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	**[44084]**
Kaushtup Bista	**[44096]**
Roshi Maharajan	**[44108]**

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

August, 2018

”PROJECT X : A BUSINESS INTELLIGENCE FOR BIKE INSURANCE”

Submitted by:

Amit Dhoju **[44084]**

Kaushtup Bista **[44096]**

Roshi Maharajan **[44108]**

Supervised by:

Ajay Mani Paudel

Executive Director

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

August, 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

We would like to thank to the head of the department of **Computer and Electronics** for providing this opportunity to undertake this project and for arranging hardware and software materials for the research.

We cannot remain thanking, **Er. Ajay Mani Poudel** for sharing knowledge regarding Artificial Intelligence, Data pre-processing and provided valuable guidance for Documentation and guided each and every step towards the successful end of the project.

We would also like to thank our teachers **Er. Rabindra Khati, Er. Ravi Chandra Koirala, Er. Dipesh Shrestha, Er. Tirtha Acharya, Er. Tek Nath Adhikari, Er. Ankit Shekhar Acharya, Er. Bishal Thapa, Er. Srijal Joshi** for their valuable suggestions and wishes towards the further development of the project.

Our thanks and appreciation to our fellow classmate **Mr. Amit Shrestha** for helping us during the difficult phase of our project development.

We would like to express our gratitude to the authors of various papers that helped us gain the in-depth knowledge of various aspects. Finally, we would like to thank all the people who are directly or indirectly related for the successful completion of this project.

Amit Dhoju	[44084]
Kaushtup Bista	[44096]
Roshi Maharajan	[44108]

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

BI Business Intelligence

CC Cubic Centimeter

CSV Comma Seprated Value

NN Neural Network

ODBC Open Database Connectivity

OS Operating System

SMOTE Synthetic Minority Over Sampling Technique

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.

1.2 Insurance

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

1.2.1 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.

1.2.2 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

1.2.3 Insurance Risk

- Some examples of insurance risk:
  1. Legal - litigation
  2. Operational - mistakes, errors, etc.
  3. Pricing - inadequate premiums
  4. Regulatory - new requirements
  5. Reputation - negative press

A schematic overview over the different quantitative risk factors for non-life insurance can be given as:

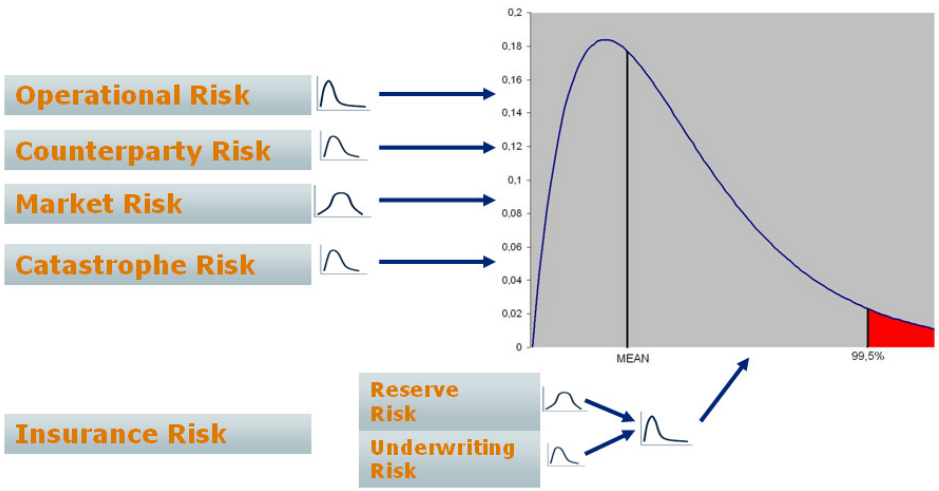

Figure 1.1: Insurance Risk
[2]

Some of the type non-life insurance arelisted below:[2]

1. **Operational Risk**

   Operational risk is usually considered as risks connected to the people, systems and processes in a business. This is a very broad group of risk and includes fraud, system failures, terrorism and employee compensation claims. The model for operational risk may be very complex with detailed models for each element influencing the risk, but simplified models based on some appropriate exposure

figure is also common. For non-life insurance a factor of earned premium is commonly used.

2. **Counter-party Risk**

   The risk of not receiving payment as agreed is often called counterparty or credit risk and the event is often called default. For non-life insurance the counterparty is often a reinsurer. The probability of default (PD) is often given by credit rating agencies like S&S and Moodys. The value at risk is simply the amount at stake multiplied by the probability of default given by the rating agencies. An example of rating and default probabilities are: A further sophistication of this model is to give a distribution of the loss given default.

3. **Market Risk**

   When modelling market risk we want to find how much the company assets may decrease due to change in the market factors. The four standard market risk factors are stock prices, interest rates, foreign exchange rates, and commodity prices. The actual portfolio of the company decides which factors are to be taken into consideration. Usually one wants to model these factors as time series and a lot of work is done in this field. However, most insurance companies will rely on Economic Scenario Generators (ESG) as a part of a risk management tool rather than develop own methods for market risk.

4. **Insurance Risk**

   Insurance risk is the risk arising from the process of transferring risk from persons or companies to the insurance company and it is the fundamental business idea of an insurance company. It is usually divided in two main categories: underwriting risk and reserve risk. Underwriting risk is the risk arising from claims incurring in future accounting periods, while reserve risk is the risk arising from previous accounting periods.

5. **Underwriting Risk**

   To be able to predict the claims for a future accounting period we need to know about the claims in previous accounting periods. We will show an example on how to model next periods claims based on historical claims. To give an example we have accumulated periodical claims data from two lines of business, health insurance and workmans compensation insurance.

1.3 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 objects 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 loops 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.

1.4 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.

1.5 Objectives

The objectives of this project are:

- **To predict the possibility to claim of bike insurance**
  We have created a classifier model with inputs CC of Bike, Bike Manufacturer,Year Manufacture Lote, Zone, and Type Cover. And the claim status is given to determine the claim of the insurance. Hence, our system with the above mentioned parameter will be able to predict the possible claim.
- **To visualize the insurance performance**
  We have used bar chart to visualize the claim and un-claim data for each individual attributes. This visualization will show total insurance of particular field of particular attributes, claim count of that attribute un-claim count of that attribute. Hence our system visualize the overall performance of the insurance
- **To generate the claimed and unclaimed report** Report is another objective of our system. Our system allows user to select the particular attribute, then total claimed and unclaimed data count of particular field of an attribute is displayed simultaneously. This will report to the management level for the decision for overall performance of the insurance. System is able to print the report.

1.6 Application

Project-X is a web application which analyze and visualize the user data. This project is specially 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.

1.7 Project Features

- Insurance performance can be visualized.
- Business Intelligence Report (Crystal Report).

1.8 Feasibility Analysis

1.8.1 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.

1.8.2 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.

1.8.3 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 its 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.

1.9 System Requirement

System Requirements are categorized into two field:

1.9.1 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.

1.9.2 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. [3] 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 dataset. 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 web-application 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.[4] 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.[5]

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.[5] 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. **Discriminative 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. [6] It shows the independence assumption among all features in a data instance.

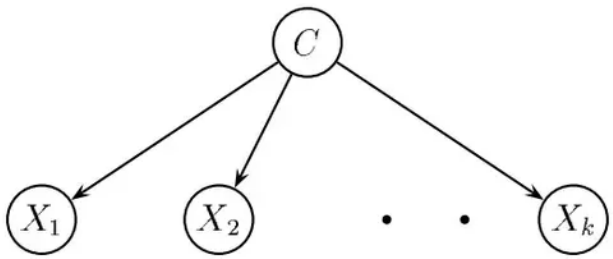

Figure 3.1: Structure of Naïve Bayesian Network

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 Laplace Smoothing

Laplace smoothing is an algorithm to smooth a polygonal mesh. For each vertex in a mesh, a new position is chosen based on local information (such as the position of neighbors) and the vertex is moved there. While checking the probabilities of event if it occur none then the probability of this event is low, but it is not zero. Further, we multiply all the probabilities during inference, even one such zero probability term will lead to the entire process failing. So, Laplace smoothing came to increase the zero probability values to a small positive number

Laplace Smoothing can be achieved by:

$$P(W_s) = \frac{C(W_s) + 1}{N + V},$$

Where,

1. P(w) is the probability that attributes is zero,
2. C(w) is the count of attributes w,
3. N is the total number of attributes,
4. V is the Constant

3.5.5 SMOTE 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.[7]

Data smoothing is usually performed using R languages. SMOTE is used 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 impacts the accuracy of the predictive model. By increasing its around accuracy by 20 %.

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 .

Terminology and derivatives from confusion matrix.

1. **Accuracy (ACC):**

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{Total}$$

2. **Error (ERR):**

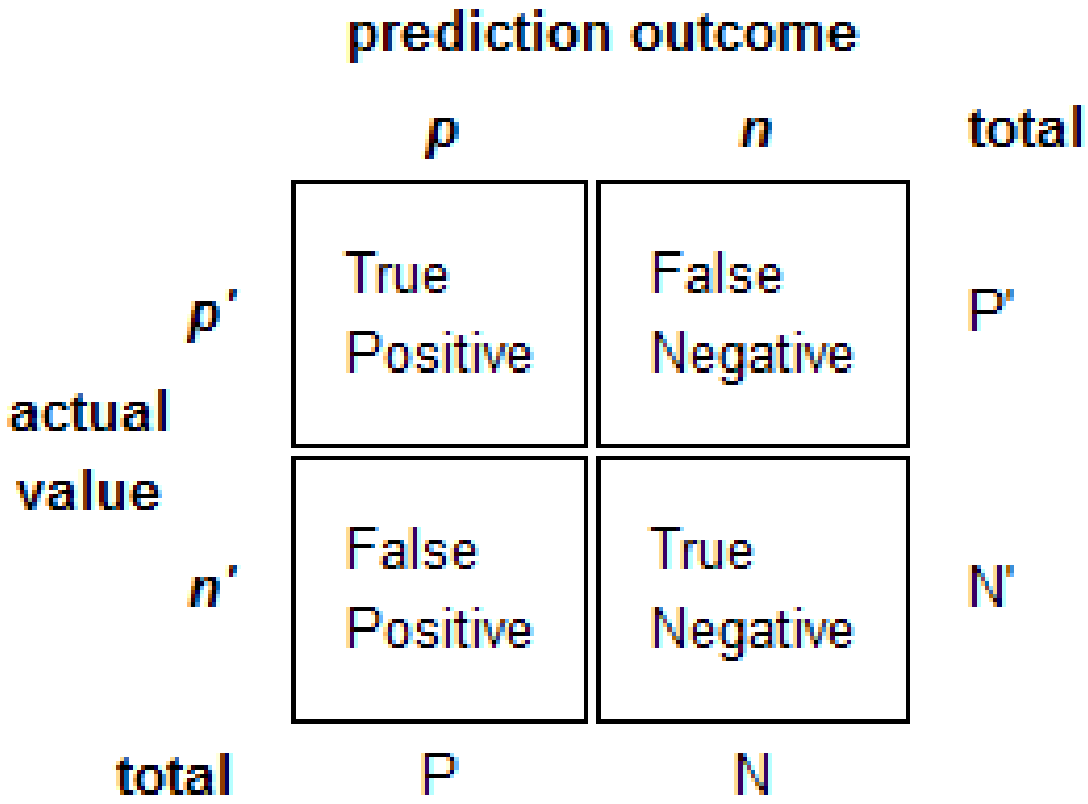

Figure 3.2: Confusion Matrix

$https://rasbt.github.io/mlxtend/user\_guide/evaluate/confusion\_matrix/$

$$ERR = \frac{FP + FN}{P + N} = \frac{FP + FN}{Total}$$

3. **Sensitivity or True Positive Rate (TPR):**

$$TPR = \frac{TP}{P} = \frac{TP}{TP + TN}$$

4. **False Positive Rate (FPR):**

$$fPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

5. **Specificity or True Negative Rate(TNR):**

$$TNR = \frac{TN}{N} = \frac{TN}{FP + TN}$$

6. **Positive Predictive Value(PPV):**

$$PPV = \frac{TP}{TP + FP}$$

7. **Negative Positive Value(NPV):**

$$NPV = \frac{TN}{TN + FN}$$

8. **Negative Predictive Value(NPV):**

$$ACC = \frac{TP + TN}{P + N}$$

9. **F1 Score:**

$$F1 = \frac{2TP}{P + P'}$$

CHAPTER 4
METHODOLOGY

4.0.1 Data Collection

As our project is business intelligence on bike insurance, we needed the real data to explore the reality on the bike insurance. We have collected data from Nepal Insurance Company Limited,Kamaladi , Kathmandu.[**?**] We have information about 1.6 Lakhs record from company.

4.0.2 Data Collection in Company

In this (Nepal Insurance company limited) company when user arrives at the company for the insurance in any branch, he/she submits the Xerox of blue book and motorcycle purchase bill and one photo of motorcycle owner. One then fills the proposal form. Proposal form contains all the related information. Then staff of insurance company entry the data in the software. When client makes claim, the research team makes research at the central corporate office located at Kamaladi, Kathmandu. And if the claim is responded positive the claimed data is again entered by the staff. This process goes on every branch of the company. Now the database data is written in DVD and send to the central office to combine data of the entire branch which is done by the IT staff of the central corporate company.

4.0.3 Overview of data

From company we are provided with data in excel format with claimed and unclaimed data. The column in excel have lots of fields and were in improper format. For example some data were saved in nepali language. Some were saved in combination format. Some are with different being same data-kind. Hence for better performance, we have to limit the size and data. Hence, we have extracted only six fields of the data. For modelling the data we have used six attributes and result is stored as new data. Namely the attributes are Zone, CC of bike, Bike Manufacturer, Type cover , Lote, Manufacture

Year and the Claim status. The data were too random and dirty. Hence we have to make them understandable and reliable which we have done in the data pre-processing phase.

4.0.4 Data Pre-processing

The entire data was in excel format before using it for system, We selected the required field in excel and extracted it.The nepali format data were manually converted into the data in english form and in desire format like two letter for Zone(BA for Bagmati,SE for Seti and so on),CC of bike in form of interval. The missing and empty data were rejected. We converted excel file to CSV for easy access of required attributes. After selecting certain number of records we found that there is huge variation of between claimed and unclaimed data. Finally, the data was reduced to 3000.

4.0.5 Data Smoothing

To keep the variation between two type of data low , we use data smoothing techniques called SMOTE.It avoids overfitting which occurs when exact replicas of minority instances are added to the main data-set. After this process the data was balanced between claimed and unclaimed and was increased to 6100.

4.0.6 Model by Naïve Bayesian Classifier

During implementation of Naïve Bayesian Classifier , firstly training set was classified into two parts i.e claimed data-set and unclaimed data-set.Since we had six attributes , we calculated the probability for both the claimed and unclaimed data for all respective values of six attributes.For e.g:For company name , probability of Claiming and Unclaiming of Honda , Yamaha, Suzuki,etc were calculated.The result was then stored in txt format.In this way we calculated the probability for the respective data and latter used it to predict whether the customer will claim or unclaim.

4.0.7 Training model

All together there were 6100 data after filtering.We had saved the data in the excel file in the *.csv format.These data were then stored to the MSSQL server database.Then the data were divided as claimed and unclaimed data and were stored in different tables namely tblClaimData and tblUnClaimedData. Then the probability of distinct data of all six attributes were calculated by Naïve Bayesian Classifier by using Bayesian Theorem:

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

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

4.0.8 Testing model

We selected 10% of data from the training set which included almost equal ratio of claimed and unclaimed data. Then for individual data their result was calculated i.e either 'yes' or 'no'.This result was compared with their expected output. Then, we checked the value for true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

We have constructed the confusion matrix and calculated the accuracy as follows:

$$ConfusionMatrix = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$$

4.0.9 Final System Result

After training and testing the data and finally selecting the required attributes,our system will be able to predict the risk policy. If there is risk, message will be appear as ”There is no Risk on Claiming” or ”Yes There is Risk”.

4.0.10 Reporting

Our system is able to produce report based on selected attributes.It will depict the claim count, unclaimed count for six different attributes of mentioned type.

4.0.11 Visualization

Our system is able to produce visualization bar chart based on the selected type of the attribute.It will depict the claim count,unclaimed count and total count of attribute of mentioned type.We have used javascript to make the bar chart more interactive and understandable.

4.1 Algorithms and Flowchart

4.1.1 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.

4.1.2 Flow Chart

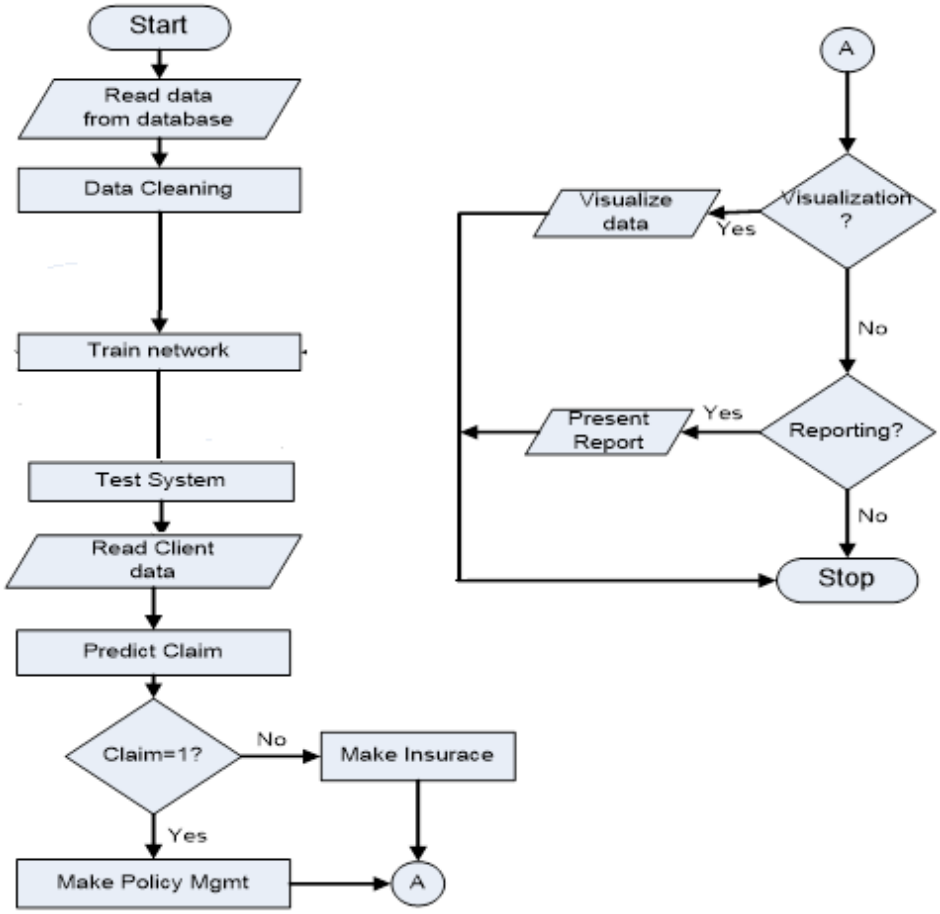

Figure 4.1: Flow Chart Of Project

4.2 Diagrams

4.2.1 E-R Diagram

The Entity Relationship Diagram of the project:

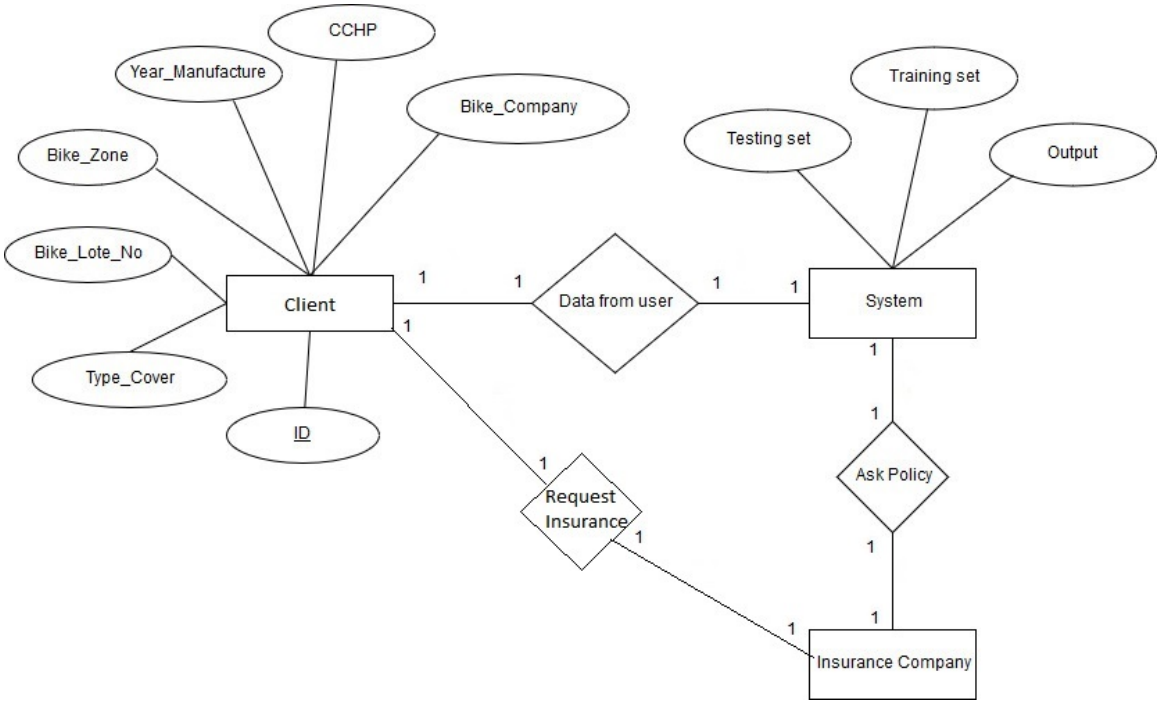

Figure 4.2: Entity Relationship Diagram

4.2.2 Use-Case

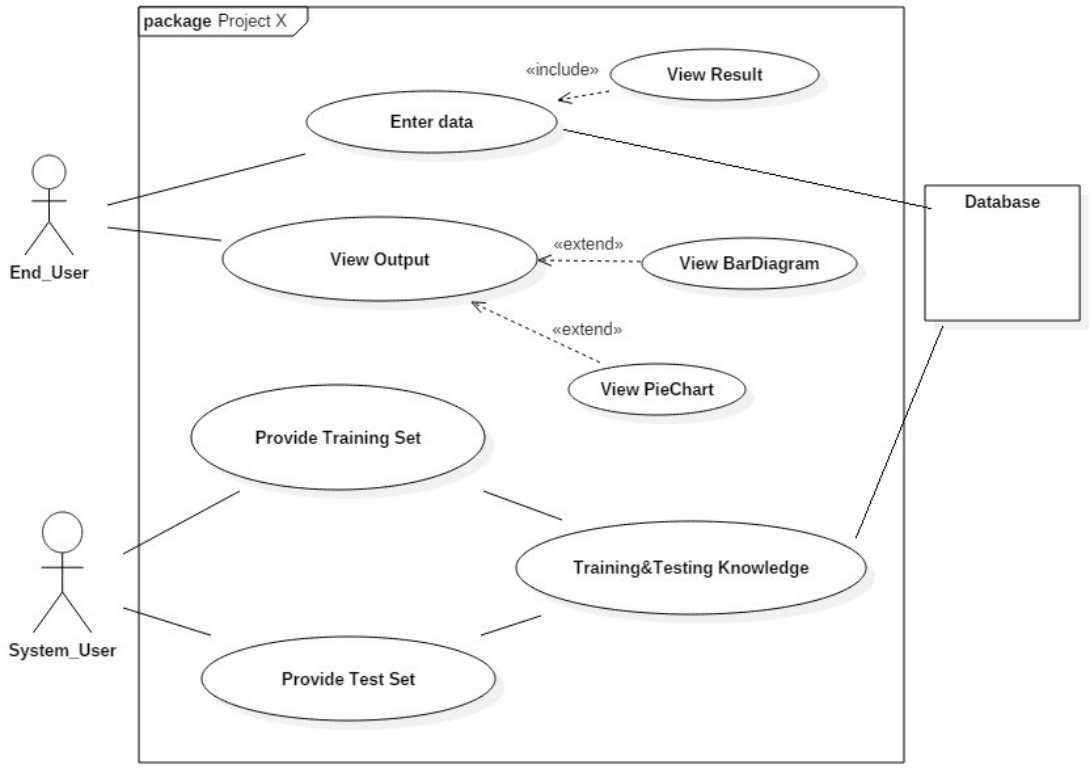

Figure 4.3: Use-case Diagram

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.

4.2.3 Class Diagram

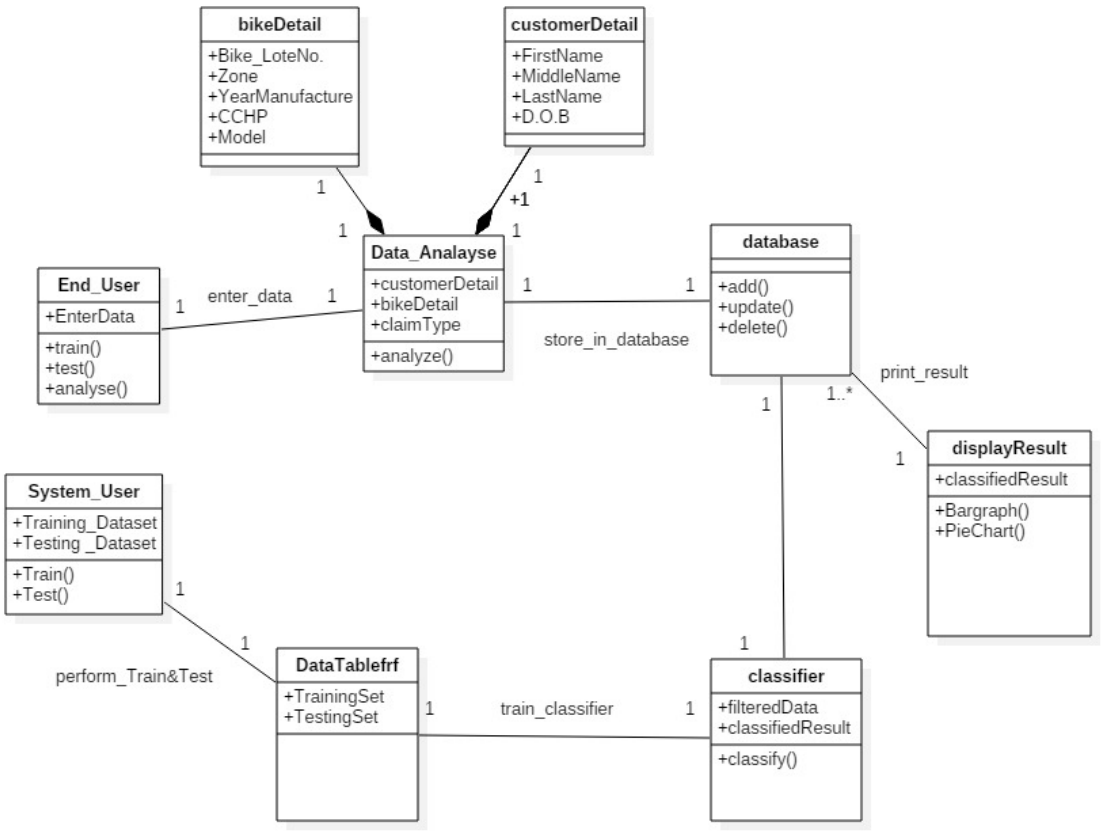

Figure 4.4: Class Diagram

4.2.4 Sequence Diagram

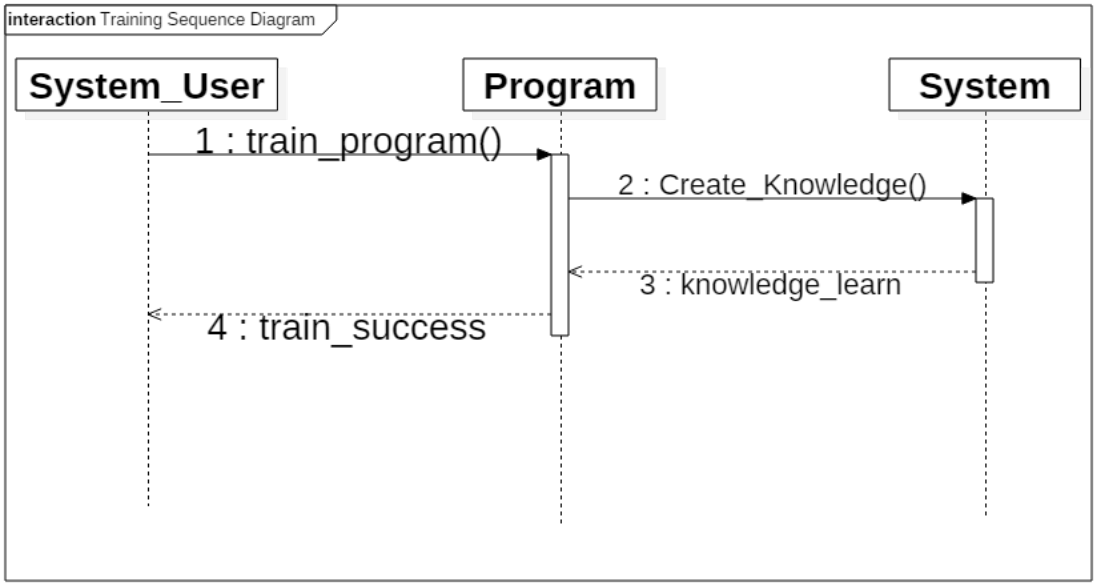

Figure 4.5: Sequence Diagram: Training

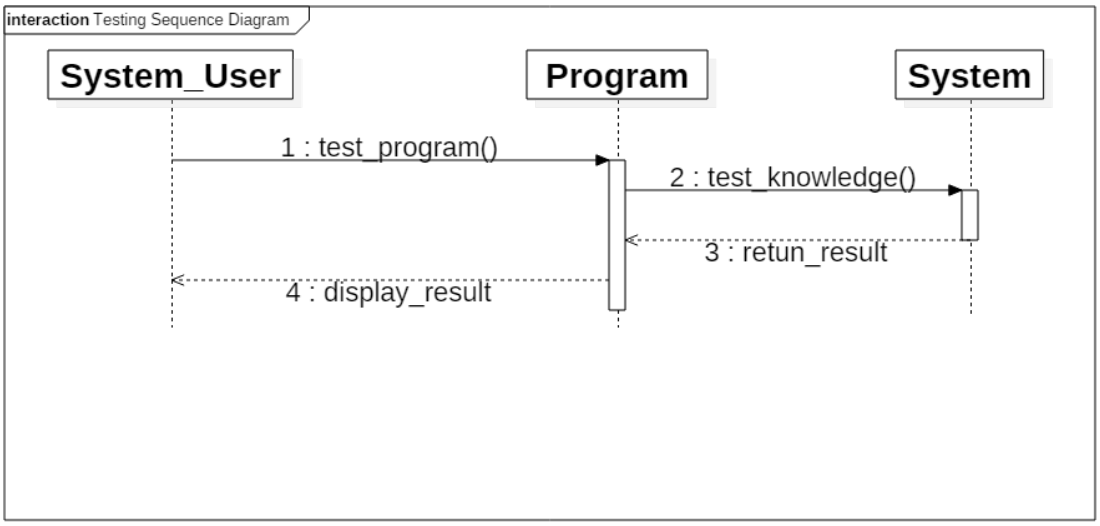

Figure 4.6: Sequence Diagram: Testing

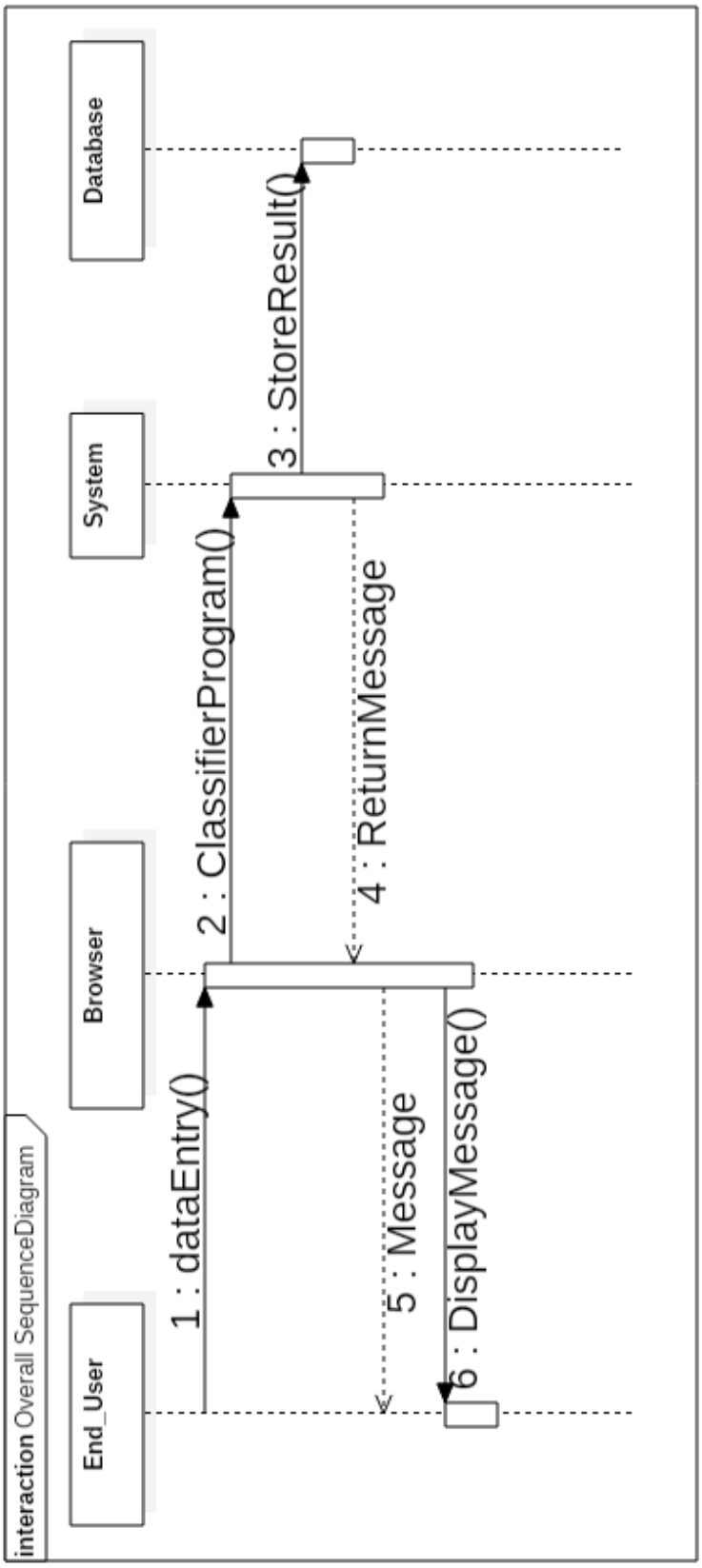

Figure 4.7: Sequence Diagram

4.2.5 Collaboration Diagram

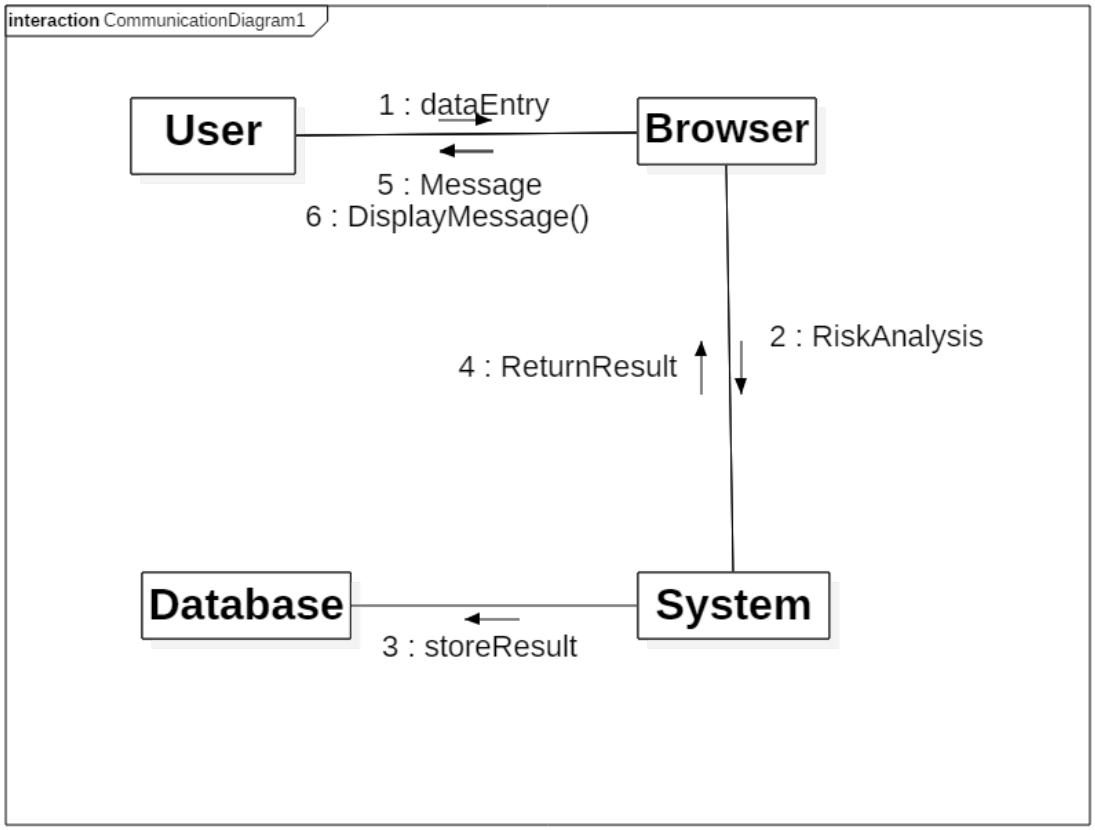

Figure 4.8: Collaboration Diagram

4.3 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.

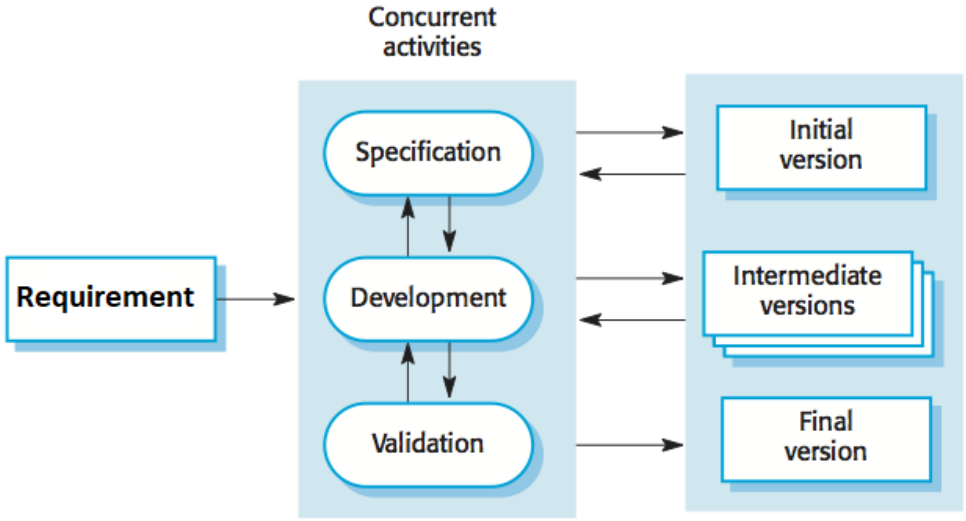

Figure 4.9: Incremental Development Model Chart

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. At this, point we were able make the system 60% accurate. The accuracy remain low due to large variation in claimed and unclaimed data.

3. **3rd Incremental Phase**

   We completed all the system designing part. We tested our program with 82% accuracy and crystal report. We were able to increase the accuracy by almost 20 % by using the Data smoothing techniques. We used SMOTE process to decrease the variation of data of claimed and unclaimed data. We generated the visual result for easy analysis.

CHAPTER 5
RESULT AND DISCUSSION

5.1 Outputs

We successfully completed the Project-X an online platform to check whether to selected or rejected as per data entered. We have used a Naïve Bayesian Classifier for creation of model on the basis of data we feed to the classifier that will predict the either to claimed or unclaimed the data.

Data have been collected from Nepal Insurance Company, Kamaladi, Kathmandu and we have pre-processed and normalized data such that 1.6 lakhs data have been reduced to 3000. We have extract only six set of data from the bunch, We extracted the data of 6 attributes namely Bike Manufacturer,Manufacture year,Bike CC, Bike zone, Bike Lote no and insurance type cover. These data were strings or the range of numeric data. These data are normalized to increase the accuracy and prevent from diverging. The normalized data are used to built up the model. Due to large variation of the claimed and unclaimed data, we use smoothing techniques to maintain the ratio.

Using confusion matrix we have obtained 82% accuracy. Using the same pre-processed and normalized data we have prepared report that can be printed and provided to individual member of management using Crystal Reporting and visualize through bar chart using feature of Visual Studio such that it depicts overall performance of the insurance of bike based on selected attributes.

5.2 Problems Faced

- Data filtering through various process.
- Limited no of claimed data.
- Implementing the Algorithm.
- Generating report.

5.3 Limitation

Using Naïve Bayesian Classifier, we created the model based upon the data of bike provided by company. Even though we have created a full working system it have some limitations.

1. Limited to the Bike Insurance only.
2. Data Based on single company only.
3. use of existing data only.

5.4 Work Schedule

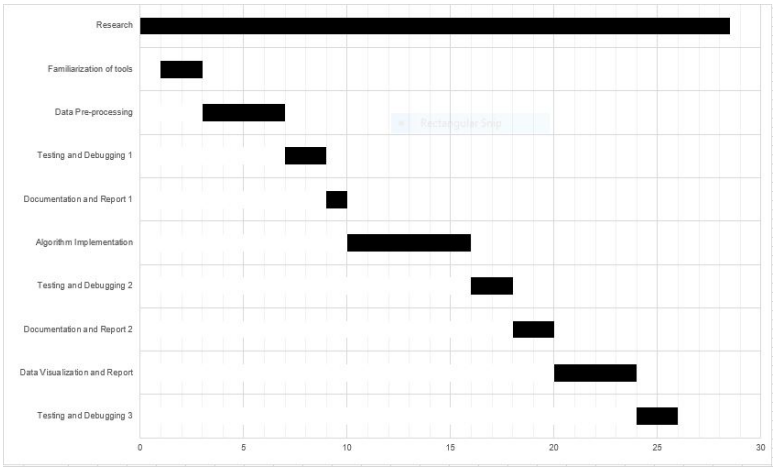

Figure 5.1: Gantt Chart

5.5 Screenshots

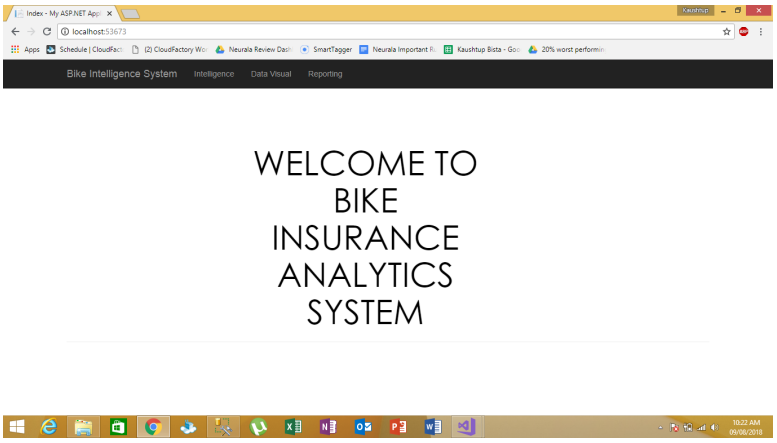

Figure 5.2: landing Page

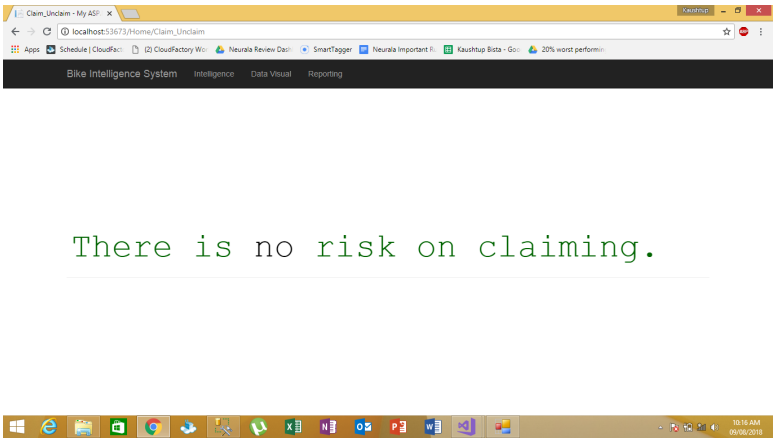

Figure 5.3: System Output

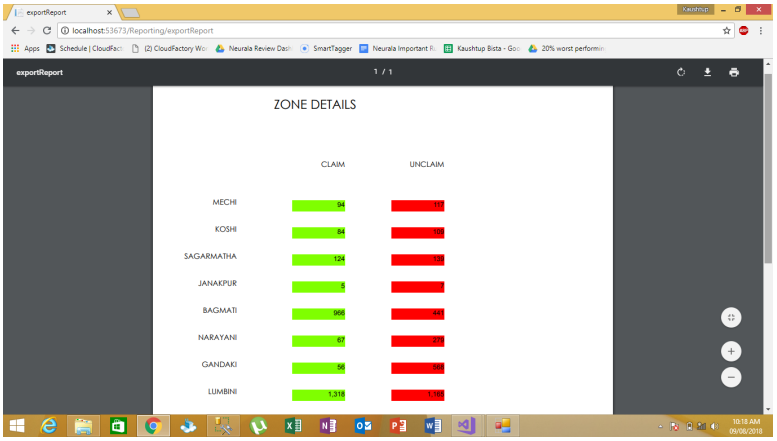

Figure 5.4: Crystal Report

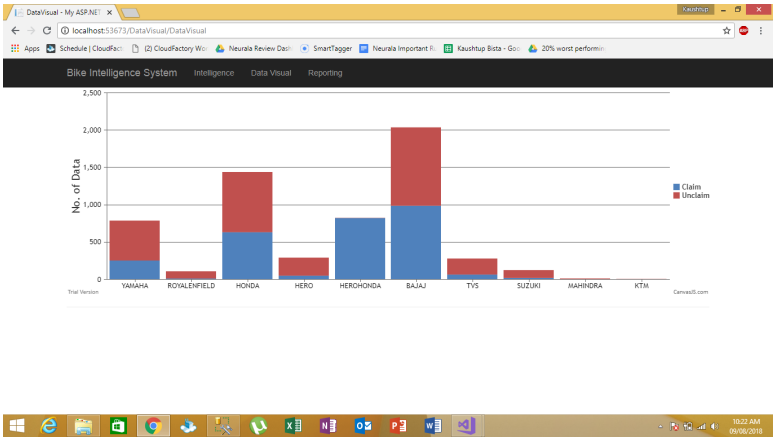

Figure 5.5: Bar Diagram

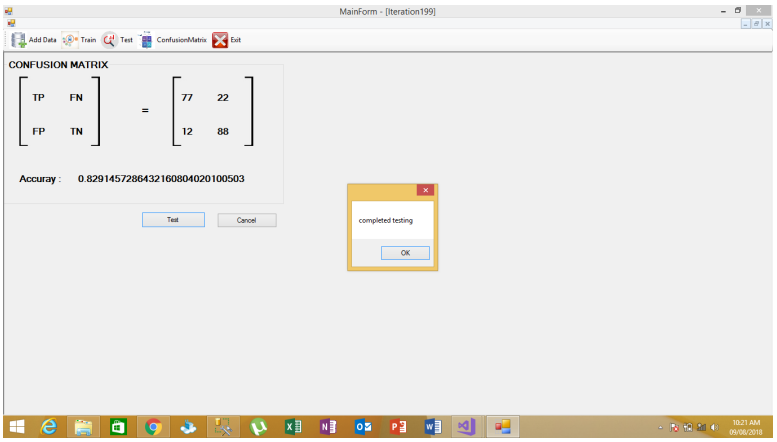

Figure 5.6: Confusion matrix

CHAPTER 6
CONCLUSION AND SCOPE FOR FUTURE ENHANCEMENT

6.1 Conclusion

This **"Project-X: Business Intelligence for Bike Insurance"** is business intelligence system as suggest by name.Business Intelligence(BI) comprises the strategies and technologies used by enterprises for the data analysis of business information. It use for analysis of data to get a result for business strategy.

Most of big and giant company like Google, Microsoft etc. have their business intelligence namely Google Cloud BI solution of Google and Microsoft Power BI, they are also for public use you can it for your purpose. Other like Sisense and QlikSense are publically open. In all these system, all of then have a common point of view that they use customer data and analyse the data and represent the output in the form chart to analyse the result easily. These tools helps the company with giant data to visualize their business status and to develop new strategy for better enhance.

Hence, this intelligence system would be great to the currently growing world with full of data with generating GB of data every hours. Project-X is just small step in business intelligence world. We were able to predict whether the enter data of bike had insurance risk or not. It was a great experience doing this Project as all member have collaborated with each other to make this possible and specially thank to **Mr.Ajay Mani Paudel** for guidance as supervisor.

6.2 Future Enhancement

1. Increase domain to accommodate all automobiles
2. Collects data from many insurance company as far as possible
3. Connecting the system with the company server and update data simultaneously
4. Prediction on the persona basis

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. Grevskott, *Quantitative requirements for non-life insurance under Solvency 2*. SAS Institute Inc, USA, 2011.

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

[4] 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.

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

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

[7] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.