Predicting Time-to-Failure of Industrial Machines with Temporal Data Mining

Jean Nakamura

A dissertation submitted in partial fulfillment of the requirement for the degree of

Masters of Science

University of Washington

2007

Committee:
Committee Chair: Isabelle Bichindaritz, Ph.D.
Committee Member: Don McLane

Program Authorized to Offer Degree: Institute of Technology - Tacoma
The purpose of this project is to perform analysis of temporal vibration data results to predict the time until a machine failure. The project performs temporal data mining, which is a method of choice to predict future events based on known past events. The difficulty in determining time-to-failure (TTF) of industrial machines is that the failure mode is not a linear progression. The progression of a severity of a fault increases at a higher rate as the machine approaches failure. Through experience, it is known that discrete frequencies in the vibration spectra are associated with machine faults and will reach expected amplitudes at the point of machine failure. Currently, there are no commercially available tools to predict machine time to failure. This project determined that it is possible to analyze a machine’s temporal vibration data results to produce an estimated time to a failure based on the progression of identified faults when there is a repetition of instances with good accuracy (80%) using discretized data, but not using raw continuous data. Results of the data mining project are provided to support this conclusion.
# TABLE OF CONTENTS

Chapter 1  INTRODUCTION .................................................................................................................. 1

Chapter 2  BACKGROUND INFORMATION ......................................................................................... 3
  2.1  Machinery Maintenance ............................................................................................................ 3
  2.2  Machine Condition Analysis software ..................................................................................... 3
  2.3  Data Mining ............................................................................................................................. 4
  2.4  Case-Based Reasoning .............................................................................................................. 4

Chapter 3  PROBLEM STATEMENT .................................................................................................... 6

Chapter 4  DESCRIPTION OF DATA ................................................................................................ 7

Chapter 5  DATA PRE-PROCESSING ................................................................................................. 10
  5.1  Invalid Data ............................................................................................................................ 10
  5.2  Grouping ................................................................................................................................. 11
  5.3  Pre-Processing Results ............................................................................................................ 12

Chapter 6  TIME TO FAILURE SYSTEM ............................................................................................ 14
  6.1  Case-Based Reasoning System ............................................................................................... 14
  6.2  Generalization ........................................................................................................................ 15
  6.3  CBR – retaining ....................................................................................................................... 16
  6.4  CBR – retrieval ......................................................................................................................... 16
  6.5  CBR – reusing ........................................................................................................................ 17
  6.6  CBR – revising ........................................................................................................................ 18

Chapter 7  ANALYSIS ........................................................................................................................ 19
  7.1  CBR ......................................................................................................................................... 19
  7.2  SPSS Linear Regression .......................................................................................................... 21
  7.3  Weighting Parameters .............................................................................................................. 25
  7.4  WEKA Algorithms and dataset ............................................................................................ 28
7.5 WEKA data discretized into 3 groups ................................................................. 29
7.6 WEKA data discretized into 3 groups with Sev and Diff only ......................... 30
7.7 WEKA data discretized into 6 groups ................................................................. 33
7.8 Progression of Severity .................................................................................... 34
7.9 WEKA Associations - data discretized into 3 groups ....................................... 37
7.10 Summary of Results ....................................................................................... 39

Chapter 8 DISCUSSION ......................................................................................... 41

Chapter 9 FUTURE WORK .................................................................................. 42

Chapter 10 EDUCATIONAL STATEMENT ......................................................... 43

Chapter 11 CONCLUSION .................................................................................... 44

BIBLIOGRAPHY ................................................................................................... 45

Appendix A DATABASE ....................................................................................... 47
    A.1 Partial Database Schema ............................................................................. 47

Appendix B PRE-PROCESSING TOOL ................................................................. 48
    B.1 MID Grouping ............................................................................................ 48
    B.2 Diagnosis Grouping .................................................................................. 48
    B.3 Selection .................................................................................................... 50

Appendix C TTF SYSTEM .................................................................................... 51
    C.1 VibrationCase – raw data (partial) ............................................................ 51
    C.2 VibrationCaseTest – raw data ................................................................. 52
    C.3 Case Evaluation ....................................................................................... 53
    C.4 Case Evaluation Distribution .................................................................. 55
    C.5 Training Results ...................................................................................... 56
    C.6 Test knowing TTF .................................................................................... 56
    C.7 Test knowing TTF-Case selection breakdown ........................................ 57
    C.8 Initial test ................................................................................................. 57
C.9 Initial test – case selection breakdown................................................................. 58
C.10 2-Nearest Neighbor ............................................................................................. 59
C.11 3-Nearest Neighbor ............................................................................................. 59

Appendix D TTF USER INTERFACE ........................................................................... 60
D.1 Determining TTF User Interface .......................................................................... 60
D.2 Machine fault trend ............................................................................................... 61
D.3 Case Library .......................................................................................................... 62

Appendix E Power Point Presentation ......................................................................... 63
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Case base program flow</td>
<td>15</td>
</tr>
<tr>
<td>7.1</td>
<td>Increasing number of cases in library</td>
<td>21</td>
</tr>
<tr>
<td>7.2</td>
<td>Severity distribution graph</td>
<td>35</td>
</tr>
<tr>
<td>7.3</td>
<td>Severity progression graph</td>
<td>36</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>VibrationCase table</td>
<td>8</td>
</tr>
<tr>
<td>4.2</td>
<td>VibrationCaseTest table</td>
<td>8</td>
</tr>
<tr>
<td>4.3</td>
<td>Case example</td>
<td>9</td>
</tr>
<tr>
<td>5.1</td>
<td>DiagnosisGroups table</td>
<td>11</td>
</tr>
<tr>
<td>5.2</td>
<td>Diagnosis group example</td>
<td>11</td>
</tr>
<tr>
<td>5.3</td>
<td>VibrationStandardEquipmentGroups table</td>
<td>12</td>
</tr>
<tr>
<td>5.4</td>
<td>MID group example</td>
<td>12</td>
</tr>
<tr>
<td>5.5</td>
<td>Pre-processing results</td>
<td>13</td>
</tr>
<tr>
<td>6.1</td>
<td>Case base logic</td>
<td>16</td>
</tr>
<tr>
<td>6.2</td>
<td>Case base retrieval logic</td>
<td>17</td>
</tr>
<tr>
<td>7.1</td>
<td>Case Library</td>
<td>19</td>
</tr>
<tr>
<td>7.2</td>
<td>SPSS Correlations</td>
<td>22</td>
</tr>
<tr>
<td>7.3</td>
<td>SPSS Model Summary</td>
<td>23</td>
</tr>
<tr>
<td>7.4</td>
<td>SPSS ANOVA</td>
<td>23</td>
</tr>
<tr>
<td>7.5</td>
<td>SPSS Coefficients</td>
<td>24</td>
</tr>
<tr>
<td>7.6</td>
<td>Result summary</td>
<td>25</td>
</tr>
<tr>
<td>7.7</td>
<td>Weighting summary</td>
<td>26</td>
</tr>
<tr>
<td>7.8</td>
<td>SPSS Multistep regression – See predictors</td>
<td>27</td>
</tr>
<tr>
<td>7.9</td>
<td>Three group data distribution</td>
<td>29</td>
</tr>
<tr>
<td>7.10</td>
<td>J48 tree – three groups</td>
<td>29</td>
</tr>
<tr>
<td>7.11</td>
<td>Multilayer Perceptron – three groups</td>
<td>30</td>
</tr>
<tr>
<td>7.12</td>
<td>Logistic – three groups</td>
<td>30</td>
</tr>
<tr>
<td>7.13</td>
<td>K* - three groups</td>
<td>30</td>
</tr>
<tr>
<td>7.14</td>
<td>J48 tree – Sev and Diff only</td>
<td>31</td>
</tr>
<tr>
<td>7.15</td>
<td>Multilayer Perceptron – Sev and Diff only</td>
<td>31</td>
</tr>
<tr>
<td>7.16</td>
<td>Multilayer Perceptron – Sev and Diff only – modified options</td>
<td>32</td>
</tr>
<tr>
<td>7.17</td>
<td>Logistic – Sev and Diff only</td>
<td>32</td>
</tr>
<tr>
<td>7.18</td>
<td>K* - Sev and Diff only</td>
<td>32</td>
</tr>
<tr>
<td>Table Number</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>7.19</td>
<td>ID3 tree – Sev and Diff only</td>
<td>33</td>
</tr>
<tr>
<td>7.20</td>
<td>Six group data distribution</td>
<td>33</td>
</tr>
<tr>
<td>7.21</td>
<td>J48 tree – six groups</td>
<td>34</td>
</tr>
<tr>
<td>7.22</td>
<td>Severity distribution</td>
<td>34</td>
</tr>
<tr>
<td>7.23</td>
<td>Severity progression</td>
<td>35</td>
</tr>
<tr>
<td>7.24</td>
<td>J48 tree – progression</td>
<td>36</td>
</tr>
<tr>
<td>7.25</td>
<td>Multilayer Perceptron – modified options, progression</td>
<td>37</td>
</tr>
<tr>
<td>7.26</td>
<td>Logistic – progression</td>
<td>37</td>
</tr>
<tr>
<td>7.27</td>
<td>K* - progression</td>
<td>37</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

My gratitude extends to those at DLI Engineering for their professional experience and knowledge and to DLI Engineering for providing a database and engineering expertise for this project. Special thanks go to Dr. Isabelle Bichindaritz for guiding and assisting me throughout the project. Her knowledge and commitment were essential in putting this project together.
Chapter 1
INTRODUCTION

Machine Condition Analysis software is a predictive maintenance tool that analyzes vibration data. The process includes the collection of vibration data, the analysis of data resulting in generating faults, and the prescription of repair recommendations to avoid machine failure. “A fault is an abnormal state of a machine or system such as dysfunction or malfunction of a part, an assembly, or the whole system” [9]. Each fault is associated with a severity indicating its degree of seriousness. An increase in severity indicates a progression to machine failure. Predicting machine failure allows for the repairing of the machine before it breaks, saving cost and minimizing machine downtime.

Experts in the field are able to conclude that machine failure can occur in days to weeks, weeks to months, or months to years. Experts know rules and have an accumulation of experiences [1]. “When a new experience takes place, it isn't simply added to the data base of prior experiences. Learning from experience means changing what you know to accommodate the new knowledge” [1].

A piece of information that is lacking in machine predictive maintenance is a good estimation of Time to Failure (TTF) [7]. This project will determine TTF of a single machine by reviewing a progression of its own developing faults. This project will also use case-based reasoning (CBR) to determine TTF of similar machine configurations. The aspects that can affect TTF are the severity of the specific faults generated, the Machine Identification (MID) which is the machine configuration, and other external influences.

Machine faults start with a very gradual slope, and over time, as the severity of these faults increase, the slope rises sharply. Machine failure progresses at very different rates. Machine TTF is not linear [3]. This project will determine the rate of severity change for faults and adjust the rates accordingly as the dataset changes. The system will learn over time to predict the best estimation of TTF of a specific machine based on early identification of a machine anomaly.
Another aspect that can assist in determining TTF is the analysis of data based on similar machinery. The DLI Engineering database contains MIDs which group machines with common model numbers and applications. With this feature, MID grouped machine faults can be analyzed together to provide more experience to the TTF analysis.

External influences also affect machine TTF where the information is not available for consideration. For example, machines not maintained properly or that run seven days a week, 24 hours a day will reach failure at a much faster rate than a machine that runs a few hours a day and is maintained properly.

DLI Engineering is sponsoring this project by providing the database and engineering expertise. DLI Engineering is a marine and mechanical engineering company performing vibration and acoustic consulting.
Chapter 2

BACKGROUND INFORMATION

This project affects machinery maintenance, utilizes results from Machine Condition Analysis software, and applies data mining and CBR concepts.

2.1 Machinery Maintenance

Machinery maintenance practices have changed greatly over the years. Originally, a machine would fail (run-to-failure) before maintenance is performed. This type of maintenance is sometimes called “crisis maintenance” [12]. Then machines with no problems had preventive maintenance performed according to some schedule improved machine uptime. Now, with predictive maintenance, early identification of machine faults results in maintenance being performed before failure [12]. With a TTF estimate, maintenance can be scheduled at the most efficient and convenient time.

The most important goal of any maintenance program is the elimination of machine breakdowns [12]. Very often, a catastrophic breakdown will cause significant peripheral damage to the machine, greatly increasing the cost of the repair. The second goal is to anticipate and plan for maintenance needs [12]. This enables planning for down time, ordering of parts, and scheduling appropriate staff. The third goal is to increase plant production by reducing machine breakdowns during operations [12]. Predicting TTF can assist in achieving these goals.

2.2 Machine Condition Analysis software

The Machine Condition Analysis software includes an Expert System that is a “Rule-based” system that generates a multitude of information. The Expert System is run against a “test,” a set of vibration data associated with a data collection timestamp. Each test may have
multiple Expert System runs where a vibration engineer may choose to change machine setup parameters. The last Expert System run is the record of choice by the engineer. In other words, all prior runs are ignored. This information used in this TTF System includes messages, severity, and diagnosis. The messages returned may indicate problems with data collection and therefore with enabling the TTF system to flag those records as “Invalid”. The severity is the degree of the diagnosis. A severity < 101 is “Slight,” 101-300 is “Moderate,” 301-600 is “Serious,” and > 600 is “Extreme.” In this TTF System, an Extreme severity is considered a machine failure. The diagnosis describes the machine’s problem. Throughout this document, the words diagnosis and fault are used interchangeably.

2.3 Data Mining

Like many areas of industry, data is collected but very little is converted to useful information. Data mining is extracting or “mining” knowledge from large amounts of data [8]. Data mining is very important with the ever-growing amounts of stored data. Data mining automates the analysis of large volumes of data to produce useful information [2]. Data mining is an application or interface of statistics and pattern technology and concerned with secondary analysis of data to predict future behavior [2] [4]. There are two kinds of models, predictive and descriptive. The predictive model makes predictions based on historical data. This model may be able to determine which customers would be interested in a new product. The descriptive model summarizes data illustrating subgroups of data. Data Mining consists of different approaches and algorithms. The approach is the algorithm that ties together the relevant data. The algorithm is the technique used to associate data: statistics, clustering, trees, neural networks, instance-based learning (case-based), and so forth.

2.4 Case-Based Reasoning

CBR is a “…model of reasoning that incorporates problem solving, understanding, and learning, and integrates all of them with memory process” [10]. “These tasks are performed using typical situations, called cases, already experienced by a system” [10]. Cases may also
be atypical, rare, or exceptional. CBR solves new problems by adapting solutions to older problems. When there is a new problem and an identical old case is found, the solution of the old case can be applied as a solution to the new problem. If an identical case is not found, an adaptation occurs and a solution is retrieved. “In case-based reasoning, the retrieval usually provides just the most similar case whose solution has to be adapted to fit the query course. As in compositional adaptation we take the solutions of a couple of similar cases into account” [11].

The four parts of a CBR system are retrieving, reusing, revising, and retaining. Retrieving is the part that returns an old case that is determined to be identical or similar to the new problem. “To carry out effective case retrieval, there must be selection criteria that determine how a case is judged to be appropriate for retrieval and a mechanism to control how the case base is searched” [10]. Reusing is the part that applies the solution of the retrieved old case, and adapts the retrieved solutions to solve the new problem. Revising is the step that corrects the adapted solution after user feedback. Lastly, retaining is the storing of valid confirmed cases.
Chapter 3

PROBLEM STATEMENT

This project will determine if it is possible to data mine temporal vibration data to produce an estimated time to a machine failure based on identified machine faults. There is a need in the predictive maintenance industry to develop a mechanism to estimate TTF. Temporal data mining is the analysis of a sequence of specific values changing over time. Data exists at irregular intervals. The goal is to use CBR to predict a future event. The REDI-PRO system, a naval application, used Figure of Merit (FOM) to monitor a system and determine remaining life of mechanical systems [5]. FOM is not available in the database evaluated in this system.
Chapter 4

DESCRIPTION OF DATA

The database contains vibration data analyzed through the Expert System resulting in severities and their associated diagnoses used in this system. It contains nearly 10,000 machines and over 142,000 tests. The database includes tests dating back to 1981. A partial database schema is included in Appendix A.1.

Two new tables, VibrationCase and VibrationCasetest, were created to store the case library. Each case is uniquely identified by a CaseID in the new VibrationCase table. This table represents each unique case in the system. Each case is associated with at least three test records stored in the new VibrationCaseTest table. VibStandardEquipmentID is a nominal value representing the machine configuration. VibStandardDiagnosisID is a nominal value representing the diagnosis. DiagnosisGroupID is a nominal value representing the diagnosis group. VibStandardEquipmentID is a nominal value representing the machine configuration group. TotalDaystoFailure is a real value that represents the total days to failure for the case. The CaseTestID is a unique identifier for each test in a case. DayPosition is a real value representing the day the test occurred in the case. VibStandardSeverityIndex is a real value that represents the severity of the diagnosis in the test. VesselID is a unique identifier for a site, EquipmentID is a unique identifier for a Machine, and TestResultID is a unique identifier for a test. VesselID, EquipmentID, and TestResultID are nominal values that associate the test back to the rest of the database. These values are keys that are not used in the analysis of the CBR system. CaseIsActive and CaseTestIsActive are nominal value that flags a case or case test that should or should not be used in the CBR system. These tables are displayed in table 4.1 and 4.2.
Table 4.1 VibrationCase table

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Key</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>CaseID</td>
<td>Primary Key</td>
<td>Integer</td>
</tr>
<tr>
<td>VibStandardEquipmentID</td>
<td>Integer</td>
<td>Nominal</td>
</tr>
<tr>
<td>VibStandardDiagnosisID</td>
<td>Integer</td>
<td>Nominal</td>
</tr>
<tr>
<td>DiagnosisGroupID</td>
<td>Integer</td>
<td>Nominal</td>
</tr>
<tr>
<td>VibStandardEquipmentGroupID</td>
<td>Integer</td>
<td>Nominal</td>
</tr>
<tr>
<td>TotalDaysToFailure</td>
<td>Integer</td>
<td>Real</td>
</tr>
<tr>
<td>CaseType</td>
<td>Integer</td>
<td>Nominal</td>
</tr>
<tr>
<td>CaseIsActive</td>
<td>Small Int</td>
<td>Nominal</td>
</tr>
</tbody>
</table>

Table 4.2 VibrationCaseTest table

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Key</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>TestCaseID</td>
<td>Primary Key</td>
<td>Integer</td>
</tr>
<tr>
<td>VesselID</td>
<td>Foreign Key</td>
<td>Integer</td>
</tr>
<tr>
<td>EquipmentID</td>
<td>Foreign Key</td>
<td>Integer</td>
</tr>
<tr>
<td>TestResultID</td>
<td>Foreign Key</td>
<td>Integer</td>
</tr>
<tr>
<td>CaseID</td>
<td>Foreign Key</td>
<td>Integer</td>
</tr>
<tr>
<td>DayPosition</td>
<td>Integer</td>
<td>Real</td>
</tr>
<tr>
<td>VibStandardSeverityIndex</td>
<td>Integer</td>
<td>Real</td>
</tr>
<tr>
<td>CaseTestIsActive</td>
<td>Small Int</td>
<td>Nominal</td>
</tr>
</tbody>
</table>
An example case is as follows:

```plaintext
@attribute CaseID nominal  
@attribute VibStandardEquipmentID nominal  
@attribute VibstandardDiagnosisID nominal  
@attribute Casetype nominal  
@attribute TotalDaysToFailure real  
@attribute DiagnosisGroupID nominal  
@attribute VibStandardEquipmentGroupID nominal  
@attribute VesselID nominal  
@attribute EquipmentID nominal  
@attribute DayPosition real  
@attribute VibDiagnosisSeverityIndex real  
@attribute TestCaseID nominal  
```

```
33,879,395,0,267,4,82,6,6253,0,0,147  
33,879,395,0,267,4,82,6,6253,139,130,148  
33,879,395,0,267,4,82,6,6253,267,640,149  
```

Table 4.3 Case Example

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CaseID</td>
<td>33</td>
<td>Unique key for case</td>
</tr>
<tr>
<td>VibStandardEquipmentID</td>
<td>879</td>
<td>Cargo Refrigeration Compressor</td>
</tr>
<tr>
<td>VibstandardDiagnosisID</td>
<td>395</td>
<td>Motor Bearing Looseness</td>
</tr>
<tr>
<td>CaseType</td>
<td>0</td>
<td>Type of case</td>
</tr>
<tr>
<td>TotalDaysToFailure</td>
<td>267</td>
<td>Number of days to failure for the case</td>
</tr>
<tr>
<td>DiagnosisGroupID</td>
<td>4</td>
<td>Looseness</td>
</tr>
<tr>
<td>VibStandardEquipmentGroupID</td>
<td>82</td>
<td>Motor driven reciprocating compressor no ball bearings</td>
</tr>
<tr>
<td>VesselID</td>
<td>6</td>
<td>Confidential</td>
</tr>
<tr>
<td>EquipmentID</td>
<td>6253</td>
<td>No.1 Cargo refrigeration compressor</td>
</tr>
<tr>
<td>DayPosition</td>
<td>0, 139, 267</td>
<td>Day position of the test within the case</td>
</tr>
<tr>
<td>VibDiagnosisSeverityIndex</td>
<td>0, 130, 640</td>
<td>The degree of the diagnosis</td>
</tr>
<tr>
<td>TestCaseID</td>
<td>147, 148, 149</td>
<td>Unique key for each test included in the case</td>
</tr>
</tbody>
</table>
Chapter 5

DATA PRE-PROCESSING

The database requires modifications to incorporate TTF functionality. Preprocessing includes new tables, new columns to existing tables, grouping associations of MIDs and Diagnoses, and identifying bad test data. Appendix B includes images of the tool used to assist in this processing.

The Mid Grouping tab displays all MIDs and MID groups in the database. Different icons are used to indicate whether an MID is already included in a group. To create a new MID group, the user enters a group name and associates which MIDs to include in the group (see Appendix B.1). This same methodology applies to diagnosis groups (see Appendix B.2). The Selection tab is simply a query tool that assists the user in determining which diagnosis or MID belongs to a specific group (see Appendix B.3). The last Tab is for miscellaneous functions. It currently only includes a button to ‘exclude’ bad tests.

5.1 Invalid Data

Data collected improperly results in invalid data being stored in the database. The Expert System produces the following messages when run on invalid data: “Questionable data at pickup…,” “Manual review of the spectra is warranted,” and “Unavailable pickups…”. The invalid records are excluded in this system’s analysis. A column VibResultInvalid (small integer) was added to the VibrationAnalysisExpertSystem table. The VibResultInvalid column contains “-1” for invalid data and “0” otherwise.
5.2 Grouping

Diagnoses and MIDs have been grouped where the diagnoses basically define the same problem and where the MIDs are basically describing the same machine configuration. The database contains 656 diagnoses, some of which are very similar. These similar diagnoses have been grouped and stored in a new table, DiagnosisGroups. This table contains two columns: DiagnosisGroupID and VibStandardEquipmentGroupDesc as seen in table 5.1. A column, DiagnosisGroupID was added to table VibrationStandardDiagnosis to associate similar diagnoses. An example of a diagnosis group is displayed in table 5.2.

Table 5.1 DiagnosisGroups table

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Key</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiagnosisGroupID</td>
<td>Primary Key</td>
<td>Integer</td>
</tr>
<tr>
<td>DiagnosisGroupDesc</td>
<td></td>
<td>Varchar(80)</td>
</tr>
</tbody>
</table>

Table 5.2 Diagnosis Group Example

<table>
<thead>
<tr>
<th>Diagnosis Group 1</th>
<th>Diagnosis Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coupling Wear</td>
<td>Ball Bearing Noise</td>
</tr>
<tr>
<td>Coupling Wear or Looseness</td>
<td>Ball Bearing Wear</td>
</tr>
<tr>
<td></td>
<td>Bearing Wear or Defect</td>
</tr>
</tbody>
</table>

The database contains over 1900 MIDs some of which are very similar. These similar MIDs have been grouped and stored in a new table, VibrationStandardEquipmentGroups. The table contains two columns: VibrationStandardEquipmentGroupID PK integer and VibrationStandardEquipmentGroupDesc varchar(80) as seen in table 5.3. A column, VibrationStandardEquipmentGroupID, was added to table VibrationStandardEquipment to associate similar MIDs. An example of a MID group is displayed in table 5.4.
Table 5.3 VibrationStandardEquipmentGroups table

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Key</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>VibrationStandardEquipmentGroupID</td>
<td>Primary Key</td>
<td>Integer</td>
</tr>
<tr>
<td>VibrationStandardEquipmentGroupDesc</td>
<td></td>
<td>Varchar(80)</td>
</tr>
</tbody>
</table>

Table 5.4 MID Group Example

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC Chill Water Pump</td>
<td>Air Conditioning Salt Water Circulating Pump</td>
</tr>
<tr>
<td>A/C Chill Water Pump</td>
<td>Air Conditioning Sea Water Service Pump</td>
</tr>
<tr>
<td>Air Conditioning Chill Water Pump</td>
<td>A/C S/W CIRC PUMP</td>
</tr>
<tr>
<td>Air Conditioning Chilled Water Pmp</td>
<td></td>
</tr>
</tbody>
</table>

MIDs are basically grouped in terms of the driver components: motors, turbines, diesels, and so forth and how it is connected to the driven component: geared or belted, or driven (direct drive). The driven component consists of rotary screw, centrifugal, piston, and so forth and the driven component consists of pump, compressor, fan, blower, and so forth. The groupings are also based on the type of fluid pushed: hydraulics, lube oil, fuel oil, water, and so forth, and speed where high speed is anything around 3000 – 3600 rpm.

There are some other separators; such as if the motor has no fan or the driven machine has no ball bearings. They are noted in the group title.

5.3 Pre-Processing Results

Two copies of the database were preprocessed. One copy, Database 1 contains tests up through 11/28/2005. A second copy of the database, Database 2 contains tests up through
09/13/2006. Database 2 contains about 17% more tests and cases used for evaluation. Table 5.5 displays the details of these counts in the two databases.

The pre-processing resulted in about 49% of the test with valid Expert System runs with faults, greater than 96% of the MIDs belong to a MID group, and greater than 57% of the Diagnosis belongs to a Diagnosis group. Database 2 contains 16.8 more valid Expert System runs with faults. Detailed results are displayed in table 5.5.

Table 5.5 Pre-processing results

<table>
<thead>
<tr>
<th></th>
<th>Database 1</th>
<th>Database 2</th>
<th>Difference</th>
<th>Percent of Database 1 to Database 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tests</td>
<td>142,324</td>
<td>169,713</td>
<td>27,389</td>
<td>83.86%</td>
</tr>
<tr>
<td>Valid Expert System runs</td>
<td>128,965</td>
<td>154,808</td>
<td>25,843</td>
<td>83.31%</td>
</tr>
<tr>
<td>Valid Expert System runs with Faults</td>
<td>69,228 (69228 / 142324 = 48.64%)</td>
<td>83,207 (83207 / 169713 = 49.03%)</td>
<td>13,979</td>
<td>83.20%</td>
</tr>
<tr>
<td>Total MIDs</td>
<td>1917</td>
<td>2080</td>
<td>163</td>
<td>92.16%</td>
</tr>
<tr>
<td>MID Groups</td>
<td>114</td>
<td>116</td>
<td>2</td>
<td>98.28%</td>
</tr>
<tr>
<td>MIDs in Group</td>
<td>1852</td>
<td>2058</td>
<td>206</td>
<td>90.00%</td>
</tr>
<tr>
<td>Total Diagnoses</td>
<td>656</td>
<td>659</td>
<td>3</td>
<td>99.54%</td>
</tr>
<tr>
<td>Diagnosis Groups</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Diagnoses in Group</td>
<td>380</td>
<td>380</td>
<td>0</td>
<td>100%</td>
</tr>
</tbody>
</table>

(380 / 656 = 57.9%)
Chapter 6

TIME TO FAILURE SYSTEM

This project predicts TTF based on analysis of existing temporal data. The most important features of this system include the TTF Database Setup Tool to perform data preprocessing, the Build Case Tool to build cases, the TTF user interface that displays TTF on the selected machine and the database for storage of the case library.

This system does not determine remaining life of a machine running in a normal state. If a machine has no faults, the analysis will result in “Insufficient data to determine TTF.” The system begins its analysis from the first fault identified. If a machine is repaired and returns to a normal running state, the system will return “Insufficient data to determine TTF.”

This system does not specifically indicate machine failure. An extreme fault will be the indicator of imminent machine failure.

6.1 Case-Based Reasoning System

This project uses CBR to provide solutions to new problems from previous cases (see figure 6.1). In this project, a case will be defined as a machine with tests beginning from a normal state and running to a critical state. When solving a new problem, the system will first attempt to make a direct match by searching for a previous case from the case library, for a machine with the same MID as the current problem. If one is found, the solution is applied and the problem is solved. If a direct match is not found, a generalization step is performed to try to find an indirect match.
6.2 Generalization

When the database was initially implemented, an MID associated with machines was site specific. This site specificity caused duplication of MIDs within the database. These MIDs across sites are very similar but may contain slight variations. These MIDs across sites are grouped for use in finding indirect case matches. For example, the MIDs “AC Chill Water Pump,” “A/C Chill Water Pump,” and “Air Conditioning Chill Water Pmp” have been grouped to belong to one MID group.

A variety of diagnoses generated by the Expert System can imply the same problem. These similar diagnoses are grouped and used in finding of indirect case matches. For example, the diagnoses “Ball bearing wear,” “Ball bearing noise” and “Ball bearing wear or defect” belongs to one diagnosis group.
6.3 CBR – retaining

Each case consists of a minimum of three consecutive tests. The tests in a case consist of one in which the machine had a severe fault, one in which the fault did not exist and one in which the fault exists but is less than severe. There are four types of cases. Case type 0 consists of a minimum of three consecutive tests on a machine where the last test has an extreme fault and the first test does not have the fault. Case type 1 consists of a minimum of three consecutive tests on a machine where at last test has an extreme grouped fault and the first test does not have the fault. Case type two and three are built on the same criteria with the exceptions listed in table 6.1.

Table 6.1 Case base logic

<table>
<thead>
<tr>
<th>Case type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 0</td>
<td>Same diagnosis and same machine</td>
</tr>
<tr>
<td>Type 1</td>
<td>Grouped diagnosis and same machine</td>
</tr>
<tr>
<td>Type 2</td>
<td>Normalize same diagnosis within the grouped MID</td>
</tr>
<tr>
<td>Type 3</td>
<td>Normalize grouped diagnosis within the grouped MID</td>
</tr>
</tbody>
</table>

6.4 CBR – retrieval

Cases are retrieved based on the different types of cases stored in the case library. An instance of a direct case is a case on the same machine with the same fault but does not include the current test instance. An instance of an indirect case is one of the following: same diagnosis on one of the grouped MIDs, grouped fault on the same MID, or grouped fault with one of the grouped MIDs. The specific instances of the case retrieval types are listed in table 6.2.

The system first attempts to find a direct match (1). If a case is not found, the system attempts to find an indirect match (2), and continues through the match types until a case is found. When a case is found, the solution is applied to the new problem.
If multiple cases are retrieved, the system selects the best case by finding a case that is failing at the closest rate as the new problem. It takes the severity and normalized date and selects the case with the closest severity to the normalized date. This is an attempt to find a case that is failing at the same rate as the new problem.

With an attempt to improve accuracy, the algorithm was tweaked to retrieve 2-nearest neighbor and 3-nearest neighbor. The system would average the TTF values of the two or three nearest neighbor cases.

Table 6.2 Case Base retrieval logic

<table>
<thead>
<tr>
<th>Match type</th>
<th>Case type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Direct</td>
<td>0       Same diagnosis and same machine</td>
</tr>
<tr>
<td>2</td>
<td>Indirect</td>
<td>0       Same diagnosis, same MID and different machine</td>
</tr>
<tr>
<td>3</td>
<td>Indirect</td>
<td>0       Same diagnosis, MID group and different machine</td>
</tr>
<tr>
<td>4</td>
<td>Indirect</td>
<td>0       Diagnosis group, MID group and different machine</td>
</tr>
<tr>
<td>5</td>
<td>Indirect</td>
<td>1       Grouped diagnosis and same machine</td>
</tr>
<tr>
<td>6</td>
<td>Indirect</td>
<td>1       Grouped diagnosis, same MID and different machine</td>
</tr>
<tr>
<td>7</td>
<td>Indirect</td>
<td>1       Grouped diagnosis, MID group and different machine</td>
</tr>
<tr>
<td>8</td>
<td>Indirect</td>
<td>2       Same diagnosis, MID group and different machine</td>
</tr>
<tr>
<td>9</td>
<td>Indirect</td>
<td>2       Diagnosis group, MID group and different machine</td>
</tr>
<tr>
<td>10</td>
<td>Indirect</td>
<td>3       Diagnosis group, MID group and different machine</td>
</tr>
</tbody>
</table>

### 6.5 CBR – reusing

After the system retrieves a case, this known solution is applied to the new problem. The test date of the new problem is normalized, applied to the previous solution and the new solution, TTF, is calculated. Normalization of date is done by calculating the number of days from the current test to the first prior test where the diagnosis in question is not present. For example, let $D_1$ equal to the date of the current test where a specific diagnosis exists. Let $D_2$ be the date of the first prior test where the specific diagnosis does not exist and let $D_{\text{result}}$ be the calculated normalized date, in days. $D_{\text{result}} = D_1 - D_2$. TTF is calculated by subtracting $D_{\text{result}}$
from the TTF from the previous solution, $\text{TTF}_{\text{case}}$. The solution to the new problem is $\text{TTF}_{\text{result}} - \text{D}_{\text{result}}$.

### 6.6 CBR – revising

A user may review all cases in the case library and evaluate for correctness. The user may decide to flag invalid cases or individual case tests. This allows the user control in excluding invalid cases or individual case tests from being used during TTF determination. This is done in the user interface displayed in Appendix D.3.
Chapter 7

ANALYSIS

This section includes the analysis of the CBR system as well data analysis using SPSS 15.0 for Windows by Lead Technologies, Inc. and WEKA 3.4.10, the University of Waikato. Both SPSS and WEKA are data mining tools providing a variety of data mining algorithms.

7.1 CBR

The system was evaluated on two copies of the database. One copy, Database 1 contains tests up through 11/28/2005. A second copy of the database, Database 2 contains tests up through 09/13/2006. Database 2 contains about 17.5% more cases in the case library. Table 7.1 displays the details of these counts in the two databases. Raw data results from the VibrationCase and VibrationCase table are listed in Appendix C.1 and C.2.

Table 7.1 Case Library

<table>
<thead>
<tr>
<th>Case type</th>
<th>Database 1 (training)</th>
<th>Database 2 (test)</th>
<th>Difference (test)</th>
<th>Percent Database 2 Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case type 0</td>
<td>355</td>
<td>430</td>
<td>85</td>
<td>17.44%</td>
</tr>
<tr>
<td>Case type 1</td>
<td>39</td>
<td>40</td>
<td>1</td>
<td>2.5%</td>
</tr>
<tr>
<td>Case type 2</td>
<td>53</td>
<td>85</td>
<td>32</td>
<td>37.65%</td>
</tr>
<tr>
<td>Case type 3</td>
<td>39</td>
<td>61</td>
<td>22</td>
<td>36.07%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>486</td>
<td>616</td>
<td>130</td>
<td>21.10%</td>
</tr>
</tbody>
</table>

The TTF accuracy was based on two (<=62 days) and three months (<=93 days). The case library was analyzed to determine the TTF precision of each MID and diagnosis combination. A comparison was done on Case Type 0 records. The difference between the minimum TTF and the maximum TTF with the same MID and same diagnosis ranged from 2...
to 1785 days and an average difference of 488 days. The majority of the cases have a standard
deviation less than 250 days. This is a very broad range for the same machine. There were
only 7 instances where the difference was less than 93 days. Some of these cases have very
long TTF values. These values are displayed in Appendix C.3 and a standard deviation
distribution chart is displayed in Appendix C.4. This may be due to some repairs being
performed on a machine and postponing TTF resulting abnormally large TTF values. The
repairs may affect the cases built and would affect the TTF values.

The first run of the TTF system was on the training set which produced results as
expected. It generated an accuracy rate of 98.15%. The training results are displayed in
Appendix C.5. The next was run based on the system knowing the TTF. If the system
retrieved multiple cases it chose the case based on the known TTF. This run produced the best
results the system can produce at less than 50%. See appendix C.6. The run resulted in 192 of
the 478 new problems found no case matches, 193 new problems found one case match and 93
new problems found multiple cases. The accuracy of each case type selected showed type 1 at
100%. This is promising, but it only contained two matches so it was truly difficult to evaluate
(see Appendix C.7).

The first real test of the TTF system was quite disappointing. It produced only a 25%
accuracy rate (see Appendix C.8). The run resulted in 213 of the 413 new problems found no
case matches, 78 new problems found one case match and 182 new problems found multiple
cases matches. The breakdown of the accuracy based on case type selected did not show any
promising results (see Appendix C.9). In an attempt to improve the accuracy, runs were
performed using 2-nearest neighbor and 3-nearest neighbor. The accuracy of 2-nearest
neighbor showed no significant difference at 24% accuracy (see Appendix C.10). The
accuracy of 3-nearest neighbor decreased to 15% (see Appendix C.11). This shows the nearest
neighbor is retrieving cases that are not similar enough to improve the accuracy.

The database contains only a small number of cases for each MID-diagnosis combination.
The database contains 270 cases which have a unique MID-diagnosis combination, and 115
cases make up 41 MID-diagnosis combinations where the MID-diagnosis combination is
included of more than one case. This results in about 50% of the new cases which have no
match.
A test was derived to determine if increasing the number of cases in the case library would improve accuracy. The first test used 10 cases, the second used 20 cases, and so on. This test displayed an increase in accuracy up until 50 cases were utilized. Then the accuracy plateaued between 20 and 30%.

Figure 7.1 Increasing number of cases in library

### 7.2 SPSS Linear Regression

SPSS linear regression was used to evaluate the relationship of the independent variables, VibrationStandardEquipmentGroupID, DiagnosisGroupID, VibrationStandardEquipmentID, VibrationStandardDiagnosisID and VibDiagnosisSeverityIndex that best predicts the value of the dependent variable, DayPosition. The analysis of the regression results was done using the Results Coach in SPSS to interpret the significance of the values. The Pearson correlation coefficient is a measure of linear association between two variables. The results range from 1 to -1 and the larger the absolute value, the stronger the relationship. The strongest relationship was with VibStandardSeverityIndex at .298 which is still not very strong a relationship. The
significance level (p-value) of <0.05 is significant and the two variables are linearly related. This is true for the relationships between DayPosition and DiagnosisGroupID, DayPosition and VibStandardEquipmentGroupID, VibDiagnosisSeverityIndex and DayPosition, VibStandardEquipmentID and DiagnosisGroupID, VibStandardDiagnosisID and DiagnosisGroupID, VibDiagnosisSeverityIndex and DiagnosisGroupID, VibStandardEquipmentGroupID and DiagnosisGroupID, VibDiagnosisSeverityIndex. SPSS Correlations displayed in table 7.2

Table 7.2  SPSS Correlations

<table>
<thead>
<tr>
<th></th>
<th>DayPosition</th>
<th>VibStandardEquipmentID</th>
<th>C. VibstandardDiagnosisID</th>
<th>DiagnosisGroupID</th>
<th>VibStandardEquipmentGroupID</th>
<th>VibDiagnosisSeverityIndex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation</td>
<td>1.000</td>
<td>-.024</td>
<td>.015</td>
<td>-.068</td>
<td>.089</td>
<td>.298</td>
</tr>
<tr>
<td></td>
<td>-.024</td>
<td>1.000</td>
<td>.043</td>
<td>-.110</td>
<td>-.045</td>
<td>-.021</td>
</tr>
<tr>
<td></td>
<td>.015</td>
<td>.043</td>
<td>1.000</td>
<td>-.249</td>
<td>-.019</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td>-.068</td>
<td>-.110</td>
<td>-.249</td>
<td>1.000</td>
<td>-.228</td>
<td>.053</td>
</tr>
<tr>
<td></td>
<td>.089</td>
<td>-.045</td>
<td>-.019</td>
<td>-.228</td>
<td>1.000</td>
<td>-.128</td>
</tr>
<tr>
<td></td>
<td>.298</td>
<td>-.021</td>
<td>.025</td>
<td>.053</td>
<td>-.128</td>
<td>1.000</td>
</tr>
<tr>
<td>Sig. (1-tailed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.195</td>
<td>.062</td>
<td>.007</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>.298</td>
<td>.062</td>
<td>.000</td>
<td>.251</td>
<td>.186</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>.007</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.030</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>.001</td>
<td>.053</td>
<td>.251</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>.000</td>
<td>.228</td>
<td>.186</td>
<td>.030</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>1268</td>
<td>1268</td>
<td>1268</td>
<td>1268</td>
<td>1268</td>
<td>1268</td>
</tr>
</tbody>
</table>

R values range from 0 to 1 with larger R values indicating stronger relationships. This model produced and R value of only .330. R squared values range from 0 to 1 with larger values indicating that the model fits the data well. This system only produced a value of .109 which shows that the model does not fit the data well. See table 7.3 for the SPSS Model Summary.
Table 7.3 SPSS Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
<th>R Square Change</th>
<th>F Change</th>
<th>df1</th>
<th>df2</th>
<th>Sig. F Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.330</td>
<td>.109</td>
<td>.105</td>
<td>.105</td>
<td>.105</td>
<td>30.818</td>
<td>5</td>
<td>1262</td>
<td>.000</td>
</tr>
</tbody>
</table>

*a.* Predictors: (Constant), VibDiagnosisSeverityIndex, VibStandardEquipmentID, C.VibstandardDiagnosisID, VibStandardEquipmentGroupID, DiagnosisGroupID

*b.* Dependent Variable: DayPosition

The table 7.4 summarizes the results of an analysis of variance. The output for Regression displays information about the variation accounted for by this model and the output for Residual displays information about the variation that is not accounted for by this model. A very high residual sum of squares indicates that the model fails to explain a lot of the variation in the dependent variable.

Table 7.4 SPSS ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>20306194</td>
<td>5</td>
<td>4061238.886</td>
<td>30.818</td>
<td>.000a</td>
</tr>
<tr>
<td>Residual</td>
<td>1.7E+008</td>
<td>1262</td>
<td>131779.459</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1.9E+008</td>
<td>1267</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a.* Predictors: (Constant), VibDiagnosisSeverityIndex, VibStandardEquipmentID, C.VibstandardDiagnosisID, VibStandardEquipmentGroupID, DiagnosisGroupID

*b.* Dependent Variable: DayPosition
Using the coefficients in the table 7.5, regression was applied using the formula:

\[
\text{DayPosition} = 133.508 + (-.011 \times \text{VibStandardEquipmentID}) \\
+ (-.015 \times \text{VibStandardDiagnosisID}) \\
+ (-8.004 \times \text{DiagnosisGroupID}) \\
+ (1.311 \times \text{VibStandardEquipmentGroupID}) \\
+ (.266 \times \text{VibDiagnosisSeverityIndex})
\]

### Table 7.5 SPSS Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>T</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>133.508</td>
<td>46.836</td>
<td>2.851</td>
</tr>
<tr>
<td></td>
<td>VibStandardEquipmentID</td>
<td>-.011</td>
<td>.015</td>
<td>-.019</td>
</tr>
<tr>
<td></td>
<td>C.VibstandardDiagnosisID</td>
<td>-.015</td>
<td>.071</td>
<td>-.006</td>
</tr>
<tr>
<td></td>
<td>DiagnosisGroupID</td>
<td>-8.004</td>
<td>3.636</td>
<td>-.063</td>
</tr>
<tr>
<td></td>
<td>VibStandardEquipmentGroupID</td>
<td>1.311</td>
<td>.319</td>
<td>.114</td>
</tr>
<tr>
<td></td>
<td>VibDiagnosisSeverityIndex</td>
<td>.266</td>
<td>.023</td>
<td>.316</td>
</tr>
</tbody>
</table>

*a  Dependent Variable: DayPosition
Note: Only a partial table is displayed

This regression produced 133 of 478 correctly predicted problems, within three months, for a 27.8% accuracy. This is only a 2% improvement over the case base method. Summary of results is listed in table 7.6.
Table 7.6 Result summary

<table>
<thead>
<tr>
<th>Test – Database 1</th>
<th>Three months - Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Test</td>
<td>25.39%</td>
</tr>
<tr>
<td>2-Nearest Neighbor</td>
<td>24.24%</td>
</tr>
<tr>
<td>3-Nearest Neighbor</td>
<td>15.00%</td>
</tr>
<tr>
<td>SPSS Regression</td>
<td>27.8%</td>
</tr>
</tbody>
</table>

7.3 Weighting Parameters

Another analysis of the data was done by changing the weighting of the attributes to determine the importance of these attributes on the accuracy of the system. The calculation was done based on the following formula where $W_i$ is the weight for the matching attributes and $W_{total_i}$ is the weight of the attribute evaluated:

$$\frac{\sum W_i F_i}{\sum W_{total_i}}$$

The first test included the following attributes with a weight of 1:

- VibStandardSeverityIndex (converted to nominal values of None, Slight, Moderate, Severe, or Extreme)
- VibStandardEquipmentID
- VibStandardEquipmentGroupID
- VibStandardDiagnosisID
- DiagnosisGroupID
- DayPosition (converted to nominal values of <1month, 1-2months, 2-3months, 3-4months, 4-5months, 5-6months, 6-7months, 8-9months, 10-11months, 11-12months, and >12months)

The first test resulted in a 22.38% accuracy. The second test changed the DayPosition to nominal values of 3 month intervals, <3month, 3-6months, 6-9months, 9-12months, and
>12 months. The 3 month interval test resulted in 25.31% accuracy. The third test changed the VibStandardSeverityIndex to None, Slight-Moderate, and Severe-Extreme, and retaining the 3 month intervals of DayPosition from the second test. This third test resulted in 25.73% accuracy. The next test, the DayPosition interval was changed to <3months and >3months and resulted in a 19.25% accuracy. With the third test having the best accuracy rate, these values were used to evaluate different combinations of weighting with the attributes evaluated. The best accuracy of 26.57% resulted when all attributes were used except for VibStandardEquipmentGroupID. The worst accuracy of 11.92% resulted when only VibStandardSeverityIndex and DayPosition were considered in the weighting. Summary of the weighting results are shown in table 7.7.

Table 7.7 Weighting Summary

<table>
<thead>
<tr>
<th>Num</th>
<th>Severity</th>
<th>DayPosition</th>
<th>MID</th>
<th>MIDGrp</th>
<th>Dx</th>
<th>DxGrp</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>25.73%</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11.92%</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>24.89%</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>23.64%</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>25.10%</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>26.57%</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>20.50%</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>20.08%</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17.99%</td>
</tr>
</tbody>
</table>

Note:
Num = Weighted test number
Severity = VibStandardSeverityIndex
MID = VibStandardEquipmentID
MIDGrp = VibStandardEquipmentGroupID
Dx = VibStandardDiagnosisID
DxGrp = DiagnosisGroupID

The results in table 7.8 were produced by SPSS linear regression applying multi-step regression. It shows the low R value with the addition of each predictor.
Table 7.8 SPSS Multi-step regression – See Predictors

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.154a</td>
<td>.024</td>
<td>.023</td>
<td>312.827</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), VibDiagnosisSeverityIndex

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.146a</td>
<td>.021</td>
<td>.021</td>
<td>313.856</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), VibStandardEquipmentGroupID, VibDiagnosisSeverityIndex

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.139a</td>
<td>.019</td>
<td>.019</td>
<td>300.691</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), DiagnosisGroupID, VibDiagnosisSeverityIndex, VibStandardEquipmentGroupID

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.266a</td>
<td>.071</td>
<td>.068</td>
<td>408.944</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), VibStandardEquipmentID, VibDiagnosisSeverityIndex, VibStandardEquipmentGroupID, DiagnosisGroupID

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.327a</td>
<td>.107</td>
<td>.103</td>
<td>365.374</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), VibstandardDiagnosisID, VibDiagnosisSeverityIndex, VibStandardEquipmentID, VibStandardEquipmentGroupID, DiagnosisGroupID
7.4 WEKA Algorithms and dataset

All numeric attributes included in the analysis were discretized to nominal values. The dataset was modified where each row contains each test severity and days to failure at the current test. Sev2 represents the severity value for the second test, Sev3 represents the severity value for test 3, Sev4 represents the severity value for test 4, if available, Diff2 is the number of days to TTF calculated from the actual TTF minus the current day position of the test, and so forth. The dataset includes DiagnosisGroupID (DxGrpID), VibStandardEquipmentGroupID (MIDGrpID), Sev2, Sev3, Sev4, Diff2, Diff3, Diff4.

The different test options used are the training set, 10-fold cross validation, and ratio validation of 50%, 66%, 75%, and 80% training. Training set as a test set provides optimal classification accuracy. 10 fold cross-validation technique averages the classification results on ten different random samples generated from the dataset. It provides more robust results when there is only one dataset available. Cross-validation tries to diversify the samples in order to get a better estimate with fewer samples. It uses averages so the method is not as accurate. The ratio validation uses a different percentage of records for the training and test set.

The algorithms used and defined by WEKA are the J48 tree, ID3 tree, Multilayer Perceptron, Logistic regression, Apriori, Predictive Apriori, and K*. J48 generates a pruned or un-pruned C4 tree where values may be missing, attributes may be numeric, and can deal with noisy data. ID3 generates an un-pruned decision tree. Attributes must be nominal and there cannot be any missing values. Empty leaves may result in unclassified instances. Multilayer Perceptron is a neural network that uses back propagation to train. "Back propagation learns by iteratively processing a data set of training tuples, comparing the network’s prediction for each tuple with the actual known target value" [8]. Logistic regression is used for building and using a multinomial logistic regression model with a ridge estimator. Apriori generates association rules from frequent item sets and Predictive Apriori finds association rules sorted by predictive accuracy. K* is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function.
7.5 **WEKA data discretized into 3 groups**

The data set was discretized into three groups by equi-depth, meaning each group has the same number of cases. The range of values in each group was quite large as seen in table 7.9. The range for Diff3 in Group2 is 268 days or almost nine months.

Table 7.9 Three group data distribution

<table>
<thead>
<tr>
<th></th>
<th>Diff2</th>
<th>Diff3</th>
<th>Diff4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>-inf-118</td>
<td>-inf-.5</td>
<td>-inf-1</td>
</tr>
<tr>
<td>Group 2</td>
<td>118-340.5</td>
<td>0.5-268.5</td>
<td>1-237.5</td>
</tr>
<tr>
<td>Group 3</td>
<td>340.5-inf</td>
<td>268.5-inf</td>
<td>237.5-inf</td>
</tr>
</tbody>
</table>

Diff3 produced the best results with all algorithms tested; J48, K*, Multilayer Perceptron, and Logistic regression. Diff2 is significantly less accurate with all algorithms because with Diff2 there is only one data point analyzed, and the analysis is based on just the initial test. Of the algorithms tested, comparing the 10-fold Cross Validation results, J48 tree produced the best results for Diff4 and Diff3. Multilayer Perceptron produced the best result for Diff2. See table 7.10, 7.11, 7.12, and 7.13.

Table 7.10 J48 tree – three groups

<table>
<thead>
<tr>
<th>J48</th>
<th>Diff4</th>
<th>Diff3</th>
<th>Diff2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
</tr>
<tr>
<td>Training</td>
<td>221</td>
<td>75.1131 %</td>
<td>430</td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>72.8507 %</td>
<td>430</td>
</tr>
<tr>
<td>50% train</td>
<td>107</td>
<td>71.9626 %</td>
<td>215</td>
</tr>
<tr>
<td>66% train</td>
<td>77</td>
<td>70.1299 %</td>
<td>147</td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>77.7778 %</td>
<td>108</td>
</tr>
<tr>
<td>80% train</td>
<td>41</td>
<td>78.0488 %</td>
<td>86</td>
</tr>
</tbody>
</table>
Table 7.11 Multilayer Perceptron – three groups

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th></th>
<th>Diff3</th>
<th></th>
<th>Diff2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of</td>
<td></td>
<td>Correctly</td>
<td></td>
<td>Total # of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Instances</td>
<td></td>
<td>Classified</td>
<td></td>
<td>Instances</td>
<td></td>
</tr>
<tr>
<td>training</td>
<td>221</td>
<td>99.095%</td>
<td>430</td>
<td>96.0456%</td>
<td>430</td>
<td>74.6512%</td>
</tr>
<tr>
<td>50% train</td>
<td>107</td>
<td>65.6109%</td>
<td>215</td>
<td>71.6279%</td>
<td>215</td>
<td>34.8837%</td>
</tr>
<tr>
<td>66% train</td>
<td>77</td>
<td>61.039%</td>
<td>147</td>
<td>78.2313%</td>
<td>147</td>
<td>33.3333%</td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>62.963%</td>
<td>108</td>
<td>77.7778%</td>
<td>108</td>
<td>27.7778%</td>
</tr>
<tr>
<td>80% train</td>
<td>41</td>
<td>65.8537%</td>
<td>86</td>
<td>77.907%</td>
<td>86</td>
<td>34.8837%</td>
</tr>
</tbody>
</table>

Table 7.12 Logistic – three groups

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th></th>
<th>Diff3</th>
<th></th>
<th>Diff2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of</td>
<td></td>
<td>Correctly</td>
<td></td>
<td>Total # of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Instances</td>
<td></td>
<td>Classified</td>
<td></td>
<td>Instances</td>
<td></td>
</tr>
<tr>
<td>training</td>
<td>221</td>
<td>90.9502%</td>
<td>430</td>
<td>88.3721%</td>
<td>430</td>
<td>57.4419%</td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>62.4434%</td>
<td>430</td>
<td>70.9302%</td>
<td>430</td>
<td>37.4419%</td>
</tr>
<tr>
<td>50% train</td>
<td>107</td>
<td>54.2056%</td>
<td>215</td>
<td>74.4186%</td>
<td>215</td>
<td>35.814%</td>
</tr>
<tr>
<td>66% train</td>
<td>77</td>
<td>62.3377%</td>
<td>147</td>
<td>70.068%</td>
<td>147</td>
<td>34.0136%</td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>70.3704%</td>
<td>108</td>
<td>70.3704%</td>
<td>108</td>
<td>30.5556%</td>
</tr>
<tr>
<td>80% train</td>
<td>41</td>
<td>65.8537%</td>
<td>86</td>
<td>74.4186%</td>
<td>86</td>
<td>33.7209%</td>
</tr>
</tbody>
</table>

Table 7.13 K* – three groups

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th></th>
<th>Diff3</th>
<th></th>
<th>Diff2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of</td>
<td></td>
<td>Correctly</td>
<td></td>
<td>Total # of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Instances</td>
<td></td>
<td>Classified</td>
<td></td>
<td>Instances</td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td>221</td>
<td>97.7376%</td>
<td>430</td>
<td>95.814%</td>
<td>430</td>
<td>76.0465%</td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>65.1584%</td>
<td>430</td>
<td>73.7209%</td>
<td>430</td>
<td>37.4419%</td>
</tr>
<tr>
<td>50% train</td>
<td>107</td>
<td>57.0093%</td>
<td>215</td>
<td>73.4884%</td>
<td>215</td>
<td>37.2093%</td>
</tr>
<tr>
<td>66% train</td>
<td>77</td>
<td>55.8442%</td>
<td>147</td>
<td>72.1088%</td>
<td>147</td>
<td>39.4558%</td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>72.2222%</td>
<td>108</td>
<td>76.8519%</td>
<td>108</td>
<td>38.8889%</td>
</tr>
<tr>
<td>80% train</td>
<td>41</td>
<td>70.7317%</td>
<td>86</td>
<td>76.8519%</td>
<td>86</td>
<td>36.0465%</td>
</tr>
</tbody>
</table>

7.6 **WEKA data discretized into 3 groups with Sev and Diff only**

This dataset used the attributes normalized into 3 groups but also excluded DxGrpID and MIDGrpID and included the Diff2, Diff3, Diff4, Sev2, Sev3, and Sev4. Of the algorithms
tested, comparing the 10-fold Cross Validation results, Logistic regression produced the best results for Diff4 and Diff3. Multilayer Perceptron continued to produce the best result for Diff2. See table 7.14 and 7.15.

Table 7.14 J48 tree – Sev and Diff only

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th></th>
<th></th>
<th>Diff3</th>
<th></th>
<th></th>
<th>Diff2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td>221</td>
<td>75.1131 %</td>
<td>430</td>
<td>80 %</td>
<td>430</td>
<td>40.2326 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>73.3032 %</td>
<td>430</td>
<td>79.0698 %</td>
<td>430</td>
<td>36.0465 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50% train</td>
<td>107</td>
<td>71.9626 %</td>
<td>215</td>
<td>80 %</td>
<td>215</td>
<td>38.6047 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66% train</td>
<td>77</td>
<td>70.1299 %</td>
<td>147</td>
<td>77.551 %</td>
<td>147</td>
<td>33.3333 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>77.7778 %</td>
<td>108</td>
<td>80.5556 %</td>
<td>108</td>
<td>34.2593 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80% train</td>
<td>41</td>
<td>82.9268 %</td>
<td>86</td>
<td>83.7209 %</td>
<td>86</td>
<td>36.0465 %</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.15 Multilayer Perceptron – Sev and Diff only

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th></th>
<th></th>
<th>Diff3</th>
<th></th>
<th></th>
<th>Diff2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td></td>
<td></td>
</tr>
<tr>
<td>training</td>
<td>221</td>
<td>83.2579%</td>
<td>430</td>
<td>81.3953 %</td>
<td>430</td>
<td>33.4884 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>70.1357%</td>
<td>430</td>
<td>78.8372 %</td>
<td>430</td>
<td>38.8372 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50% train</td>
<td>107</td>
<td>60.7477%</td>
<td>215</td>
<td>78.6047 %</td>
<td>215</td>
<td>38.6047 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66% train</td>
<td>77</td>
<td>66.2338%</td>
<td>147</td>
<td>78.2313 %</td>
<td>147</td>
<td>35.3741 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>61.1111%</td>
<td>108</td>
<td>80.5556%</td>
<td>108</td>
<td>37.037 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80% train</td>
<td>41</td>
<td>85.3659%</td>
<td>41</td>
<td>85.3659%</td>
<td>86</td>
<td>36.0465 %</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multilayer Perceptron options were modified in an attempt to improve results: learningRate changed to 0.1 and momentum changed to 0 rather than 0.3 and 0.2 respectively, nominalToBinaryFilter set to False rather than True, and set trainingTime to 600 rather than 500. These changes produced improved accuracy but Logistic regression still performed slightly better for Diff4 and Diff3. See table 7.16 and table 7.17.

Kappa statistics were recorded on the best results as displayed in tables 7.16 and 7.17. Kappa statistic compares the agreement against what is predicted and what is observed, correcting for agreement occurring by chance. The values range from 1 representing perfect agreement, 0 representing no agreement and 1 representing total disagreement. The Kappa statistic of 60-70% for Diff3 and Diff4 are much better than random classifier. Multilayer
Perceptron’s Diff2 Kappa statistic was very low at only 0.086. Logistic regression Diff4 and Diff3 Kappa statistic was .6358 and .6697 respectively. Table 7.18 displays the results from K* but the results are not as good as that produced by Multilayer Perceptron and Logistic regression.

Table 7.16 Multilayer Perceptron – Sev and Diff only – modified options

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th></th>
<th></th>
<th></th>
<th>Diff3</th>
<th></th>
<th></th>
<th></th>
<th>Diff2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Kappa</td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Kappa</td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Kappa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>training</td>
<td>221</td>
<td>78.733 %</td>
<td>0.6768</td>
<td>430</td>
<td>79.5349 %</td>
<td>0.6671</td>
<td>430</td>
<td>39.0698 %</td>
<td>0.0849</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>75.5656 %</td>
<td>0.6288</td>
<td>430</td>
<td>79.5349 %</td>
<td>0.6642</td>
<td>430</td>
<td>39.0698 %</td>
<td>0.086</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>79.6296 %</td>
<td>0.6935</td>
<td>108</td>
<td>81.4815 %</td>
<td>0.6842</td>
<td>108</td>
<td>37.037 %</td>
<td>0.091</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.17 Logistic – Sev and Diff only

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th></th>
<th></th>
<th></th>
<th>Diff3</th>
<th></th>
<th></th>
<th></th>
<th>Diff2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Kappa</td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Kappa</td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Kappa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>training</td>
<td>221</td>
<td>77.3756 %</td>
<td>.6561</td>
<td>430</td>
<td>80.4651 %</td>
<td>.678</td>
<td>430</td>
<td>40.2326 %</td>
<td>.1037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>76.0181 %</td>
<td>.6358</td>
<td>430</td>
<td>80 %</td>
<td>.6697</td>
<td>430</td>
<td>36.0465 %</td>
<td>.0409</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50% train</td>
<td>107</td>
<td>71.028 %</td>
<td>.5604</td>
<td>215</td>
<td>80 %</td>
<td>.6672</td>
<td>215</td>
<td>38.6047 %</td>
<td>.084</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66% train</td>
<td>77</td>
<td>80.5195 %</td>
<td>.7064</td>
<td>147</td>
<td>79.5918 %</td>
<td>.664</td>
<td>147</td>
<td>35.3741 %</td>
<td>.0428</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>87.037 %</td>
<td>.8037</td>
<td>108</td>
<td>80.5556 %</td>
<td>.67</td>
<td>108</td>
<td>34.2593 %</td>
<td>.0373</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80% train</td>
<td>41</td>
<td>85.3659 %</td>
<td>.7796</td>
<td>86</td>
<td>81.3953 %</td>
<td>.6764</td>
<td>86</td>
<td>36.0465 %</td>
<td>.0705</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.18 K* – Sev and Diff only
ID3 tree does not allow missing values. Due to missing values in Diff4, only Diff3 and Diff2 could be analyzed. ID3 did not produce any improved results (see table 7.19).

Table 7.19 ID3 tree –Sev and Diff only

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th></th>
<th>Diff3</th>
<th></th>
<th>Diff2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
</tr>
<tr>
<td>training</td>
<td>221</td>
<td>81.9005 %</td>
<td>430</td>
<td>80 %</td>
<td>430</td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>71.4932 %</td>
<td>430</td>
<td>79.0698 %</td>
<td>430</td>
</tr>
<tr>
<td>50% train</td>
<td>107</td>
<td>64.486 %</td>
<td>215</td>
<td>79.0698 %</td>
<td>215</td>
</tr>
<tr>
<td>66% train</td>
<td>77</td>
<td>59.7403 %</td>
<td>147</td>
<td>77.551 %</td>
<td>147</td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>74.0741 %</td>
<td>108</td>
<td>81.4815 %</td>
<td>108</td>
</tr>
<tr>
<td>80% train</td>
<td>41</td>
<td>80.4878 %</td>
<td>86</td>
<td>82.5581 %</td>
<td>86</td>
</tr>
</tbody>
</table>

7.7 **WEKA data discretized into 6 groups**

The dataset is identical with the prior dataset except Sev2, Sev3, Sev4, Diff2, Diff3, and Diff4 was discretized into six groups rather than three. The range is smaller but is still quite large as seen in table 7.20.
Table 7.20 Six group data distribution

<table>
<thead>
<tr>
<th></th>
<th>Diff2</th>
<th>Diff3</th>
<th>Diff4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>-inf-72.5</td>
<td>-inf-.5</td>
<td>-inf-1</td>
</tr>
<tr>
<td>Group 2</td>
<td>72.5-119.5</td>
<td>0.5-93.5</td>
<td>1-93.5</td>
</tr>
<tr>
<td>Group 3</td>
<td>119.5-194.5</td>
<td>93.5-186</td>
<td>93.5-196.5</td>
</tr>
<tr>
<td>Group 4</td>
<td>194.5-340.5</td>
<td>186-337</td>
<td>196.5-321</td>
</tr>
<tr>
<td>Group 5</td>
<td>340.5-572</td>
<td>337-642.5</td>
<td>321-689</td>
</tr>
<tr>
<td>Group 6</td>
<td>572-inf</td>
<td>642.5-inf</td>
<td>689-inf</td>
</tr>
</tbody>
</table>

This test was only run using J48 against Diff4 because the results were significantly lower than when three groups were used so no further testing was performed.

Table 7.21 J48 tree – six groups

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total Number of Instances</td>
<td>Ignored Instances</td>
<td>Correctly Classified Instances</td>
</tr>
<tr>
<td>Training data</td>
<td>221</td>
<td>209</td>
<td>66.9683 %</td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>209</td>
<td>59.276 %</td>
</tr>
<tr>
<td>Split 50% train</td>
<td>107</td>
<td>108</td>
<td>50.4673 %</td>
</tr>
<tr>
<td>Split 66% train</td>
<td>77</td>
<td>70</td>
<td>49.3506 %</td>
</tr>
<tr>
<td>Split 80% train</td>
<td>41</td>
<td>45</td>
<td>60.9756 %</td>
</tr>
</tbody>
</table>

7.8 Progression of Severity

In order to determine if the progression of severity assists in determining TTF, another attribute was added to demonstrate a decrease, increase or stable severity. Sev_1-2 is the severity change between Sev1, Sev2, Sev_2-3 is the severity change between Sev2 and Sev3, and so forth. The data was analyzed to show the distribution of each type of severity in Sev2, Sev3, and Sev4. Sev2 contains the most “slight” records as compared to that in Sev3 and Sev4. Sev2 includes a comparable number of “moderate”, “serious”, and “extreme” records. Sev3
includes the most “extreme” records as compared to that in Sev2 and Sev4. Sev4 contains the most “extreme” records as compared to other records within Sev4. See table 7.22 and figure 7.2

Table 7.22 Severity distribution

<table>
<thead>
<tr>
<th></th>
<th>none</th>
<th>slight</th>
<th>moderate</th>
<th>serious</th>
<th>extreme</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sev2</td>
<td>2</td>
<td>52</td>
<td>118</td>
<td>125</td>
<td>133</td>
<td>430</td>
</tr>
<tr>
<td>Sev3</td>
<td>0</td>
<td>16</td>
<td>59</td>
<td>68</td>
<td>287</td>
<td>430</td>
</tr>
<tr>
<td>Sev4</td>
<td>1</td>
<td>10</td>
<td>39</td>
<td>42</td>
<td>129</td>
<td>221</td>
</tr>
</tbody>
</table>

Fig 7.2 Severity distribution graph

The data shows that the severity increases between Sev1 and Sev2 in almost all records and about 50% of the time between Sev2 and Sev3. Sev _3-4 has the highest percent of stable records. See table 7.23 and figure 7.3

Table 7.23 Severity progression

<table>
<thead>
<tr>
<th></th>
<th>decrease</th>
<th>Stable</th>
<th>increase</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sev 1-2</td>
<td>0</td>
<td>2</td>
<td>428</td>
<td>430</td>
</tr>
<tr>
<td>Sev 2-3</td>
<td>39</td>
<td>184</td>
<td>207</td>
<td>430</td>
</tr>
</tbody>
</table>
Figure 7.3 Severity progression graph

![Severity Progression Graph](image)

The dataset includes the attributes: Sev2, Sev3, Sev4, Diff2, Diff3, Diff4, Sev_1-2, Sev_2-3, and Sev_3-4. With J48 tree, Diff4 was better and Diff3 and Diff2 remained the same. With Multilayer Perceptron, modified options applied, Diff4 was worse, Diff3 was better, and Diff2 was identical when comparing results in table 7.16 Multilayer Perceptron – Sev and Diff only – modified results. With Logistic, Diff4 and Diff3 were worse and Diff2 was just slightly better when comparing with table 7.17 Logistic – Sev and Diff only. With K*, Diff4 and Diff3 was worse and Diff2 remained the same when compared to table 7.18 K* - Sev and Diff only. With this dataset, J48 had the best results for Diff4, Multilayer Perceptron and Logistic regression had the best identical results for Diff3 and Multilayer Perceptron had the best results for Diff2. The dataset with Sev and Diff, tables 7.16 and 7.17, produced the best overall results for Diff4 and Diff3 with the Logistic regression and Diff2 with the Multilayer Perceptron. The results from the progression analysis are displayed in table 7.24, 7.25, 7.26, and 7.27.

Table 7.24 J48 tree – progression

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th></th>
<th></th>
<th>Diff3</th>
<th></th>
<th></th>
<th>Diff2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td></td>
<td></td>
</tr>
<tr>
<td>training</td>
<td>221</td>
<td>78.2805 %</td>
<td>430</td>
<td>80 %</td>
<td>430</td>
<td>40.2326 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>74.6606 %</td>
<td>430</td>
<td>79.0698 %</td>
<td>430</td>
<td>36.0465 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>77.7778 %</td>
<td>108</td>
<td>80.5556 %</td>
<td>108</td>
<td>34.2593 %</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7.25 Multilayer Perceptron – modified options – progression

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th>Diff3</th>
<th>Diff2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
</tr>
<tr>
<td>training</td>
<td>221</td>
<td>80.0905 %</td>
<td>430</td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>71.4932 %</td>
<td>430</td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>71.4932 %</td>
<td>108</td>
</tr>
</tbody>
</table>

Table 7.26 Logistic – progression

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th>Diff3</th>
<th>Diff2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
</tr>
<tr>
<td>training</td>
<td>221</td>
<td>77.8281 %</td>
<td>430</td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>71.9457 %</td>
<td>430</td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>77.7778 %</td>
<td>108</td>
</tr>
</tbody>
</table>

Table 7.27 K* – progression

<table>
<thead>
<tr>
<th></th>
<th>Diff4</th>
<th>Diff3</th>
<th>Diff2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total # of Instances</td>
<td>Correctly Classified Instances</td>
<td>Total # of Instances</td>
</tr>
<tr>
<td>training</td>
<td>221</td>
<td>83.2579 %</td>
<td>430</td>
</tr>
<tr>
<td>10-fold</td>
<td>221</td>
<td>68.7783 %</td>
<td>430</td>
</tr>
<tr>
<td>75% train</td>
<td>54</td>
<td>68.7783 %</td>
<td>108</td>
</tr>
</tbody>
</table>

7.9 WEKA Associations - data discretized into 3 groups

The dataset includes the following attributes: Sev2, Sev3, Sev4, Diff2, Diff3, and Diff4. The rules produced by Apriori showed an association between Diff4 having a value in “-inf-1” and Sev4 being “extreme”, which is a fact that when a Diff# is zero, Sev# is “extreme”. The attribute Diff2 having a value in “118-340.5”, the attribute Diff3 having a value in “-inf-.5” and Sev3 being extreme, shows there is a decrease in value from Diff# to Diff#-1 progressing to a Severity of “extreme”. Rule 3 showed that if Sev2 is “extreme” and Diff3 is zero, Sev3 is
“extreme”, Rule 4 shows that if Diff4 is large, Diff2 is also large, and Rule 5 shows that if Diff3 and Diff4 are large, Diff2 is also large. Predictive Apriori produced similar results listed below.

Apriori - Best rules found:

1. Diff4='(-inf-1]' 83 ==> Sev4=extreme 83  conf:(1)
2. Diff2='(118-340.5]' Diff3='(-inf-0.5]' 75 ==> Sev3=extreme 75  conf:(1)
3. Sev2=extreme Diff3='(-inf-0.5]' 73 ==> Sev3=extreme 73  conf:(1)
4. Diff4='(237.5-inf)' 69 ==> Diff2='(340.5-inf)' 69  conf:(1)
5. Diff3='(268.5-inf)' Diff4='(237.5-inf)' 68 ==> Diff2='(340.5-inf)' 68  conf:(1)
6. Diff3='(0.5-268.5]' Diff4='(-inf-1]' 65 ==> Sev4=extreme 65  conf:(1)
7. Diff3='(-inf-0.5]' 214 ==> Sev3=extreme 212  conf:(0.99)
8. Diff4='(237.5-inf)' 69 ==> Diff2='(340.5-inf)' Diff3='(268.5-inf)' 68  conf:(0.99)
9. Diff2='(340.5-inf)' Diff4='(237.5-inf)' 69 ==> Diff3='(268.5-inf)' 68  conf:(0.99)
10. Diff4='(237.5-inf)' 69 ==> Diff3='(268.5-inf)' 68  conf:(0.99)

PredictiveApriori – Best rules found:

1. Diff3='(-inf-0.5]' 214 ==> Sev3=extreme 212  acc:(0.99495)
2. Diff4='(-inf-1]' 83 ==> Sev4=extreme 83  acc:(0.99494)
3. Diff4='(237.5-inf)' 69 ==> Diff2='(340.5-inf)' 69  acc:(0.99489)
4. Diff4='(237.5-inf)' 69 ==> Diff2='(340.5-inf)' Diff3='(268.5-inf)' 68  acc:(0.99403)
5. Diff3='(0.5-268.5]' 20 ==> Sev4=extreme 20  acc:(0.99337)
6. Sev3=moderate Sev4=moderate Diff2='(340.5-inf)' 16 ==> Diff3='(268.5-inf)' 16  acc:(0.99223)
7. Sev3=moderate Sev4=moderate Diff3='(268.5-inf)' 16 ==> Diff2='(340.5-inf)' 16  acc:(0.99223)
8. Sev4=significant Diff3='(0.5-268.5]' 15 ==> Diff4='(1-237.5]' 15  acc:(0.99176)
9. Sev2=slight Sev3=extreme Diff2='(118-340.5]' 15 ==> Diff3='(-inf-0.5]' 15  acc:(0.99176)
10. Sev2=moderate Sev4=moderate Diff2='(340.5-inf)' 15 ==> Diff3='(268.5-inf)' 15  acc:(0.99176)
11. Sev2=moderate Sev4=moderate Diff3='(268.5-inf)' 15 ==> Diff2='(340.5-inf)' 15  acc:(0.99176)
12. Sev2=significant Diff2='(118-340.5]' Diff4='(-inf-1]' 14 ==> Sev4=extreme Diff3='(0.5-268.5]' 14  acc:(0.99116)
13. Sev4=significant Diff2='(118-340.5]' 12 ==> Diff4='(1-237.5]' 12  acc:(0.98939)
14. Sev2=extreme Sev4=extreme Diff4='(237.5-inf)' 12 ==> Sev3=extreme Diff2='(340.5-inf)' 12  acc:(0.98939)
15. Sev2=extreme Sev4=extreme Diff4='(237.5-inf)' 12 ==> Sev3=extreme Diff3='(268.5-inf)' 12  acc:(0.98939)
16. Sev3=significant Diff2='(118-340.5]' Diff4='(-inf-1]' 12 ==> Sev4=extreme Diff3='(0.5-268.5]' 12 acc:(0.98939)
17. Sev4=moderate Diff2='(340.5-inf)' 32 ==> Diff3='(268.5-inf)' 31 acc:(0.98741)
18. Sev2=significant Diff3='(268.5-inf)' 30 ==> Diff2='(340.5-inf)' 29 acc:(0.98585)
19. Sev2=moderate Sev3=extreme Diff2='(-inf-118]' 24 ==> Diff3='(-inf-0.5]' 23 acc:(0.97697)
20. Sev3=extreme Sev4=extreme Diff3='(268.5-inf)' 24 ==> Diff2='(340.5-inf)' 23 acc:(0.97697)
21. Sev2=significant Sev3=extreme Diff2='(-inf-118]' 40 ==> Diff3='(-inf-0.5]' 38 acc:(0.96093)
22. Sev2=extreme Diff2='(-inf-118]' 61 ==> Sev3=extreme 58 acc:(0.95323)
23. Sev2=extreme Diff3='(0.5-268.5]' 34 ==> Sev4=extreme 32 acc:(0.94227)
24. Sev3=moderate Diff3='(268.5-inf)' 33 ==> Diff2='(340.5-inf)' 31 acc:(0.93793)
25. Sev4=moderate Diff3='(268.5-inf)' 33 ==> Diff2='(340.5-inf)' 31 acc:(0.93793)
26. Diff3='(268.5-inf)' 108 ==> Diff2='(340.5-inf)' 100 acc:(0.92740)
27. Diff2='(-inf-118]' 143 ==> Sev3=extreme 132 acc:(0.9247)
28. Sev3=extreme Diff2='(-inf-118]' 132 ==> Diff3='(-inf-0.5]' 121 acc:(0.91709)
29. Sev2=significant Diff2='(-inf-118]' 43 ==> Diff3='(-inf-0.5]' 40 acc:(0.9155)
30. Sev2=extreme 133 ==> Sev3=extreme 120 acc:(0.89766)
31. Sev2=none 2 ==> Diff2='(118-340.5]' 2 acc:(0.85914)

### 7.10 Summary of Results

The initial concept of determining TTF for VibrationStandardEquipmentID, VibrationStandardEquipmentGroupID, VibrationStandardDiagnosisGroupID and VibrationStandardDiagnosisID and using continuous, numerical values for the Severity, DaysToFailure, and TTF were much too specific for the data available. Through extensive data analysis, it is apparent there are not enough cases when analysis was performed on VibrationStandardEquipmentID, VibrationStandardEquipmentGroupID, VibrationStandardDiagnosisID, and VibrationStandardDiagnosisGroupID. There are too few cases specific to each of these attributes. Even when attempting to generalize the VibrationStandardEquipmentID and VibrationStandardDiagnosisID into groups, the groups were still too specific. This was shown when the best results were generated from the dataset that only included the Sev# and Diff# discretized into equi-depth groups.

In order to gain accuracy, precision would be lost. When the dataset was discretized to six groups rather than three groups in an attempt to improve precision, the group range was much
smaller but the accuracy dropped significantly. With the data available, high precision and accuracy is not possible. With the addition of the progression of severity, there was no improvement in accuracy but it produced just slightly worse accuracy.

When applying the progression of severity, the results were quite similar but showed no improvements. The progression of severity also predicts well for Diff3 and Diff4 but not for Diff2.

Since the best results were produced by the Logistic regression algorithm, a system should use this type of classification. Multilayer Perceptron would be the second best algorithm of choice to develop. The system should not base an estimate on Diff2 because it produced less than 40% correctly classified instances. Diff3 was the best attribute to determine TTF.
Chapter 8
DISCUSSION

Greitzer and Ferryman studied the “investigation of a generalized statistical method for characterizing and predicting system degradation” [5]. This project predicted ‘failure’ using Figure of Merit (FOM), the system’s quantified current degree of fault, progression of fault, and the level of fault that will produce a failure of the system. “The specification of these factors, which is necessary to perform diagnostics and prognostics, is typically done through engineering/analytical studies such as Failure Modes and Effects Analysis. These analyses and expert judgments yield descriptions of how the system fails, what faults can be measured given available sensors and the values expected for these sensors when these failures occur” [5]. The issue with the use of FOM is that some databases, including the database used in this project, does not monitor or store FOM.

Greitzer, Stahlman, Ferryman, Wilson, Kangas, and Sisk studied “health monitoring of complex mechanical systems for diagnostics, prognostics, and maintenance scheduling” [6]. The project, the Pacific Northwest National Laboratory’s LEAP project, used linear regression to predict the time to failure. To improve this time to failure, they estimated the “candidate statistical methods include multivariate regression, Bayesian regression methods, time-series analysis, and discrimination or clustering analysis” [6]. The project focused on two different types of applications. One application used real-time data for gas turbine engines and another application used oil analysis data on large diesel electric engines. This system is quite limiting since it is applied to only these two application areas.
Chapter 9

FUTURE WORK

Begin by converting the numerical values, Sev# and Diff# to discretized nominal values. Use just the Sev# and Diff# values and apply the Logistic regression or Multilayer Perceptron algorithm in predicting TTF.

When evaluating the cases shown in Appendix C.3, the Machines with the same MID showed a broad range of TTF values. This may be due to a machine having some maintenance performed and postponing TTF. Incorporating repair history while determining cases can potentially improve the accuracy of the TTF values.

Another aspect to investigate is to have a broader acceptable range when the fault is first detected and narrowing the range as the fault severity increases and TTF approaches. It is not as important to know what day a machine will fail when the system is estimating a TTF that is still over six months in the future.
Chapter 10

EDUCATIONAL STATEMENT

This project drew upon the knowledge obtained from several courses in the Computing and Software Systems Master’s Curriculum and my work experience. Work experience built the foundation to work with databases and understand database structure. This project had a practical application of efficiently accessing data. TCSS543 Advanced Algorithms built the foundation for the mathematical aspects of this project. Algorithms and statistical analysis were studied and researched to achieve the goal of this project. TCSS598 Master’s Seminar greatly helped in writing, researching, and reading technical papers. TCSS560 Software Engineering was very helpful in designing the project. TCSS555 Data Mining and TCSS435 Artificial Intelligence stated the need to extract knowledge from databases and presented the methodologies for attaining new knowledge. The project allowed me to apply many concepts introduced throughout the Master’s program.
Chapter 11

CONCLUSION

This paper has shown that determining TTF using CBR and regression with the initial generalization groupings and continuous values produced low accuracy rates. Good results were obtained by discretizing in small number of groups with equi-depths. Due to the low numbers of cases for each specific MID and Diagnoses, it is best to determine TTF based on Sev# and Diff# only. Many databases at DLI Engineering do not include repair history information but the database used in this project does include repair history information. Future work will determine if incorporating repair history can improve accuracy.
BIBLIOGRAPHY


Appendix A
DATABASE

A.1 Partial Database Schema
Appendix B
PRE-PROCESSING TOOL

B.1 MID Grouping

MID Grouping tab is used to assist the user in grouping MIDs.

B.2 Diagnosis Grouping

Diagnosis Grouping tab is used to assist the user in grouping Diagnoses.
B.3 Selection

The Selection tab assists the user in selecting certain diagnoses or MIDs based on different criteria entered.
Appendix C
TTF SYSTEM

C.1 VibrationCase – raw data (partial)

<table>
<thead>
<tr>
<th>CaseID</th>
<th>VibStandardEquipmentID</th>
<th>VibStandardDiagnosisID</th>
<th>DiagnosisGroupID</th>
<th>VibStandardEquipmentGroupID</th>
<th>TotalDaysToFailure</th>
<th>CaseType</th>
<th>CaseIsActive</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,483,238</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>14,491,459</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>16,528,398</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>18,306,481</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>24,882,397</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>26,817,484</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>28,306,481</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>30,743,397</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>32,845,605</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>356,528</td>
<td></td>
<td></td>
<td>1</td>
<td>1390</td>
<td>1,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>357,743</td>
<td></td>
<td></td>
<td>1</td>
<td>344</td>
<td>1,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>358,727</td>
<td></td>
<td></td>
<td>1</td>
<td>864</td>
<td>1,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>359,275</td>
<td></td>
<td></td>
<td>10</td>
<td>566</td>
<td>1,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>360,571</td>
<td></td>
<td></td>
<td>1</td>
<td>1518</td>
<td>1,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>361,1148</td>
<td></td>
<td></td>
<td>1</td>
<td>643</td>
<td>1,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>362,632</td>
<td></td>
<td></td>
<td>1</td>
<td>1361</td>
<td>1,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>363,202</td>
<td></td>
<td></td>
<td>1</td>
<td>403</td>
<td>1,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>364,214</td>
<td></td>
<td></td>
<td>1</td>
<td>985</td>
<td>1,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>365,217</td>
<td></td>
<td></td>
<td>1</td>
<td>828</td>
<td>1,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>366,79</td>
<td></td>
<td></td>
<td>10</td>
<td>484</td>
<td>1,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>398,17</td>
<td></td>
<td></td>
<td>11</td>
<td>1725</td>
<td>2,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>399,132</td>
<td></td>
<td></td>
<td>13</td>
<td>159</td>
<td>2,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>400,151</td>
<td></td>
<td></td>
<td>18</td>
<td>811</td>
<td>2,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>401,180</td>
<td></td>
<td></td>
<td>21</td>
<td>506</td>
<td>2,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>402,294</td>
<td></td>
<td></td>
<td>22</td>
<td>715</td>
<td>2,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>403,204</td>
<td></td>
<td></td>
<td>31</td>
<td>449</td>
<td>2,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>404,398</td>
<td></td>
<td></td>
<td>35</td>
<td>314</td>
<td>2,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>405,106</td>
<td></td>
<td></td>
<td>37</td>
<td>1019</td>
<td>2,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>406,475</td>
<td></td>
<td></td>
<td>38</td>
<td>175</td>
<td>2,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>448,1</td>
<td></td>
<td></td>
<td>1</td>
<td>1018</td>
<td>3,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>449,1</td>
<td></td>
<td></td>
<td>1</td>
<td>1257</td>
<td>3,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>450,9</td>
<td></td>
<td></td>
<td>9</td>
<td>1280</td>
<td>3,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>451,1</td>
<td></td>
<td></td>
<td>1</td>
<td>13,972</td>
<td>3,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>452,10</td>
<td></td>
<td></td>
<td>21</td>
<td>451</td>
<td>3,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>453,1</td>
<td></td>
<td></td>
<td>1</td>
<td>35,403</td>
<td>3,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>454,9</td>
<td></td>
<td></td>
<td>37</td>
<td>229</td>
<td>3,1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>455,1</td>
<td></td>
<td></td>
<td>40</td>
<td>1016</td>
<td>3,1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
### C.2 VibrationCaseTest – raw data

Column Names: TestCaseID, VesselID, EquipmentID, TestResultID, CaseID, DayPosition, VibDiagnosisSeverityIndex, CaseTestIsActive

<table>
<thead>
<tr>
<th>TestCaseID</th>
<th>VesselID</th>
<th>EquipmentID</th>
<th>TestResultID</th>
<th>CaseID</th>
<th>DayPosition</th>
<th>VibDiagnosisSeverityIndex</th>
<th>CaseTestIsActive</th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>2</td>
<td>2613</td>
<td>10224</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>62</td>
<td>2</td>
<td>2613</td>
<td>10323</td>
<td>10</td>
<td>275</td>
<td>260</td>
<td>1</td>
</tr>
<tr>
<td>63</td>
<td>2</td>
<td>2613</td>
<td>10561</td>
<td>10</td>
<td>560</td>
<td>630</td>
<td>1</td>
</tr>
<tr>
<td>74</td>
<td>3</td>
<td>6636</td>
<td>12922</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>75</td>
<td>3</td>
<td>6636</td>
<td>12921</td>
<td>14</td>
<td>414</td>
<td>276</td>
<td>1</td>
</tr>
<tr>
<td>76</td>
<td>3</td>
<td>6636</td>
<td>13187</td>
<td>14</td>
<td>1095</td>
<td>714</td>
<td>1</td>
</tr>
<tr>
<td>84</td>
<td>3</td>
<td>7384</td>
<td>15635</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>85</td>
<td>3</td>
<td>7384</td>
<td>16485</td>
<td>16</td>
<td>212</td>
<td>312</td>
<td>1</td>
</tr>
<tr>
<td>86</td>
<td>3</td>
<td>7384</td>
<td>16752</td>
<td>16</td>
<td>281</td>
<td>333</td>
<td>1</td>
</tr>
<tr>
<td>87</td>
<td>3</td>
<td>7384</td>
<td>17284</td>
<td>16</td>
<td>567</td>
<td>617</td>
<td>1</td>
</tr>
<tr>
<td>2012</td>
<td>71</td>
<td>935</td>
<td>4367</td>
<td>399</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2009</td>
<td>71</td>
<td>935</td>
<td>4843</td>
<td>399</td>
<td>33</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>2008</td>
<td>71</td>
<td>949</td>
<td>4844</td>
<td>399</td>
<td>35</td>
<td>45</td>
<td>1</td>
</tr>
<tr>
<td>2011</td>
<td>67</td>
<td>935</td>
<td>3243</td>
<td>399</td>
<td>60</td>
<td>85</td>
<td>1</td>
</tr>
<tr>
<td>2010</td>
<td>67</td>
<td>949</td>
<td>3166</td>
<td>399</td>
<td>69</td>
<td>45</td>
<td>1</td>
</tr>
<tr>
<td>2006</td>
<td>72</td>
<td>949</td>
<td>2745</td>
<td>399</td>
<td>70</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>2007</td>
<td>72</td>
<td>935</td>
<td>1125</td>
<td>399</td>
<td>114</td>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td>2013</td>
<td>71</td>
<td>935</td>
<td>4428</td>
<td>399</td>
<td>159</td>
<td>995</td>
<td>1</td>
</tr>
<tr>
<td>2060</td>
<td>13</td>
<td>734</td>
<td>17693</td>
<td>402</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2058</td>
<td>59</td>
<td>865</td>
<td>1728</td>
<td>402</td>
<td>155</td>
<td>144</td>
<td>1</td>
</tr>
<tr>
<td>2059</td>
<td>41</td>
<td>11165</td>
<td>15167</td>
<td>402</td>
<td>538</td>
<td>234</td>
<td>1</td>
</tr>
<tr>
<td>2061</td>
<td>13</td>
<td>734</td>
<td>18901</td>
<td>402</td>
<td>715</td>
<td>1050</td>
<td>1</td>
</tr>
<tr>
<td>2066</td>
<td>76</td>
<td>657</td>
<td>2380</td>
<td>403</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2067</td>
<td>76</td>
<td>657</td>
<td>2477</td>
<td>403</td>
<td>81</td>
<td>760</td>
<td>1</td>
</tr>
<tr>
<td>2062</td>
<td>76</td>
<td>758</td>
<td>2484</td>
<td>403</td>
<td>256</td>
<td>450</td>
<td>1</td>
</tr>
<tr>
<td>2063</td>
<td>76</td>
<td>757</td>
<td>2483</td>
<td>403</td>
<td>276</td>
<td>443</td>
<td>1</td>
</tr>
<tr>
<td>2064</td>
<td>41</td>
<td>668</td>
<td>14882</td>
<td>403</td>
<td>405</td>
<td>293</td>
<td>1</td>
</tr>
<tr>
<td>2065</td>
<td>36</td>
<td>768</td>
<td>12860</td>
<td>403</td>
<td>449</td>
<td>257</td>
<td>1</td>
</tr>
<tr>
<td>2109</td>
<td>20</td>
<td>11918</td>
<td>14760</td>
<td>406</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2103</td>
<td>74</td>
<td>11905</td>
<td>660</td>
<td>406</td>
<td>1</td>
<td>343</td>
<td>1</td>
</tr>
<tr>
<td>2106</td>
<td>50</td>
<td>11905</td>
<td>1485</td>
<td>406</td>
<td>100</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>2110</td>
<td>20</td>
<td>11918</td>
<td>15154</td>
<td>406</td>
<td>100</td>
<td>789</td>
<td>1</td>
</tr>
<tr>
<td>2104</td>
<td>52</td>
<td>11930</td>
<td>2165</td>
<td>406</td>
<td>121</td>
<td>77</td>
<td>1</td>
</tr>
<tr>
<td>2107</td>
<td>34</td>
<td>11905</td>
<td>17013</td>
<td>406</td>
<td>141</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>2105</td>
<td>52</td>
<td>11905</td>
<td>2502</td>
<td>406</td>
<td>164</td>
<td>174</td>
<td>1</td>
</tr>
<tr>
<td>2108</td>
<td>20</td>
<td>11905</td>
<td>18135</td>
<td>406</td>
<td>175</td>
<td>486</td>
<td>1</td>
</tr>
<tr>
<td>2132</td>
<td>37</td>
<td>970</td>
<td>12818</td>
<td>409</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2133</td>
<td>37</td>
<td>970</td>
<td>12921</td>
<td>409</td>
<td>11</td>
<td>2194</td>
<td>1</td>
</tr>
<tr>
<td>2130</td>
<td>17</td>
<td>957</td>
<td>15727</td>
<td>409</td>
<td>77</td>
<td>1551</td>
<td>1</td>
</tr>
<tr>
<td>2129</td>
<td>17</td>
<td>970</td>
<td>15720</td>
<td>409</td>
<td>89</td>
<td>2191</td>
<td>1</td>
</tr>
<tr>
<td>2128</td>
<td>37</td>
<td>957</td>
<td>18637</td>
<td>409</td>
<td>233</td>
<td>311</td>
<td>1</td>
</tr>
<tr>
<td>2131</td>
<td>14</td>
<td>970</td>
<td>14254</td>
<td>409</td>
<td>265</td>
<td>231</td>
<td>1</td>
</tr>
</tbody>
</table>
### C.3 Case Evaluation

<table>
<thead>
<tr>
<th>Case Type</th>
<th>MID</th>
<th>Dx</th>
<th>Case Count</th>
<th>TTF Values</th>
<th>Mean</th>
<th>Difference</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>229</td>
<td>489</td>
<td>2</td>
<td>818, 820</td>
<td>819.00</td>
<td>2</td>
<td>1.00</td>
</tr>
<tr>
<td>0</td>
<td>548</td>
<td>398</td>
<td>2</td>
<td>217, 223</td>
<td>220.00</td>
<td>6</td>
<td>3.00</td>
</tr>
<tr>
<td>0</td>
<td>1239</td>
<td>395</td>
<td>2</td>
<td>178, 193</td>
<td>185.50</td>
<td>15</td>
<td>7.48</td>
</tr>
<tr>
<td>0</td>
<td>920</td>
<td>211</td>
<td>2</td>
<td>528, 552</td>
<td>540.00</td>
<td>24</td>
<td>12.00</td>
</tr>
<tr>
<td>0</td>
<td>229</td>
<td>461</td>
<td>2</td>
<td>191, 254</td>
<td>222.50</td>
<td>63</td>
<td>31.50</td>
</tr>
<tr>
<td>0</td>
<td>1685</td>
<td>459</td>
<td>2</td>
<td>128, 192</td>
<td>160.00</td>
<td>64</td>
<td>32.00</td>
</tr>
<tr>
<td>0</td>
<td>1192</td>
<td>395</td>
<td>2</td>
<td>276, 368</td>
<td>322.00</td>
<td>92</td>
<td>46.00</td>
</tr>
<tr>
<td>0</td>
<td>1886</td>
<td>363</td>
<td>2</td>
<td>1065, 753</td>
<td>909.00</td>
<td>312</td>
<td>156.00</td>
</tr>
<tr>
<td>0</td>
<td>637</td>
<td>486</td>
<td>2</td>
<td>311, 427</td>
<td>369.00</td>
<td>116</td>
<td>58.00</td>
</tr>
<tr>
<td>0</td>
<td>621</td>
<td>598</td>
<td>2</td>
<td>154, 281</td>
<td>217.50</td>
<td>127</td>
<td>63.50</td>
</tr>
<tr>
<td>0</td>
<td>886</td>
<td>395</td>
<td>2</td>
<td>341, 470</td>
<td>405.50</td>
<td>129</td>
<td>64.50</td>
</tr>
<tr>
<td>0</td>
<td>65</td>
<td>508</td>
<td>3</td>
<td>674, 732, 867</td>
<td>757.67</td>
<td>135</td>
<td>80.86</td>
</tr>
<tr>
<td>0</td>
<td>365</td>
<td>238</td>
<td>2</td>
<td>504, 666</td>
<td>585.00</td>
<td>162</td>
<td>81.00</td>
</tr>
<tr>
<td>1</td>
<td>79</td>
<td>10</td>
<td>2</td>
<td>484, 653</td>
<td>568.50</td>
<td>169</td>
<td>84.50</td>
</tr>
<tr>
<td>0</td>
<td>1173</td>
<td>36</td>
<td>2</td>
<td>247, 419</td>
<td>333.00</td>
<td>172</td>
<td>86.00</td>
</tr>
<tr>
<td>0</td>
<td>1314</td>
<td>459</td>
<td>3</td>
<td>372, 553</td>
<td>497.67</td>
<td>181</td>
<td>89.07</td>
</tr>
<tr>
<td>0</td>
<td>202</td>
<td>34</td>
<td>2</td>
<td>110, 289</td>
<td>199.50</td>
<td>179</td>
<td>89.50</td>
</tr>
<tr>
<td>0</td>
<td>1596</td>
<td>198</td>
<td>5</td>
<td>151, 92, 74, 316, 227</td>
<td>172.00</td>
<td>242</td>
<td>89.67</td>
</tr>
<tr>
<td>0</td>
<td>204</td>
<td>317</td>
<td>2</td>
<td>174, 393</td>
<td>283.50</td>
<td>219</td>
<td>109.50</td>
</tr>
<tr>
<td>0</td>
<td>222</td>
<td>54</td>
<td>2</td>
<td>64, 98</td>
<td>366.00</td>
<td>272</td>
<td>136.00</td>
</tr>
<tr>
<td>0</td>
<td>1939</td>
<td>113</td>
<td>2</td>
<td>131, 418</td>
<td>274.50</td>
<td>287</td>
<td>143.50</td>
</tr>
<tr>
<td>0</td>
<td>59</td>
<td>398</td>
<td>2</td>
<td>333, 624</td>
<td>478.50</td>
<td>291</td>
<td>145.50</td>
</tr>
<tr>
<td>0</td>
<td>91</td>
<td>398</td>
<td>3</td>
<td>626, 939, 956</td>
<td>840.33</td>
<td>330</td>
<td>151.72</td>
</tr>
<tr>
<td>0</td>
<td>2091</td>
<td>198</td>
<td>6</td>
<td>359, 259, 642, 254, 624, 218</td>
<td>392.67</td>
<td>424</td>
<td>175.32</td>
</tr>
<tr>
<td>0</td>
<td>1453</td>
<td>459</td>
<td>2</td>
<td>467, 827</td>
<td>647.00</td>
<td>360</td>
<td>180.00</td>
</tr>
<tr>
<td>0</td>
<td>1182</td>
<td>481</td>
<td>2</td>
<td>194, 559</td>
<td>376.50</td>
<td>365</td>
<td>182.50</td>
</tr>
<tr>
<td>MID</td>
<td>Dx</td>
<td>Case Count</td>
<td>TTF Values</td>
<td>Mean</td>
<td>Difference</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>----</td>
<td>------------</td>
<td>------------</td>
<td>------</td>
<td>------------</td>
<td>----</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1979</td>
<td>2</td>
<td>171, 547</td>
<td>359.00</td>
<td>376</td>
<td>188.00</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1587</td>
<td>459</td>
<td>2</td>
<td>15, 710, 386, 788</td>
<td>587.00</td>
<td>402</td>
<td>201.00</td>
</tr>
<tr>
<td>0</td>
<td>2091</td>
<td>206</td>
<td>5</td>
<td>222, 213, 764, 439, 219</td>
<td>371.40</td>
<td>551</td>
<td>214.17</td>
</tr>
<tr>
<td>0</td>
<td>1310</td>
<td>459</td>
<td>2</td>
<td>567, 1012</td>
<td>789.50</td>
<td>445</td>
<td>222.50</td>
</tr>
<tr>
<td>0</td>
<td>204</td>
<td>395</td>
<td>2</td>
<td>375, 841</td>
<td>608.00</td>
<td>466</td>
<td>233.00</td>
</tr>
<tr>
<td>0</td>
<td>366</td>
<td>459</td>
<td>2</td>
<td>834, 1321</td>
<td>1077.50</td>
<td>862</td>
<td>243.50</td>
</tr>
<tr>
<td>0</td>
<td>483</td>
<td>238</td>
<td>4</td>
<td>460, 1059, 1170, 1422</td>
<td>1052.75</td>
<td>862</td>
<td>313.42</td>
</tr>
<tr>
<td>0</td>
<td>920</td>
<td>198</td>
<td>4</td>
<td>148, 267, 930, 157</td>
<td>375.50</td>
<td>782</td>
<td>323.55</td>
</tr>
<tr>
<td>1</td>
<td>2091</td>
<td>10</td>
<td>2</td>
<td>642, 1494</td>
<td>1068.00</td>
<td>852</td>
<td>426.00</td>
</tr>
<tr>
<td>0</td>
<td>1311</td>
<td>459</td>
<td>2</td>
<td>272, 1323</td>
<td>797.50</td>
<td>1051</td>
<td>525.50</td>
</tr>
<tr>
<td>0</td>
<td>1155</td>
<td>398</td>
<td>7</td>
<td>173, 181, 369, 389, 612, 1262, 1723</td>
<td>672.71</td>
<td>1550</td>
<td>550.10</td>
</tr>
<tr>
<td>0</td>
<td>186</td>
<td>395</td>
<td>2</td>
<td>206, 1322</td>
<td>764.00</td>
<td>1116</td>
<td>558.00</td>
</tr>
<tr>
<td>0</td>
<td>169</td>
<td>198</td>
<td>9</td>
<td>166, 230, 253, 309, 500, 540, 718, 1748, 1752</td>
<td>690.67</td>
<td>1586</td>
<td>589.62</td>
</tr>
<tr>
<td>0</td>
<td>359</td>
<td>398</td>
<td>4</td>
<td>320, 414, 635, 2007</td>
<td>844.00</td>
<td>1687</td>
<td>681.13</td>
</tr>
<tr>
<td>0</td>
<td>79</td>
<td>282</td>
<td>4</td>
<td>357, 609, 670, 2142</td>
<td>944.50</td>
<td>1785</td>
<td>701.26</td>
</tr>
</tbody>
</table>

Note:
MID: VibStandardEquipmentID
Dx: VibrationStandardDiagnosisID
Case Count: Number of cases
TTF Values: TTF for each case
Mean: Mean TTF
Difference: Maximum TTF minus Minimum TTF
SD: Standard deviation
C.4 Case Evaluation Distribution
### C.5 Training Results

<table>
<thead>
<tr>
<th>Case</th>
<th>Training - Database 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Three months</td>
</tr>
<tr>
<td>Type 0</td>
<td>280/280</td>
</tr>
<tr>
<td>Type 1</td>
<td>33/33</td>
</tr>
<tr>
<td>Type 2</td>
<td>37/39</td>
</tr>
<tr>
<td>Type 3</td>
<td>23/27</td>
</tr>
<tr>
<td>Total</td>
<td>373/379</td>
</tr>
</tbody>
</table>

### C.6 Test knowing TTF

<table>
<thead>
<tr>
<th>Case</th>
<th>Test – Database 2</th>
<th>Test (Database 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New tests in db</td>
<td>Three months</td>
</tr>
<tr>
<td>Type 0</td>
<td>23/49</td>
<td>49.94%</td>
</tr>
<tr>
<td>Type 1</td>
<td>1/2</td>
<td>50.00%</td>
</tr>
<tr>
<td>Type 2</td>
<td>4/11</td>
<td>36.36%</td>
</tr>
<tr>
<td>Type 3</td>
<td>8/20</td>
<td>40.00%</td>
</tr>
<tr>
<td>Total</td>
<td>36/82</td>
<td>43.90%</td>
</tr>
</tbody>
</table>
### C.7 Test knowing TTF-Case selection breakdown

<table>
<thead>
<tr>
<th>Case Selected</th>
<th>3 months</th>
<th>2 months</th>
<th>Out of range</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Percent</td>
<td>Count</td>
<td>Percent</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>100.0%</td>
<td>1</td>
<td>50.00%</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>43.48%</td>
<td>6</td>
<td>26.09%</td>
</tr>
<tr>
<td>3</td>
<td>47</td>
<td>44.76%</td>
<td>30</td>
<td>28.57%</td>
</tr>
<tr>
<td>4</td>
<td>47</td>
<td>58.75%</td>
<td>34</td>
<td>42.50%</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>75.00%</td>
<td>6</td>
<td>50.00%</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>45.45%</td>
<td>9</td>
<td>40.90%</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>50.00%</td>
<td>8</td>
<td>40.00%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>286</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### C.8 Initial test

<table>
<thead>
<tr>
<th>Case</th>
<th>Test (Database 1)</th>
<th>Three months</th>
<th>Percent</th>
<th>Two months</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 0</td>
<td>49/207</td>
<td>23.67%</td>
<td>39/207</td>
<td>18.84%</td>
<td></td>
</tr>
<tr>
<td>Type 1</td>
<td>6/24</td>
<td>25.00%</td>
<td>5/24</td>
<td>20.83%</td>
<td></td>
</tr>
<tr>
<td>Type 2</td>
<td>6/16</td>
<td>37.50%</td>
<td>4/16</td>
<td>25.00%</td>
<td></td>
</tr>
<tr>
<td>Type 3</td>
<td>4/9</td>
<td>44.44%</td>
<td>4/9</td>
<td>44.44%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>65/256</td>
<td>25.39%</td>
<td>49/256</td>
<td>20.31%</td>
<td></td>
</tr>
</tbody>
</table>
C.9 *Initial test – case selection breakdown*

<table>
<thead>
<tr>
<th>Case Selected</th>
<th>3 months</th>
<th></th>
<th>2 months</th>
<th></th>
<th>Out of range</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Percent</td>
<td>Count</td>
<td>Percent</td>
<td>Count</td>
<td>Percent</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0.0%</td>
<td>1</td>
<td>100.0%</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.0%</td>
<td>2</td>
<td>11.76%</td>
<td>15</td>
<td>88.24%</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>22.11%</td>
<td>14</td>
<td>14.74%</td>
<td>74</td>
<td>77.89%</td>
<td>95</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>29.73%</td>
<td>17</td>
<td>22.98%</td>
<td>52</td>
<td>70.27%</td>
<td>74</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
<td>0.00%</td>
<td>9</td>
<td>100.0%</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>15</td>
<td>100.0%</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
<td>0.00%</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0.00%</td>
<td>3</td>
<td>50.00%</td>
<td>3</td>
<td>50.00%</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>38.10%</td>
<td>7</td>
<td>33.33%</td>
<td>13</td>
<td>61.90%</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>29.63%</td>
<td>6</td>
<td>27.27%</td>
<td>14</td>
<td>63.63%</td>
<td>22</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>260</strong></td>
</tr>
</tbody>
</table>
### C.10 2-Nearest Neighbor

<table>
<thead>
<tr>
<th>Case</th>
<th>Test (Database 1)</th>
<th>Three months</th>
<th>Percent</th>
<th>Two months</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 0</td>
<td>51/229</td>
<td>22.27%</td>
<td>36/229</td>
<td>15.72%</td>
<td></td>
</tr>
<tr>
<td>Type 1</td>
<td>8/33</td>
<td>24.24%</td>
<td>6/33</td>
<td>18.18%</td>
<td></td>
</tr>
<tr>
<td>Type 2</td>
<td>7/18</td>
<td>38.89%</td>
<td>6/18</td>
<td>33.33%</td>
<td></td>
</tr>
<tr>
<td>Type 3</td>
<td>6/17</td>
<td>35.29%</td>
<td>6/17</td>
<td>35.29%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>72/297</td>
<td>24.24%</td>
<td>54/297</td>
<td>18.18%</td>
<td></td>
</tr>
</tbody>
</table>

### C.11 3-Nearest Neighbor

<table>
<thead>
<tr>
<th>Case</th>
<th>Test (Database 1)</th>
<th>Three months</th>
<th>Percent</th>
<th>Two months</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 0</td>
<td>27/233</td>
<td>11.59%</td>
<td>39/233</td>
<td>16.74%</td>
<td></td>
</tr>
<tr>
<td>Type 1</td>
<td>6/32</td>
<td>18.75%</td>
<td>5/32</td>
<td>15.63%</td>
<td></td>
</tr>
<tr>
<td>Type 2</td>
<td>6/18</td>
<td>33.33%</td>
<td>7/18</td>
<td>38.89%</td>
<td></td>
</tr>
<tr>
<td>Type 3</td>
<td>6/17</td>
<td>35.29%</td>
<td>6/17</td>
<td>35.29%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>45/300</td>
<td>15%</td>
<td>58/300</td>
<td>19.33%</td>
<td></td>
</tr>
</tbody>
</table>
Appendix D

TTF USER INTERFACE

D.1 Determining TTF User Interface
D.2 Machine fault trend
D.3 Case Library
Appendix E

Power Point Presentation

Slide 1

![Predicting Time-to-Failure of Industrial Machines with Temporal Data Mining](image)

Summer 2007
Jean Nakamura
Committee Chair: Isabelle Bichindaritz
Committee Member: Don McLane

Slide 2

Outline
- Overview
- Background
- Case Base System
- Results
- Additional Analysis
- Summary and Conclusion

Slide 3

Overview
- Analyze temporal vibration data results to predict machine failure
- Difficulties:
  - All information affecting failure is not available: i.e., running time, repairs, etc.
  - Machine failure is not linear
- Apply case based reasoning (CBR) to determine time-to-failure (TTF)
- CBR solves new problem by adapting old solutions
Slide 4

Overview - CBR

- The four parts of a CBR system
  - Retrieving – returns an old case that is identical or similar to the new problem
  - Reusing – applies the solution of the retrieved old case
  - Revising – adapts the retrieved solutions to resolve the new problem
  - Retaining – storing of valid cases

Slide 5

Overview

Slide 6

Background

- DLI Engineering
- Machine Condition Analysis software that analyzes vibration data for predictive maintenance
- The software generates:
  - Faults and a corresponding severity.
  - Recommendations and a corresponding priority, etc.

Slide 7

Background

- A FAULT is a description of a problem with the machine
- SEVERITY is the degree of the Fault
- A RECOMMENDATION is suggested repair action to resolve the Fault
- A PRIORITY is the degree of the Recommendation
Slide 8

**Background**

- Predictive Maintenance
  - Perform tests to determine what is starting to need a repair
  - Fixing the right problem rather than guessing what should be fixed
- Preventive Maintenance – repairs done on a schedule

Slide 9

**Background – Expert Report**

- Machine: Main Service Pump #1
- MID: 6
- Acquired: 7/27/2004 07:36 AM
- Figure of Merit = 197.
- Maximum level: 111 (+14) VdB at 1.00x on 2A
- RECOMMENDATIONS:
  - IMPORTANT: INSPECT COUPLING AND CHECK SHAFT ALIGNMENT
  - DIAGNOSTICS:
    - SERIOUS ANGULAR MISALIGNMENT
      - 111 (+14) VdB at 1.00xM on 2A in low range
    - 107 (+11) VdB at 1.00xM on 2T in low range
    - SLIGHT PUMP FREE BALL BEARING WEAR
      - 96 (+8.7) VdB at 22.6xP on 4A in high range
      - 91 (+1.5) VdB at 19.8xP on 4A in high range
      - 90 (+21) VdB at 5.88xP on 4A in low range

Slide 10

**Background**

- Machine Identification (MID)
- Machinery configuration
- Nominal speed, orientation, fault frequencies, etc.

Slide 11

**Background - Machine**
Background - MID

Data Pre-processing
- Modify database to include TTF system
- Flag invalid data
  - Incorrectly collected data
  - Expert System reports: "Manual review of Spectra is warranted", "Unavailable pickups at..." or "Questionable data at..."

Data Pre-processing
- Group MIDs
  - Example group:
    - AC Chill Water Pump
    - A/C Chill Water Pump
    - Air Conditioning Chill Water Pump
    - Air Conditioning Chilled Water Pmp

Data Pre-processing
- Group Faults
  - Example group:
    - Ball Bearing Noise
    - Ball Bearing Wear
    - Bearing Wear or Defect
Slide 16

**Databases**

- Two copies of the database were preprocessed.
- One copy, Database 1 contains tests up through 11/28/2005.
- A second copy of the database, Database 2 contains tests up through 09/13/2006.

Slide 17

**Pre-Processing Results**

<table>
<thead>
<tr>
<th></th>
<th>DB1</th>
<th>DB2</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Diagnoses</td>
<td>2058</td>
<td>1917</td>
<td>141</td>
</tr>
<tr>
<td>% of DB1 to DB2</td>
<td>98.9%</td>
<td>96.6%</td>
<td>-2.3%</td>
</tr>
<tr>
<td>Total MIDs</td>
<td>83207</td>
<td>142324</td>
<td>59017</td>
</tr>
<tr>
<td>% of DB1 to DB2</td>
<td>49.03%</td>
<td>48.64%</td>
<td>-0.39%</td>
</tr>
<tr>
<td>Valid Expert System runs</td>
<td>25843</td>
<td>128965</td>
<td>100122</td>
</tr>
<tr>
<td>% of DB1 to DB2</td>
<td>83.86%</td>
<td>83.31%</td>
<td>-0.55%</td>
</tr>
<tr>
<td>Tests</td>
<td>27389</td>
<td>169713</td>
<td>95324</td>
</tr>
<tr>
<td>% of DB1 to DB2</td>
<td>83.31%</td>
<td>83.86%</td>
<td>0.55%</td>
</tr>
</tbody>
</table>

Slide 18

**Case Library Definition - Retaining**

- Direct case – same fault, same machine (case type 0)
- Indirect case – grouped fault, same machine (case type 1)
- Indirect case – Normalize same fault within the grouped MID (case type 2)
- Indirect case – Normalize grouped fault within the grouped MID (case type 3)

Slide 19

**Case Definition - Direct Case**

- Case type 0
- Consists of a minimum of 3 consecutive tests on a machine with a specific fault:
  - Fault with an Extreme severity (considered machine failure)
  - Fault with any severity
  - Test where the fault does not exist.
Case Definition – Direct Case

<table>
<thead>
<tr>
<th>Date</th>
<th>Severity</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>01/14/04</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>04/07/04</td>
<td>Moderate</td>
<td>Ball Bearing Wear</td>
</tr>
<tr>
<td>09/03/04</td>
<td>Extreme</td>
<td>Ball Bearing Wear</td>
</tr>
</tbody>
</table>

Case Definition – Indirect case

Case type 1
- Consists of a minimum of 3 consecutive tests on a machine with a group fault:
  - Group Fault with an Extreme severity
  - Group Fault with any severity
  - Test where the group fault does not exist.

Case Definition – Indirect case

<table>
<thead>
<tr>
<th>Date</th>
<th>Severity</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>01/14/04</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>04/07/04</td>
<td>Moderate</td>
<td>Ball Bearing Noise</td>
</tr>
<tr>
<td>09/03/04</td>
<td>Extreme</td>
<td>Ball Bearing Wear</td>
</tr>
</tbody>
</table>

Case type 2
- Consists of a minimum of 3 consecutive tests on a MID group with a fault:
  - Fault with an Extreme severity
  - Fault with any severity
  - Test where the fault does not exist.
Case Definition – Indirect case

- Fault: "Indication of Engine Overload or Injector Timing Problem"
- MID Group 22: Diesel Engine
- MID: 223 Main Propulsion Diesel Engine
- MID: 456 Auxiliary Propulsion Engine

<table>
<thead>
<tr>
<th>Date</th>
<th>Day</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>01/14/04</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>09/03/04</td>
<td>233</td>
<td>Extreme</td>
</tr>
<tr>
<td>01/22/06</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>04/14/06</td>
<td>82</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

Case Definition – Indirect case

- Case type 3
- Consists of a minimum of 3 consecutive tests on a MID group with a fault group:
  - Group Fault with an Extreme severity
  - Group Fault with any severity
  - Test where the group fault does not exist.

Case Library Retrieval

1. Direct match (case type 0) – same diagnosis, same MID, same machine
2. Indirect match (case type 0) – same diagnosis, same MID, different machine
3. Indirect match (case type 0) – same diagnosis, MID group
4. Indirect match (case type 0) – diagnosis group, MID group
5. Indirect match (case type 1) – grouped diagnosis, same machine
6. Indirect match (case type 1) – grouped diagnosis, same MID, different machine
7. Indirect match (case type 1) – grouped diagnosis, MID group, different machine
8. Indirect match case (case type 2) - same diagnosis within the grouped MIDs
9. Indirect match case (case type 2) – diagnosis within the diagnosis group in the grouped MIDs
10. Indirect match case (case type 3) – grouped diagnosis within the grouped MIDs
Revising Case Library

- Removal of entire case
- Removal of individual test in case

Case-Library Summary

<table>
<thead>
<tr>
<th>Case Type</th>
<th>DB1 (training)</th>
<th>DB2 (test)</th>
<th>Difference</th>
<th>% DB2 Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case type 0</td>
<td>355</td>
<td>460</td>
<td>85</td>
<td>27.44%</td>
</tr>
<tr>
<td>Case type 1</td>
<td>59</td>
<td>68</td>
<td>1</td>
<td>3.56%</td>
</tr>
<tr>
<td>Case type 2</td>
<td>63</td>
<td>67</td>
<td>32</td>
<td>37.65%</td>
</tr>
<tr>
<td>Case type 3</td>
<td>59</td>
<td>62</td>
<td>32</td>
<td>36.67%</td>
</tr>
<tr>
<td>Total</td>
<td>886</td>
<td>916</td>
<td>30</td>
<td>31.18%</td>
</tr>
</tbody>
</table>

Case-Library Analysis

<table>
<thead>
<tr>
<th>Case</th>
<th>Case Count</th>
<th>DB1 Value</th>
<th>DB2 Value</th>
<th>Difference</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>355</td>
<td>233,234</td>
<td>232,123</td>
<td>-111</td>
<td>0.46%</td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td>345,346</td>
<td>344,234</td>
<td>-111</td>
<td>0.46%</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
<td>456,457</td>
<td>455,346</td>
<td>-111</td>
<td>0.46%</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>546,547</td>
<td>545,346</td>
<td>-111</td>
<td>0.46%</td>
</tr>
<tr>
<td>5</td>
<td>63</td>
<td>656,657</td>
<td>655,346</td>
<td>-111</td>
<td>0.46%</td>
</tr>
</tbody>
</table>

Case-Library Analysis – cont.

<table>
<thead>
<tr>
<th>Case</th>
<th>Case Count</th>
<th>DB1 Value</th>
<th>DB2 Value</th>
<th>Difference</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>59</td>
<td>745,746</td>
<td>744,346</td>
<td>-111</td>
<td>0.46%</td>
</tr>
<tr>
<td>7</td>
<td>63</td>
<td>856,857</td>
<td>855,346</td>
<td>-111</td>
<td>0.46%</td>
</tr>
<tr>
<td>8</td>
<td>59</td>
<td>965,966</td>
<td>964,346</td>
<td>-111</td>
<td>0.46%</td>
</tr>
<tr>
<td>9</td>
<td>63</td>
<td>1076,1077</td>
<td>1075,346</td>
<td>-111</td>
<td>0.46%</td>
</tr>
<tr>
<td>10</td>
<td>59</td>
<td>1186,1187</td>
<td>1185,346</td>
<td>-111</td>
<td>0.46%</td>
</tr>
<tr>
<td>11</td>
<td>63</td>
<td>1296,1297</td>
<td>1295,346</td>
<td>-111</td>
<td>0.46%</td>
</tr>
</tbody>
</table>

Slide 28

Slide 29

Slide 30

Slide 31
Slide 32

Case Evaluation Distribution

Slide 33

TTF - Criteria

- The TTF accuracy was based on:
  - two months (<=62 days)
  - three months (<=93 days)

Slide 34

CBR Results - training

<table>
<thead>
<tr>
<th>Type</th>
<th>Case Evaluation Distribution</th>
<th>Two months</th>
<th>Three months</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 3</td>
<td>309/269</td>
<td>90.60%</td>
<td>95/95</td>
<td>98.94%</td>
</tr>
<tr>
<td>Type 1</td>
<td>55/30</td>
<td>18.33%</td>
<td>15/15</td>
<td>100.00%</td>
</tr>
<tr>
<td>Type 2</td>
<td>35/37</td>
<td>94.59%</td>
<td>27/27</td>
<td>96.30%</td>
</tr>
<tr>
<td>Type 5</td>
<td>20/20</td>
<td>100.00%</td>
<td>10/10</td>
<td>100.00%</td>
</tr>
<tr>
<td>Total</td>
<td>390/352</td>
<td>90.92%</td>
<td>137/137</td>
<td>99.86%</td>
</tr>
</tbody>
</table>

Slide 35

CBR Results - Knowing TTF

<table>
<thead>
<tr>
<th>Type</th>
<th>Case Evaluation Distribution</th>
<th>Two months</th>
<th>Three months</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 3</td>
<td>95/286</td>
<td>33.22%</td>
<td>134/286</td>
<td>46.85%</td>
</tr>
<tr>
<td>Type 1</td>
<td>6/15</td>
<td>40.00%</td>
<td>6/15</td>
<td>40.00%</td>
</tr>
<tr>
<td>Type 2</td>
<td>29/82</td>
<td>35.37%</td>
<td>36/82</td>
<td>43.90%</td>
</tr>
<tr>
<td>Type 5</td>
<td>7/20</td>
<td>35.00%</td>
<td>8/20</td>
<td>40.00%</td>
</tr>
<tr>
<td>Total</td>
<td>136/352</td>
<td>38.92%</td>
<td>136/352</td>
<td>38.92%</td>
</tr>
</tbody>
</table>

New tests in database
CBR Results – Initial Test

<table>
<thead>
<tr>
<th>Case</th>
<th>Test (Database 1)</th>
<th>Three months</th>
<th>Percent</th>
<th>Two months</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 0</td>
<td>6/24</td>
<td>20.83%</td>
<td>4/24</td>
<td>25.00%</td>
<td></td>
</tr>
<tr>
<td>Type 1</td>
<td>6/36</td>
<td>18.84%</td>
<td>3/36</td>
<td>25.00%</td>
<td></td>
</tr>
<tr>
<td>Type 2</td>
<td>6/16</td>
<td>25.00%</td>
<td>4/16</td>
<td>25.00%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>18/94</td>
<td>20.00%</td>
<td>8/94</td>
<td>25.00%</td>
<td></td>
</tr>
</tbody>
</table>

CBR Results – 2 Nearest Neighbor

<table>
<thead>
<tr>
<th>Case</th>
<th>Test (Database 1)</th>
<th>Three months</th>
<th>Percent</th>
<th>Two months</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 0</td>
<td>6/24</td>
<td>18.18%</td>
<td>6/24</td>
<td>24.24%</td>
<td></td>
</tr>
<tr>
<td>Type 1</td>
<td>6/36</td>
<td>18.18%</td>
<td>6/36</td>
<td>24.24%</td>
<td></td>
</tr>
<tr>
<td>Type 2</td>
<td>6/18</td>
<td>18.18%</td>
<td>6/18</td>
<td>24.24%</td>
<td></td>
</tr>
<tr>
<td>Type 3</td>
<td>6/17</td>
<td>15.63%</td>
<td>6/17</td>
<td>23.53%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>26/175</td>
<td>15.00%</td>
<td>10/175</td>
<td>18.00%</td>
<td></td>
</tr>
</tbody>
</table>

CBR Results – 3 Nearest Neighbor

<table>
<thead>
<tr>
<th>Case</th>
<th>Test (Database 1)</th>
<th>Three months</th>
<th>Percent</th>
<th>Two months</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 0</td>
<td>6/24</td>
<td>15.72%</td>
<td>6/24</td>
<td>22.73%</td>
<td></td>
</tr>
<tr>
<td>Type 1</td>
<td>6/36</td>
<td>18.18%</td>
<td>6/36</td>
<td>24.24%</td>
<td></td>
</tr>
<tr>
<td>Type 2</td>
<td>6/18</td>
<td>18.18%</td>
<td>6/18</td>
<td>24.24%</td>
<td></td>
</tr>
<tr>
<td>Type 3</td>
<td>6/17</td>
<td>15.63%</td>
<td>6/17</td>
<td>23.53%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>26/175</td>
<td>15.00%</td>
<td>10/175</td>
<td>18.00%</td>
<td></td>
</tr>
</tbody>
</table>

SPSS

- Analytical software specializing in data mining
  - [www.spss.com](http://www.spss.com)
To evaluate the relationship of the independent variables:
- VibrationStandardEquipmentGroupID
- DiagnosisGroupID
- VibrationStandardEquipmentID
- VibrationStandardDiagnosisID
- VibDiagnosisSeverityIndex

To the dependant variable: DayPosition

R values range from 0 to 1 with larger R values indicating stronger relationships.
This model produced and R value of only .330.
R squared values range from 0 to 1 with larger values indicating that the model fits the data well.
This system only produced a value of .109 which shows that the model does not fit the data well.

Using the coefficients in the prior slide, apply to equation in determining DayPosition:
DayPosition = 133.508 + (-.011 * VibStandardEquipmentID) + (-.015 * VibStandardDiagnosisID) + (-8.004 * DiagnosisGroupID) + (1.311 * VibrationStandardEquipmentGroupID) + (.266 * VibDiagnosisSeverityIndex)
Slide 44

<table>
<thead>
<tr>
<th>SPSS Coefficient - Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>Initial Test</td>
</tr>
<tr>
<td>2-Nearest Neighbor</td>
</tr>
<tr>
<td>3-Nearest Neighbor</td>
</tr>
<tr>
<td>SPSS Regression</td>
</tr>
</tbody>
</table>

Slide 45

**Weighting Parameters**

- Changing the weighting of the attributes to determine the importance of the attributes on the accuracy of the system.
- The calculation was done based on the following formula where $W_i$ is the weight for the matching attribute and $W_{total}$ is the weight of the attribute evaluated.

$$
\frac{\sum W_i F_i}{\sum W_{total}}
$$

Slide 46

**Weighting Dataset**

- VibStandardSeverityIndex (converted to nominal values: None, Slight, Moderate, Serious, or Extreme)
- VibStandardEquipmentID
- VibStandardEquipmentGroupID
- VibStandardDiagnosisID
- DiagnosisGroupID
- DayPosition (converted to nominal values: <1month, 1-2months, 2-3months, 3-4months, 5-6months, 6-7months, 8-9months, 10-11months, 11-12months, and >12months)

Slide 47

**Abbreviation used**

- Num = Weighted test number
- Severity = VibStandardSeverityIndex
- MID = VibStandardEquipmentID
- MIDGrp = VibStandardEquipmentGroupID
- DX = VibStandardDiagnosisID
- DXGrp = DiagnosisGroupID
### Weighting Result Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>Severity</th>
<th>Day Positive</th>
<th>MID</th>
<th>HillTop</th>
<th>Bs</th>
<th>Email</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>17.99%</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20.08%</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>20.50%</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>26.57%</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>25.10%</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>23.64%</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>24.89%</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>11.92%</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>25.73%</td>
</tr>
</tbody>
</table>

### WEKA

- Collection of machine learning algorithms for data mining tasks.
- [www.cs.waikato.ac.nz/ml/weka/](http://www.cs.waikato.ac.nz/ml/weka/)
- Open source under GPL
- Written in Java

### Weka Algorithms

- **J48 tree, ID3 tree, Multilayer Perceptron, Logistic regression, Apriori, Predictive Apriori, and K**
- **J48** - generates a pruned or un-pruned C4 tree where values may be missing, attributes may be numeric, and can deal with noisy data.
- **ID3** generates an un-pruned decision tree. Attributes must be nominal and there cannot be any missing values. Empty leaves may result in unclassified instances.
- **Multilayer Perceptron** is a neural network that uses back-propagation to train.
- **Logistic regression** is used for building and using a multinomial logistic regression model with a ridge estimator.
- **Apriori** generates association rules from frequent item-sets
- **Predictive Apriori** finds association rules sorted by predictive accuracy.
- **K** is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function.
WEKA – Test Options
- Training set - provides optimal classification accuracy
- 10-fold cross validation
  - averages the classification results on ten different random samples generated from the dataset.
  - It provides more robust results when there is only one dataset available.
  - Tries to diversify the samples in order to get a better estimate with fewer samples.
  - It uses averages so the method is not as accurate.
- Ratio validation of 50%, 66%, 75%, and 80% training - uses a different percentage of records for the training and test set

WEKA – Dataset
- Modified so each row contains a case: each test severity and DayPosition
- Sev2 represents the severity value of the 2nd test, Sev3 represents the severity value of test 3, etc.
- Diff2 is the number of days to TTF calculated from the actual TTF minus the current day position of the test, and so forth.
- The dataset includes DiagnosisGroupID (DxGrpID), VfStndRdEquipmentGroupID (MIDGrpID), Sev2, Sev3, Sev4, Diff2, Diff3, Diff4.

Discretized to 3 groups
- Discretized into three equi-depth groups
- Each group has same number of cases
- Produced large ranges
- The range for Diff3 in Group2 is 268 days or almost nine months.

Sample Dataset
- DiagGrpID
- Sev2
- Sev3
- Sev4
- Diff2
- Diff3
- Diff4
- DayPosition
- DxGrpID
### 3 Groups - Summary

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Diff 2</th>
<th>Diff 3</th>
<th>Diff 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly Classified Instances</td>
<td>59.276</td>
<td>56.968</td>
<td>68.870</td>
</tr>
<tr>
<td>Total # of Instances</td>
<td>222</td>
<td>222</td>
<td>222</td>
</tr>
<tr>
<td>Ignored Class Instances</td>
<td>208</td>
<td>208</td>
<td>208</td>
</tr>
<tr>
<td>J48</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>10 Fold Cross Validation</td>
<td>59.556</td>
<td>67.100</td>
<td>81.028</td>
</tr>
<tr>
<td><strong>Multilayer Perceptron</strong></td>
<td>66.968</td>
<td>72.467</td>
<td>84.048</td>
</tr>
</tbody>
</table>

### Discretized to 6 groups

- Discretized into 6 groups
- Produced smaller ranges but still large

<table>
<thead>
<tr>
<th>Group</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Group</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Group</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>-inf</td>
<td>-1.000</td>
<td>Group 3</td>
<td>93.5-196.5</td>
<td>93.5-186.0</td>
<td>Group 5</td>
<td>1-93.5</td>
<td>0.5-93.5</td>
</tr>
<tr>
<td>Group 2</td>
<td>0.5-93.5</td>
<td>72.5-119.5</td>
<td>Group 4</td>
<td>196.5-321</td>
<td>194.5-340.5</td>
<td>Group 6</td>
<td>321-689</td>
<td>337-642.5</td>
</tr>
<tr>
<td>Group 3</td>
<td>321-689</td>
<td>340.5-572</td>
<td>Group 5</td>
<td>572-inf</td>
<td>642.5-inf</td>
<td>Group 1</td>
<td>642.5-inf</td>
<td>689-inf</td>
</tr>
</tbody>
</table>

### 6 Groups - Results

<table>
<thead>
<tr>
<th>J48 tree</th>
<th>Diff 2</th>
<th>Ignored Class Instances</th>
<th>Correctly Classified Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>training data</td>
<td>222</td>
<td>208</td>
<td>68.870 %</td>
</tr>
<tr>
<td>80% train</td>
<td>187</td>
<td>108</td>
<td>95.875 %</td>
</tr>
<tr>
<td>80% train</td>
<td>108</td>
<td>27</td>
<td>95.358 %</td>
</tr>
<tr>
<td>80% train</td>
<td>45</td>
<td>45</td>
<td>66.876 %</td>
</tr>
</tbody>
</table>

### Progression of Failure

- Does the progression of severity assist in determining TTF?
- Another attribute was added to demonstrate a decrease, increase or stable severity.
- Sev_1-2 is the severity change between Sev1, Sev2, Sev_2-3 is the severity change between Sev2 and Sev3, and so forth.
Slide 60

Severity Distribution

<table>
<thead>
<tr>
<th>Sev</th>
<th>0%</th>
<th>20%</th>
<th>40%</th>
<th>60%</th>
<th>80%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sev2</td>
<td>2</td>
<td>12</td>
<td>19</td>
<td>89</td>
<td>120</td>
<td>319</td>
</tr>
<tr>
<td>Sev3</td>
<td>6</td>
<td>16</td>
<td>48</td>
<td>68</td>
<td>287</td>
<td>690</td>
</tr>
<tr>
<td>Sev4</td>
<td>18</td>
<td>72</td>
<td>120</td>
<td>228</td>
<td>330</td>
<td>319</td>
</tr>
</tbody>
</table>

Severity Progression

<table>
<thead>
<tr>
<th>Sev</th>
<th>Decrease</th>
<th>Stable</th>
<th>Increase</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sev 1-2</td>
<td>8</td>
<td>95</td>
<td>42</td>
<td>65</td>
</tr>
<tr>
<td>Sev 2-3</td>
<td>33</td>
<td>184</td>
<td>287</td>
<td>420</td>
</tr>
<tr>
<td>Sev 3-4</td>
<td>26</td>
<td>111</td>
<td>31</td>
<td>168</td>
</tr>
</tbody>
</table>

Severity Progression-Summary

<table>
<thead>
<tr>
<th>Method</th>
<th>10-fold Cross Validation</th>
<th>Multilayer Perceptron - modified options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Confusion Matrix</td>
<td>Confusion Matrix</td>
</tr>
<tr>
<td>Confusion Matrix</td>
<td>Confusion Matrix</td>
<td>Confusion Matrix</td>
</tr>
</tbody>
</table>

WEKA Associations

- Apriori and Predictive Apriori
- The dataset includes the following attributes: Sev2, Sev3, Sev4, Diff2, Diff3, and Diff4.
Apriori
1. Diff4='(-inf-1]' 83 ==> Sev4=extreme 83
   conf:(1)
2. Diff2='(118-340.5]' Diff3='(-inf-0.5]' 75 ==> Sev3=extreme 75
   conf:(1)
3. Sev2=extreme Diff3='(-inf-0.5]' 73 ==> Sev3=extreme 73
   conf:(1)
4. Diff3='[237.5-inf]' 69 ==> Diff2='[340.5-inf]' 69
   conf:(1)
5. Diff3='[268.5-inf]'

Apriori – cont.
6. Diff3='[0.5-268.5]'

Predictive Apriori
1. Diff3='(-inf-0.5]' 214 ==> Sev3=extreme 212
   acc:(0.99495)
2. Diff4='(-inf-1]' 83 ==> Sev4=extreme 83
   acc:(0.99494)
3. Diff4='(237.5-inf)' 69 ==> Diff2='(340.5-inf)' 69
   acc:(0.99489)
4. Diff4='(237.5-inf)' 69 ==> Diff2='(340.5-inf)' Diff3='(268.5-inf)' 68
   acc:(0.99403)
5. Diff2='(-inf-118]' Diff3='[0.5-268.5]' 15 ==> Sev4=extreme 16
   acc:(0.99337)
6. Sev3=moderate Sev4=moderate Diff3='[1340.5-inf]' 16 ==>

Predictive Apriori – cont.
6. Sev3=moderate Sev4=moderate Diff3='[340.5-inf]' 16 ==>
   Diff3='[268.5-inf]' 16 acc:(0.99223)
7. Sev3=moderate Sev4=moderate Diff3='[268.5-inf]' 16 ==>
   Diff4='[1-237.5]' 15 acc:(0.99223)
8. Sev4=significant Diff3='[0.5-268.5]' 15 ==>
   Diff4='[1-237.5]' 15 acc:(0.99223)
9. Sev2=slight Sev3=extreme Diff3='[118-340.5]' 15
   ==> Diff3='[inf-0.5]' 15 acc:(0.99176)
10. Sev2=moderate Sev4=moderate Diff3='[340.5-inf]' 15 ==>
    Diff3='[268.5-inf]' 15 acc:(0.99176)
Summary of Results

- Poor results from continuous numeric values.
- Discretize into three equi-depth groups
- Best results (80% accuracy) from Logistic regression and secondly from Multilayer Perceptron
- Do not determine TTF based on Diff2
- Progression of severity comparably good results but no improvements.

Future Work

- Use discretized attributes Sev# and Diff# only
- Develop a system using the Logistic regression or Multilayer Perceptron algorithm
- See if repair history can eliminate those long TTF cases

Conclusion

- Determining TTF using CBR with the generalization groupings and continuous data produced low accuracy rates.
- Due to the low numbers of cases for each specific MID and Diagnoses, it is best to determine TTF based on Sev# and Diff# only.
- Include repair history information in an attempt to improve accuracy.
- Equi-depth groups with Logistic regression or Multilayer Perceptron algorithms produced good results.