# Data Mining Application for CH-47D Aft Swashplate Bearing Fault Detection

**Team Members:**

**Surya Selvaraj - 679003533**

**Akash Niruban Narayanan - 662268909**

## INTRODUCTION:

In October 2002, a swashplate bearing failed in the aft rotor head of a CH-47D during a ground run. The post-accident investigation determined that failure of the cage of the duplex ball bearing between the rotating and non-rotating swashplates caused the accident. For the incident, it is suspected that one end of the cage was displaced out from its position between the races and un-caged the balls, eventually resulting in a bearing failure. In the analysis**,** four faulted bearings were tested on a special test rig where vibration measurements were acquired at several simulated flight conditions and at a steady flight condition over a 24-hour run. The four chosen faulted bearings were: (a) corroded, (b) spalled, (c) popped (raised) cage, and (d) overlapped cage. The main goal of the project is to apply data mining approach to detect the bearing faults (separate one type of bearing from others) using the minimum number of features (inputs).

## ALGORITHM OUTPUT:

### K-NN Algorithm – CONSIDERING THE FULL DATA SET

Data Points (6 specimens) – 120

# Of Features – 255

Data set partitioning (%) – 70 : 30 ( Training : validation)

**K = 10**

**MODEL ACCURACY = 100 %**

We achieved a 100 % accuracy when the full data set were considered **using K-NN, Naïve Bayes and Logistic Regression.** We proceeded with the K-NN algorithm for our further analysis.

## MODEL FITTING OUTPUT ON THE TEST DATA: RESULTS

We used K-NN algorithm with K=10 on the test data (**full Data set**) to get the following predictions.

| UNKNOWN SPECIMENS | PREDICTED SPECIMENS |
|---|---|
| T1 | S6 |
| T2 | S6 |
| T3 | S6 |
| T4 | S6 |
| T5 | S6 |
| T6 | S6 |
| T7 | S6 |
| T8 | S6 |
| T9 | S6 |
| T10 | S6 |
| T11 | S5 |
| T12 | S5 |
| T13 | S5 |
| T14 | S5 |
| T15 | S5 |
| T16 | S5 |
| T17 | S5 |
| T18 | S5 |
| T19 | S5 |
| T20 | S5 |
| T21 | S4 |
| T22 | S4 |
| T23 | S4 |
| T24 | S4 |
| T25 | S4 |
| T26 | S4 |
| T27 | S4 |
| T28 | S4 |
| T29 | S4 |
| T30 | S4 |
| T31 | S3 |
| T32 | S3 |
| T33 | S3 |
| T34 | S3 |
| T35 | S3 |
| T36 | S3 |
| T37 | S3 |
| T38 | S3 |

| | |
|---|---|
| T39 | S3 |
| T40 | S3 |
| T41 | S2 |
| T42 | S2 |
| T43 | S2 |
| T44 | S2 |
| T45 | S2 |
| T46 | S2 |
| T47 | S2 |
| T48 | S2 |
| T49 | S2 |
| T50 | S2 |
| T51 | S1 |
| T52 | S1 |
| T53 | S1 |
| T54 | S1 |
| T55 | S1 |
| T56 | S1 |
| T57 | S1 |
| T58 | S1 |
| T59 | S1 |
| T60 | S1 |

**SCREEN SHOT OF THE OUTPUT PAGE:**

```
mydata (120, 256)
mydata_data (120, 255)
mydata_target(120,)
train (84, 256)
test (36, 256)
train_data (84, 255)
train_target (84,)
test_data (36, 255)
test_target (36,)
['S6' 'S6' 'S6' 'S6' 'S6' 'S6' 'S6' 'S6' 'S6' 'S6' 'S5' 'S5' 'S5' 'S5' 'S5'
 'S5' 'S5' 'S5' 'S5' 'S5' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4'
 'S3' 'S3' 'S3' 'S3' 'S3' 'S3' 'S3' 'S3' 'S3' 'S3' 'S2' 'S2' 'S2' 'S2' 'S2'
 'S2' 'S2' 'S2' 'S2' 'S2' 'S1' 'S1' 'S1' 'S1' 'S1' 'S1' 'S1' 'S1' 'S1' 'S1']
Predicted model accuracy: 100.0%
```

**FEATURE SELECTION:**

Main goal of feature selection is to select the main features out of initial 255 features and try to predict the unknown test specimens with the given 6 specimens with minimum # of features

**Method used for feature selection** – ExtraTreesClassifier (uses F-Test to show the features importance on the data set)

## # OF FEATURES SIGNIFICANT FROM THE ANALYSIS - 53

| RANKING | SELECTED FEATURE | IMPORTANCE SCORE |
|---|---|---|
| 1 | 49 | 0.04 |
| 2 | 111 | 0.04 |
| 3 | 135 | 0.04 |
| 4 | 147 | 0.04 |
| 5 | 61 | 0.031077 |
| 6 | 30 | 0.0238 |
| 7 | 153 | 0.0218 |
| 8 | 180 | 0.02135 |
| 9 | 60 | 0.02 |
| 10 | 55 | 0.02 |
| 11 | 1 | 0.02 |
| 12 | 254 | 0.02 |
| 13 | 52 | 0.02 |
| 14 | 208 | 0.02 |
| 15 | 178 | 0.02 |
| 16 | 65 | 0.02 |
| 17 | 53 | 0.02 |
| 18 | 71 | 0.02 |
| 19 | 73 | 0.02 |
| 20 | 22 | 0.02 |
| 21 | 203 | 0.02 |
| 22 | 90 | 0.02 |
| 23 | 117 | 0.02 |
| 24 | 121 | 0.02 |
| 25 | 235 | 0.02 |
| 26 | 122 | 0.02 |
| 27 | 218 | 0.02 |
| 28 | 106 | 0.02 |
| 29 | 172 | 0.02 |
| 30 | 168 | 0.02 |
| 31 | 138 | 0.02 |
| 32 | 86 | 0.02 |
| 33 | 174 | 0.02 |
| 34 | 129 | 0.02 |
| 35 | 249 | 0.019559 |

| | | |
|---|---|---|
| 36 | 98 | 0.0195 |
| 37 | 87 | 0.018969 |
| 38 | 222 | 0.01853 |
| 39 | 14 | 0.0185 |
| 40 | 159 | 0.0185 |
| 41 | 130 | 0.0177 |
| 42 | 110 | 0.0162 |
| 43 | 242 | 0.0159 |
| 44 | 216 | 0.0102 |
| 45 | 228 | 0.0098 |
| 46 | 84 | 0.0097 |
| 47 | 185 | 0.0077 |
| 48 | 80 | 0.0052 |
| 49 | 85 | 0.0051 |
| 50 | 20 | 0.00357 |
| 51 | 81 | 0.00342 |
| 52 | 173 | 0.0019 |
| 53 | 184 | 0.00152 |

Therefore, total number of features that are significant from the ExtraTreesClassifier (F-Test) is **53.**

**ALGORITHM OUTPUT – (Only for the selected features)**

**K-NN Algorithm – CONSIDERING THE 53 SIGNIFICANT FEATURES**

Data Points (6 specimens) – 120

# Of Features – 53

Data set partitioning (%) – 70 : 30 ( Training : validation)

**K = 10**

**MODEL ACCURACY = 100 %**

**MODEL FITTING OUTPUT ON THE TEST DATA: RESULTS**

We used K-NN algorithm with K=10 on the test data (**for selected features**) to get the following predictions.

| UNKNOWN SPECIMENS | PREDICTED SPECIMENS (53 FEATURES CONSIDERED) |
|---|---|
| T1 | S6 |
| T2 | S6 |
| T3 | S4 |
| T4 | S4 |
| T5 | S6 |
| T6 | S4 |
| T7 | S6 |
| T8 | S6 |
| T9 | S4 |
| T10 | S6 |
| T11 | S5 |
| T12 | S5 |
| T13 | S5 |
| T14 | S5 |
| T15 | S5 |
| T16 | S5 |
| T17 | S5 |
| T18 | S5 |
| T19 | S3 |
| T20 | S3 |
| T21 | S4 |
| T22 | S4 |
| T23 | S4 |
| T24 | S4 |
| T25 | S4 |
| T26 | S4 |
| T27 | S4 |
| T28 | S4 |
| T29 | S4 |
| T30 | S4 |
| T31 | S3 |
| T32 | S3 |
| T33 | S3 |
| T34 | S3 |
| T35 | S3 |
| T36 | S3 |
| T37 | S3 |

| | |
|---|---|
| T38 | S3 |
| T39 | S3 |
| T40 | S3 |
| T41 | S2 |
| T42 | S2 |
| T43 | S2 |
| T44 | S2 |
| T45 | S2 |
| T46 | S2 |
| T47 | S2 |
| T48 | S2 |
| T49 | S2 |
| T50 | S2 |
| T51 | S4 |
| T52 | S4 |
| T53 | S4 |
| T54 | S4 |
| T55 | S4 |
| T56 | S4 |
| T57 | S4 |
| T58 | S4 |
| T59 | S4 |
| T60 | S1 |

**SCREEN SHOT OF THE OUTPUT PAGE:**

```
mydata (120, 56)
mydata_data (120, 55)
mydata_target(120,)
train (84, 56)
test (36, 56)
train_data (84, 55)
train_target (84,)
test_data (36, 55)
test_target (36,)
['S6' 'S6' 'S4' 'S4' 'S6' 'S4' 'S6' 'S6' 'S4' 'S6' 'S5' 'S5' 'S5' 'S5' 'S5'
 'S5' 'S5' 'S5' 'S3' 'S3' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4'
 'S3' 'S3' 'S3' 'S3' 'S3' 'S3' 'S3' 'S3' 'S3' 'S3' 'S2' 'S2' 'S2' 'S2' 'S2'
 'S2' 'S2' 'S2' 'S2' 'S2' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S4' 'S1']
Predicted model accuracy: 100.0%
```

**CONCLUSION:**

- The raw data was fitted using three models - KNN, Logistic Regression and Naive Bayes. The accuracy was 100%
- Since the accuracy of any predictive model cannot be 100% (Accuracy Paradox - An accuracy of 100% is considered harmful), we resorted to Feature Selection by Feature Importance using ExtraTreesClassifier and RandomForestClassifier. It was observed to be difficult to identify an optimal number of estimators for the RandomForestClassifier (no change in accuracy); hence the choice of ExtraTreesClassifier over RandomForestClassifier
- After multiple runs, the feature importance was narrowed down to a range of 51-55. On an average, 53 features were found to be the most significant ones
- The data with only the most significant 53 features was fitted in the same three models It was found that **the results of KNN and Logistic Regression was quite similar** when compared to Naive Bayes. But the accuracy was still 100%
- However, the outputs of KNN and Logistic Regression can be considered as the most closest prediction since both the models gave similar output
- But an **accuracy of 100%** implies that the **model was an overfit due to a smaller dataset**
- Implementing neural networks for such a small dataset would not guarantee proper results. Neural networks work best on large data, so that we can identify the number of optimal layers for prediction. Since these models gave 100% accuracy, it clearly proves that the dataset is not really sufficient to make guaranteed predictions