KANTIPUR ENGINEERING COLLEGE

(Affiliated to Tribhuvan University)

Dhapakhel, Lalitpur

[Subject Code: CT755]

A MAJOR PROJECT FINAL REPORT ON

”PROJECT X : A BUSINESS INTELLIGENCE FOR BIKE INSURANCE”

Submitted by:

Amit Dhoju	**[091/BCT/2071]**
Kaushtup Bista	**[103/BCT/2071]**
Roshi Maharajan	**[117/BCT/2071]**

A MAJOR PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER ENGINEERING

Submitted to:

Department of Computer and Electronics Engineering

July, 2018

”PROJECT X : A BUSINESS INTELLIGENCE FOR BIKE INSURANCE”

Submitted by:

Amit Dhoju **[091/BCT/2071]**

Kaushtup Bista **[103/BCT/2071]**

Roshi Maharajan **[117/BCT/2071]**

A MAJOR PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER ENGINEERING

Submitted to:

Department of Computer and Electronics Engineering

Kantipur Engineering College

Dhapakhel, Lalitpur

July, 2018

ABSTRACT

There is no alternative for intelligence. What if intelligence used in gambling? the result is success. In short this is our project, title "**Project X:A Business Intelligence For Bike Insurance**". Our project describe a predict policy for bike insurance by visualization towards the history of bike insurance data. Our system will train for our intelligence and will be able to predict the possible claim. The insurance industry worldwide is facing the challenges of deregulation, consolidation and convergence of financial services. There is today a pressing demand for cutting edge services of insurance business management and enriched customer experiences at a significantly lower cost. Insurance is gambling and we are trying to give technique to win this gamble. Our system will determine the possible accident of bike so that management of insurance in-charge policy for new client whether to insurance or not. If system predicts the possible claim then system will suggest the insurance company not to make that insurance otherwise insurance is made. Based upon CC of Bike, Bike Manufacturer, Lote, Zone, and Type Cover, our system will predict the possible claim.

Keywords– Business Intelligence, Prediction, Analysis, Insurance

ACKNOWLEDGMENT

Write Acknowledgment Here. Ea his munere torquatos, quidam essent luptatum cu pro. Ei duo scaevola

Second para of Acknowledgment. Sed veri aeque persecuti ut. Ut accusam mediocrem accusamus eos, quis

to display members name under Acknowledgement

Amit Dhoju	[091/BCT/2071]
Kaushtup Bista	[103/BCT/2071]
Roshi Maharajan	[117/BCT/2071]

TABLE OF CONTENTS

Abstract i

Acknowledgment ii

List of Figures v

List of Abbreviations vi

1 Introduction 1

1.1 Background . . . 1

1.1.1 Insurance . . . 2

1.1.2 Definition of Non-Life Insurance . . . 2

1.1.3 Principle of Insurance . . . 2

1.1.4 Insurance Risk . . . 3

1.2 Crystal Report . . . 3

1.3 Problem Statement . . . 4

1.4 Objectives . . . 4

1.5 Application . . . 4

1.6 Project Features . . . 4

1.7 Feasibility Analysis . . . 5

1.7.1 Economic Feasibility . . . 5

1.7.2 Technical Feasibility . . . 5

1.7.3 Operational Feasibility . . . 5

1.8 System Requirement . . . 6

1.8.1 Software Requirement . . . 6

1.8.2 Hardware Requirement . . . 6

2 Literature Review 7

2.1 Sisense . . . 7

2.1.1 How Sisense work . . . 8

2.2 Qlik Sense . . . 9

2.2.1 How does Qlik sense work . . . 9

3 Theory 10

3.1 Bayes' Theorem . . . 10

3.2 Classification . . . 11

3.3 Bayesian Classification 11
3.3.1 Learning Models 12
3.4 Naïve Bayesian Classifier 12
3.5 Data Pre-Processing 13
3.5.1 Measure of data Quality 13
3.5.2 Major Task in Data Pre-processing 14
3.5.3 Data Smoothing 14
3.5.4 SOMTE Technique 15
3.6 Confusion Matrix 15
3.7 Implementing Algorithm 16
4 Methodology **18**
4.1 E-R Diagram 18
4.2 Flow Chart 19
4.3 Use-Case 20
4.4 Class Diagram 21
4.5 Sequence Diagram 22
4.6 Collaboration Diagram 24
4.7 Software Development Process 25
5 Result and Discussion **27**
5.1 Works Completed 27
5.2 Works Remaining 27
5.3 Problems Faced 27
5.4 Work Schedule 28
References **29**

LIST OF FIGURES

1.1 Insurance Risk 3
3.1 Structure of Naïve Bayesian Network 12
3.2 Confusion Matrix 16
4.1 Entity Relationship Diagram 18
4.2 Flow Chart Of Project 19
4.3 Use-case Diagram 20
4.4 Class Diagram 21
4.5 Sequence Diagram: Training 22
4.6 Sequence Diagram: Testing 22
4.7 Sequence Diagram 23
4.8 Collaboration Diagram 24
4.9 Incremental Development Model Chart 25
5.1 Gantt Chart 28

LIST OF ABBREVIATIONS

BI Business Intelligence

CC Cubic Centimeter

CSV Comma Seprated Value

NN Neural Network

ODBC Open Database Connectivity

OS Operating System

SQL Server Query Language

SVM Support Vector Machine

CHAPTER 1
INTRODUCTION

1.1 Background

Business Intelligence(BI) is defined as the ability for an organization to take all its capabilities and convert them into knowledge. This produces large amounts of information which can lead to the development of new opportunities for the organization. When these opportunities have been identified and a strategy has been effectively implemented, they can provide an organization with a competitive advantage in the market, and stability in the long run.[1]

Business Intelligence in Insurance provides the information to insurance company on claim management of motorbike which helps in gain insight and visibility on the cost of claims; analyze claims breakdown by CC of Bike, Bike Manufacturer, Lote , Zone and Type Cover. To predict and analyze the insurance historical data based on the claims, data can identify suitable polices for risk modelling , reinsurance, profitability analysis, loss analysis, claim analysis, claim estimation by using the Back propagation algorithm to find the claim pattern according to CC of Bike, Bike Manufacturer, Lote, Zone, and Type Cover. Some insurers have gone for non-scalable temporary solutions, which often fail to leverage the ever-increasing volumes of data. Hence, recognizing the need for an effective business a data warehouse cannot be the answer to all the information requirements hence it is also very important to set clear business objectives for the business intelligence solution with total top management support.

1.1.1 Insurance

A practice by which a company provides a guarantee of compensation for specified loss, damage, illness, or death in return for payment.

1.1.2 Definition of Non-Life Insurance

Non-life insurance, also called property and casualty insurance, is a type of coverage that is very common and covers businesses and individuals. It protects them, monetarily, from disaster by providing money in the event of a financial loss. Before you purchase this type of insurance or if you already own any kind of non-life insurance, you should understand what it is.

1.1.3 Principle of Insurance

1. Insurance transfers the financial consequences of an existing risk.
2. Law of Large Numbers
3. The more predictable the outcome
4. Pricing and risk management
5. Considerations in Setting Premium Rates

1.1.4 Insurance Risk

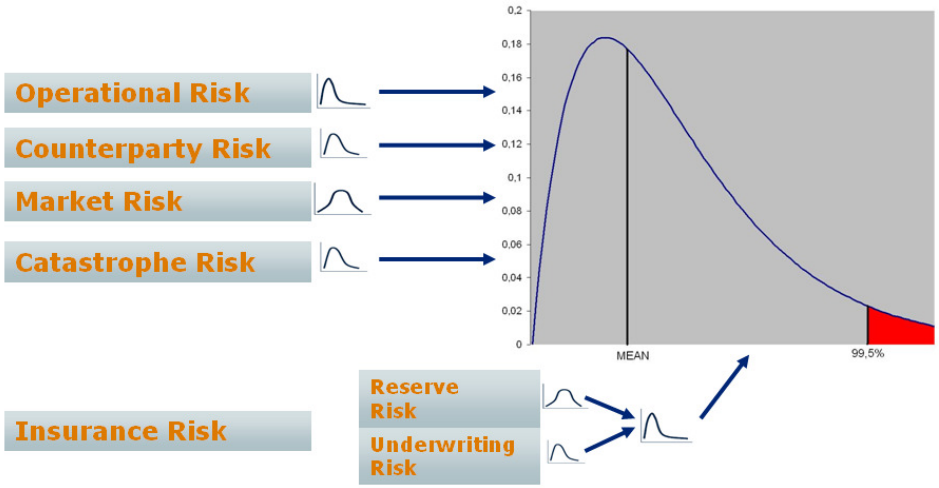

Figure 1.1: Insurance Risk

Source: https://www.pinterest.co.uk/pin/534521049520207946/

1.2 Crystal Report

Crystal Reports is a popular Windows-based report writer (report generation program) that allows a programmer to create reports from a variety of data sources with a minimum of written code. Developed by Seagate Software, Crystal Reports can access data from most widely-used databases and can integrate data from multiple databases within one report using Open Database Connectivity (ODBC). Crystal Reports uses an ActiveX control called Crystal Report to establish a connection with another program. A programmer can set properties of the Crystal Report control during design time or at run time.

The programmer can use automation tools called Experts to be guided through common tasks, such as linking and embedding reports. Crystal Reports treats all text, graphics, and database fields as object s that a programmer can place, arrange, and format on forms. The program also generates a record set object and code needed to perform programming tasks such as loop s or mathematical calculations. Crystal Reports can create a report on the fly from user-defined variables and can convert it to HTML and publish it to the Web automatically.

1.3 Problem Statement

The insurance industry worldwide is facing the challenges of deregulation, consolidation and convergence of financial services. There is today a pressing demand for cutting edge services of insurance business management and enriched customer experiences at a significantly lower cost. Our project aims to predict the risk behind bike insurance. It relates the information of the current customer and bike to the past data determining whether to make insurance or not.

1.4 Objectives

The objectives of this project are:

- To predict the possibility to claim of bike insurance.
- To visualize the insurance performance.
- To generate the claimed and unclaimed report.

1.5 Application

Project-X is a web application which analyze and visualize the user data. This project is specoally developed for insurance company. The application of this is help the management level for policy making in insurance company. It provides visualization to the company. The visualization generate the report for claimed and unclaimed report separately with respective attributes. The report provides the management level with overall glance of the insurance to claim policy. The generated report is printable. Due to this , it required less human resource. In simple meaning it is decision support system for bike insurance company.

1.6 Project Features

- Claimed and unclaimed report generation based on various attributes.
- Insurance performance can be visualized.

- Business Intelligence Report (Crystal Report).

1.7 Feasibility Analysis

1.7.1 Economic Feasibility

The economic feasibility of the software refers to the cost of its creation, cost required to launch the system and its efficiency in terms of cost after its launch to real platform. It determine whether the project is economically and logically possible or not. This project reduces the operations and maintenance cost of IT infrastructures

1.7.2 Technical Feasibility

This feasibility includes the software and hardware used for project.As this project will be developed using 'C sharp' as programming language and simply using the logical concept, it won't require sophisticated equipment. A PC with any Operating System with a browser and network will be able to run the software. Almost all current devices support the surfing the network. Thus watching overall ,this project is technically feasible.

1.7.3 Operational Feasibility

A feasibility study aims to objectively and rationally uncover the strengths and weaknesses of a project. As a webpage, it can be anywhere at any time after it deployment to the server. Our project provides user friendly platform, people with simple knowledge of computer can operate easily and the system can perform the rest operation.

1.8 System Requirement

System Requirements are categorized into two field:

1.8.1 Software Requirement

- **For System Development**
    1. Operating System : Windows OS
    2. IDE : Visual Studio 2017
    3. Framework : Microsoft .net framework
    4. Database : Microsoft SQL Server 2014
- **For End Users**
    1. Any Operating System with browser and internet.

1.8.2 Hardware Requirement

- **For System Development**
    1. PC with 1.8 GHz 4 GB RAM
- **For End Users**
    1. Any Device with supporting latest browser and internet

CHAPTER 2
LITERATURE REVIEW

Insurance is a practice by which a company provides a guarantee of compensation for specified loss, damage, illness, or death in return for payment.Bike Insurance is insurance for bike, motorcycle. Its primary use is to provide financial protection against physical damage or bodily injury resulting from traffic collisions and against liability that could also arise there from. Bike Insurance also provide financial support in case of bike stolen, damaged to bike from events other than traffic collisions , and keying and collision with stationary object. [2] Some Previously developed platform:

2.1 Sisense

Sisense is Business Intelligence software by Sisense Inc.,the industry in for complex data-easily prepare analyze and explore growing data from multiple sources. Sisense is awarded with **Best Business Intelligence Software for 2016**. It is the business analytics software that lets you easily prepare and analyze big, scattered datasets. it have 3 types of user

i. Admin : Can access all features of Application.
ii. Designer : Mainly to design the elastic cubes and all admin function except Serve and user management.
iii. Viewer : Can only view data result and visualization.

The unique about Sisense are :

1. Single-Stack architecture: A single tool that helps you collect, prepare, organize and analyze data.
2. In-chip Engine and Proprietary technology.
3. Optimal use of computational resources.
4. A 90-minutes real test-drive for prospective client.

2.1.1 How Sisense work

Sisense is webapp that work on both Computer or in smartphones. It works on the dashboard system with Elastic Cubes which is the proprietary analytical tools database that enables to connect multiple data sources and run complex queries in split of second. The data sources that can be used with sisense are file-based data sources (Excel ,CSV), Traditional database (SQL, MySQL, Oracle) and Online Web-Services(Google adwords). Data sources are added in elastic cube manager. Different types of data sources can be added in same cube. Any two field with same type of data on different data sources and be connected by drag and drop facility of sisense. At last, by building elastic cubes the data will be pull form data sources into elastic cube. When the result is ready it is ready for calculation and the result can be visualize in multiple form.

2.2 Qlik Sense

QlikSense is software launch by Qlik software company. QlikSense lets you discover insights that query-based BI tools simply miss. One can freely search and explore across all his data, instantly pivoting your analysis when new ideas surface.[3] The unique about QlikSense are :

1. Full spectrum of visual analytic.
2. Externally governed assets and data modules.
3. Data compression.
4. Policy-based security rules

2.2.1 How does Qlik sense work

Qlik Sense is platform of data analysis. It analyze data and make discovery on your own.One can easily share knowledge and data in organization using Qlik sense. It generates views on information for the user. It does not require predefined and static reports. One do not have to dependent on other resources just click and learn. Each click sends instant responds updating every visualizations and view in the app with new data and visualizations. App model help by asking question and user can answer the next question on their own without going back to expert for new report or visualizations for the data .

CHAPTER 3
THEORY

3.1 Bayes' Theorem

Bayes Theorem is a statement from probability theory that allows for the calculation of certain conditional probabilities. Conditional probabilities are those probabilities that reflect the influence of one event on the probability of another event. The term generally used in Bayes theorem are prior probability and posterior probability. The prior probability of a hypothesis or events the original probability obtained before any additional information is obtained. The posterior probability is the revised probability of the hypothesis using some additional information or evidence obtained.[4]

Bayes Theorem can be written as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

1. P(A) is the prior probability of A ,
2. P(B) is the prior probability of B ,
3. $P(A|B)$ is the posterior probability of A given B &
4. $P(B|A)$ is the posterior probability of B given A

Bayes' Theorem for n independent variable:

$$P(B|A) = P(B_1|A) * P(B_2|A) * P(B_3|A) * ... * P(B_n|A)$$

3.2 Classification

Classification in data analysis is the task of assigning a class to instances of data described by a set of attributes. Classification or supervised classification includes the construction of a classifier which is trained on a set of training data that already has the correct class assigned to each data point. This builds a concise model of the distribution of class labels. It is then used to classify new data where the values of features are known but the class is unknown.

Many algorithms have been developed for supervised classification based on

1. Artificial Intelligence
2. Perceptron-based Techniques (Single and Multi-Layered Perceptron)
3. Statistical Learning Techniques (Bayesian networks,Instance based techniques)

3.3 Bayesian Classification

Bayesian Network provides a powerful graphical method for encoding the probabilistic relationship among a set of variables and hence and naturally be used for classification. It learned by using likelihood to achieve classification accuracy. It learned by supervisied learning, where a training data set of instances with labels representing instance of classes is used to train as classifier.[4] Likelihood is calculated by:

$$CLL(G|D) = \sum_{i=1}^{N} \log P(C_i|V_i)$$

3.3.1 Learning Models

Types of Learning models:

1. **Generative Model**

   Generative models summarize data probabilistically and are more flexible,since the user can bring in conditional independence assumptions, priors, and hidden variables. Generative classifiers learn a model of the joint probability of the variables and the related class label, and use Bayes theorem to compute the posterior probability of the class variable and make predictions.

2. **Discrimative Model**

   Discriminative models (e.g.NN and SVM) only learn from data to make accurate predictions by directly estimating the class posterior probability or via discriminant functions, and thus offer the user less flexibility in data representation and inference.

.

3.4 Naïve Bayesian Classifier

A Naïve Bayes Classifier is a probabilistic classifier based on applying Bayes theorem with strong independence assumptions. When represented as a Bayesian network, a Naïve Bayes classifier has the structure depicted in Figure below. [5] It shows the independence assumption among all features in a data instance.

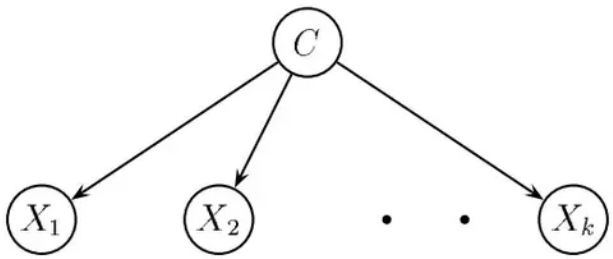

Figure 3.1: Structure of Naïve Bayesian Network

Source: http://www.mdpi.com/1099-4300/19/6/247

Naïve Bayes Classifier is simple probabilistic classifier that calculates a set of probability by counting the frequency and combinations of value in a given set of data. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. Naïve Bayesian classifier is based on Bayes theorem and the theorem of total probability.

3.5 Data Pre-Processing

- Pre-process Steps
  - Data Cleaning
  - Data Integration and Transformation
  - Data Reduction
- Data in real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining result
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data
  - Reduced Accuracy

3.5.1 Measure of data Quality

Well-accepted multidimensional view:

- Accuracy
- Completeness
- Consistency
- Reliable
- Interpretability
- Accessibility

3.5.2 Major Task in Data Pre-processing

1. Data Cleaning
    - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
    - Data cleaning tasks
        - Fill in missing values
        - Identify outliers and smooth out noisy data
        - Correct inconsistent data
2. Data integration
    - Combine data from multiple data source.
    - Detecting and resolving data value conflicts
    - possible reasons: different representations, different name
3. Data Transformation
    - Normalization and aggregation
    - Data smoothing
    - Data summarization
    - Data Reduction

3.5.3 Data Smoothing

When data collected over time displays random variation, smoothing techniques can be used to reduce or cancel the effect of these variations. When properly applied, these techniques smooth out the random variation in the time series data to reveal underlying trends. The variation in data cause the program to produce unreliable result and inaccurate. Smoothing is technique to remove the variation in data.

Different ways of data smoothing technique are:

1. Data Level approach: Resampling Techniques
    (a) Random Under-Sampling
    (b) Random Over-Sampling

(c) Cluster-Based Over Sampling

(d) Informed Over Sampling:Synthetic Minority Over Sampling Technique (SMOTE)

(e) Modified synthetic oversampling technique(MSMOTE)

2. Algorithmic Ensemble Techniques

(a) Bagging based

(b) Boosting Based

- Adaptive Boosting-Ada Boost
- Gradient Tree Boosting
- XG Boost

3.5.4 SOMTE Technique

SMOTE is short hand for Synthetic Minority Over Sampling Technique. This techniques is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data taken from a minority class as an example and then synthetic similar instances are created. The synthetic instances are then added to the original dataset. Then the new data set is used to train the classifier models.

We have seperately perform data smoothing using R languages. We used SMOTE to balance the data of claimed and unclaimed data to maintain low variation and train the balanced data using Gradient boosting algorithm in R.

This process significantly imapcats the accuracy of the predictive model. By increasing its around accuracy by 20 percent .

3.6 Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or ”classifier”) on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

Confusion matrix is 2x2 matrix with 4 variables with Actual true and false and Predicated true and false class .

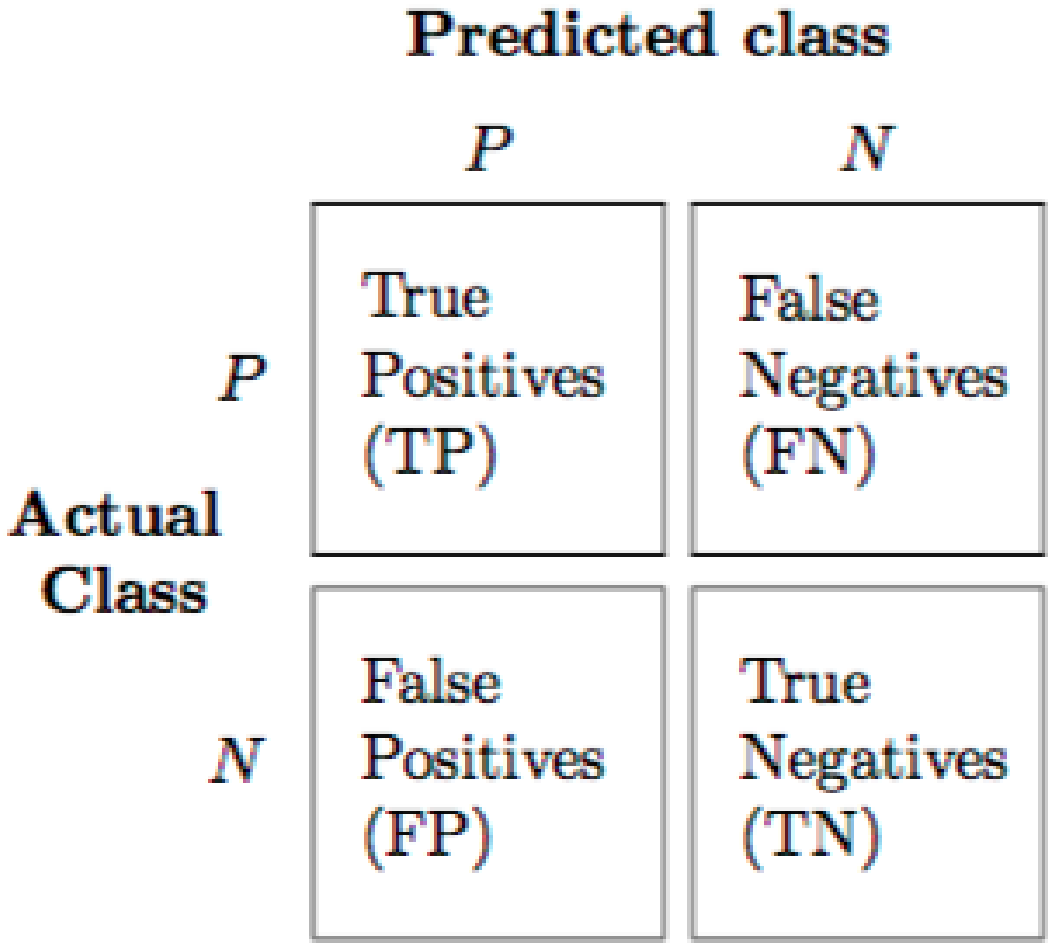

Figure 3.2: Confusion Matrix

From figure:

1. $Accuracy = \frac{TP+TN}{P+N}$

3.7 Implementing Algorithm

- Problem Statement
  - Given Features : $X_1, X_2, X_3.....X_n$.
  - Predict a label Y.
- Consider each attribute and class label as random variables.
- Given a record with attributes $(A_1, A_2, A_3.....A_n)$.

  - Goal is to predict class C.
  - Specifically find value that maximizes P(C|$A_1, A_2, A_3.....A_n$)
- Compute posterior probability P(C|$A_1, A_3, A_3.....A_n$) for all value of C using Bayes theorem.

$$P(C|B) = \frac{P(A_1 * A_2 * A_3 * ..... * A_n|C)P(C)}{P(A_1 * A_2 * A_3 * ..... * A_n)}$$

- Choose value that maximizes P(C|$A_1, A_2, A_3.....A_n$)
- Equivalent to choosing value of C that maximizes. $P(A_1, A_2, A_3.....A_n|C)P(C)$
- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, A_3.....A_n|C)P(C) = P(A_1|C_j)P(A_2|C_j)P(A_3|C_j)...P(A_n|C_j)$.
  - Estimate $P(A_i|C_j)$ for all A_i and C_j.
  - New point is classified to C_j if

    $P(C_j) * P(A_i|C_j)$ is maximum.

CHAPTER 4
METHODOLOGY

4.1 E-R Diagram

The Entity Relationship Diagram of the project:

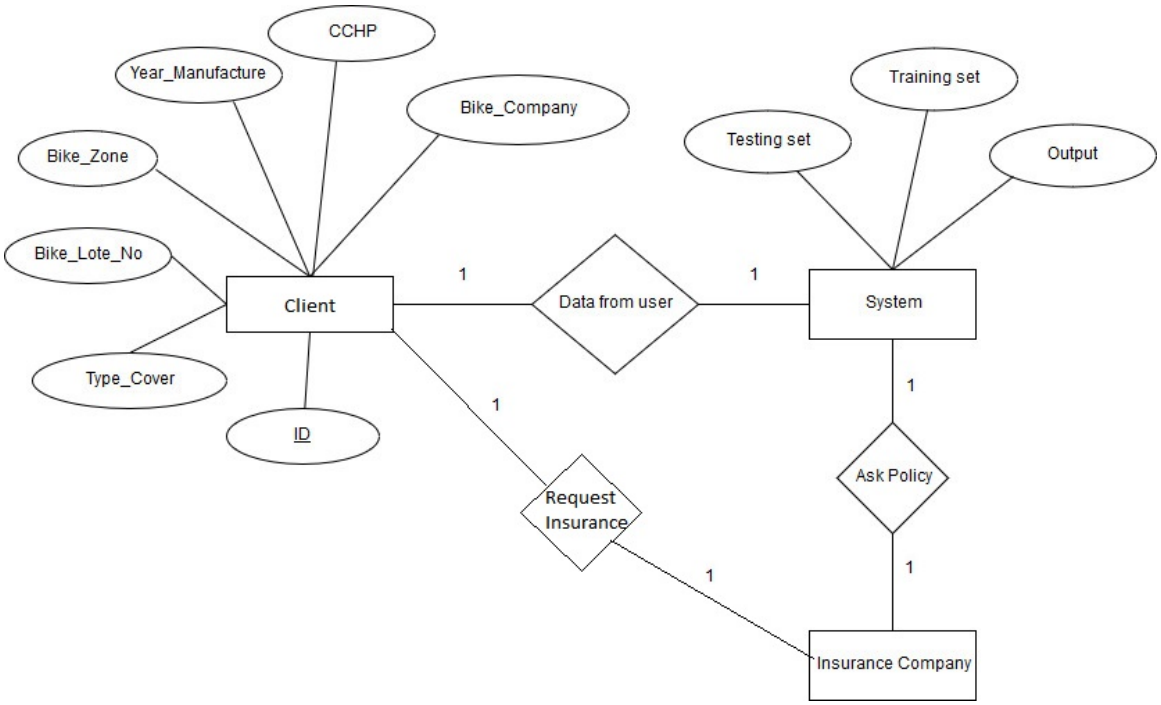

Figure 4.1: Entity Relationship Diagram

4.2 Flow Chart

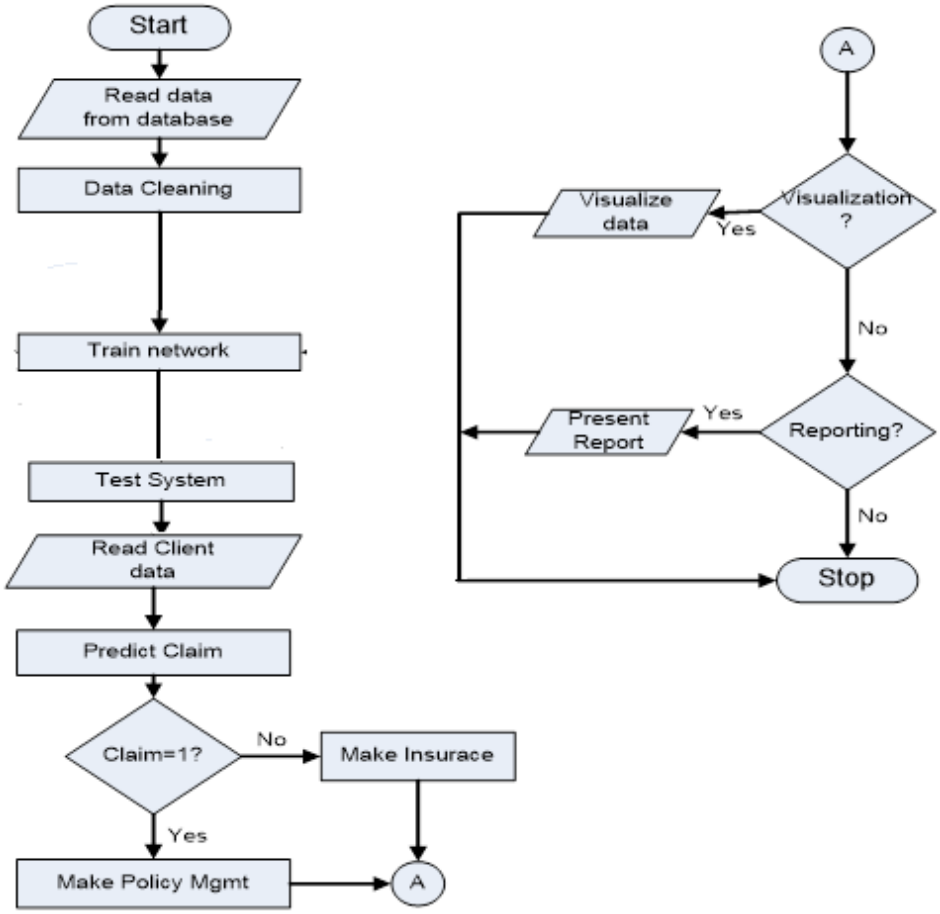

Figure 4.2: Flow Chart Of Project

4.3 Use-Case

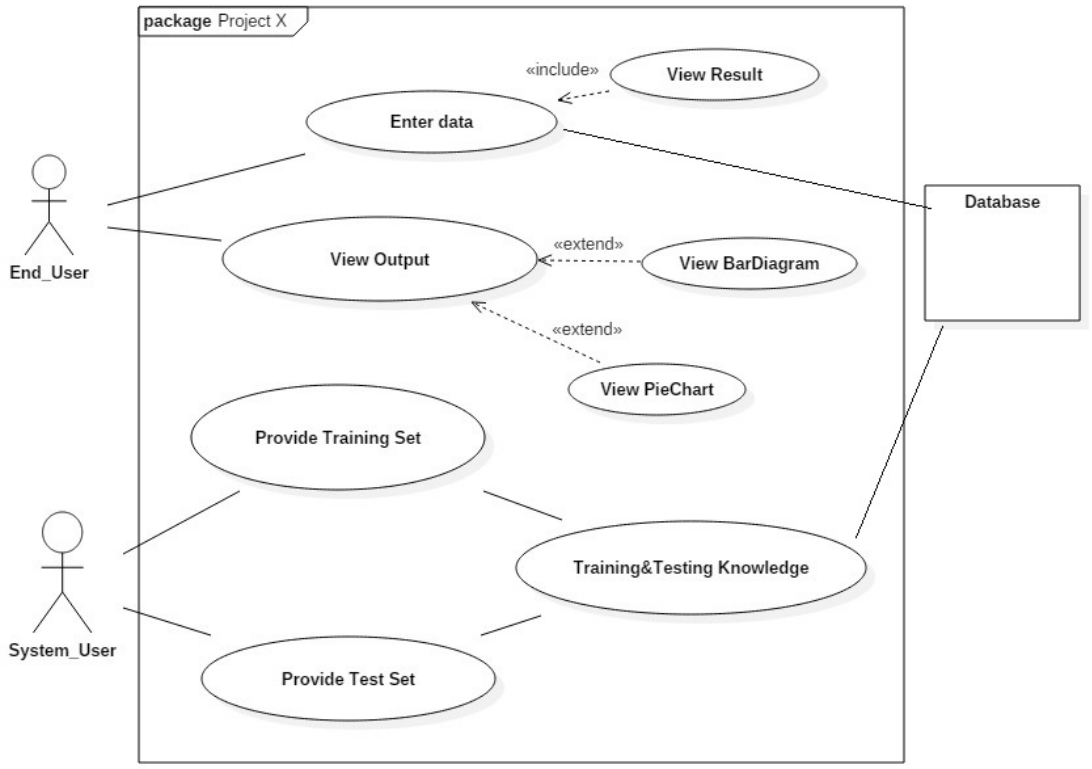

Figure 4.3: Use-case Diagram

In this project there will be three options. The customers information will be taken and based on which insurance to be done is decided. Report and Visualization of data based on various attributes can be done which helps to know the progress of an insurance company.

4.4 Class Diagram

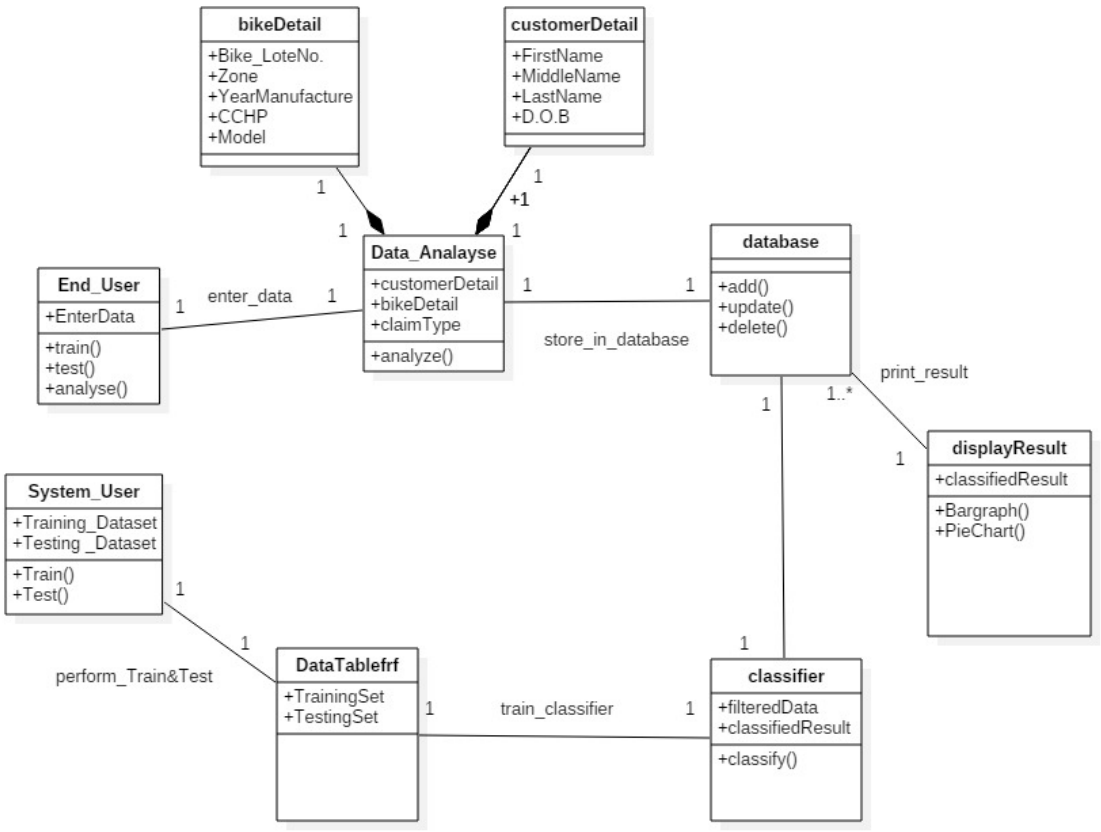

Figure 4.4: Class Diagram

4.5 Sequence Diagram

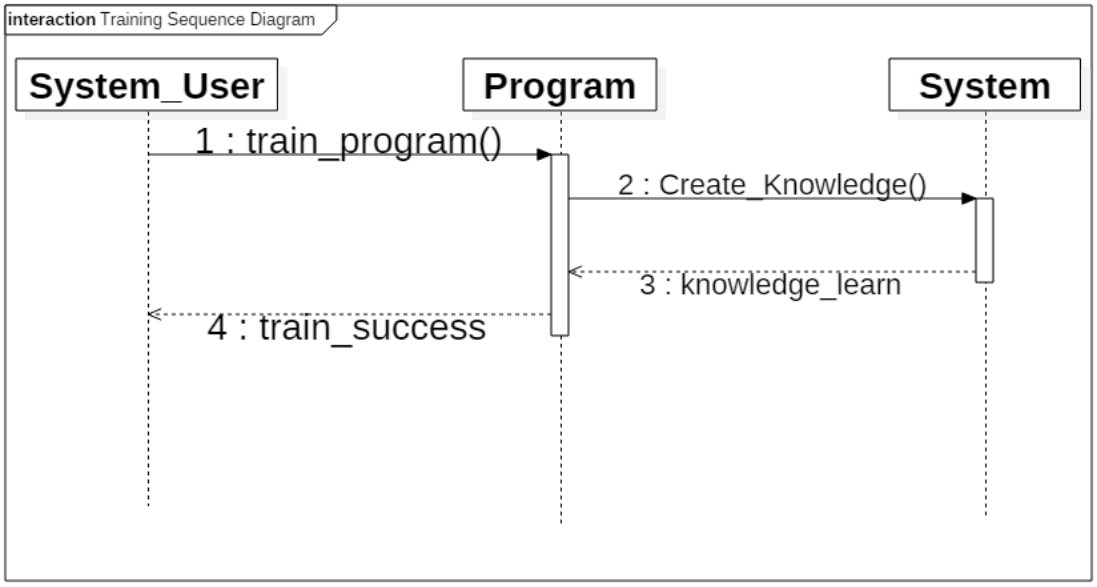

Figure 4.5: Sequence Diagram: Training

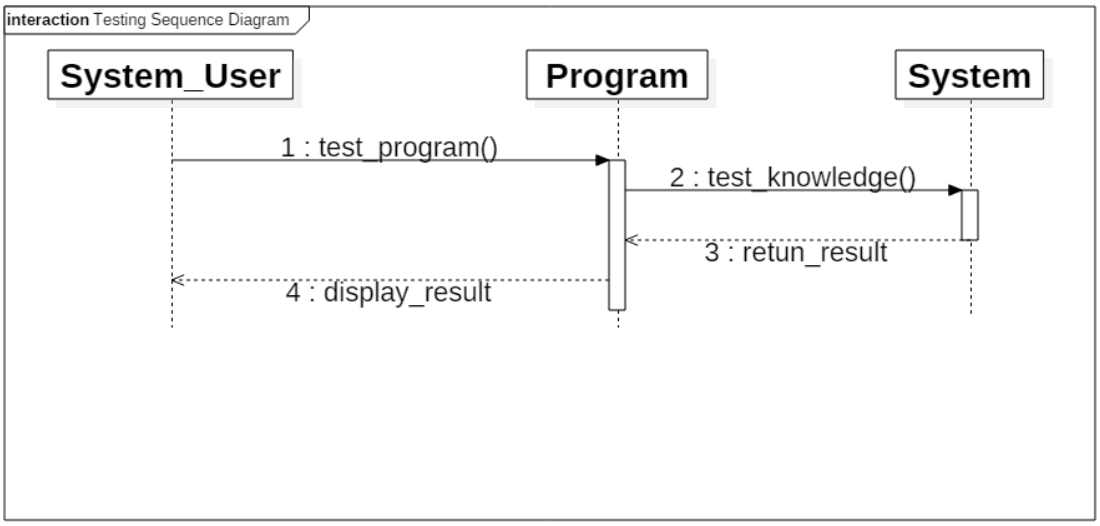

Figure 4.6: Sequence Diagram: Testing

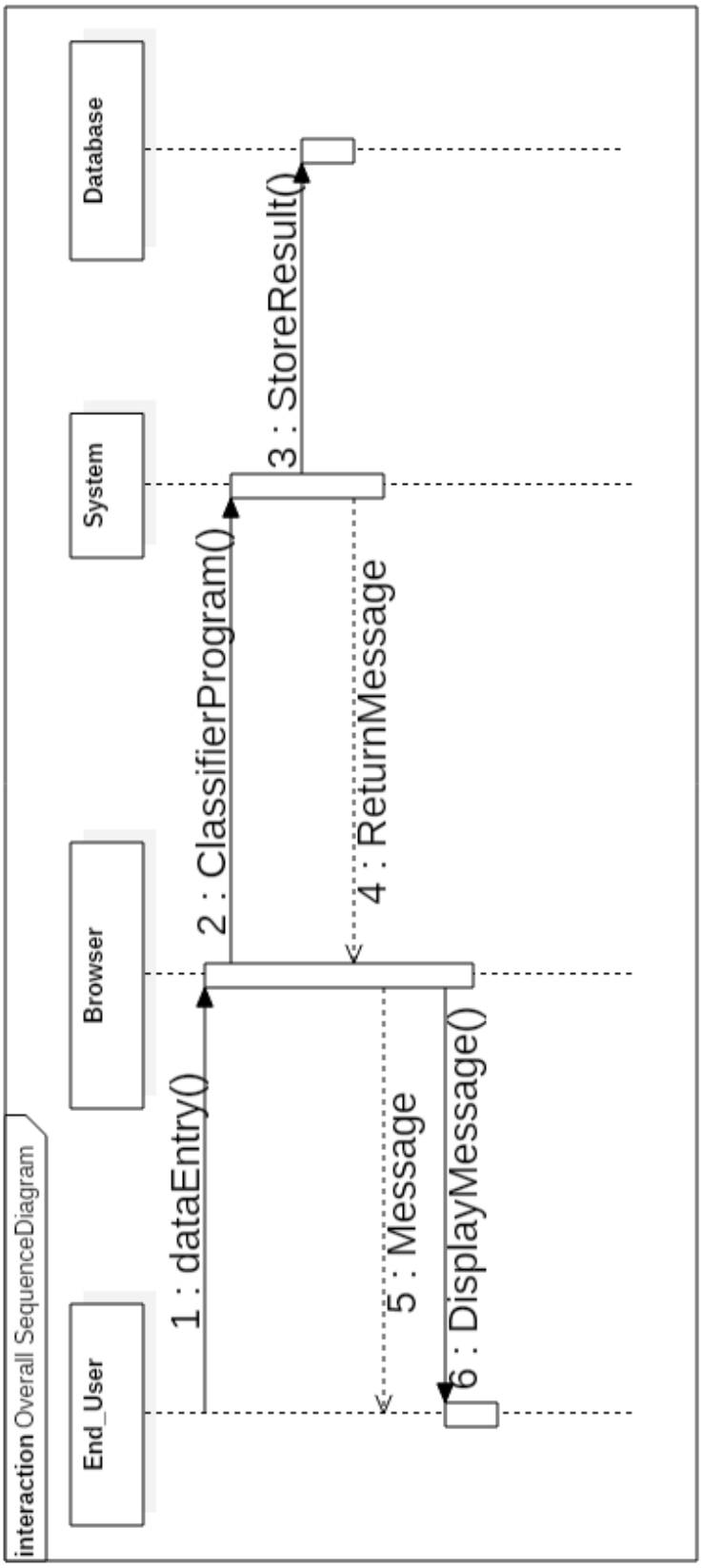

Figure 4.7: Sequence Diagram

4.6 Collaboration Diagram

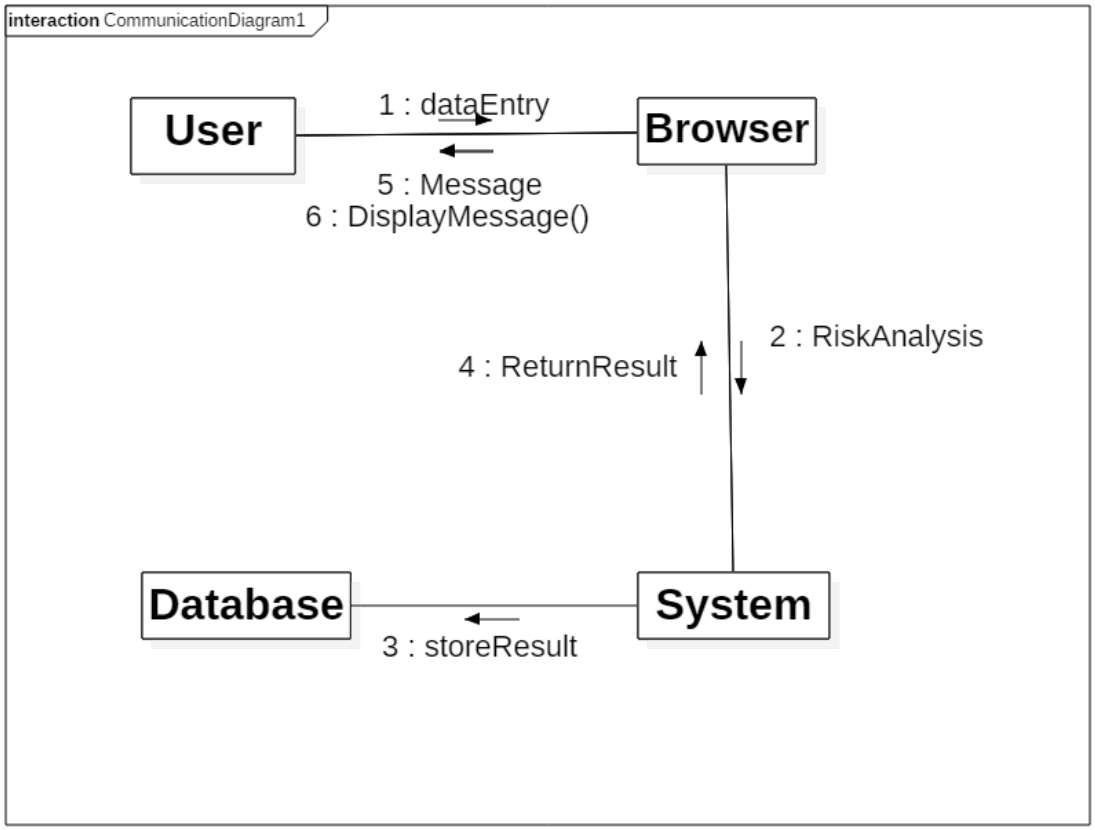

Figure 4.8: Collaboration Diagram

4.7 Software Development Process

Project-X is prepared by following approaches based on incremental software development model. The objective of following incremental model, it provides flexibility of starting project by available requirements, such that requirements can be added on the basis of later added projects objectives. One of the main feature of the model, it provides designing, testings phase in each increment.

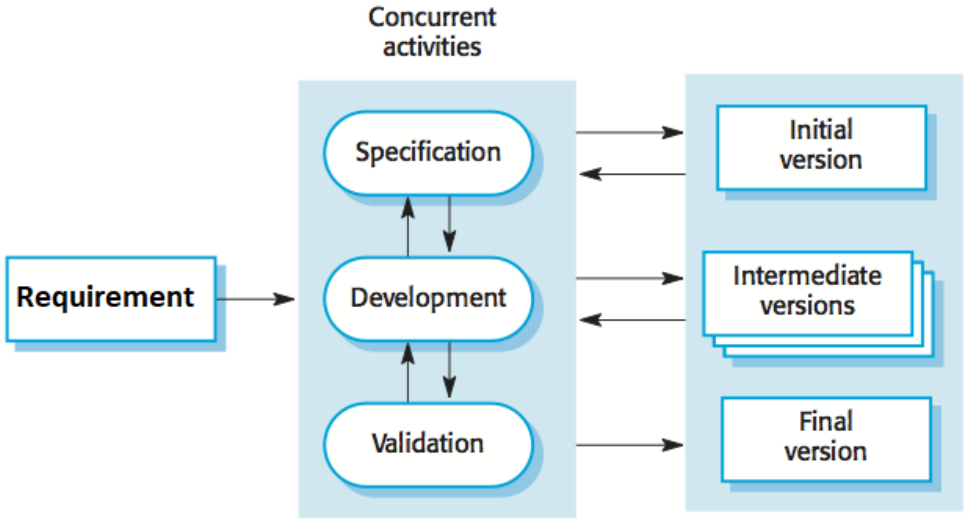

Figure 4.9: Incremental Development Model Chart

Source: https://istqbexamcertification.com

Project-X is application based on Naïve Bayesian Classifier for classifier data and predict the result for user. We mainly divide our project into 3 phase.

1. **1st Incremental Phase**

   As Initial phase, we mainly involved in researching about our project basic idea. We also visit company to get data insurance data that is vital for our project. We have done most of front end.

2. **2nd Incremental Phase**

   We started with implementing of algorithm and training and testing of data was done to find accuracy of data from collected data. We also get more data from company for further analysis of program.

3. **3rd Incremental Phase**

   We completed almost of front end and back end. We tested our program with 70% accuracy and crystal report

CHAPTER 5
RESULT AND DISCUSSION

5.1 Outputs

5.2 Limitation

5.3 Problems Faced

- Data filtering through various process.
- Limited data.
- Implementing the Algorithm.

5.4 Work Schedule

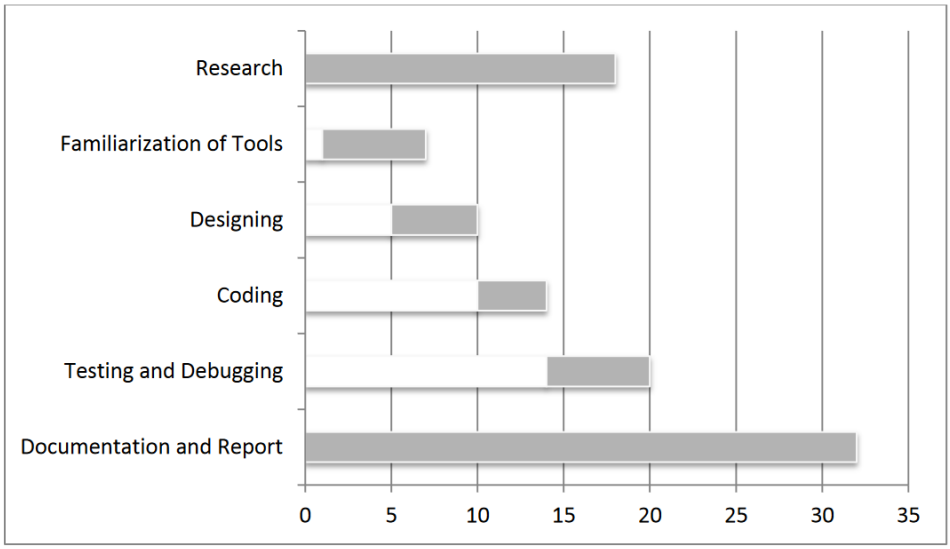

Figure 5.1: Gantt Chart

CHAPTER 6
CONCLUSION

6.1 Conclusion

6.2 Future Enhancement

REFERENCES

[1] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact." *MIS quarterly*, vol. 36, no. 4, 2012.

[2] S. R. Insurance, "Bike insurance," 2015.

[3] O. Troyansky, T. Gibson, and C. Leichtweis, *QlikView Your Business: An Expert Guide to Business Discovery with QlikView and Qlik Sense*. John Wiley & Sons, 2015.

[4] B. Garg, "Design and development of naive bayes classifier," 2013.

[5] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[6] N. I. Company, "Insurance data," Kamaladi,Kathmandu, Tech. Rep., 2017.