Contents

List of Figures

List of Tables

Chapter 1

Question

Ph.D. Candidacy Examination for Ashutosh Nandeshwar From: Gopala, IMSE, 06/05/09

Consider a job shop that has metal cutting machines such as lathes, milling machines, drilling machines, grinders etc. The raw material comes in as cast metal pieces of various shapes and sizes. Process plans are generated based on the product that needs to be produces and the machining operations are sequenced on one or more machines. After the final machining operation, each part is sent to a heat treatment furnace for annealing and then sent to a steam driven degreaser before shipping. In this manufacturing scenario, consider all relevant manufacturing parameters on production equipment as well as parameters on support equipment such as material handling equipment. Also consider the utility system (electricity, natural gas) and the parameters on equipment that provide energy for operating production machinery of various types. Answer the following questions.

Design a data mining system for the manufacturing facility that would enable the plant personnel to identify the key aspects that would affect quality, production throughput, energy, inventory levels, and Overall Equipment Effectiveness (OEE). I would specifically you to elaborate on:

1. The data that you would want to collect on the shop floor and on support equipment.
2. Where the sensors would be placed and the nature of data that they would collect and at what frequency.
3. Data acquisition protocols you would consider using.
4. The data mining strategy and specific algorithms you would use.
5. How the results will be presented to the plant engineer/manager in a manner that is easy to understand.
6. The use of predictive rule induction and its use for this production scenario.
7. How the plant engineer/manager could use your system on a frequent basis and the types of benefits that they would derive.

Provide a typical (you make up the data set) data set showing the typical variables and data that you may have collected using your system and show data mining results. Submit in a well written report.

Due June 26, 2009

Chapter 2

Introduction

2.1 What is Data mining?

Although data mining definitions change with the area of the researcher, the definitions by some of the well-known researchers are apt for this research. Hand et al. (2001) defined data mining as "the science of extracting useful information from large data sets or databases." Witten and Frank (2005) defined data mining as "the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic." Berry and Linoff (1997) defined data mining as "the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules."

Data mining is also known as knowledge discovery in databases (KDD), and this discovery process is shown in Figure 2.1. Enterprise Resource Planning (ERP) systems hold massive amounts of data, which usually consists of information, such as, machines, orders, supply, financial, payroll, others. The data entry people working in each functional area enter this information in ERP systems. Database administrators load this information in databases using Extract, Transform, and Load (ETL) tools. Data analysts or miners analyze these databases, understand the data or work with the domain experts, develop prediction, classification, or clustering models, evaluate the models, and implement them; using this approach, data miners transform information into tangible knowledge for decision-making.

Areas of computer science, statistics, database technologies, machine learning, and others form the field of data mining. Statistics influenced the field of data mining tremendously; so much that Kuonen (2004) asked whether data mining is "statistical déjà vu." Amalgamation of statistics and computer science started data mining; however, data mining as a field is evolving on its own. Han and Kamber (2006) described the overlap of multiple disciplines as shown in Figure 2.2.

2.1.1 Decision Trees

Decisions tree are a collection of nodes, branches, and leaves. Each node represents an attribute; this node is then split into branches and leaves. Decision trees work on the "divide and conquer" approach; each node is divided, using purity information criteria, until the data are classified to meet a stopping

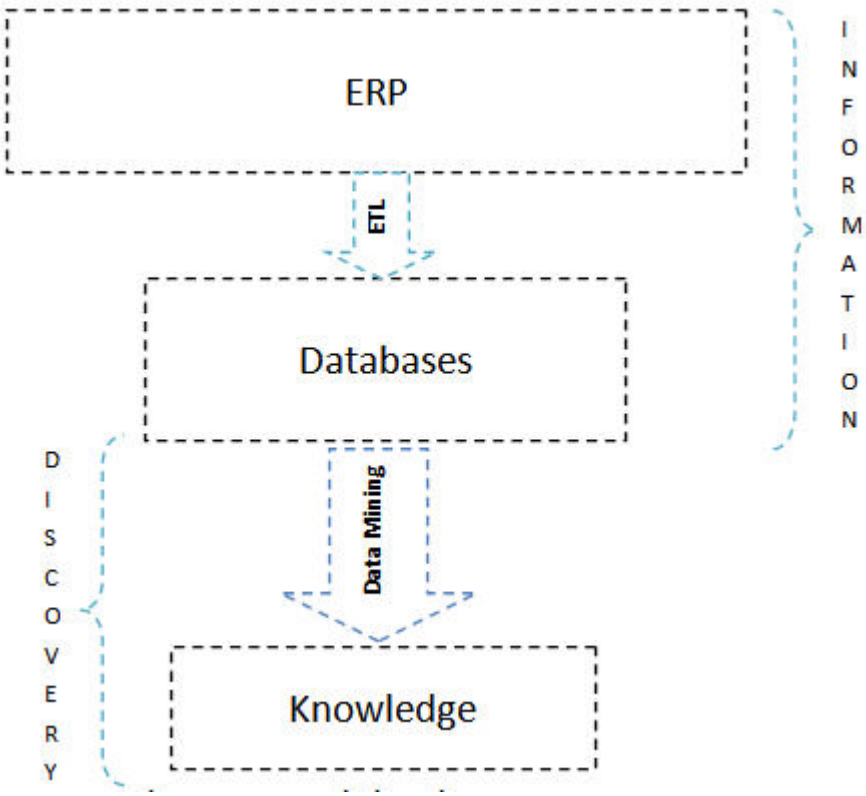

Figure 2.1: Data to Knowledge

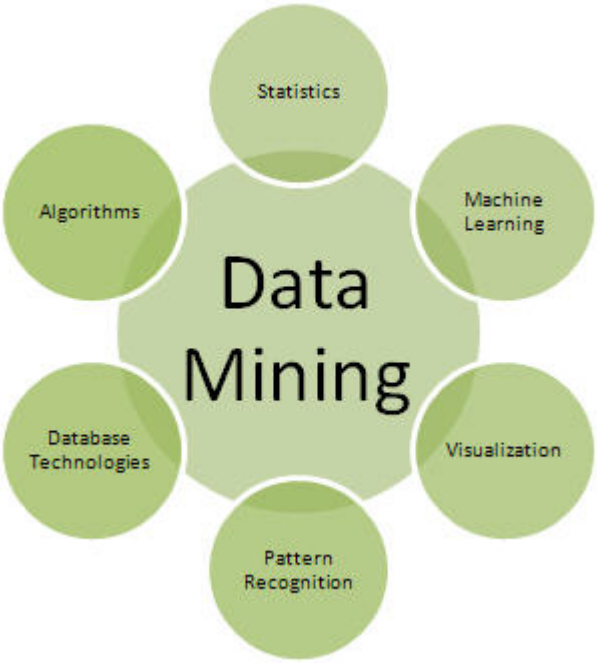

Figure 2.2: Data Mining-Confluence of Multiple Disciplines

condition. Gini index and information gain ratio are two common purity measurement criteria; Classification and Regression Tree (CART) algorithm uses Gini index, and C4.5 algorithm uses the information gain ratio (Quinlan, 1986, 1996). The Gini index is given by Equation 2.1, and the information gain is given by Equation 2.2.

$$I_G(i) = 1 - \sum_{j=1}^{m} f(i,j)^2 = \sum_{j \neq k} f(i,j) f(i,k) \tag{2.1}$$

$$I_E(i) = -\sum_{j=1}^{m} f(i,j) \log_2 f(i,j) \tag{2.2}$$

where, m is the number of values an attribute can take, and $f(i,j)$ is the proportion of class in i that belong to the j^th class.

2.1.2 Rules

Construction of rules is quite similar to the construction of decision trees; however, rules first cover all the instances for each class, and exclude the instances, which do not have class in it. Therefore, these algorithms are called as covering algorithms, and pseudocode of such algorithm is given in Figure 2.4 reproduced from Witten and Frank (2005).

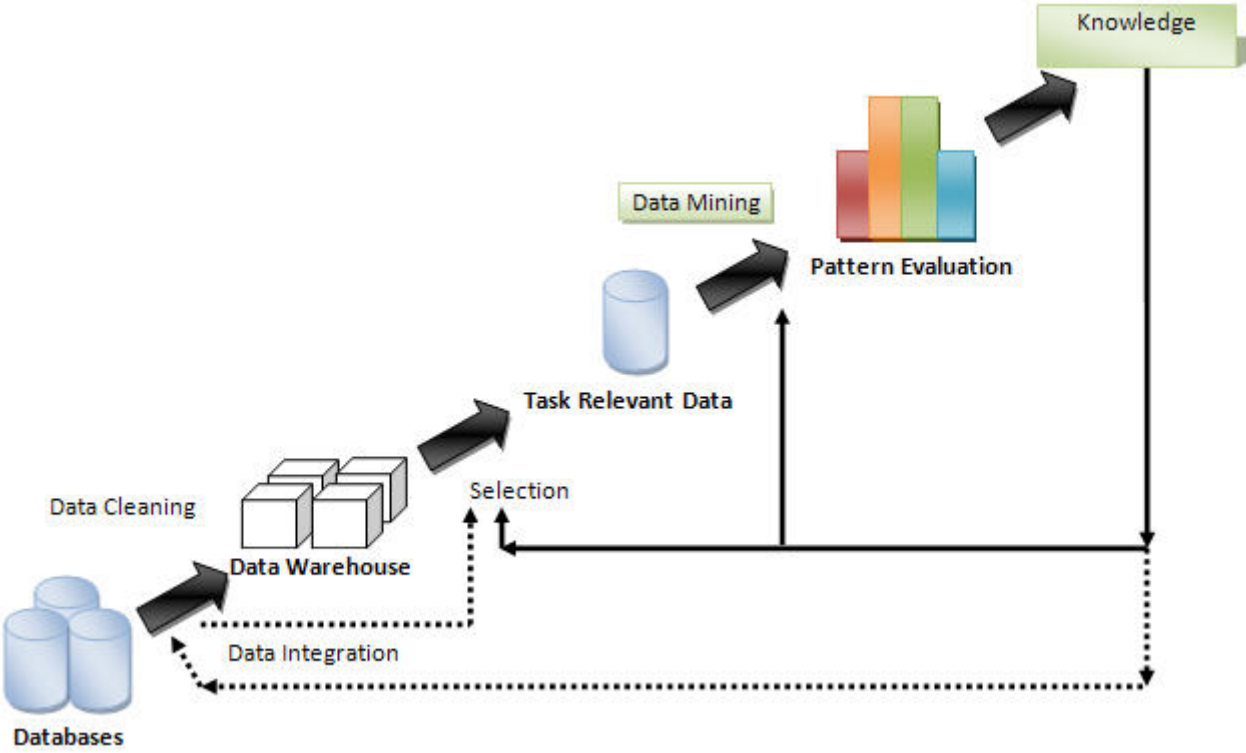

Figure 2.3: Knowledge Discovery Process

```
For each class C
  Initialize E to the instance set
  While E contains instances in class C
  Create a rule R with an empty left-hand side that predicts class C
  Until R is perfect (or there are no more attributes to use) do
  For each attribute A not mentioned in R, and each value v,
   Consider adding the condition A=v to the LHS of R
   Select A and v to maximize the accuracy p/t
   (break ties by choosing the condition with the largest p)
   Add A=v to R
   Remove the instances covered by R from E
```

Figure 2.4: Pseudocode for a Basic Rule Learner

Feature Subset Selection (FSS)

Feature subset selection is a method to select relevant attributes (or features) from the full set of attributes as a measure of dimensionality reduction. Although some of the data mining techniques, such as decision trees, select relevant attributes, their performance can be improved, as the experiments have shown(Witten and Frank, 2005, p. 288).

Two main approaches of feature or attribute selection are the filters and the wrappers (Witten and Frank, 2005). A filter is an unsupervised attribute selection method, which conducts an independent assessment on general characteristics of the data. It is called as a filter because the attributes are filtered before the learning procedure starts. A wrapper is a supervised attribute selection method, which uses data mining algorithms to evaluate the attributes. It is called as a wrapper because the learning method is wrapped in the attribute selection technique. In an attribute selection method, different search algorithms are employed, such as, genetic algorithm, greedy step-wise, rank search, and others.

2.1.3 Discretization

Some of the classifiers work well with discretized variables, such as tree and rule learners, therefore, discretizing numerical attributes is a very important preprocessing step. In addition, methods often produce better results (or run faster) , if the attributes are prediscretized(Witten and Frank, 2005, p. 287). There are two types of discretizers: unsupervised and supervised.

Unsupervised Discretization

Similar to unsupervised learning, unsupervised discretization works without the knowledge of the class attribute. Although unsupervised discretization is easy to understand and arguably fast, it risks the danger of excluding some important information (for the learners) as a result of discrete intervals being too short or too long(Witten and Frank, 2005, p. 298). Some of the unsupervised discretization methods are:

1. Equal Interval Binning: as the name says, this discretization method divides the attribute in equal (predetermined arbitrary) intervals.

2. Equal Frequency Binning: this method is also called as histogram equalization, because the attributes are discretized in such a manner so that each intervals gets equal number of instances.

3. Proportional k-interval Discretization (PKID) (Yang and Webb, 2001): Yang and Webb (2003) warned that proportional k-interval discretization worked better for larger datasets, and suggested weighted proportional k-interval discretization. The proportional k-intervals are calculated using the Equation 2.3.

$$k = \sqrt{N} \tag{2.3}$$

   where, N is the number of instances.

Supervised Discretization

One of the best and state of the art supervised discretization method is Fayyad and Irani's (1992) minimum description length (MDL) criterion and entropy-based discretization. This discretization method is based on the idea of reducing the impurity by splitting (*cut point*) the intervals where the information value is smallest. The numeric attribute values are sorted in the ascending order, and a split is created where the subintervals are as pure as possible.

2.2 What is Manufacturing Execution Systems (MES)?

The MESA International ((MESA, 1997)) provided this definition of MES: "Manufacturing Execution Systems (MES) deliver information that enables the optimization of production activities from order launch to finished goods. Using current and accurate data, MES guides, initiates, responds to, and reports on plant activities as they occur. The resulting rapid response to changing conditions, coupled with a focus on reducing non value-added activities, drives effective plant operations and processes. MES improves the return on operational assets as well as on-time delivery, inventory turns, gross margin, and cash flow performance. MES provides mission-critical information about production activities across the enterprise and supply chain via bi-directional communications." The functional model with all the components and the links to other systems is shown in Figure 2.5 (MESA, 1997).

The manufacturing execution system (MES) is central system to store manufacturing information such as resource allocation, manufacturing planning, supply/demand, quality assurance data, measured data acquisition, staff work time logging, and others (Kletti, 2006). MES along with ERP systems can bridge the gap between personnel data information and production planning information. This integration, therefore, makes MES very useful for gathering real-time production data (work in-progress, equipment availability, scheduling, inventory management, material movement, WIP tracking, etc), which essentially supports efficient reporting and predictive modeling to support decision making (Chen and Wu, 2005). Figure 2.6 shows the dependence of control requirements on planning time using ERP and MES.

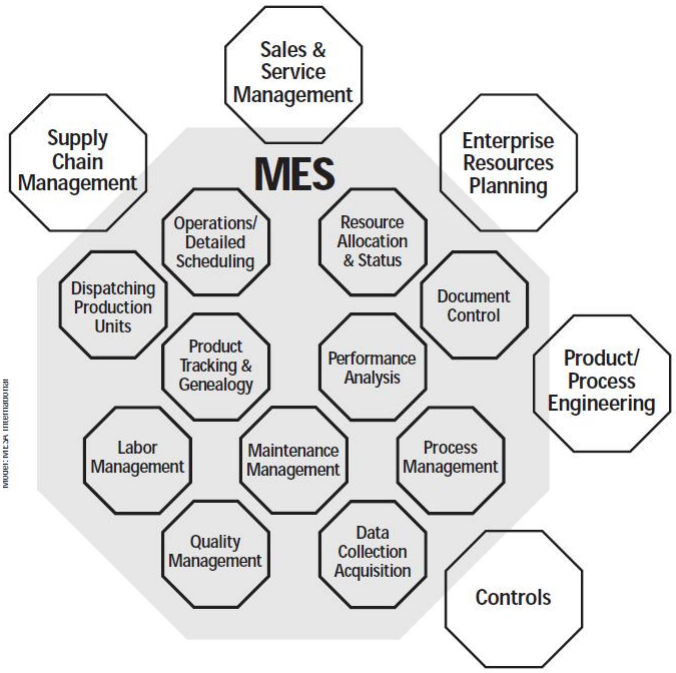

Figure 2.5: MES Functional Model (MESA, 1997)

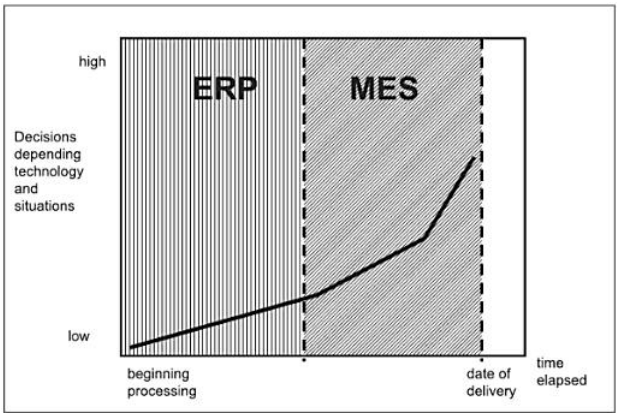

Figure 2.6: Dependence of Control Requirements on Planning Time (Kletti, 2006)

Eleven principal functions of MES are (MESA, 1997):

1. Operations/Detail Scheduling - sequencing and timing activities for optimized plant performance based on finite capacities of the resources;
2. Resource Allocation and Status - guiding what people, machines, tools, and materials should do, and tracking what they are currently doing or have just done;
3. Dispatching Production Units - giving the command to send materials or orders to certain parts of the plant to begin a process or step;
4. Document Control - managing and distributing information on products, processes, designs, or orders, as well as gathering certification statements of work and conditions;
5. Product Tracking and Genealogy - monitoring the progress of units, batches, or lots of output to create a full history of the product;
6. Performance Analysis - comparing measured results in the plant to goals and metrics set by the corporation, customers, or regulatory bodies;
7. Labor Management - tracking and directing the use of operations personnel during a shift based on qualifications, work patterns, and business needs;
8. Maintenance Management - planning and executing appropriate activities to keep equipment and other capital assets in the plant performing to goal;
9. Process Management - directing the flow of work in the plant based on planned and actual production activities;
10. Quality Management - recording, tracking, and analyzing product and process characteristics against engineering ideals;
11. Data Collection/Acquisition - monitoring, gathering, and organizing data about the processes, materials, and operations from people, machines, or Controls.

One of the main important functions, at least for this project, is the "Data Collection/Acquisition" function. MES enables a linkage to obtain timely operational and production data to populate the forms and records of various tables; it can do that automatically (we can set the frequency) for various equipment with the help of programmable logic controllers (PLCs) via sensors, or it can be updated manually for the operator data.

2.3 Problem Process

In this project, we have a typical job floor, where raw material is sent to various machines based on the operations (from the process plan), then sent for finalizing operations to a heat furnace and a steam degreaser. As with any other

manufacturing process, there would be material moving equipment between each process and a queue at each stage. In addition, energy would be consumed by equipment, machines, and furnace/degreaser. This process is depicted in Figure 2.7.

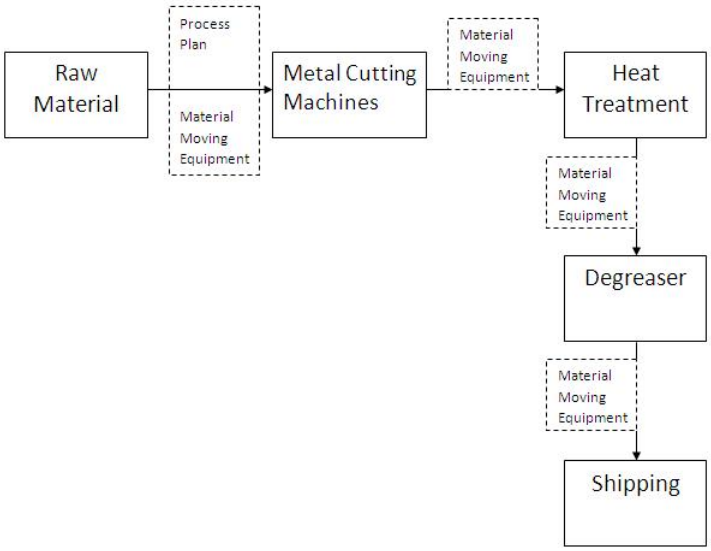

Figure 2.7: Manufacturing Process of the Problem

2.4 Problem Solving

As the problem objective is very generic, a solution to the problem should also be generic and flexible. In this problem, there are various inputs and various outputs. Usually, a data mining problem is solved once we know the set of input variables and the associated target variable(s), or at least, we are aware of the problem we are trying to solve; in this case, however, we do not know the complete set of input variables, and neither do we know the exact output variables, nor do we know the exact problem we are attempting to solve. With these attributes, the challenge of solving the current problem using data mining becomes very difficult, but not impossible with the help of solid data warehouse foundation, extract, transform, and load (ETL) processes, use of on-line analytical processing (OLAP) tools, business intelligence tools, and the implementation of on/off-line data mining techniques.

For this problem, I suggest implementing an MES, or a similar system, which I will discuss in the following sections. This MES will acquire data, manually or automatically, from various machines, equipment, and personnel at various stages of production. It will keep track of inventory, work-in-progress, supply and demand, scraps, down times and its causes, etc, which will enable data miners to derive the desired aspects, such as, quality, production throughput, energy, inventory levels, and Overall Equipment Effectiveness (OEE).

As a part of this MES, the sensors would be placed on all the machines, furnaces, and degreasers. These sensors would be activated using programmable logic controls (PLCs) (as shown in Figure 2.5) — they play a big role in the MES and are the primary source of data acquisition — then these sensors would feed in to the databases. When running the initial data mining experiments, the data should be collected at start of the operation, at the end of the operation, and equally divided intervals between the start and the end. Once initial data mining experiments yield some indicators, these frequencies can be adjusted.

Manual entrance of data should be completed whenever there is a need to enter that data, such as in the cases of new raw material coming in (if bar code system could be developed to track demand and supply, that would be even better), machine/equipment down times, overwriting sensor data, and others. In order to better design this system, it was critical to study the usage of data mining, data warehouses, and other techniques in the manufacturing industry.

2.5 Literature Review

In the manufacturing industry, intelligent decision-making methods using computer software are well researched and practiced. Gopalakrishnan (1990) developed expert system to select machining tool parameters. Arezoo et al. (2000) developed an knowledge-based expert system to select the cutting tools and other parameters. Kalaga and Kazic developed automated extraction system of cutting tool information from various metal cutting operations. Gupta (2005, 2007) enhanced and developed a new system to tightly integrate machining selection parameters and process planning. Mardikar (2007) developed a system to perform energy analysis in the wood industry.

Although manufacturing industry is a data-rich environment, research and practice using data mining techniques is relatively new. In the area of data mining and manufacturing, researchers have developed and discussed the need and the design of data warehouse for data mining (Chen and Wu, 2005; Chen et al., 2006). Some researchers have developed models to detect quality using rough-sets (Tseng et al., 2005), association rule-learning (Chen et al., 2004), neural networks (Wang and Feng, 2002; Feng and Wang, 2003), Bayesian networks (Correa et al., 2009), and rule-based learning (Wang et al., 2007; Wang, 2007). Other data mining examples are: in the automotive industry (Strobel and Hrycej, 2006; Chen et al.), in the semi-conductor industry (Dabbas and Chen, 2001; Chen et al., 2006), on the floor-shop (Jenkole et al., 2007), an energy management system (Doukas et al., 2007), and inventory control applications using neural network (Bansal et al., 1998).

Wang (2007) listed the following specific data mining applications in the manufacturing industry:

1. Manufacturing system modeling
2. Manufacturing process control
3. Quality control
4. Monitoring and diagnosis
5. Safety evaluation

6. Process planning and scheduling
7. Optimization of manufacturing yield
8. Assembly selection and
9. Learning robotics
10. Material requirement planning
11. Preventive machine maintenance

Chapter 3

Data Warehouse

A typical data warehouse consists of these layers: an enterprise resource planning system (ERP), which have highly normalized relational and operational databases; data marts, which are formed using extract, load, and transform (ETL) techniques to create relevant data-sets; an OLAP server, which acts as a layer between data marts and analytical engine, and reporting and analytic tools. Figure 3.1 shows a typical data warehouse (Chen and Wu, 2005).

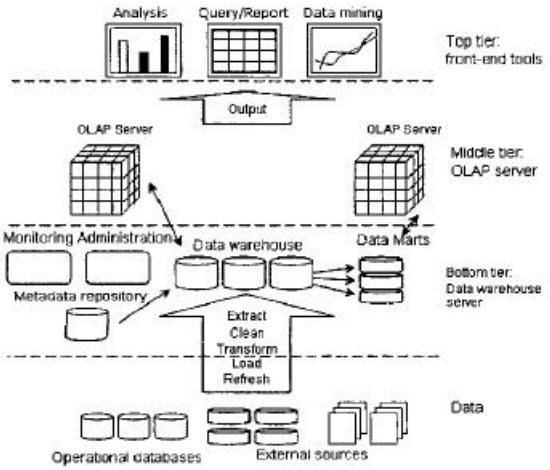

Figure 3.1: A Typical Data Warehouse (Chen and Wu, 2005)

3.1 Relational Database

A relational database is a highly normalized database, which enables the data integrity and eliminates duplicate data. Usually, these databases are the at the first-tier of a typical data warehouse, and in our case, this is the manufacturing execution system (MES) that is updated automatically (by the controls and sensors) or manually (by the operators) on the shop floor. This system has the

most detailed information in the whole data warehouse. Careful design of this database is critical to the overall successful functioning of data warehouse and its applications.

3.2 Data Marts

Data marts are created using extract, load, and transform (ETL) tools, and they still are in a relational database form, but albeit less normalized to provide more data in the same package to the users. As applying data mining techniques directly to the ERP data is a very tedious and a lengthy task, usually, data mining techniques are applied after the data are loaded in the data marts. According to my knowledge, I would suggest the schema, which has the important attributes to carry out a data mining project, given in Figure 3.2 for this project. This data mart would be automatically populated using ETL tools at specific scheduled time. As you would see, almost every table has a primary key and some foreign keys. These keys along with joins define the relationships between tables.

Please note that this schema is based on my knowledge of this problem, general information about the manufacturing process, and a brief literature review. In addition, this is a minimally working example of a data mart for this problem. Obviously, only a domain-expert can design a schema to be efficiently and effectively used for this project.

3.2.1 Table Descriptions

Coolant

This table will hold the information all the coolants, including its type and cost.

- CoolantID
- Name
- Type
- Cost
- pHStability

Degreasers

This table will hold the information on degreasers, including its type and set-up time

- DegreaserID
- DegreaserName
- Type
- SetupTime

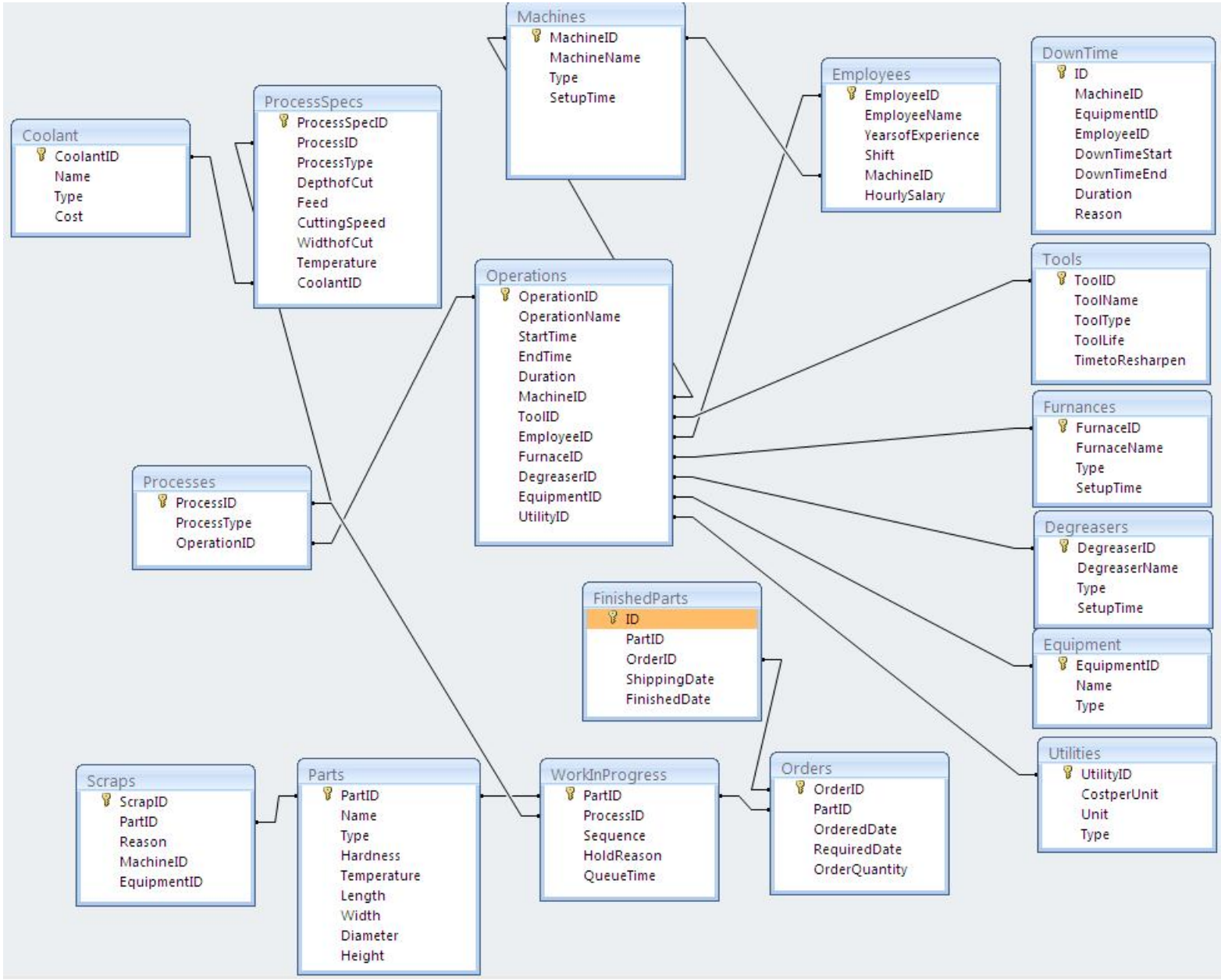

Figure 3.2: Data Mart

DownTime

All non-value added time will be logged in this table, which will list the machine, equipment, or employee, and the reason why there was some down-time and its duration.

- ID
- MachineID
- EquipmentID
- EmployeeID
- DownTimeStart
- DownTimeEnd
- Duration
- Reason

Employees

This table will hold personnel information, including hourly salary, shift, and years of experience.

- EmployeeID
- EmployeeName
- YearsofExperience
- Shift
- MachineID
- HourlySalary

Equipment

This table will list out all the equipment the shop floor has and its type.

- EquipmentID
- Name
- Type

FinishedParts

This table will useful for tracking demand and supply, and for quality testing results, as it will hold the information on finished parts and shipping info.

- ID
- PartID
- OrderID
- QualityTest
- ShippingDate
- FinishedDate

Furnances

This table will hold the furnaces information, including furnace type and its set-up time.

- FurnaceID
- FurnaceName
- Type
- SetupTime

Machines

This table will hold the information about all the machines, its type, and its set-up time.

- MachineID
- MachineName
- Type
- SetupTime

Operations

This is the most important table in this data warehouse, as this table connects all the different table by their IDs. This table will list all the operations that need to be carried on a part and the schedule of these operations.

- OperationID
- OperationName
- StartTime
- EndTime
- Duration

- MachineID
- ToolID
- EmployeeID
- FurnaceID
- DegreaserID
- EquipmentID
- UtilityID
- PartID

Orders

This table is part of supply and demand tracking, because it will hold all the order information, including quantity and required time.

- OrderID
- PartID
- OrderedDate
- RequiredDate
- OrderQuantity

Parts

This table will list all the information about the raw material and its properties. As these properties will play an important role in quality and energy estimation, this table is critical for this project.

- PartID
- Name
- Type
- Hardness
- Temperature
- Length
- Width
- Diameter
- Height

Processes

This table links the operations table with the work-in-progress table.

- ProcessID
- ProcessType
- OperationID

ProcessSpecs

This table will hold the actual specifications of a machining process, including depth of cut, feed rate, cutting speed, and furnace temperature.

- ProcessSpecID
- ProcessID
- ProcessType
- DepthofCut
- Feed
- CuttingSpeed
- WidthofCut
- Temperature
- CoolantID
- CoolantPressure

Scraps

This table will hold information about all scraps along with the machine and equipment id to track down the quality issues.

- ScrapID
- PartID
- Reason
- MachineID
- EquipmentID

Tools

This table will hold the information about tools available for machining and its type, life, and time to re-sharpen.

- ToolID
- ToolName
- ToolType
- ToolLife
- TimetoResharpen

Utilities

This table will hold the different utility types that floor-shop uses, and the associated costs per unit .

- UtilityID
- CostperUnit
- Unit
- Type

WorkInProgress

This table will hold information about work-in-progress parts, including queue time and hold reason.

- PartID
- ProcessID
- Sequence
- HoldReason
- QueueTime

3.3 OLAP and Data Cubes

Based on this data mart, OLAP tools could be used to allow managers perform reporting and quick analysis. Business intelligence tools, especially reporting tools, such as Cognos, Hyperion, Discoverer, and others, could be used to disseminate information faster about orders, work-in-progress, operations, employees, and others.

Experienced SQL users can also write SQL queries to get basic data. For example, if a user wanted to know the set-up times for different machines and the duration that machine was used, he can write a query like this:

```
SELECT Machines.MachineName, Operations.Duration,
Machines.SetupTime
FROM Machines
INNER JOIN
Operations ON Machines.MachineID = Operations.MachineID;
```

Or, let's say that the user wants to find electricity costs by different operations:

```
SELECT Operations.OperationName, Operations.Duration,
Utilities.CostperUnit,
 [Duration]*[CostperUnit] AS EnergyCost, Utilities.Type
FROM Utilities
INNER JOIN
(Machines INNER JOIN Operations ON
Machines.MachineID = Operations.MachineID)
ON Utilities.UtilityID = Operations.UtilityID
WHERE (((Utilities.Type)='Electricity'));
```

I have also used SQL queries to pull the data files that are needed for the different data mining objectives.

Chapter 4

Data Mining

For successful, efficient, and routine data mining projects for this floor shop, I would suggest the framework shown in Figure 4.1 for intelligent decision-making. Chen et al. (2006) used similar framework for semi-conductor manufacturing, and I have modified it to suit the needs of this project. As discussed in previous section, this framework has all those layers with the addition of off-line data mining modeling and early warning/tracking system.

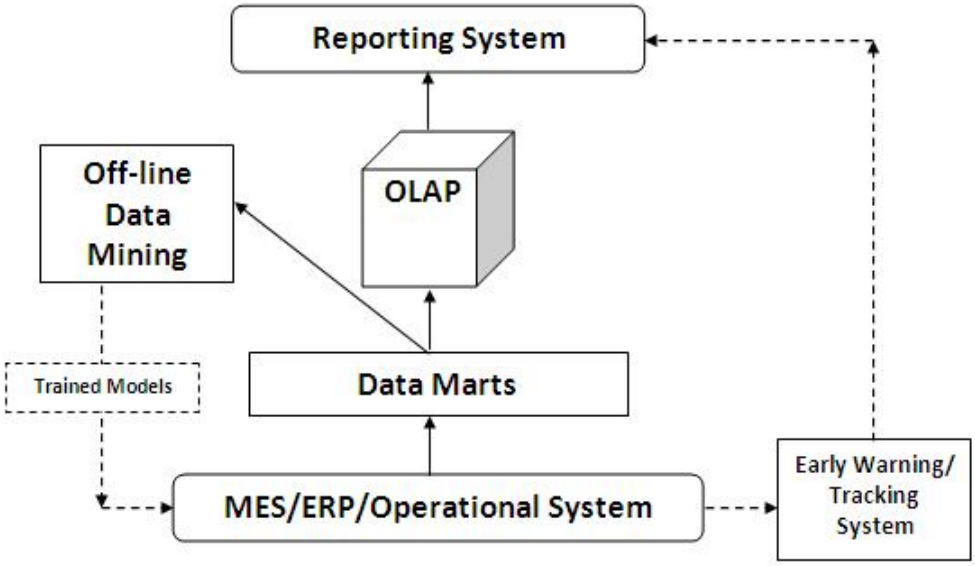

Figure 4.1: A Framework for Intelligent Decision-Making and Analysis

After the data were pulled from data marts based on different objectives, I have applied data mining modeling techniques and develop models to be used in the production scenario. These trained models will be applied on-line and real-time on the production data to populate early warning/tracking system for the desired targets. This system will feed to the reporting system again to

create formal reports of the results of data mining models and early warning system.

In this project, I have not only provided data mining solutions to the shop, but also I have provided a complete framework for data driven decision making. Home-brew reporting applications can also be developed based on the OLAP and/or early warning system data, where simple visualization dashboards can be built for busy managers instead of complicated data mining results.

4.1 Data Generation

I used Excel and Access to generate random values and populate all the tables. I used some details given in Ezugwu et al. (2005) and internet resources as a guideline to generate random values. There might be some rubbish data as a result of random generation of values; however, it could have not been avoided as there were quite a few tables to populate and data structure was completely unrelated to any other publication. Although the knowledge gained from the data mining models using this dataset is more or less meaningless, this dataset should suffice for demonstration purposes.

I have created two different datasets for running data mining algorithms on for two different objectives. As we are interested in quality and energy, I have created two datasets with quality test and energy costs as target variables. Obviously, depending on the objective more datasets can be created and data mining models can be created.

4.1.1 Quality Test Dataset

In this dataset, I have tried to find the reasons why a finished part failed quality testing. Figure 4.2 shows the query to pull the quality test data, and Figure 4.3 shows the SQL query of this dataset.

This dataset includes all the fields from the data-mart that could affect quality, such as, machines, depth of cut, feed, cutting speed, width of cut, coolant type, coolant pressure, coolant pH stability, raw material hardness, length, height, temperature, employee shift, employee experience, tool type used, tool life, and others. Summary of this dataset is shown in Figure 4.4.

4.1.2 Energy Cost Dataset

In this dataset, I have computed energy costs for all the utilities by multiplying the duration of the operation by the cost per unit of that utility. For example, if a steam degreasing operation lasted for 10 mins and if gas was used to generate the steam that was \$5/min, then the energy cost for that operation would be \$50. Of course, this is a over-simplification of the energy cost calculation, but I do not have a background in energy, so I would ask plant engineer's for assistance to come up with a cost/unit for each utility type. Figure 4.5 shows the query that was used to pull the data, and Figure 4.6 shows the SQL query.

This dataset includes all the fields from the data-mart that could affect energy costs, such as, machines, depth of cut, feed, cutting speed, width of cut, coolant type, coolant pressure, coolant pH stability, raw material hardness, length, height, temperature, employee shift, employee experience, tool

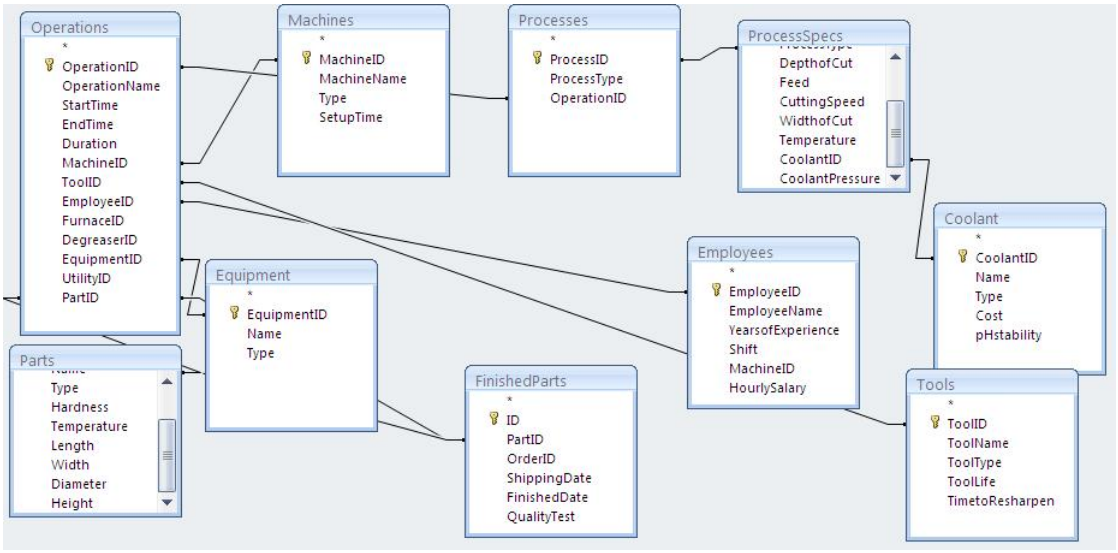

Figure 4.2: Query to Pull Quality Data

type used, tool life, and others with a derived energy cost field. Summary of this dataset is shown in Figure 4.7.

4.1.3 Down Time Dataset

In this dataset, the operators have entered the various reasons, equipment, machines, and employee information and the time that was wasted because of some failure and the reasons of failure. This dataset could be used to find out various reasons of failures by association rule learning (without a target variable) or classification rule learning on the duration of failure. I have used the duration of failure as a target variable for classification. Figure 4.8 shows the SQL query used to pull this dataset. Summary of this dataset is given in Figure 4.9.

```
SELECT Operations.Duration, Operations.MachineID,
Operations.ToolID,
Operations.EmployeeID, Operations.FurnaceID,
Operations.DegreaserID,
Operations.EquipmentID, Operations.UtilityID,
Machines.Type, Machines.SetupTime,
ProcessSpecs.ProcessType, ProcessSpecs.DepthofCut,
ProcessSpecs.Feed,
ProcessSpecs.CuttingSpeed, ProcessSpecs.WidthofCut,
ProcessSpecs.Temperature,
ProcessSpecs.CoolantID, ProcessSpecs.CoolantPressure,
Parts.Hardness, Parts.Temperature, Parts.Length,
Parts.Width, Parts.Height,
Equipment.Type,
Employees.YearsofExperience, Employees.Shift, Employees.HourlySalary,
Coolant.Type, Coolant.pHstability,
Tools.ToolType, Tools.ToolLife,
FinishedParts.QualityTest
FROM Coolant INNER JOIN (Tools INNER JOIN (Parts INNER JOIN
(FinishedParts INNER JOIN (Employees INNER JOIN
(Equipment INNER JOIN (((Machines INNER JOIN Operations ON
Machines.MachineID = Operations.MachineID)
INNER JOIN Processes ON
 Operations.OperationID = Processes.OperationID)
INNER JOIN ProcessSpecs ON
 Processes.ProcessID = ProcessSpecs.ProcessID)
ON Equipment.EquipmentID = Operations.EquipmentID) ON
Employees.EmployeeID = Operations.EmployeeID) ON
FinishedParts.PartID = Operations.PartID) ON
(Parts.PartID = Operations.PartID) AND
(Parts.PartID = FinishedParts.PartID)) ON
Tools.ToolID = Operations.ToolID) ON
Coolant.CoolantID = ProcessSpecs.CoolantID;
```

Figure 4.3: SQL Query to Pull Quality Data

```
   Duration       MachineID  ToolID   EmployeeID FurnaceID DegreaserID  EquipmentID
Min.   : 2.00   3      :11   1:14   10     : 7   1:46      1:47        1      :13
1st Qu.: 9.50   4      :11   2: 8   12     : 7   2:45      2:44        10     :10
Median :20.00   6      :11   3:12   13     : 7                       5      :10
Mean   :19.88   1      :10   4:15   18     : 7                       7      :10
3rd Qu.:29.00   11     :10   5:14   3      : 7                       4      : 9
Max.   :40.00   9      : 9   6:16   14     : 6                       8      : 9
                (Other):29   7:12   (Other):50                       (Other):30
  Machines_Type   SetupTime     UtilityID  ProcessType    DepthofCut         Feed
CNC-Lathe: 9    Min.   :10.00   1:31      Min.   : NA   Min.   :2.000   Min.   :0.2500
Drilling :18    1st Qu.:10.00   2:36      1st Qu.: NA   1st Qu.:3.000   1st Qu.:0.2600
Grinder  :20    Median :15.00   3:24      Median : NA   Median :4.000   Median :0.2800
Lathe    :26    Mean   :14.95             Mean   :NaN   Mean   :3.560   Mean   :0.2769
Milling  :18    3rd Qu.:19.00             3rd Qu.: NA   3rd Qu.:5.000   3rd Qu.:0.2900
                Max.   :20.00             Max.   : NA   Max.   :5.000   Max.   :0.3000
                                          NA's   : 91
 CuttingSpeed     WidthofCut    ProcessSpecs_Temperature CoolantID CoolantPressure
Min.   :20.00   Min.   :1.000   Min.   :100.0            10:24     Min.   :110.0
1st Qu.:20.00   1st Qu.:1.000   1st Qu.:100.0            4 : 6     1st Qu.:110.0
Median :40.00   Median :2.000   Median :300.0            5 : 7     Median :150.0
Mean   :34.84   Mean   :2.044   Mean   :323.1            6 :13     Mean   :157.5
3rd Qu.:40.00   3rd Qu.:3.000   3rd Qu.:500.0            7 :16     3rd Qu.:203.0
Max.   :50.00   Max.   :3.000   Max.   :500.0            8 :14     Max.   :203.0
                                                         9 :11
   Hardness     Parts_Temperature     Length          Width          Diameter       Height
Min.   :20.00   Min.   :1000      Min.   :2.000   Min.   :2.000   Min.   : NA   Min.   : 1.000
1st Qu.:28.00   1st Qu.:1098      1st Qu.:2.000   1st Qu.:2.000   1st Qu.: NA   1st Qu.: 3.500
Median :33.00   Median :1185      Median :3.000   Median :3.000   Median : NA   Median : 5.000
Mean   :32.15   Mean   :1167      Mean   :3.187   Mean   :3.308   Mean   :NaN   Mean   : 4.879
3rd Qu.:37.50   3rd Qu.:1250      3rd Qu.:4.500   3rd Qu.:4.000   3rd Qu.: NA   3rd Qu.: 6.000
Max.   :40.00   Max.   :1300      Max.   :5.000   Max.   :5.000   Max.   : NA   Max.   :10.000
                                                                  NA's   : 91
    Equipment_Type YearsofExperience     Shift     HourlySalary   Coolant_Type  pHstability
Crane      : 9     Min.   : 2.00     Day    :36   Min.   : 20.0   Liquid:51    Min.   :8.000
Forklift   :16     1st Qu.: 9.00     Evening:16   1st Qu.: 90.0   Pastes:40    1st Qu.:8.000
HandTruck  :27     Median :12.00     Night  :39   Median :120.0                Median :8.000
Lift       :21     Mean   :11.71                  Mean   :117.1                Mean   :8.462
PalletTruck:18     3rd Qu.:15.00                  3rd Qu.:150.0                3rd Qu.:9.000
                   Max.   :17.00                  Max.   :170.0                Max.   :9.000

   ToolType     ToolLife     QualityTest
Cutter :37   Min.   :15.00   FAILED:18
Drill  :26   1st Qu.:18.00   OK    :73
EndMill:28   Median :20.00
             Mean   :35.89
             3rd Qu.:50.00
             Max.   :90.00
```

Figure 4.4: Summary of Quality Dataset

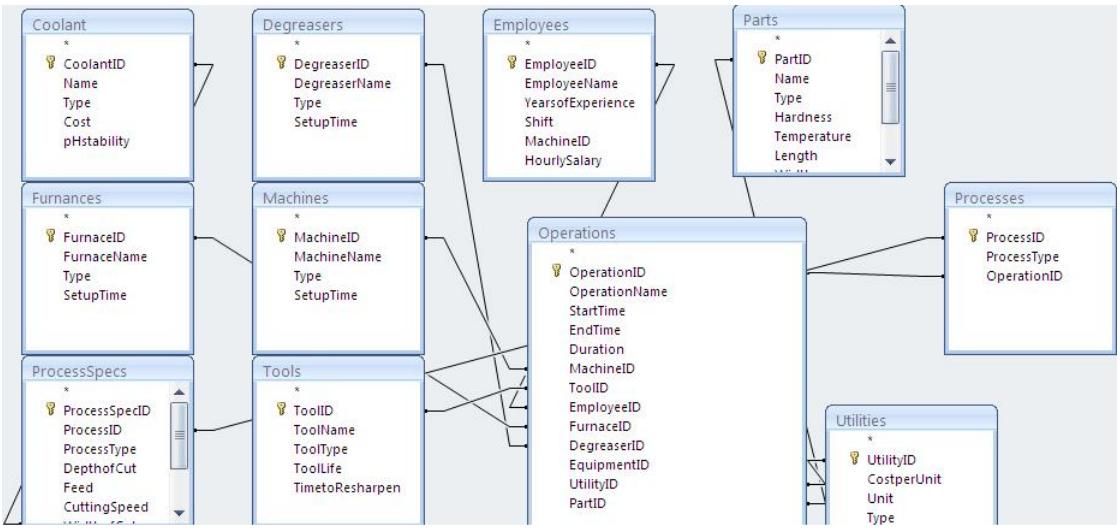

Figure 4.5: Query to Pull Energy Cost Data

```
SELECT Operations.MachineID, Operations.ToolID, Operations.EmployeeID,
Operations.FurnaceID, Operations.DegreaserID, Operations.EquipmentID,
Operations.UtilityID, Operations.PartID,
 Coolant.Type, Coolant.pHstability,
Degreasers.Type, Degreasers.SetupTime,
Employees.YearsofExperience, Employees.Shift, Employees.HourlySalary,
Parts.Type, Parts.Hardness, Parts.Temperature, Parts.Length, Parts.Width,
Parts.Diameter, Parts.Height,
Tools.ToolType, Tools.ToolLife, Tools.TimetoResharpen,
 ProcessSpecs.DepthofCut, ProcessSpecs.Feed,
ProcessSpecs.CuttingSpeed, ProcessSpecs.WidthofCut,
ProcessSpecs.Temperature, ProcessSpecs.CoolantID,
ProcessSpecs.CoolantPressure,
[Duration]*[CostPerUnit] AS EnergyCost

FROM Utilities INNER JOIN (Tools INNER JOIN
(((Parts INNER JOIN (Machines INNER JOIN
 (Furnances INNER JOIN (Employees INNER JOIN
(Degreasers INNER JOIN Operations ON Degreasers.DegreaserID =
Operations.DegreaserID) ON
Employees.EmployeeID = Operations.EmployeeID)
ON Furnances.FurnaceID = Operations.FurnaceID)
ON Machines.MachineID = Operations.MachineID)
ON Parts.PartID = Operations.PartID) INNER JOIN
Processes ON Operations.OperationID = Processes.OperationID)
INNER JOIN (Coolant INNER JOIN ProcessSpecs
ON Coolant.CoolantID = ProcessSpecs.CoolantID)
ON Processes.ProcessID = ProcessSpecs.ProcessID)
ON Tools.ToolID = Operations.ToolID)
ON Utilities.UtilityID = Operations.UtilityID;
```

Figure 4.6: SQL Query to Pull Energy Cost Data

```
  MachineID  ToolID   EmployeeID FurnaceID DegreaserID  EquipmentID UtilityID     PartID
3      :11   1:14   10     : 7   1:46      1:47        1      :13   1:31      13     : 8
4      :11   2: 8   12     : 7   2:45      2:44        10     :10   2:36      2      : 7
6      :11   3:12   13     : 7                          5      :10   3:24      18     : 6
1      :10   4:15   18     : 7                          7      :10             26     : 6
11     :10   5:14   3      : 7                          4      : 9             12     : 5
9      : 9   6:16   14     : 6                          8      : 9             25     : 5
(Other):29   7:12   (Other):50                          (Other):30             (Other):54
Coolant_Type  pHstability    Degreasers_Type   SetupTime      YearsofExperience     Shift
Liquid:51     Min.   :8.000   Steam:91        Min.   :20.00   Min.   : 2.00     Day    :36
Pastes:40     1st Qu.:8.000                  1st Qu.:20.00   1st Qu.: 9.00     Evening:16
              Median :8.000                  Median :30.00   Median :12.00     Night  :39
              Mean   :8.462                  Mean   :25.16   Mean   :11.71
              3rd Qu.:9.000                  3rd Qu.:30.00   3rd Qu.:15.00
              Max.   :9.000                  Max.   :30.00   Max.   :17.00

 HourlySalary   Parts_Type    Hardness     Parts_Temperature     Length          Width
Min.   : 20.0   Alloy:56   Min.   :20.00   Min.   :1000      Min.   :2.000   Min.   :2.000
1st Qu.: 90.0   Steel:35   1st Qu.:28.00   1st Qu.:1098      1st Qu.:2.000   1st Qu.:2.000
Median :120.0              Median :33.00   Median :1185      Median :3.000   Median :3.000
Mean   :117.1              Mean   :32.15   Mean   :1167      Mean   :3.187   Mean   :3.308
3rd Qu.:150.0              3rd Qu.:37.50   3rd Qu.:1250      3rd Qu.:4.500   3rd Qu.:4.000
Max.   :170.0              Max.   :40.00   Max.   :1300      Max.   :5.000   Max.   :5.000

   Diameter       Height          ToolType     ToolLife     TimetoResharpen   DepthofCut
Min.   : NA   Min.   : 1.000   Cutter :37   Min.   :15.00   Min.   :10.00   Min.   :2.000
1st Qu.: NA   1st Qu.: 3.500   Drill  :26   1st Qu.:18.00   1st Qu.:12.00   1st Qu.:3.000
Median : NA   Median : 5.000   EndMill:28   Median :20.00   Median :20.00   Median :4.000
Mean   :NaN   Mean   : 4.879                Mean   :35.89   Mean   :22.53   Mean   :3.560
3rd Qu.: NA   3rd Qu.: 6.000                3rd Qu.:50.00   3rd Qu.:30.00   3rd Qu.:5.000
Max.   : NA   Max.   :10.000                Max.   :90.00   Max.   :40.00   Max.   :5.000
NA's   : 91
     Feed          CuttingSpeed     WidthofCut    ProcessSpecs_Temperature CoolantID
Min.   :0.2500   Min.   :20.00   Min.   :1.000   Min.   :100.0            10:24
1st Qu.:0.2600   1st Qu.:20.00   1st Qu.:1.000   1st Qu.:100.0            4 : 6
Median :0.2800   Median :40.00   Median :2.000   Median :300.0            5 : 7
Mean   :0.2769   Mean   :34.84   Mean   :2.044   Mean   :323.1            6 :13
3rd Qu.:0.2900   3rd Qu.:40.00   3rd Qu.:3.000   3rd Qu.:500.0            7 :16
Max.   :0.3000   Max.   :50.00   Max.   :3.000   Max.   :500.0            8 :14
                                                                          9 :11
CoolantPressure   EnergyCost
Min.   :110.0   Min.   :  2.0
1st Qu.:110.0   1st Qu.: 28.5
Median :150.0   Median : 96.0
Mean   :157.5   Mean   :136.7
3rd Qu.:203.0   3rd Qu.:154.0
Max.   :203.0   Max.   :600.0
```

Figure 4.7: Summary of Energy Cost Dataset

```
SELECT DownTime.MachineID, DownTime.EquipmentID, DownTime.EmployeeID,
 Machines.Type, Machines.SetupTime, Equipment.Type,
 Employees.YearsofExperience, Employees.Shift, Employees.HourlySalary,
DownTime.Duration
FROM ((DownTime INNER JOIN Machines ON
DownTime.MachineID = Machines.MachineID)
INNER JOIN Equipment ON
DownTime.EquipmentID = Equipment.EquipmentID)
INNER JOIN Employees ON
DownTime.EmployeeID = Employees.EmployeeID;
```

Figure 4.8: SQL Query to Pull Down Time Data

```
  MachineID  EquipmentID   EmployeeID   Machines_Type   SetupTime         Equipment_Type
10      :6   3      :4    16     : 5   CNC-Lathe: 2    Min.   :10.00   Crane      :4
4       :6   4      :4    11     : 4   Drilling : 1    1st Qu.:15.00   Forklift   :8
11      :4   5      :4    2      : 4   Grinder  :15    Median :18.00   HandTruck  :7
5       :3   6      :4    12     : 2   Lathe    : 9    Mean   :16.23   Lift       :6
6       :3   8      :4    14     : 2   Milling  : 4    3rd Qu.:20.00   PalletTruck:6
8       :3   1      :3    15     : 2                   Max.   :20.00
(Other):6   (Other):8    (Other):12
YearsofExperience     Shift     HourlySalary       Duration
Min.   : 2.000    Day    : 8   Min.   : 20.00   Min.   : 1.00
1st Qu.: 2.000    Evening: 4   1st Qu.: 20.00   1st Qu.:16.00
Median : 8.000    Night  :19   Median : 80.00   Median :24.00
Mean   : 8.935                 Mean   : 89.35   Mean   :28.29
3rd Qu.:14.000                 3rd Qu.:140.00   3rd Qu.:43.50
Max.   :17.000                 Max.   :170.00   Max.   :61.00
```

Figure 4.9: Summary of Down Time Dataset

```
For each dataset
   Discretize all numeric attributes
   For each discretized data-set
      Apply FSS techniques
         For each FSS'ed data-set
             Apply classification algorithms
   Compare all data set combinations
```

Figure 4.10: Pseudo-code of the Methodology

4.2 Methodology

I have used the methodology given in Figure 4.11 for this project. After creating the data set needed for the data mining modeling, I have applied discretization techniques (discussed in Section 2.1.3), then I have used feature subset selection (discussed in Section 2.1.2) techniques to select relevant attributes, then I have developed different classification models - rules and trees (discussed in Section 2.1.1 and 2.1.2). After developing these different models I compared these models using quartile charts and win-loss tables, and I chose the best model for production to go in the early-warning system. This methodology is given in given in the pseudo-code shown in Figure 4.10.

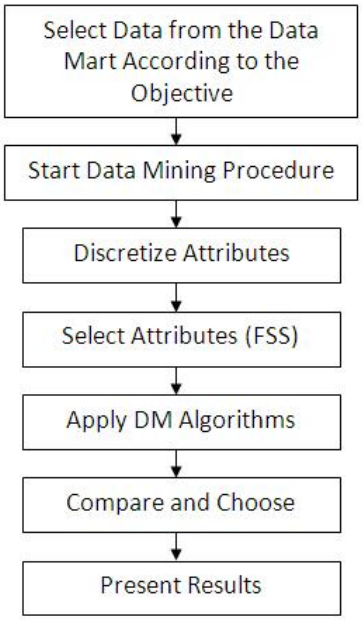

Figure 4.11: Methodology of Data Mining Modeling

After some basic manipulation of data (the plant manager decided to energy cost at $100, and the duration of down time at 30 minutes; these were critical breaking points for the plant engineer), I had three datasets to run the

data mining algorithms on: qryQualityData, which has the quality test data; EnergyCostData_2, which has the energy cost data with the target variable as energy costs greater than $100 or less than equal to $100; and, DownTimeData_2, which has the down time data with the target variable as the duration of down time either greater than 30 minutes or less than equal to 30 minutes.

On all these three datasets, I ran two discretization techniques: Fayyad and Irani's (1992) supervised discretizer and equal-frequency un-supervised discretizer, which attempts to discretize numeric variables in equal frequency bins.

I used three different feature subset selection (FSS) and search method techniques: Correlation-based feature selection (Hall, 1999) with best-first ranking and genetic algorithm ranking (Goldberg, 1989), and attribute evaluator with ranker search (Kira and Rendell, 1992).

I had applied four different algorithms: J48, a tree learner, which is based on Quinlan's C4.5 tree learner (Quinlan, 1993); JRip, a fast effective rule learner (Cohen, 1995); RiDor, a ripple down learner (Gaines and Compton, 1995); and naive Bayes learner, which is based on Bayesian estimator (John and Langley, 1995). I chose these learners, because they are simple to explain and work very fast. In addition, these are some of the standard learners used in the data mining domain.

All these combinations were run using 10-fold cross-validation, hence, generating 10-folds $20 \times$ 3-datasets $\times$ 2-discretizers $\times$ 3-FSS $\times$ 4-algorithms $= 1440$ runs. Following are the legends of these datasets, discretizers, feature subset selectors, and algorithms:

- Datasets
    1. Quality dataset: qryQualityData
    2. Energy cost dataset: EnergyCostData_2
    3. Down time dataset: DownTimeData_2
- Discretizers
    1. Fayyad and Irani's (1992) supervised discretizer: fayyadIrani
    2. Equal-frequency un-supervised discretizer: disceqfreq
- Feature subset selection (FSS)
    1. Correlation-based feature selection (Hall, 1999) with best-first ranking: attrselectCfsBF
    2. Correlation-based feature selection (Hall, 1999) with genetic algorithm ranking (Goldberg, 1989): attrselectCfsGA
    3. attribute evaluator with ranker search (Kira and Rendell, 1992): attrselectReliefRnkr
- Algorithms
    1. J48 tree learner (Quinlan, 1993): j48
    2. JRip rule learner (Cohen, 1995):jrip
    3. RiDor ripple down rule learner (Gaines and Compton, 1995): ridor
    4. Bayesian estimator learner(John and Langley, 1995): bayes

4.3 Results

After running all these combinations, I ran a script to create quartile charts or box plots, which compared each combination with other combinations on the accuracy of that combination, and listed the best combination in ascending order along with the min, max, median, and standard deviation. These charts are given in Table 4.1.

method	sd	max	Chart		
DownTimeData_2.ridor.fayyadIrani.attrselectReliefRnkr	9.7	40.0	-100%		100%
DownTimeData_2.ridor.disceqfreq.attrselectReliefRnkr	9.7	33.3	-100%		100%
qryQualityData.ridor.disceqfreq.attrselectReliefRnkr	11.1	33.3	-100%		100%
qryQualityData.jrip.disceqfreq.attrselectCfsGA	11.1	33.3	-100%		100%
qryQualityData.jrip.disceqfreq.attrselectCfsBF	11.1	33.3	-100%		100%
qryQualityData.ridor.disceqfreq.attrselectCfsGA	11.2	26.7	-100%		100%
qryQualityData.ridor.disceqfreq.attrselectCfsBF	11.2	26.7	-100%		100%
EnergyCostData_2.ridor.fayyadIrani.attrselectCfsGA	10.1	42.2	-100%		100%
EnergyCostData_2.jrip.disceqfreq.attrselectCfsGA	10.1	42.2	-100%		100%
EnergyCostData_2.ridor.fayyadIrani.attrselectReliefRnkr	10.1	40.0	-100%		100%
EnergyCostData_2.bayes.fayyadIrani.attrselectCfsGA	10.1	40.0	-100%		100%
qryQualityData.bayes.disceqfreq.attrselectCfsGA	11.2	20.0	-100%		100%
qryQualityData.bayes.disceqfreq.attrselectCfsBF	11.2	20.0	-100%		100%
DownTimeData_2.bayes.disceqfreq.attrselectReliefRnkr	9.8	13.3	-100%		100%
EnergyCostData_2.ridor.disceqfreq.attrselectCfsGA	10.1	37.8	-100%		100%
EnergyCostData_2.jrip.fayyadIrani.attrselectCfsGA	10.1	37.8	-100%		100%
EnergyCostData_2.bayes.fayyadIrani.attrselectCfsBF	10.1	37.8	-100%		100%
EnergyCostData_2.bayes.disceqfreq.attrselectCfsBF	10.1	37.8	-100%		100%
qryQualityData.bayes.disceqfreq.attrselectReliefRnkr	11.2	17.8	-100%		100%
EnergyCostData_2.bayes.disceqfreq.attrselectReliefRnkr	10.1	35.6	-100%		100%
EnergyCostData_2.ridor.fayyadIrani.attrselectCfsBF	10.1	33.3	-100%		100%
EnergyCostData_2.ridor.disceqfreq.attrselectCfsBF	10.1	33.3	-100%		100%
EnergyCostData_2.jrip.fayyadIrani.attrselectCfsBF	10.1	33.3	-100%		100%
EnergyCostData_2.jrip.disceqfreq.attrselectCfsBF	10.1	33.3	-100%		100%
EnergyCostData_2.j48.fayyadIrani.attrselectReliefRnkr	10.1	33.3	-100%		100%
EnergyCostData_2.j48.fayyadIrani.attrselectCfsGA	10.1	33.3	-100%		100%
EnergyCostData_2.j48.fayyadIrani.attrselectCfsBF	10.1	33.3	-100%		100%
EnergyCostData_2.j48.disceqfreq.attrselectReliefRnkr	10.1	33.3	-100%		100%
EnergyCostData_2.j48.disceqfreq.attrselectCfsGA	10.1	33.3	-100%		100%
EnergyCostData_2.j48.disceqfreq.attrselectCfsBF	10.1	33.3	-100%		100%
EnergyCostData_2.bayes.fayyadIrani.attrselectReliefRnkr	10.1	33.3	-100%		100%
EnergyCostData_2.bayes.disceqfreq.attrselectCfsGA	10.1	33.3	-100%		100%
qryQualityData.jrip.disceqfreq.attrselectReliefRnkr	11.2	13.3	-100%		100%
qryQualityData.j48.disceqfreq.attrselectReliefRnkr	11.2	13.3	-100%		100%
qryQualityData.j48.disceqfreq.attrselectCfsGA	11.2	13.3	-100%		100%
qryQualityData.j48.disceqfreq.attrselectCfsBF	11.2	13.3	-100%		100%
DownTimeData_2.jrip.fayyadIrani.attrselectReliefRnkr	9.8	6.7	-100%		100%
DownTimeData_2.jrip.fayyadIrani.attrselectCfsGA	9.8	6.7	-100%		100%

Table 4.1 continued on next page

method	sd	max		Chart	
DownTimeData_2.jrip.fayyadIrani.attrselectCfsBF	9.8	6.7	-100%		100%
DownTimeData_2.jrip.disceqfreq.attrselectReliefRnkr	9.8	6.7	-100%		100%
DownTimeData_2.jrip.disceqfreq.attrselectCfsGA	9.8	6.7	-100%		100%
DownTimeData_2.jrip.disceqfreq.attrselectCfsBF	9.8	6.7	-100%		100%
DownTimeData_2.bayes.fayyadIrani.attrselectReliefRnkr	9.8	6.7	-100%		100%
DownTimeData_2.bayes.fayyadIrani.attrselectCfsGA	9.8	6.7	-100%		100%
DownTimeData_2.bayes.fayyadIrani.attrselectCfsBF	9.8	6.7	-100%		100%
DownTimeData_2.bayes.disceqfreq.attrselectCfsGA	9.8	6.7	-100%		100%
DownTimeData_2.bayes.disceqfreq.attrselectCfsBF	9.8	6.7	-100%		100%
DownTimeData_2.ridor.fayyadIrani.attrselectCfsGA	9.8	0.0	-100%		100%
DownTimeData_2.ridor.fayyadIrani.attrselectCfsBF	9.8	0.0	-100%		100%
DownTimeData_2.ridor.disceqfreq.attrselectCfsGA	9.8	0.0	-100%		100%
DownTimeData_2.ridor.disceqfreq.attrselectCfsBF	9.8	0.0	-100%		100%
DownTimeData_2.j48.fayyadIrani.attrselectReliefRnkr	9.8	0.0	-100%		100%
DownTimeData_2.j48.fayyadIrani.attrselectCfsGA	9.8	0.0	-100%		100%
DownTimeData_2.j48.fayyadIrani.attrselectCfsBF	9.8	0.0	-100%		100%
DownTimeData_2.j48.disceqfreq.attrselectReliefRnkr	9.8	0.0	-100%		100%
DownTimeData_2.j48.disceqfreq.attrselectCfsGA	9.8	0.0	-100%		100%
DownTimeData_2.j48.disceqfreq.attrselectCfsBF	9.8	0.0	-100%		100%
qryQualityData.ridor.fayyadIrani.attrselectReliefRnkr	11.2	4.5	-100%		100%
qryQualityData.ridor.fayyadIrani.attrselectCfsGA	11.2	4.5	-100%		100%
qryQualityData.ridor.fayyadIrani.attrselectCfsBF	11.2	4.5	-100%		100%
qryQualityData.jrip.fayyadIrani.attrselectReliefRnkr	11.2	4.5	-100%		100%
qryQualityData.j48.fayyadIrani.attrselectReliefRnkr	11.2	4.5	-100%		100%
qryQualityData.j48.fayyadIrani.attrselectCfsGA	11.2	4.5	-100%		100%
qryQualityData.j48.fayyadIrani.attrselectCfsBF	11.2	4.5	-100%		100%
qryQualityData.bayes.fayyadIrani.attrselectReliefRnkr	11.2	4.5	-100%		100%
EnergyCostData_2.ridor.disceqfreq.attrselectReliefRnkr	10.1	20.0	-100%		100%
qryQualityData.jrip.fayyadIrani.attrselectCfsGA	11.2	0.0	-100%		100%
qryQualityData.jrip.fayyadIrani.attrselectCfsBF	11.2	0.0	-100%		100%
qryQualityData.bayes.fayyadIrani.attrselectCfsGA	11.2	0.0	-100%		100%
qryQualityData.bayes.fayyadIrani.attrselectCfsBF	11.2	0.0	-100%		100%
EnergyCostData_2.jrip.disceqfreq.attrselectReliefRnkr	10.0	4.5	-100%		100%
EnergyCostData_2.jrip.fayyadIrani.attrselectReliefRnkr	10.0	0.0	-100%		100%

Table 4.1: Quartile Charts of All Runs

Along with the quartile charts, I ran a script to create a win-loss table, which tests each combination with all other combinations using Mann-Whitney U Test at 95% confidence level. If a combination is better than the other combination, that combination is termed as a win and a loss for the other combination, and it is a tie if they are not better than each other. Table 4.2 lists the results of this test in a form of a win-loss table, which is sorted descending for wins.

method	ties	wins	losses
qryQualityData.jrip.disceqfreq.attrselectCfsGA	5	66	0
qryQualityData.bayes.disceqfreq.attrselectReliefRnkr	5	66	0

Table 4.2 continued on next page

method	ties	wins	losses
qryQualityData.jrip.disceqfreq.attrselectReliefRnkr	7	64	0
qryQualityData.jrip.disceqfreq.attrselectCfsBF	9	62	0
qryQualityData.bayes.disceqfreq.attrselectCfsBF	11	60	0
qryQualityData.bayes.disceqfreq.attrselectCfsGA	13	58	0
qryQualityData.j48.disceqfreq.attrselectCfsBF	16	49	6
qryQualityData.j48.disceqfreq.attrselectReliefRnkr	16	49	6
qryQualityData.ridor.disceqfreq.attrselectCfsBF	20	49	2
qryQualityData.j48.disceqfreq.attrselectCfsGA	16	49	6
qryQualityData.ridor.disceqfreq.attrselectCfsGA	19	48	4
qryQualityData.ridor.disceqfreq.attrselectReliefRnkr	20	48	3
qryQualityData.j48.fayyadIrani.attrselectCfsGA	18	48	5
qryQualityData.j48.fayyadIrani.attrselectReliefRnkr	21	48	2
qryQualityData.j48.fayyadIrani.attrselectCfsBF	18	48	5
qryQualityData.ridor.fayyadIrani.attrselectCfsGA	20	45	6
qryQualityData.ridor.fayyadIrani.attrselectCfsBF	20	45	6
qryQualityData.ridor.fayyadIrani.attrselectReliefRnkr	20	45	6
qryQualityData.jrip.fayyadIrani.attrselectCfsGA	25	43	3
qryQualityData.jrip.fayyadIrani.attrselectCfsBF	29	38	4
qryQualityData.bayes.fayyadIrani.attrselectReliefRnkr	28	33	10
qryQualityData.bayes.fayyadIrani.attrselectCfsBF	33	32	6
EnergyCostData_2.bayes.disceqfreq.attrselectCfsBF	27	29	15
EnergyCostData_2.j48.disceqfreq.attrselectCfsBF	24	29	18
EnergyCostData_2.bayes.fayyadIrani.attrselectCfsBF	28	28	15
EnergyCostData_2.j48.disceqfreq.attrselectCfsGA	24	28	19
qryQualityData.jrip.fayyadIrani.attrselectReliefRnkr	37	28	6
EnergyCostData_2.j48.disceqfreq.attrselectReliefRnkr	25	28	18
EnergyCostData_2.j48.fayyadIrani.attrselectCfsGA	24	28	19
EnergyCostData_2.bayes.fayyadIrani.attrselectCfsGA	28	28	15
EnergyCostData_2.j48.fayyadIrani.attrselectReliefRnkr	25	27	19
EnergyCostData_2.jrip.fayyadIrani.attrselectCfsBF	25	27	19
qryQualityData.bayes.fayyadIrani.attrselectCfsGA	39	26	6
EnergyCostData_2.j48.fayyadIrani.attrselectCfsBF	26	25	20
EnergyCostData_2.jrip.fayyadIrani.attrselectCfsGA	27	25	19
EnergyCostData_2.jrip.disceqfreq.attrselectCfsBF	26	25	20
EnergyCostData_2.ridor.fayyadIrani.attrselectCfsBF	26	25	20
EnergyCostData_2.jrip.fayyadIrani.attrselectReliefRnkr	26	25	20
EnergyCostData_2.ridor.disceqfreq.attrselectCfsBF	26	25	20
EnergyCostData_2.bayes.disceqfreq.attrselectCfsGA	25	24	22
EnergyCostData_2.jrip.disceqfreq.attrselectCfsGA	26	24	21
EnergyCostData_2.bayes.disceqfreq.attrselectReliefRnkr	15	24	32
EnergyCostData_2.ridor.fayyadIrani.attrselectCfsGA	25	24	22
EnergyCostData_2.bayes.fayyadIrani.attrselectReliefRnkr	17	24	30
EnergyCostData_2.ridor.disceqfreq.attrselectCfsGA	23	24	24
EnergyCostData_2.ridor.fayyadIrani.attrselectReliefRnkr	25	24	22
EnergyCostData_2.ridor.disceqfreq.attrselectReliefRnkr	8	24	39
EnergyCostData_2.jrip.disceqfreq.attrselectReliefRnkr	14	24	33
DownTimeData_2.ridor.disceqfreq.attrselectReliefRnkr	14	9	48
DownTimeData_2.ridor.fayyadIrani.attrselectReliefRnkr	14	9	48
DownTimeData_2.ridor.fayyadIrani.attrselectCfsBF	18	5	48
DownTimeData_2.ridor.fayyadIrani.attrselectCfsGA	18	5	48
DownTimeData_2.ridor.disceqfreq.attrselectCfsBF	19	4	48
DownTimeData_2.ridor.disceqfreq.attrselectCfsGA	19	4	48
DownTimeData_2.bayes.disceqfreq.attrselectReliefRnkr	23	0	48
DownTimeData_2.bayes.fayyadIrani.attrselectCfsGA	21	0	50
DownTimeData_2.j48.fayyadIrani.attrselectCfsGA	23	0	48
DownTimeData_2.j48.disceqfreq.attrselectCfsBF	23	0	48
DownTimeData_2.jrip.fayyadIrani.attrselectCfsBF	17	0	54
DownTimeData_2.jrip.disceqfreq.attrselectCfsGA	17	0	54
DownTimeData_2.bayes.disceqfreq.attrselectCfsGA	19	0	52
DownTimeData_2.bayes.fayyadIrani.attrselectCfsBF	21	0	50
DownTimeData_2.j48.fayyadIrani.attrselectCfsBF	23	0	48

Table 4.2 continued on next page

method	ties	wins	losses
DownTimeData_2.jrip.disceqfreq.attrselectReliefRnkr	21	0	50
DownTimeData_2.bayes.fayyadIrani.attrselectReliefRnkr	23	0	48
DownTimeData_2.jrip.fayyadIrani.attrselectReliefRnkr	23	0	48
DownTimeData_2.j48.disceqfreq.attrselectReliefRnkr	23	0	48
DownTimeData_2.j48.disceqfreq.attrselectCfsGA	23	0	48
DownTimeData_2.j48.fayyadIrani.attrselectReliefRnkr	23	0	48
DownTimeData_2.jrip.disceqfreq.attrselectCfsBF	17	0	54
DownTimeData_2.bayes.disceqfreq.attrselectCfsBF	21	0	50
DownTimeData_2.jrip.fayyadIrani.attrselectCfsGA	17	0	54

Table 4.2: Win-Loss Table of All Runs

After looking at the quartile charts and win-loss table, for each dataset, I picked the best combination, and ran the models again to produce a deliverable theory. Here are the three combinations for each dataset:

1. qryQualityData.ridor.disceqfreq.attrselectReliefRnkr
2. EnergyCostData_2.ridor.fayyadIrani.attrselectCfsGA
3. DownTimeData_2.ridor.fayyadIrani.attrselectReliefRnkr

4.3.1 Model for Quality Data

Using the combination, qryQualityData.ridor.disceqfreq.attrselectReliefRnkr, the data mining model produced the following rule with an overall accuracy of 91%:

```
QualityTest = OK
  Except (Height > 4.5) and (Hardness > 37) => QualityTest = FAILED
  Except (Width > 4.5) and (Hardness > 27) => QualityTest = FAILED
```

4.3.2 Model for Energy Cost Data

Using the combination, EnergyCostData_2.ridor.fayyadIrani.attrselectCfsGA, the data mining model produce the following rule with an overall accuracy of 61%.

```
EnergyCostBin = <=100
 Except (WidthofCut = '(2.5-inf)') and (ToolLife = '(30-45]')
                      => EnergyCostBin = >100
 Except (CuttingSpeed = '(25-35]') => EnergyCostBin = >100
 Except (ToolLife = '(16.5-18.5]') => EnergyCostBin = >100
```

4.3.3 Model for Down Time Data

Using the combination, DownTimeData_2.ridor.fayyadIrani.attrselectReliefRnkr, the data mining model produced the following rule with an overall accuracy of 51%:

```
DurationBin = >30
    Except (SetupTime > 16.5) and (YearsofExperience > 10)
                          => DurationBin = <=30
    Except (EmployeeID = 11) => DurationBin = <=30
```

4.4 Usage

Apart from the models that were developed for this project, the plant engineer can reap numerous benefits from this framework, such as, building dashboard to visualize the information, identify and act instantly on problems, develop predictive models based on the data, and others. Once the off-line data mining modeling is completed, these rules can be fed in the early warning/tracking system, which could be used to generate instantaneous reports. In addition, as the MES or a similar system would automatically garner information from the operators and machines (via controls and sensors), the plant engineer will have a continuous flow of data to do more analysis and planning.

4.4.1 Usage of Model for Quality Data

As we can see from the rule generated for the quality dataset, whenever the height of the raw material exceeds 4.5 inches and material hardness exceeds 37 HRC, or the width of the raw material exceeds 4.5 and material hardness exceeds 27 HRC, the finished part failed the quality test. That means that either the machines or the tools are not working well above these specifications. The plant engineer can study these attributes (machines, tools, etc) and see why the shop is unable to produce if the raw material is of higher specification.

4.4.2 Usage of Model for Energy Cost Data

This model explained that the when the width of cut was higher than 2.5 inches and tool life was between 30 to 45 mins, or the cutting speed was between 25 to 35 minutes, or the tool life was between 16.5 minutes to 18.5 minutes, the energy cost was above $100. To minimize energy costs, the plant engineer can try to procure tools with higher tool life, but at the same time are able to handle higher cut (speed and width of cut) specifications.

4.4.3 Usage of Model for Down Time Data

Although this model did not produce rules with high accuracy, it sill provided three key factors: Set-up time, years of experience, and employee with ID of 11. These rules explain that whenever set-up time is higher than 16.5 minutes and the operator working on this job has 10 or more years of experience, the down time is less than 30 minutes; and that the employee ID 11 has down-time considerably low all the time. Here is the record of employee ID 11:

- Years of experience: 2
- Shift: night
- Hourly Salary: 20

By looking at this record, we can make an inference that this employee is working hard as he is significantly underpaid, he is working at night shift, and he does not enough work experience; therefore, although he commits mistakes often, he works hard to keep the down time minimum. The plant engineer should think of providing more training to employees with less experience and are working at night shift.

Chapter 5

Conclusions

For a typical job shop with milling, grinding, and cutting machines, and steam degreasers and heat treatment units, a data warehouse (to get the relevant fields for quality, energy, inventory levels, and others) and a framework for data mining/business intelligence tools was developed. From the data warehouse, three different datasets on quality, energy costs, and down time were extracted. Using these datasets, various combinations of data mining tools (discretizers, attribute selection methods, and data mining learners) were applied to produce predictive rules, which were very simple and easy to understand.

Using this complete framework, the plant engineer can monitor, analyze, and act on various areas of this plant, such as, supply and demand, quality, Overall Equipment Effectiveness (derived from machine/equipment down time), energy costs, employee development, and others. This framework provides the plant engineer the power of retrieving the information on the fly, and provides the much needed assistance in converting the overload of information in actionable knowledge. This knowledge will help the plant engineer to make efficient decisions and plan processes and schedules better.

To conclude this project please find this index guide to answers of the questions asked in Chapter 1:

1. The data that you would want to collect on the shop floor and on support equipment: discussed in section 3.2 on page number 14.
2. Where the sensors would be placed and the nature of data that they would collect and at what frequency: discussed in section 2.4 on page number 10.
3. Data acquisition protocols you would consider using: discussed in sections 2.2 and 2.4 on page numbers from 7 to 10.
4. The data mining strategy and specific algorithms you would use: discussed in section 4.2 on page number 31.
5. How the results will be presented to the plant engineer/manager in a manner that is easy to understand: discussed in section 4.3.1 on page number 36.
6. The use of predictive rule induction and its use for this production scenario: discussed in section 4.3.1 on page number 36.

7. How the plant engineer/manager could use your system on a frequent basis and the types of benefits that they would derive: discussed in section 4.4 on page number 37.

Bibliography

B. Arezoo, K. Ridgway, and AMA Al-Ahmari. Selection of cutting tools and conditions of machining operations using an expert system. *Computers in Industry*, 42(1):43–58, 2000.

K. Bansal, S. Vadhavkar, and A. Gupta. Brief Application Description. Neural Networks Based Forecasting Techniques for Inventory Control Applications. *Data mining and knowledge discovery*, 2(1):97–102, 1998.

M. J. Berry and G. Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Support*. John Wiley & Sons, Inc. New York, NY, USA, 1997.

Kai-Ying Chen and Teh-Chang Wu. Data warehouse design for manufacturing execution systems. pages 751–756, July 2005.

Ruey-Shun Chen, Yung-Shun Tsai, and Chan-Chine Chang. Design and implementation of an intelligent manufacturing execution system for semiconductor manufacturing industry. volume 4, pages 2948–2953, July 2006.

W.C. Chen, S.S. Tseng, and C.Y. Wang. A novel manufacturing defect detection method using data mining approach. *Lecture notes in computer science*, pages 77–86, 2004.

Y.S. Chen, C.H. Cheng, and C.J. Lai. A hybrid procedure for extracting rules of production performance in the automobile parts industry. *Journal of Intelligent Manufacturing*, pages 1–15.

W.W. Cohen. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Citeseer, 1995.

M. Correa, C. Bielza, and J. Pamies-Teixeira. Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process. *Expert Systems With Applications*, 36(3P2):7270–7279, 2009.

R.M. Dabbas and H.N. Chen. Mining semiconductor manufacturing data for productivity improvementan integrated relational database approach. *Computers in Industry*, 45(1):29–44, 2001.

H. Doukas, K.D. Patlitzianas, K. Iatropoulos, and J. Psarras. Intelligent building energy management system using rule sets. *Building and Environment*, 42(10): 3562–3569, 2007.

E.O. Ezugwu, D.A. Fadare, J. Bonney, R.B. Da Silva, and W.F. Sales. Modelling the correlation between cutting and process parameters in high-speed machining of inconel 718 alloy using an artificial neural network. *International Journal of Machine Tools and Manufacture*, 45(12-13):1375 – 1385, 2005. ISSN 0890-6955. doi: DOI:10.1016/j.ijmachtools.2005.02.004. URL http://www.sciencedirect.com/science/article/B6V4B-4FSX697-2/2/b6a3a18802079ee7eaa0c6b98b2fb369.

U.M. Fayyad and K.B. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, 8(1):87–102, 1992.

C.X. Feng and X.F. Wang. Surface roughness predictive modeling: neural networks versus regression. *IIE Transactions*, 35(1):11–27, 2003.

BR Gaines and P. Compton. Induction of ripple-down rules applied to modeling large databases. *Journal of Intelligent Information Systems*, 5(3):211–228, 1995.

D.E. Goldberg. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

B. Gopalakrishnan. Expert systems for machining parameter selection: Design aspects. *Advance Manufacturing Engineering*, 2, April 1990.

Deepak Prakash Gupta. Energy sensitive machining parameter optimization model. Master's thesis, West Virginia University, 2005. http://hdl.handle.net/10450/4406.

Deepak Prakash Gupta. *Development of an integrated model for process planning and parameter selection for machining processes*. PhD thesis, West Virginia University, 2007. http://hdl.handle.net/10450/5468.

M.A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, MA, 2001.

J. Jenkole, P. Kralj, N. Lavrac, and A. Sluga. A data mining experiment on manufacturing shop floor data. In *Proceedings of the 40th International Seminar on Manufacturing Systems (CIRP-07)*, 2007.

G.H. John and P. Langley. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann, 1995.

S. Kalaga and T. Kazic. Automated extraction of Cutting Tool information for various Metal Cutting Operations.

K. Kira and L.A. Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning table of contents*, pages 249–256. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1992.

J. Kletti. *MES-Manufacturing Execution System*. Springer, 2006.

D. Kuonen. Data mining and statistics: What is the connection?, 2004.

Yogesh Mukesh Mardikar. *Establishing Baseline Electrical Energy Consumption in Wood Processing Sawmills: A Model Based on Energy Analysis and Diagnostics*. PhD thesis, West Virginia University, 2007. http://hdl.handle.net/10450/5412.

MESA. MES Explained: A High Level Vision. *MESA international White Paper*, 1997.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

J. R. Quinlan. Improved use of continuous attributes in C4. 5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.

J.R. Quinlan. *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.

C.M. Strobel and T. Hrycej. A data mining approach to the joint evaluation of field and manufacturing data in automotive industry. *Lecture Notes in Computer Science*, 4213:625, 2006.

T.L. Tseng, Y. Kwon, and Y.M. Ertekin. Feature-based rule induction in machining operation using rough set theory for quality assurance. *Robotics and Computer Integrated Manufacturing*, 21(6):559–567, 2005.

K. Wang. Applying data mining to manufacturing: the nature and implications. *Journal of Intelligent Manufacturing*, 18(4):487–495, 2007.

K. Wang, S. Tong, B. Eynard, L. Roucoules, and N. Matta. Application of data mining in manufacturing quality data. In *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, pages 5382–5385, 2007.

X. Wang and CX Feng. Development of empirical models for surface roughness prediction in finish turning. *The International Journal of Advanced Manufacturing Technology*, 20(5):348–356, 2002.

I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, San Francisco, 2 edition, 2005.

Y. Yang and G.I. Webb. Proportional k-interval discretization for naïve-Bayes classifiers. *Proceedings of the 12th European Conference on Machine Learning*, pages 564–575, 2001.

Y. Yang and G.I. Webb. Weighted Proportional k-Interval Discretization for Naive-Bayes Classifiers. *Lecture Notes in Artificial Intelligence*, 2637:501–512, 2003.