REVIEW**

**[1] Chakrabarty, Navoneel. (2019). A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines.**

In the present scenario of domestic flights in USA, there have been numerous instances of flight delays and cancellations. In the United States, the American Airlines, Inc. have been one of the most entrusted and the world's largest airline in terms of number of destinations served. But when it comes to domestic flights, AA has not lived up to the expectations in terms of punctuality or on-time performance. Flight Delays also result in airline companies operating commercial flights to incur huge losses. So, they are trying their best to prevent or avoid Flight Delays and Cancellations by taking certain measures. This study aims at analyzing flight information of US domestic flights operated by American Airlines, covering top 5 busiest airports of US and predicting possible arrival delay of the flight using Data Mining and Machine Learning Approaches. The Gradient Boosting Classifier Model is deployed by training and hyper-parameter tuning it, achieving a maximum accuracy of 85.73%. Such an Intelligent System is very essential in foretelling flights ‘on-time performance.

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

Accurate flight delay prediction is fundamental to establish the more efficient airline business. Recent studies have been focused on applying machine learning methods to predict the flight delay. Most of the previous prediction methods are conducted in a single route or airport. This paper explores a broader scope of factors which may potentially influence the flight delay, and compares several machine learning-based models in designed generalized flight delay prediction tasks. Experimental results show that long short-term memory (LSTM) can handle the obtained aviation sequence data, but overfitting problem occurs in our limited dataset. Compared with the previous schemes, the proposed random forest-based model can obtain higher prediction accuracy (90.2% for the binary classification) and can overcome the overfitting problem.

**[3] Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5.**

As the computer technology and computer network technology are developing, the amount of data in information industry is getting higher and higher. It is necessary to analyze this large amount of data and extract useful knowledge from it. Process of extracting the useful knowledge from huge set of incomplete, noisy, fuzzy and random data is called data mining. Decision tree classification technique is one of the most popular data mining techniques. In decision tree divide and conquer technique is used as basic learning strategy. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. This paper focus on the various algorithms of Decision tree (ID3, C4.5, CART), their characteristic, challenges, advantage and disadvantage.

**Summary:** In this paper, we learn about Decision Tree, types of Decision tree (ID3, C4.5, CART etc). It also discusses about the advantages and disadvantages of Decision Tree.

**[4] Friedman, Jerome. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis. 38. 367-378. 10.1016/S0167-9473(01)00065-2.**

Gradient boosting constructs additive regression models by sequentially fitting a simple parameterized function (base learner) to current “pseudo”-residuals by least squares at each iteration. The pseudo-residuals are the gradient of the loss functional being minimized, with respect to the model values at each training data point evaluated at the current step. It is shown that both the approximation accuracy and execution speed of gradient boosting can be substantially improved by incorporating randomization into the procedure. Specifically, at each iteration a subsample of the training data is drawn at random (without replacement) from the full training data set. This randomly selected subsample is then used in place of the full sample to fit the base learner and compute the model update for the current iteration. This randomized approach also increases robustness against overcapacity of the base learner.

**Summary:** In this paper, we learn about Stochastic Gradient Boosting.

**CHAPTER 3**

**SOFTWARE AND HARDWARE SPECIFICATIONS**

**3.1 Specific Requirements**

**3.1.1 Functional Requirements:**

* Functional requirements outline the functions of computer code systems. The system's behavior is evaluated with specific inputs or conditions, calculations, knowledge manipulation and process, and specific functionality.

**3.1.2 Non-Functional Requirements:**

* Nonfunctional requirements illustrate the behavior of a system and the constraints on its functionality. Also called system quality features.
  + Reliability
  + Maintainability
  + Performance
  + Scalability
  + Flexibility
  + Usability

**3.2 Hardware and Software Requirements:**

**3.2.1 Hardware Requirements:**

# Processor - I3/Intel Processor

# RAM - 4GB (min)

# Hard Disk - 128 GB

# Key Board - Standard Windows Keyboard

# Mouse - Two or Three Button Mouse

# Monitor – Any

* + 1. **Software Requirements:**
* **Operating System:** Windows 7+
* **Server-side Script:** Python 3.6+
* **IDE:** PyCharm
* **Libraries Used:** Pandas, Numpy, Sci-Kit Learn, Matplotlib, Seaborn, Flask.
* **Dataset :** 2015 FAA Flight dataset.
* **Pandas:** Pandas provide us with many Series and DataFrames. It allows you to easily organize, explore, represent, and manipulate data. Pandas has some special features that allow you to handle missing data or value with a proper measure.
* **Numpy:** Numpy provides masked arrays along with general array objects. It also comes with functionalities such as manipulation of logical shapes, discrete Fourier transform, general linear algebra, and many more.
* **Matplotlib:** Matplotlib can create such quality figures that are really good for publication. Figures you create with Matplotlib are available in hardcopy formats across different interactive platforms.
* **Seaborn:** Seaborn is built on top of Python’s core visualization library Matplotlib. It is meant to serve as a complement, and not a replacement. However, Seaborn comes with some very important features.
* **Scikit-Learn:** Scikit Learn offers easy methods for data representation. Whether you want to present data as a table or matrix, it is all possible with Scikit Learn.
* **Flask Framework:** Python's Flask web framework provides a lightweight and flexible way to develop a wide range of web applications, from simple APIs to full-featured applications.

**CHAPTER 4**

**PROBLEM STATEMENT**

* A flight delay is said to occur when an airline lands or takes off later than its scheduled arrival or departure time respectively.
* These delays lead to customer dissatisfaction and reduce profit of airlines.
* How to tackle these problems?

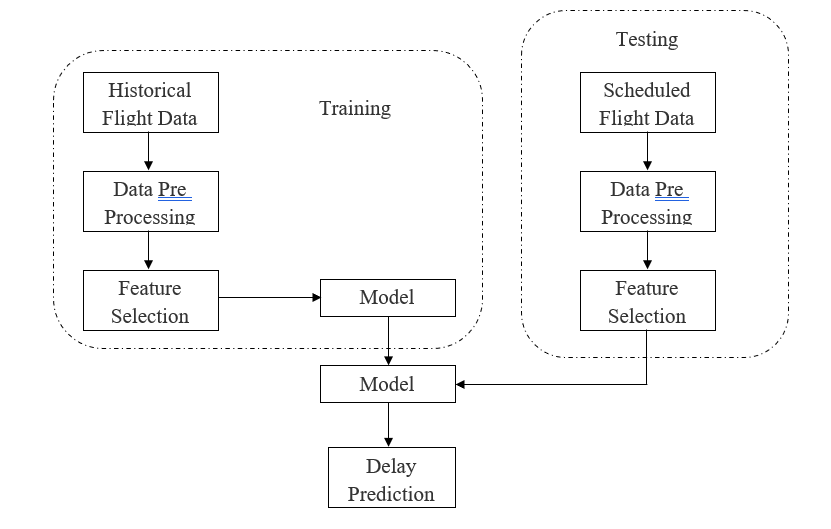
**4.1 OBJECTIVES**

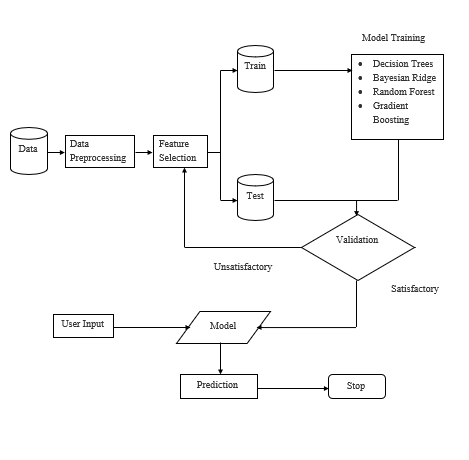
* In the view of the problem statement, we feel that providing a error detection model for detecting the flight delays.
  1. First, we structure the unstructured data.
  2. Then we segregate the data.
  3. Then, apply different algorithms to find the delays precisely.

**CHAPTER 5**

**DESIGNING**

* 1. **PROPOSED SYSTEM**

****Accurate flight delay prediction is fundamental to establish the more efficient airline business. An important business of airlines is to get customer satisfaction. The existing methods requires highly skilled people and hence is costly to implement as it requires manually selecting features for prediction. In this paper, using machine learning models such as Decision Tree Regression, Bayesian Ridge, Random Forest Regression and Gradient Boosting Regression we predict whether the arrival of a particular flight will be delayed or not.

* 1. ** SYSTEM ARCHITECTURE**

**DATA**

**Data Preprocessing:**

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. Cleaning the data refers to removing the null values, filling the null values with meaningful value, removing duplicate values, removing outliers, removing unwanted attributes. If dataset contains any categorical records means convert those categorical variables to numerical values

Here, we are removing rows with null values, selecting features.

**Encoding Dataset:**

The dataset is separated into predictors and Target variables. We perform Ordinal Encoding on the predictors and Scale them using StandardScaler.

**Model Training:**

The model is selected by the user and then the data is split into two parts namely testing and training dataset.

The models:

* **Decision Tree:**
  + A Decision tree is a flowchart like tree structure.
  + Each internal node denotes a test on an attribute.
  + Each branch represents an outcome of the test.
  + Each leaf node (terminal node) holds a class label.
  + An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute.
  + Gini method or Information Gain (entropy) is used for attribute selection.
* **Bayesian Ridge Regression:**
  + Bayesian linear regression is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference.
  + Bayesian regression allows a natural mechanism to survive insufficient data or poorly distributed data by formulating linear regression using probability distributors rather than point estimates.
  + One of the most useful types of Bayesian regression is Bayesian Ridge regression which estimates a probabilistic model of the regression problem.
* **Random Forest Regression:**
  + Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low.
  + In the case of a classification problem, the final output is taken by using the majority voting classifier.
  + In the case of a regression problem, the final output is the mean of all the outputs.
  + A Random Forest is an ensemble technique capable of performing both regression and classification tasks.
* **Gradient Boosting Regression:**
  + Boosting is a method of converting weak learners into strong learners.
  + Gradient boosting is a type of machine learning boosting.
  + It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error.
  + The key idea is to set the target outcomes for this next model in order to minimize the error.

**Tabulation**

The system takes input from the user about the target variable and displays all the 5-evaluation metrics for all the algorithms of that target variable.

**Visualization**

The system takes a particular evaluation metric as an input from the user and generates a grouped bar plot for all different models with 2 different target variables.

**USER**

**Upload Data:**

The user uploads a .csv dataset from the web application which contains FAA flight records.

**View Data:**

The user views the data in the web application after it is cleaned.

**Model Testing:**

The user tests all possible models trained by the system using the testing dataset to view all 5-evaluation metrics in a table.

**Visualization:**

The user selects a particular metric to view a comparative graph of all the algorithms in a grouped bar plot with both the target variables.

**ALGORITHM**

**Decision Tree:**

Decision trees are non-parametric supervised learning Method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn’t split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.

**Bayesian Ridge Regression:**

Ridge Regression is the name usually given to Linear Regression with an L2 regularizer. The regularizer penalizes model complexity by adding the sum of the parameter squares to the error function. You can get there using Maximum Likelihood estimation on a Gaussian likelihood model and then applying the rationale of structural risk minimization (think of an SVM).

From the Bayesian side of things, if you start with your Gaussian likelihood model, a Gaussian prior on the model parameters with mean zero and standard deviation 1, and then apply the Bayes Rule to find the posterior distribution of the model parameters given your dataset, you will find that said posterior distribution is also a Gaussian whose mean is equivalent to the Ridge Regression estimate of the model coefficients. This is known as the Maximum A Posteriori estimate of the regression model.

**Random Forest Regression:**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

**Gradient Boosting Regression:**

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

The idea of gradient boosting originated in the observation that boosting can be interpreted as an optimization algorithm on a suitable cost function. Explicit regression gradient boosting algorithms were subsequently developed simultaneously with the more general functional gradient boosting.

The boosting can be viewed as iterative functional gradient descent algorithms. That is, algorithms that optimize a cost function over function space by iteratively choosing a function (weak hypothesis) that points in the negative gradient direction. This functional gradient view of boosting has led to the development of boosting algorithms in many areas of machine learning and statistics beyond regression and classification.

**5.3 METHODOLOGY**

**Use-Case Diagram**

**Sequence Diagram**



**CHAPTER 6**

**6.1 IMPLEMENTATION**

Installing Python:

1. To download and install Python visit the official website of Python <https://www.python.org/downloads/> and choose your version.
2. Once the download is complete, run the exe for install Python. Now click on Install Now.
3. You can see Python installing at this point.
4. When it finishes, you can see a screen that says the Setup was successful. Now click on "Close".

Installing PyCharm:

1. To download PyCharm visit the website <https://www.jetbrains.com/pycharm/download/> and click the "DOWNLOAD" link under the Community Section.



1. Once the download is complete, run the exe for install PyCharm. The setup wizard should have started. Click “Next.”
2. On the next screen, Change the installation path if required. Click “Next.”
3. On the next screen, you can create a desktop shortcut if you want and click on “Next.”
4. Choose the start menu folder. Keep selected JetBrains and click on “Install.”
5. Wait for the installation to finish.
6. Once installation finished, you should receive a message screen that PyCharm is installed. If you want to go ahead and run it, click the “Run PyCharm Community Edition” box first and click “Finish.”
7. After you click on "Finish," the Following screen will appear.



9. You need to install some packages to execute your project in a proper way.

10. Open the command prompt/ anaconda prompt or terminal as administrator.

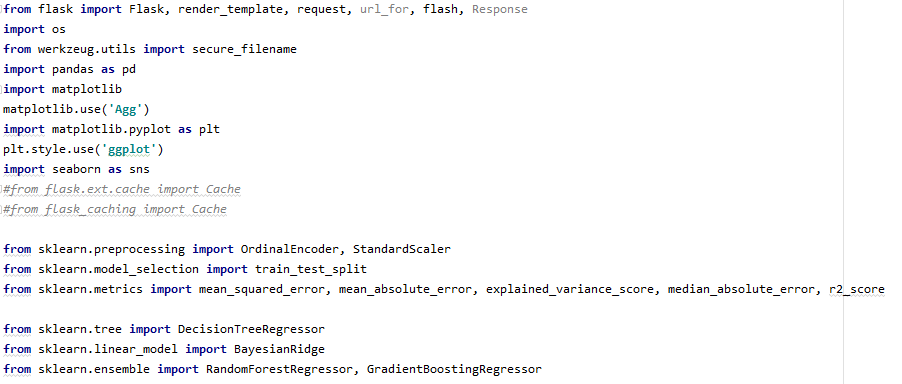
11. The prompt will get open, with specified path, type “pip install package name” which you want to install (like numpy, pandas, seaborn, scikit-learn, matplotlib.pyplot)

Ex: pip install numpy

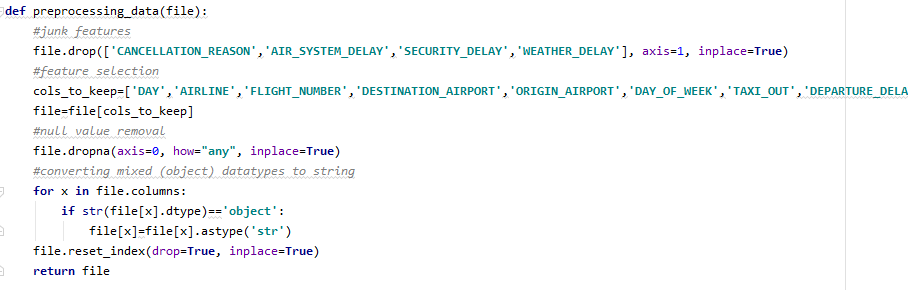
1. Import all the Libraries/packages.
2. Load the FAA Flight dataset.
3. Perform Exploratory data analysis.
4. Preprocess the datasets.
5. Remove the Null values.
6. Check for unbalanced data.
7. Scale the datasets.
8. Split the dataset.
9. Train all datasets on all classification algorithms mentioned below and record their accuracies using both Arrival Delay and Departure Delay target variable seperately.
10. The algorithms are:
    1. Decision Trees.
    2. Bayesian Ridge.
    3. Random Forest.
    4. Gradient Boosting.
11. The flight delay is predicted using the test dataset by the user.
12. The evaluation metrics used are Mean Squared error (MSE), Mean Absolute Error (MAE), Explained Variance Score, Median Absolute Error, R2 Score.
13. The evaluation metrics for different models and different target variables are tabulated and visualized and displayed in the web application.

**CODE SNIPPTS:**

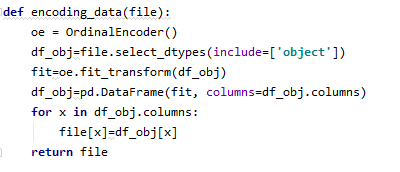
Imports:



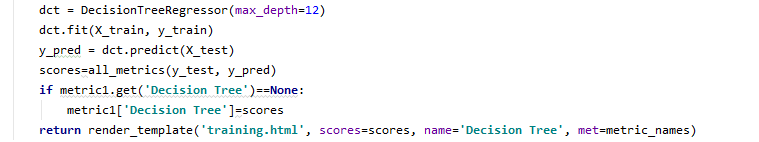
PreProcessing:



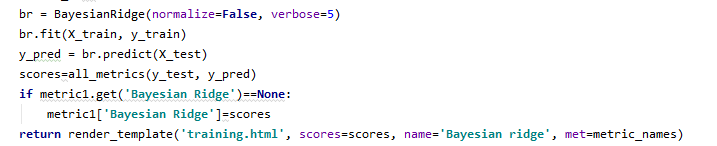
Encoding:



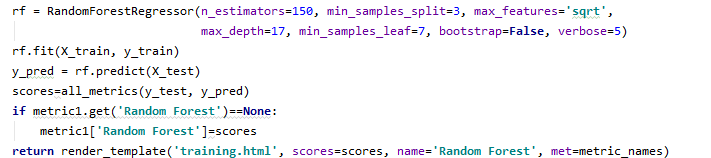
Decision Tree:



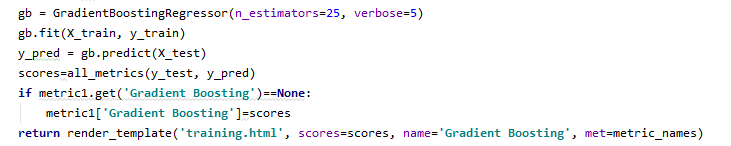
Bayesian Ridge:



Random Forest:



Gradient Boosting:



**6.2 SYSTEM STUDY**

**FEASIBILITY STUDY**

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

* Economical Feasibility
* Technical Feasibility
* Social Feasibility

**ECONOMICAL FEASIBILITY**

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

### **TECHNICAL FEASIBILITY**

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

**SOCIAL FEASIBILITY**

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

**6.3 SYSTEM TESTING**

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

**White Box Testing**

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

**Black Box Testing**

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

**6.3.1 Unit Testing:**

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

**Test strategy and approach**

Field testing will be performed manually and functional tests will be written in detail.

**Test objectives**

* All field entries must work properly.
* Pages must be activated from the identified link.
* The entry screen, messages and responses must not be delayed.

**Features to be tested**

* Verify that the entries are of the correct format
* No duplicate entries should be allowed
* All links should take the user to the correct page.

# 6.3.2 Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

**Test Results:** All the test cases mentioned above passed successfully. No defects encountered.

**6.3.3 Acceptance Testing**

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

**Test Results:** All the test cases mentioned above passed successfully. No defects encountered.

**CHAPTER 7**

**7.1 EXPERIMENTAL RESULTS**

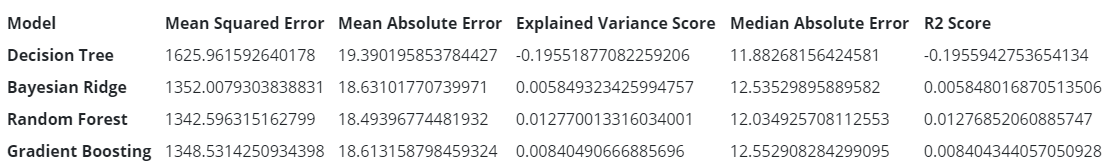
|  |  |  |
| --- | --- | --- |
| **Input** | **Output** | **Result** |
| Input | Tested for different model given by user on the different model. | Success |
| Random Forest | Tested for different input given by the user on different models are created using the different algorithms and data. | Success |
| Prediction | Prediction will be performed using the different models build from the algorithms. | Success |

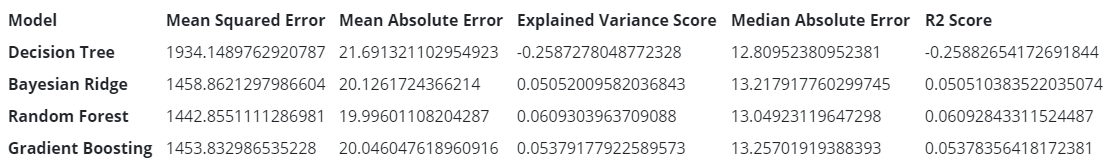
**Test cases Model building**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.NO** | **Test cases** | **I/O** | **Expected O/T** | **Actual O/T** | **P/F** |
| 1 | Read the datasets. | Dataset’s path. | Datasets need to read successfully. | Datasets fetched successfully. | It produced P. If this not F will come |
| 2 | Verifying the Heart  Disease identify stroke are not. | Input for heart  disease  classification | Output as either in the form heart  Disease or stroke are not. | Output is classified as heart  disease | It produced P. If this is not, it will undergo F |
| 3 | Verifying the heart  Disease identify stroke are not. | Input for heart  disease classification | Output as either in the form heart  Disease or stroke are not. | Output is classified as stroke are not. | It produced P. If this is not, it will undergo F |
| 4 | Verifying the heart  Disease identify stroke are not. | Input heart  disease for prediction the stroke | Need to predict the best accuracy | Model successfully predicted best accuracy | It produced P. If this is not, it will undergo F |

**7.2 OUTPUTS**

**The performance for all models are shown below:**

 For Departure Delay as Target variable:

 For Arrival Delay as Target variable:

**CHAPTER 8**

**CONCLUSION**

In this application, we have preprocessed the data by removing the null values and encoding all the variables. we have also scaled all the predictor variables.

We have used Decision Tree, Bayesian Ridge, Random Forest and Gradient Boosting regression.

The best model was the Random Forest (by a small margin) model with Hyper parameters tuning. The dataset used was the 2015 FAA Flight dataset.

**CHAPTER 9**

**FUTURE SCOPE**

We should consider flight delay prediction using boosting techniques like XgBoost which involves extreme gradient boosting. We may also model a neural network which are high in complexities but offers higher accuracy and automation of feature selection.

**CHAPTER 10**

**REFERENCES**

[1] Chakrabarty, Navoneel. (2019). A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines.

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

[3] Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5.

[4] Friedman, Jerome. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis. 38. 367-378. 10.1016/S0167-9473(01)00065-2.

[5] N. G. Rupp, "Further Investigation into the Causes of Flight Delays," in Department of Economics, East Carolina University, 2007.

[6] Navoneel, et al., Chakrabarty, "Flight Arrival Delay Prediction Using Gradient Boosting Classifier," in Emerging Technologies in Data Mining and Information Security, Singapore, 2019.

[7] A. M. Kalliguddi, Area K., Leboulluec, "Predictive Modelling of Aircraft Flight Delay," Universal Journal of Management, pp. 485 - 491, 2017.

[8] Noriko, Etani, "Development of a predictive model for on-time arrival fight of airliner by discovering correlation between fight and weather data," 2019.

[9] Chakrabarty, Navoneel. “A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines.” 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON) (2019): 102-107.

[10] Sternberg, Alice & Soares, Jorge & Carvalho, Diego & Ogasawara, Eduardo. (2017). A Review on Flight Delay Prediction.

[11] V. Venkatesh, A. Arya, P. Agarwal, S. Lakshmi and S. Balana, "Iterative machine and deep learning approach for aviation delay prediction," 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, 2017, pp. 562-567, doi: 10.1109/UPCON.2017.8251111.

[12] Yogita Borse , Dhruvin Jain , Shreyash Sharma , Viral Vora, Aakash Zaveri, 2020, Flight Delay Prediction System, International Journal Of Engineering Research & Technology (IJERT) Volume 09, Issue 03 (March 2020).