Prediction of Online Vehicle Insurance System using Decision Tree Classifier and Bayes Classifier – A Comparative Analysis

S. S. Thakur
Department of Computer Science & Engineering
MCKV Institute of Engineering
Liluah, Howrah – 711204, West Bengal, India

J. K. Sing
Department of Computer Science & Engineering
Jadavpur University
Kolkata – 700032, West Bengal, India

ABSTRACT
A classification technique (or classifier) is a systematic approach used in building classification models from an input data set. The model generated by the learning algorithm should fit both the input data well and correctly predict the class labels of records it has never seen before. Therefore, a key objective of the learning algorithm is to build models with good generalization capability i.e. models that accurately predict the class labels of previously unknown records. The accuracy or error rate computed from the test set can also be used to compare the relative performance of different classifiers on the same domain. However, the results obtained for accuracy is good and average error rate obtained is equally acceptable for the test records in which the class labels of the test records was not known , in both the classifiers. As computational complexity is concerned Bayes Classifier performs better than Decision Tree Classifier in our system. While the Decision Tree Classifier performed better than Bayes Classifier based on prediction in this system.

General Terms
Data Mining, Artificial Intelligence (AI).

Keywords
Decision tree classifier, Bayes classifier, Web services, Conditional Probability, Posterior Probability, Online Insurance etc

1. INTRODUCTION
Vehicle insurance (also known as auto insurance, GAP insurance, car insurance, or motor insurance) is insurance purchased for cars, trucks, motorcycles, and other road vehicles. Its primary use is to provide financial protection against physical damage and/or bodily injury resulting from traffic collisions and against liability that could also arise there from. The specific terms of vehicle insurance vary with legal regulations in each region. To a lesser degree vehicle insurance may additionally offer financial protection against theft of the vehicle and possibly damage to the vehicle, sustained from things other than traffic collisions.

Car Insurance is mandatory by law. Driving around without valid car insurance is illegal in India. In case of death or bodily injury to a third party or any damage to its car, the car insurance policy provides compensation of up to Rs 1 lakh. Such type of vehicle insurance is known as the third party insurance and it protects not only you but also other people or family members who may be riding / driving you car. Comprehensive car insurance protects your car from any man made or natural calamities like terrorist attacks, theft, riots, earth quake, cyclone, hurricane etc in addition to third party claims/damages. At times car insurance can be confusing and difficult to understand. There are certain guidelines that should be followed by the Car Insurance buyers while choosing the policy. Car insurance [1] acts like a great friend at the time of crisis. It covers the losses made in an accident and thus saves you from paying out the huge sum from your pocket. In many jurisdictions it is compulsory to have vehicle insurance before using or keeping a motor vehicle on public roads. Most jurisdictions relate insurance to both the car and the driver, however the degree of each varies greatly. Several jurisdictions have experimented with a "pay-as-you-drive" insurance plan which is paid through a gasoline tax (petrol tax).

2. VEHICLE INSURANCE IN INDIA - PRESENT SCENARIO
Auto Insurance in India deals with the insurance covers for the loss or damage caused to the automobile or its parts due to natural and man-made calamities. It provides accident cover for individual owners of the vehicle while driving and also for passengers and third party legal liability. There are certain general insurance companies who also offer online insurance service for the vehicle. Auto Insurance in India is a compulsory requirement for all new vehicles used whether for commercial or personal use. The insurance companies [1] have tie-ups with leading automobile manufacturers. They offer their customers instant auto quotes. Auto premium is determined by a number of factors and the amount of premium increases with the rise in the price of the vehicle. The claims of the Auto Insurance in India can be accidental, theft claims or third party claims. Certain documents are required for claiming Auto Insurance in India, like duly signed claim form, RC copy of the vehicle, Driving license copy, FIR copy, Original estimate and policy copy. There are different types of Auto Insurance in India:

1) **Private Car Insurance** – In the Auto Insurance in India, Private Car Insurance is the fastest growing sector as it is compulsory for all the new cars. The amount of premium depends on the make and value of the car, state where the car is registered and the year of manufacture.

2)**Two Wheeler Insurance** – The Two Wheeler Insurance under the Auto Insurance in India covers accidental insurance for the drivers of the vehicle [2]. The amount of premium depends on the current showroom price multiplied by the depreciation rate fixed by the Tariff Advisory Committee at the time of the beginning of policy period.

3) **Commercial Vehicle Insurance** – Commercial Vehicle Insurance under the Auto Insurance in India provides cover for all the vehicles which are not used for personal purposes, like the Trucks and HMVs. The amount of premium depends on the showroom price of the vehicle at the commencement of the insurance period, make of the vehicle and the place of registration of the vehicle.

The auto insurance generally includes:

- Loss or damage by accident, fire, lightning, self ignition, external explosion, burglary, housebreaking or theft, malicious act.
- Liability for third party injury/death, third party property and liability to paid driver
- On payment of appropriate additional premium, loss/ damage to electrical/electronic accessories

The auto insurance does not include:

- Consequential loss, depreciation, mechanical and electrical breakdown, failure or breakage
- When vehicle is used outside the geographical area
- War or nuclear perils and drunken driving.

This paper outlines the implementation of Vehicle Insurance Prediction system using Decision Tree Classifier, and Baye's Classifier.

3. PROPOSED APPROACH

Decision tree classifier [3, 4] and Bayes Classifier [8, 9] for Prediction of Online Vehicle Insurance system emphasizes some key areas. These are

- Proposed Approach for Prediction of Online Vehicle Insurance system has been dealt in section III.
- Algorithm for Vehicle Insurance Prediction system has been dealt in section IV.
- Implementation Methodology for building Decision tree classifier, its working principle and splitting of continuous attributes has been dealt in section v.
- Implementation with Bayes Classifier has been dealt in section VI.
- Experimental evaluation and Results has been dealt in section VII, discussion and conclusion has been dealt in section VIII.

Figure 1 shows the block diagram of the complete system in which the approach for solving classification problems [5, 6, 7] is explained in detail.

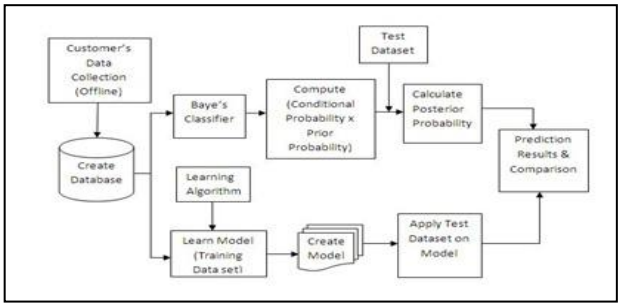

Fig 1: Block diagram of the complete system

First, a training set (Table I) consisting of records whose class labels are known must be provided. The training set is used to build a classification model, which is subsequently applied to the test set (Table II), which consist of records with unknown class labels.

Table 1. Training Dataset

Type	Binary	Categorical	Continuos	Class
Tid	**Vehicle Owner**	**Educational Qualification**	**Age**	**Online - Insurance**
1	Yes	Higher Secondary	70	Yes
2	No	Graduate	60	Yes
3	No	Higher Secondary	30	Yes
4	Yes	Graduate	65	Yes
5	No	Post Graduate	55	No
6	No	Graduate	20	Yes
7	Yes	Post Graduate	70	Yes
8	No	Higher Secondary	40	No
9	No	Graduate	35	Yes
10	No	Higher Secondary	45	No

Table 2. Test Dataset

Tid	Vehicle Owner	Educational Qualification	Age	Online - Insurance
1	No	HS	15	?
2	Yes	Graduate	37	?
3	Yes	Post Graduate	62	?
4	No	Graduate	57	?
5	No	Post Graduate	25	?

3.1 Predictive Modeling

A classification model [6, 7] is used to predict the class label of unknown records. A classification model can be treated as a black box that automatically assigns a class label when presented with the attribute set of an unknown record. Classification techniques are most suited for predicting or describing data sets with binary or nominal categories. They are less effective for ordinal categories (e.g. to classify a person as a member of high, medium, or low income group) because they do not consider the implicit order among the categories. Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model. These counts are tabulated in table known as confusion matrix.

Table 3. Confusion Matrix for a 2-Class Problem

		Predicted Class	
		Class = 1	*Class* = 0
Actual Class	*Class* = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

Table 3- depicts the confusion matrix for a binary classification problem. Each entry *f*ij in this table denotes the number of records for class *i* predicted to be of class *j*. For instance, *f*01 is the number of records from class 0 incorrectly predicted as class 1. Based on the entries in the confusion matrix, the total number of correct predictions made by the model is (*f*11+*f*00) and the total number of incorrect predictions is (*f*10+*f*01). Although a confusion matrix provides the information needed to determine how well a classification model performs, summarizing this information with a single number would make it more convenient to compare the performance of different models. This can be done using a performance metric such as *accuracy*, which is defined as follows:

Accuracy

$$Accracy = \frac{Number\ \ of\ \ Correct\ \ \Pr ediction}{Total\ \ Number\ \ of\ \ \Pr ediction}$$

$$= \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} .....(1)$$

Equivalently, the performance of a model can be expressed in terms of its error rate, which is given by the following equation:

$$Error\ \ \ Rate = \frac{Number\ \ of\ \ Wrong\ \ \Pr ediction}{Total\ \ Number\ \ of\ \ \Pr ediction}$$

$$= \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}} .....(2)$$

3.2 Bayes Theorem Preliminaries

Classification systems were developed earlier to organize a large collection of objects. Automated classification has been a subject of intensive research for many years. Many pattern recognition problems also require the discrimination of objects from different classes. Examples include speech recognition, handwritten character identification, and image classification. The subject of classification is also a major research topic in the field of neural networks, statistical learning, and machine learning. An in-depth treatment of various classification techniques is given in the book by Cherkassky and Mulier [11].

In many applications the relationship between the attribute set and the class variable is non-deterministic. In other words, the class label of a test record cannot be predicted with certainty even though its attribute set is identical to some of the training examples. The situation may arise due to noisy data or the presence of certain confounding factors that affect classification but are not included in the analysis.

In probability theory Bayes' theorem (often called Bayes' Law) relates the conditional and marginal probabilities of two random events [8]. It is often used to compute posterior probabilities given observations. For example, a patient may be observed to have certain symptoms. Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation. As a formal theorem, Bayes' theorem is valid in all interpretations of probability.

Let *X* and *Y* be a pair of random variables. Their joint probability, *P(X = x, Y = y)*, refers to the probability that variable *X* will take on the value *x* and variable *Y* will take on the value *y*. A conditional probability is the probability that a random variable will take on a particular value given that the outcome from another random variable is known. For example, the conditional probability *P(Y = y|X = x)* refers to the probability that a variable *Y* will take on the value *y*, given that the variable *X* is observed to have the value *x*. The joint and conditional probabilities for *X* and *Y* are related in the following manner as shown below:

$$P(X, Y) = P(Y \mid X) \times P(X) = P(X \mid Y) \times P(Y) \qquad (3)$$

Rearranging the last two expressions leads to the following formula, as shown in equation below:

$$P(Y \mid X) = P(X \mid Y)P(Y)/P(X \qquad (4)$$

The above formula is referred to as Bayes' theorem, and can be used to solve the prediction problem.

Before describing how the Bayes' theorem can be used for classification [8] [9] let us formalize the classification problem from a statistical perspective. Let *X* denotes the attribute set and *Y* denote the class variable. If the class variable has the non-deterministic relationship with the attributes, then we can treat *X* and *Y* as random variable and capture their relationship probabilistically using *P(Y|X)*. The conditional probability is also known as the posterior probability for *Y*, as opposed to its prior probability, *P(Y)*. During the training phase, we need to learn the posterior probabilities *P(Y|X)* for every combination of *X* and *Y* based on information gathered from the training data. By knowing these probabilities, test records *X'* can be classified by finding the class *Y'* that maximizes the posterior probability, *P(Y'|X')*.

To illustrate this approach, consider the task of predicting whether a customer i.e. a vehicle owner will go for online insurance. Table I shows a training set with the following attributes: Vehicle owner, Qualification, and Age. Customers interested for Online insurance are classified as Yes, while others are classified as No. Suppose we are given a test record with the following attribute set: X= (Vehicle owner = No, Qualification = Graduate, Age = 60 years). To classify the record, we need to compute the posterior probabilities P(Yes|X) and P(No|X) based on information available in the training data. If P(Yes|X) >P(No|X), then the record is classified as Yes, otherwise it is classified as No, as shown in Table II.

Estimating the posterior probabilities accurately for every possible combination of class label and attribute value is a difficult problem because it requires a very large training set, even for a moderate number of attributes. The Bayes theorem is useful because it allows us to express the posterior probability in terms of prior probability P(Y), the class-conditional probability *P(X|Y)*, and the evidence, *P(X)* can be computed as follows:

$$P(Y \mid X) = P(X \mid Y)P(Y)/P(X) \qquad (5)$$

When comparing the posterior probabilities for different values of Y, the denominator term, *P(X)*, is always constant, and thus can be ignored. The prior probability *P(Y)* can be easily estimated from the training set by computing the fraction of training records that belongs to each class. To estimate the class-conditional *P (X|Y),* we are using naïve Bayes Classifier in our proposed work.

Naïve Bayes Classifier: A naïve Bayes classifier [8] estimates the class-conditional probability by assuming that the attributes are conditionally independent, given the class label y. The conditional independence can be formally stated as follows:

$$P(X \mid Y = y) = \prod_{i=1}^{d} P(X_i \mid Y = y) \quad (6)$$

Where each attribute set $X = \{X_1, X_2, \ldots, X_d\}$ consists of d attributes.

4. ALGORITHM–VEHICLE INSURANCE PREDICTION SYSTEM

A skeleton decision tree classifier algorithm called Tree Growth is shown in Algorithm 4[A] and Bayes Classifier shown in 4[B]. The input to this algorithm consists of the training records E and the attribute set F. The algorithm works by recursively selecting the best attribute to split the data (Step 2) and expanding the leaf nodes of the tree (Step 6 and 7) until the stopping criterion is met (Step 1).

4.1 Algorithm: A Skeleton Decision Tree Classifier Algorithm - TreeGrowth (E, F)

Part - I

1. Create database for the training dataset.
2. Eliminate Redundancy using Normalization.

Part –II

Step 1: Check if (stopping_cond (*E, F*) = True)

then Set leaf = createNode ( )

Set leaf.label = Classify (*E*).

Return (Leaf).

Step 2: Otherwise

Set *root* = createNode ( )

Set *root.test_cond* = find_best_split (*E, F*).

Step 3: Set *V* = {*v*} /* *v* is a possible outcome of

root.test_cond/.*

Step 4: Repeat Step 5 to 7 for each $v \in V$

Step 5: Set E_v = {e | *root.test_cond* (*e*) = *v* and

$e \in E$ *}.*

Step 6: Set *child* = TreeGrowth (E_v, F)

Step 7: Add c*hild* as descendent of *root* and label

the edge *(root→child)* as *v.*

/* End of if */

Step 8: Return (*root*).

4.2 Algorithm: A Skeleton Classifier Algorithm called Bayes' Classifier

The input to this algorithm consists of the records E and the attribute set F.

1. Collect data from customers.
2. Create training records database.
3. Apply Bayes theorem, to express the posterior probability in terms of prior probability *P(Y),* the **class-conditional** probability *P(X|Y),* and the evidence *P(X)*
4. The class-conditional probability *P(X|Y),* implementation using naïve Bayes Classifier.
5. Estimate class-conditional probability, assuming that the attributes are conditionally independent.
6. To classify a **test record,** compute the posterior probability for each class **Y**, using naïve Bayes Classifier.
7. Estimate the conditional probability $P(X_i|Y)$ for categorical and continuous attributes.
8. End

5. DECISION TREE CLASSIFIER IMPLEMENTATION

The subject of classification is also a major research topic in the field of neural networks, statistical learning, and machine learning. An in-depth treatment of various classification techniques is given in the book by Cherkassky and Mulier[11].An overview of decision tree induction algorithms can be found in the survey articles by Murthy [4] and Safavian etal. [5]. Examples of some well known decision tree algorithms including CART, ID3, C4.5 and CHAID. Both ID3 and C4.5 employ the entropy measure as their splitting function. The input data for a classification task is a collection of records. Each record, also known as instance or example, is characterized by a tuple(x, y), where x is the attribute set and y is a special attribute, designated as the class label (also known as category or target attribute). Although the attribute set are mostly discrete, the attribute set can also contains continuous attributes. The class label, on the other hand must be a discrete attribute. This is a key characteristic that distinguishes classification from regression, a predictive modeling task in which y is a continuous attribute. These are the set of rules for constructing a decision tree.

The tree has three types of nodes:

- *A root node:* that has no incoming edges and zero or more outgoing edges.
- *Internal nodes:* each of which has exactly one incoming edge and two or more outgoing edges.
- *Leaf* or *terminal nodes:* each of which has exactly one incoming edge and no outgoing edges.

In a decision tree each leaf node is assigned a class label. The non terminal nodes which include the root and other internal nodes, contain attribute test condition to separate records that have different characteristics. The algorithm presented in this paper assumes that the splitting condition is specified one attribute at a time. An oblique decision tree can use multiple attributes to form the attribute test condition in the internal nodes [6].

Although oblique decision trees help to improve the expressiveness of a decision tree representation, learning the appropriate test condition at each node is computationally challenging.

5.1 Working Principle

To illustrate how the algorithm works, consider the problem of predicting whether a customer will go for online insurance or manual insurance. A training set for this problem has been constructed by data collection through customers who are owners of either 4 wheeler or 2 wheeler vehicles, at different locations in our city which consists of 465 records, which is created using Oracle 10g. After redundancy elimination 416 records are remaining in dataset. In the example shown in Table II, for simplicity we have shown only10 records, where each record contains the personal information of a customer, along with a class label indicating whether the customer has shown interest for Online insurance or not. The initial tree for the classification problem contains a single node with class label Online Insurance = Yes, (see Figure 2), which means that most of the vehicle owners irrespective of 4 wheelers or 2 wheeler owners are going for online insurance.

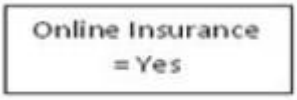

Fig 2: Initial Test Condition

The tree however needs to be redefined since the root node contains records from both classes. These records are subsequently divided into smaller subsets based on the outcomes of the Vehicle Owner of 4 wheeler test condition, as shown in Figure 3. For now, we will assume that this is the best criterion for splitting the data at this point. This algorithm is then applied recursively to each child of the root node.

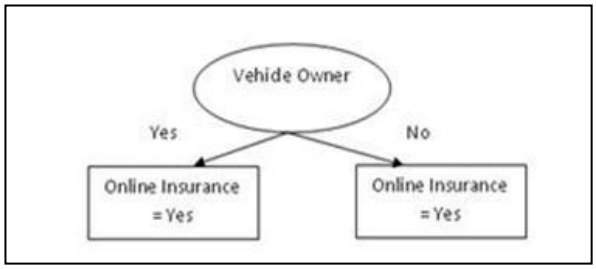

Fig 3: Test condition with 2 attributes

From the training set given in Table II, notice that all customers who are vehicle owners, are interested for online insurance. The left child of the root is therefore a leaf node labeled Online Insurance = Yes, i.e. Manual Insurance = No (see Figure 3).For the right child, we need to continue apply the recursive step of the algorithm, until all the records belong to the same class. The trees resulting from each recursive step are shown in Figure 4 and 5.

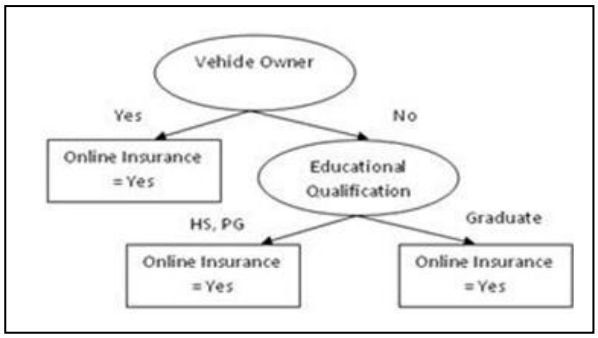

Fig 4: Detailed test condition

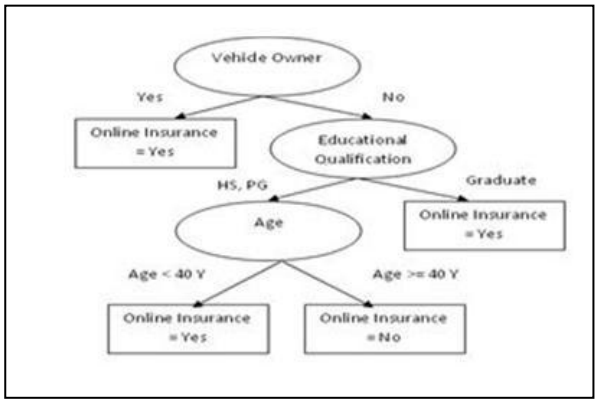

Fig 5: Final test condition

This algorithm will work if every combination of attribute values is present in the training data and each combination has a unique class label. These assumptions are too stringent for use in most practical situations.

5.2 Splitting of Continuous Attributes

Consider the example shown in Table IV, in which the test condition $Age \leq v$ is used to split the training records for the manual insurance classification problem. A brute-force method [13, 14] for finding *v* is to consider every value of the attribute in the *N* records as the candidate split position. For each candidate *v*, the data set is scanned once to count the number of records with annual income less than or greater than *v*. We then compute the Gini index for each candidate and choose the one that gives the lowest value. This approach is computationally expensive because it requires *O*(*N*) operations to compute the Gini index at each candidate split position. Since there are *N* candidates, the overall complexity of this task is *O*(*N2*).To reduce the complexity, the training records are sorted based base on their age, a computation that requires *O*(*N* log*N*) time. Candidate split positions are identified by taking the med points between two adjacent sorted values, and so on. However, unlike the brute-force approach, we do not have to examine all *N* records when evaluating the Gini index of a candidate split position.

Table 4. Splitting of Continuous Attributes

Class	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes
Sorted values →	20	25	30	35	40	45	55	60	65	70

Age

Split position →	15		22		28		33		37		43		50		58		62		67		75	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
No	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
Gini	0.364		0.393		0.346		0.364		0.372		0.340		0.274		0.364		0.346		0.393		0.364	

For simplicity we have shown only 110 records in case of splitting of continuous attributes. For the first candidate, v =15, none of the records has Age less than 15 years. As a result, the Gini index for the descendent node with Age <= 15years is zero. On the other hand, the number of records with Age greater than 15 years is 7 (for class yes) and 3 (for class No), respectively. Thus the Gini index for this node is 0.364.The overall Gini index for this candidate split position is equal to 0.364.For the second candidate, v = 22, we can determine its class distribution by updating the distribution of the previous candidate.

More specifically, the new distribution is obtained by examining the class label of the record with the lowest Age (i.e. 20 years). Since the class label for this record is Yes, the count for class Yes is increased from 0 to 1 (for Age 22 years) and is decreased from7 to 6 (for Age > 22 years). The distribution for class No remains unchanged. The new weighted-average Gini index for this candidate split position is 0.393.This procedure is repeated until the Gini index values for all candidates are computed, as shown in Table IV. The best split position corresponds to the one that produces the smallest Gini index, i.e. v = 50. This procedure is less expensive because it requires a constant amount of time to update the class distribution at each candidate split position.

It can be further optimized by considering only candidate split positions located between two adjacent records with different class labels. As shown in Table IV, the first four sorted records (with Age 20, 25, 30, 35) have identical class labels, the best split position should not reside between 20 years and 35 years and from 55 years and 70 years. Therefore, the candidate split positions at v= 15, 22, 28, 33, 58, 62, 67, 75) are ignored because they are located between two adjacent records with the same class labels. This approach allows us to reduce the number of candidate split positions from 11 to 2.

6. BAYES CLASSIFIER IMPLEMENTATION

To illustrate how the algorithm works, consider the problem of predicting whether a customer will go for manual insurance or online insurance. A training set for this problem has been constructed by data collection through customers at different locations in our city. In the example shown in Table I, each record contains the personal information of a customer, along with a class label indicating whether the customer has shown interest for online insurance or not.

With the conditional independence assumption, instead of computing the class-conditional probability for every combination of X, we only have to estimate the condition probability of each X_i, given Y. This approach is more practical because it does not require a very large training set to obtain a good estimate of the probability. To classify a test record, the naïve Bayes classifier [9] computes the posterior probability for each class Y:

$$P(Y \mid X) = P(Y) \prod_{i=1}^{d} P(X_i \mid Y) / P(X) \quad (7)$$

Since P(X) is fixed for every Y, it is sufficient to choose the class that maximizes the numerator term,

$$P(Y) \prod_{i=1}^{d} P(X_i \mid Y) / P(X) \quad (8)$$

We describe some approaches for estimating the conditional probabilities $P(X_i|Y)$ for categorical and continuous attributes.

6.1 Conditional Probability Estimation - Categorical Attributes

For a categorical attribute X_i, the conditional probability $P(X_i = x_i|Y = y)$ is estimated according to the fraction of training [10] instances in class y that take on a particular attribute value x_i. For example, in the training set given in Figure 2, three out of seven people who take online insurance also own a vehicle. As a result, the conditional probability for P(Vehicle Owner = Yes|Yes) is equal to 3/7. Similarly, the condition probability for defaulted borrowers who are single is given by P(Qualification = Graduate |Yes) = 2/3.

6.2 Conditional Probability Estimation - Continuous Attributes

There are two ways to estimate the class-conditional probabilities for continuous attributes in naïve Bayes classifiers:

1) We can discretize each continuous attribute and then replace the continuous attribute value with its corresponding discrete interval. This approach transforms the continuous attributes into ordinal attributes. The conditional probability $P(X_i|Y = y)$ is estimated by computing corresponding interval for X_i. The estimation error depends on the discretization strategy, as well as the number of discrete intervals. If the number of intervals is too large, there are too few training records in each interval to provide a reliable estimate for $P(X_i|Y)$. On the other hand, if the number of intervals is too small, then some intervals may aggregate records from different classes and we may miss the correct decision boundary.

2) We may assume a certain form of probability distribution for the continuous variable and estimate the parameters of the distribution using the training data [12]. A Gaussian distribution is usually chosen to represent the class – conditional probability for continuous attributes. The distribution is characterized by two parameters, its mean μ and variance σ^2. For each class y_j, the class-conditional probability for attribute X_i is given below.

$$P(x = v \mid c) = \frac{1}{\sqrt{2\pi\sigma^2 c}} e^{-\frac{(v-\mu_c)^2}{2\sigma^2 c}} \quad (9)$$

7. EXPERIMENTAL EVALUATION & RESULTS

The total error is obtained by summing up the errors for both runs. The k-fold cross validation method generalizes this approach by segmenting the data into k equal – sized partitions. During each run, one of the partitions is chosen for testing, while the rest of them are used for training.

This procedure is repeated *k* times, so that each partition is used for testing exactly once as shown in Table V. Again, the total error is found by summing up the errors for all k *runs* as shown in Table VI.

Table 5. Prediction Results

		Predicted Class		Predicted Class		Predicted Class		Predicted Class		Predicted Class	
		Class= 1	Class= 0	Class= 1	Class= 0	Class= 1	Class= 0	Class= 1	Class= 0	Class= 1	Class= 0
Actual class	Class = 1	267	31	255	30	230	38	224	36	212	37
	Class = 0	33	85	32	102	46	104	38	116	40	126

Table 6. Accuracy and Error Rate

Test Data Sets	**1**	**2**	**3**	**4**	**5**	**Results - Average**
Accuracy	84.23	85.41	79.41	81.88	80.94	82.34
Error rate	15.77	14.59	20.59	18.12	19.06	17.66

The average accuracy obtained for Prediction of Online Vehicle Insurance system using Decision tree classifier is 82.34, and error rate obtained is 17.66. In addition to this we run SQL queries on the dataset to validate our results as shown in Figure 6, Figure 7 and Figure 8.

- Yes – Vehicle owners of 4 wheelers interested in Online Insurance, No – Vehicle owners of 4 wheelers not interested in Online Insurance.
- Yes – Vehicle owners of 2 wheelers interested in Online Insurance, No – Vehicle owners of 2 wheelers not interested in Online Insurance.

Fig 6: Results for 4 wheelers and 2 wheelers

Fig 7: Results for 4 wheelers and 2 wheelers with graduate attributes

Fig 8: Results for 4 wheelers and 2 wheelers with graduate and age attributes

From Figure 7, we observe that Vehicle owners of 4 wheelers, who are Graduate and irrespective of Age shows an increasing trend towards Online Insurance. Similarly from Figure 8, we observe that Vehicle owners of 2 wheelers, who are Graduate and Age less than 25 years shows an increasing trend towards Online Insurance and Vehicle owners of 2 wheelers, who are Graduate and Age between 25 years and 50 years, and Age greater than 50 years shows an increasing trend towards Manual Insurance i.e. they are not interested in Online insurance. From this we can conclude that with same qualification as Graduate, Vehicle Owners of 4 wheelers and 2 wheelers are definitely dependent on the Age criteria, which is used as a splitting condition in our research work.

The parameter μ_{ij} can be estimated based on the sample mean of *Xi* (x) for all training records that belong to the class y_j. Similarly, σ^2_{ij} can be estimated from the simple variance (s^2) of such training records. For example consider the age attribute as shown in Table II. The sample mean and variance for this attribute with respect to the class Yes are given below.

Class: $P(C) = N_c / N$

e.g., P(Yes) = 7/10

P(No) = 3/10

- For discrete attributes:

$$P(A_i \mid C_k) = |A_{ik}| / N_c$$

where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k. The prior probabilities of each class can be estimated by calculating the fraction of training records that belong to each class. Since there are three records that belongs to the class Yes, and seven records that belongs to the class No, P(Yes) = 0.3 and P(No)=0.7. To predict the class label of a test record X = (Vehicle owner=No, Qualification= Graduate, Age=65), we need to compute the posterior probabilities P(No|X) and P(Yes|X). These posterior probabilities can be estimated by computing the product between the prior probability P(Y) and the class-conditional probabilities.

As computational complexity is concerned Baye's Classifier performs better than Decision Tree Classifier in our system, whereas the Decision Tree Classifier performs better than Baye's Classifier in case of results prediction. Thus the Online Vehicle Insurance system will help the customers, who logs into the system to see the Prediction results.

8. DISCUSSION AND CONCLUSION

The developed system overcomes the limitations of the manual system, as all the information related to Prediction of Vehicle insurance are available online. To access the developed system the Customer requires a PC and an Internet connection. The system is developed by making use of available tools, techniques and resources that could generate a Secure Web Service. This system is user friendly, cost effective and flexible and the performance of the system is found to be satisfactory. Compared to client-server environments, a Web service is much more dynamic and secured for such an environment poses unique challenges. In conventional systems user identity is known in advance and can be used for performing access control.

In open systems participants may not have any pre-existing relationship and share a common security domain. An important issue is development of access control models, Subject Authentication and Parameter Negotiations are required to restrict access to Web services to authorized users only. In future we will be extending our proposed work, using these features as mentioned above, to develop an online system for Accidental Claim and repair of Damaged Vehicles, using Secure Web Service Negotiation. Easy maintenance, location independent, 24 x 7 availability are some of the features of the developed system. This system is user friendly, cost effective and flexible and the performance of the system is found to be satisfactory.

9. ACKNOWLEDGMENTS

The authors are thankful to Mr. Monoj kumar Mondal, and Mr. Ritam Nath, students of Final Year, CSE Deptt, of MCKV Institute of Engineering, Liluah for their involvement in data collection for the said research work. The authors are also thankful to Prof. Puspen Lahiri, and Prof. Avisekh Saha, Assistant Professor in CSE Department, MCKVIE, Liluah for his valuable suggestions during the proposed work. The authors are also thankful to Prof. Parasar Bandyopadhyay, Principal, MCKVIE, Liluah for giving permission to use the labs. for carrying out the research work.

10. REFERENCES

[1] "What determines the price of my policy?". Insurance Information Institute Retrieved 11 May 2006.

[2] "Am I covered?". Accident compensation Corporation. Retrieved 23 December 2011.

[3] Alsabti K, Ranka S, and Singh V. "CLOUDS: A Decision Tree Classifier for Large Datasets." In Proc. of the 4th Intl. Conf. On Knowledge Discovery and Data Mining, pages 2-8, New York, NY, August 1998.

[4] Murthy S K. "Automatic Construction of Decision from Data: A Multi disciplinary Survey," *Data Mining and Knowledge Discovery*, 2(4):345-389, 1998.

[5] Safavian S R and Landgrebe D." A Survey of Decision Tree Classifier Methodology." *IEEE Trans. Systems, Man and Cybernetics*, 22:660-674, May/June 1998

[6] Utgoff P E and Brodley C E. "An incremental method for finding multivariate splits for decision trees." In *Proc. of the 7thIntl. Conf. on Machine Learning*, pages 58-65, Austin, TX, June 1990.

[7] Wang H and Zaniolo C. "CMP: A Fast Decision Tree Classifier using Multivariate Predictions." *In.Proc. of the 16th Intl. Conf.on Data Engineering*, pages 449-460, San Diego, CA, March 2000.

[8] Ramoni M, and .Sebastiani P, "Robust Bayes classifiers." Artificial Intelligence, 125:209-226, 2001.

[9] Joshi V M, "On evaluating performance of Classifiers for Rare Classes." In Proc. of the 2002 IEEE Intl. Conf. on Data Mining, Maebashi City, Japan, December 2002

[10] Joshi M V, Agarwal R C, , and Kumar V, "Mining Needles in a Haystack: Classifying Rare Classes via Two-Phase Rule Induction." In Proc. of 2001 ACM-SIGMOD Intl. Cong. on Management of Data, pages 91-102, Santa Barbara, CA, June 2001.

[11] Cherkassky V, and Mulier F, "Learning from Data: Concepts, Theory, and Methods." Wiley Interscience, 1998.

[12] Domingos P, "MetaCost: A General Method for Making Classifiers Cost-Sensitive." In Proc. of the 5th Intl. Conf. on Knowledge Discovery and Data Mining, pages 155-164, San Diego, CA, August 1999.

[13] Hastie T, and Tibshirani R, "Classification by pair wise coupling." Annals of Statistics, 26(2):451-471, 1998.

[14] Heckerman D, "Bayesian Networks for Data Mining." Data Mining and knowledge Discovery, 1(1):79-119, 1997.

[15] Kumar V, Joshi M V , Han E H, Tan P N , and Steinbach.M "High Performance Data Mining." In *High performance Computing for Computational Science (VECPAR 2002*), pages111-125, Springer, 2002.

[16] Han J,. and Kamber M, "Data Mining: Concepts and Techniques." Morgan Kaufmann Publishers, San Francisco, 2001.