IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 4, Ver. I (Jul - Aug 2018), PP 06-16
www.iosrjournals.org

Bayesian Classification Model in Predicting Tuberculosis Infection

Bukola Badeji – Ajisafe
University of Medical Sciences, Laje Rd. Ondo City, Ondo State
Corresponding Author: Bukola Badeji – Ajisafe

Abstract: *Predictive model for predicting Tuberculosis infection risk in individuals who came to receive treatment in Tuberculosis and leprosy centre (TBL) Ado – Ekiti was developed.*
The risk variables were identified and developed a predictive model based on the idenified factors. Interviewed were conducted with the staff of of TBL centre to identify risk variables, individuals that come for treatments at the TBL centres with one of the risk factors data set were generated which amounted to 699 patients data were preprocessed and 10-fold cross validation technique was used to partition the dataset into training and testing data. The model was developed using machine learning technique (Naïve Bayes' classifiers) and the result show that Naïve Bayes' classifiers was suitable in carrying out the task for predicting risk with minimum 92% accuracy of predictive model. Receiver Operating Characteristics area for the model was also 0.959 showing the level of bias was low.

Date of Submission: 25-06-2018 Date of acceptance: 10-07-2018

I. Introduction

Tuberculosis is an infectious disease that usually affects the lungs. Compared with other diseases caused by a single infectious agent, tuberculosis is the second biggest killer, globally. In 2015, 1.8 million people died from the disease, with 10.4 million falling ill. In the 18th and 19th centuries, a tuberculosis epidemic rampaged throughout Europe and North America, before the German microbiologist Robert Koch discovered the microbial causes of tuberculosis in 1882.

Following Koch's discovery, the development of vaccines and effective drug treatment led to the belief that the disease was almost defeated. Indeed, at one point, the United Nations, predicted that tuberculosis (TB) would be eliminated worldwide by 2025. However, in the mid-1980s, TB cases began to rise worldwide, so much so, that in 1993, the World Health Organization (WHO) declared that TB was a global emergency; the first time that a disease had been labeled as such.

Fast facts on tuberculosis: Here are some key points about tuberculosis. More detail and supporting information is in the main article.

1. The World Health Organization estimates that 9 million people a year get sick with TB, with 3 million of these "missed" by health systems
2. TB is among the top 3 causes of death for women aged 15 to 44
3. TB symptoms (cough, fever, night sweats, weight loss, etc.) may be mild for many months, and people ill with TB can infect up to 10-15 other people through close contact over the course of a year
4. TB is an airborne pathogen, meaning that the bacteria that cause TB can spread through the air from person to person

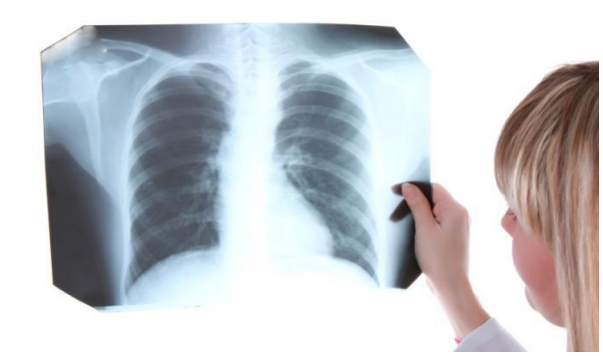

Figure1: Retrieved from http://bestpractice.bmj.com/best-practice/monograph/165/diagnosis/step-by-step.html

The picture above shows that TB usually affects the lungs, although it can spread to other organs around the body.

There are two types of tuberculosis; Doctors make a distinction between these two kinds of tuberculosis infection: which are the latent and active.

Latent TB - the bacteria remain in the body in an inactive state. They cause no symptoms and are not contagious, but they can become active.

Active TB - the bacteria do cause symptoms and can be transmitted to others.

Thomas (2006), said about one-third of the world's population is believed to have latent TB. There is a 10 percent chance of latent TB becoming active, but this risk is much higher in people who have compromised immune systems, i.e., people living with HIV or malnutrition, or people who smoke. TB affects all age groups and all parts of the world. However, the disease mostly affects young adults and people living in developing countries. In 2012, 80 percent of reported TB cases occurred in just 22 countries. These epidemics affect people of all ages both young and old.

1.1 Data Mining and Machine Learning

Data mining can be a useful tool in the health sector and healthcare. Organizations that perform data mining are better positioned to meet their long term needs. Benko and Wilson (2003) argued that data can be a great asset to healthcare organizations, but but they have to be first transformed into information. Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications. Classification is a data mining technique used to predict group membership for data instances.
This work uses data mining technique" Naïve Bayes Classifier" for construction of Tuberculosis prediction. Naïve Bayes classifier technique is mainly applicable when the dimensionality of the input is high Naïve Bayes can often outperform more sophisticated classification method (Rupali Patil, 2014). Naïve Bayes recognizes the characteristics of patients with Tuberculosis. It shows the probability of each input attributes for the predictable state.

II. Literature Review

The Mycobacterium tuberculosis bacterium causes Tuberculosis (TB) is contagious, but it is not easy to catch. The chances of catching TB from someone you live or work with are much higher than from a stranger. Most people with active TB who have received appropriate treatment for at least 2 weeks are no longer contagious. It is spread through the air when a person with TB (whose lungs are affected) coughs, sneezes, spits, laughs, or talks. Of every 100 people with TB, between five and ten people show symptoms.[3] In these people, the disease is called *active*. Tuberculosis kills more than half of the people who are infected if they do not get treatment.

World map shows the prevalence of TB, per 100.000 people, as of 2007. Countries with more cases are shown yellow; those with fewer cases are shown in blue. The most cases were recorded in Sub-Saharan Africa, many occurred in Asia as well.[4] Experts believe that one third of the world population is infected with M. tuberculosis. Peter (2005*) said n*ew infections occur at a rate of one per second. WHO (2010) record in 2007 said, about 13.7 million chronic cases were active globally. Lawn and Zumla (2011) in 2010 said about 8.8 million new cases developed and nearly 1.5 million people died from the disease, most of them in developing countries. WHO (2009) gives account of the number of tuberculosis cases has been decreasing since 2006, and new cases have decreased since 2002. Tuberculosis does not happen at the same rate around the world. About eighty percent of the population in many Asian and African countries test positive for TB, but only five to ten percent of people in the United States do. *Kumar 2007said* People usually get tuberculosis because of a weakened immune system. Many people with HIV and AIDS can also get tuberculosis WHO (2011).

Diagnosis of active TB relies on radiology, doctors often look at an X-ray of the chest and they check body fluids. These fluids have microbes in them, which are grown in cell cultures. The cell cultures are then analysed to see if the person is infected with TB. If the patient has TB, but does not show symptoms, the disease is 'latent'. Doctors use a skin test, called the Mantoux test, to detect latent TB. They often do blood tests too.

TB used to be easily treated and cured with antibiotics. However, the bacterium is now highly resistant to most antibiotics. This resistance makes treatment difficult. Many different kinds of antibiotics need to be given over a long period of time. There is a form of tuberculosis that is resistant to *all* drugs.

Also, Tuberculosis has been a major public health challenge worldwide and still remains a major public health problem in developing countries.

Due to increasing trend of tuberculosis the WHO declared it has a global emergency in 1993. Tuberculosis has been acknowledged as the leading killer of adults and youth and responsible for an estimated million deaths in 1996, majority of which occurred in developing countries.

Nigeria ranks among the top 22 countries with highest burden of tuberculosis in the world (WHO 1999). Tuberculosis is therefore a major public problem in the country. The HIV/TB co-infection rates increased from 2.2% in 1991/1992 to 13.2% in 1996, which poses further threat to the already serious TB situation. The increasing threat to HIV makes it even more imperative that delayed presentation and diagnosis of TB be minimized. The smear positive help the passive cases finding activities of health care providers. Recognizing the patient's perspective is the key to achieving good TB control.

Factors affecting the behaviour of patients and health care workers determine the outcome of case finding and the speed of Tb diagnosis.

The economic impact on patient delayed as analyzed by Sanderson et al(1995) founded that patients bear more than 60% of the total burden of TB cost and majority of time lost from work are incurred before diagnosis. This may ever deter the patient from presenting to health facilities for their TB related symptoms. Patient related cost information is relevant in assessing the assessing the economic impact of the disease on programme design.

The extent of influence by demographic variable (gender, education, residency, type of employment and income) on the likelihood of long has not been fully established (Lawn 1998, Liamet et al. 1999) however in their study how attitude and knowledge found there was a direct correlation between educational background and level of knowledge of TB. TB affects all age groups and all parts of the world. However, the disease mostly affects young adults and people living in developing countries. In 2012, 80 percent of reported TB cases occurred in just 22 countries.

Diagnosis of tuberculosis

Tuberculosis. (2016, October). Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/ TB is most commonly diagnosed via a skin test involving an injection in the forearm.

To check for TB, a doctor will use a stethoscope to listen to the lungs and check for swelling in the lymph nodes. They will also ask about symptoms and medical history as well as assessing the individual's risk of exposure to TB.

The most common diagnostic test for TB is a skin test where a small injection of PPD tuberculin, an extract of the TB bacterium, is made just below the inside forearm.

The injection site should be checked after 2-3 days, and, if a hard, red bump has swollen up to a specific size, then it is likely that TB is present.

Unfortunately, the skin test is not 100 percent accurate and has been known to give incorrect positive and negative readings. However, there are other tests that are available to diagnose TB. Blood tests, chest X-rays, and sputum tests can all be used to test for the presence of TB bacteria and may be used alongside a skin test.

2.0 What causes tuberculosis?

The Mycobacterium tuberculosis bacterium causes TB. It is spread through the air when a person with TB (whose lungs are affected) coughs, sneezes, spits, laughs, or talks.

TB is contagious, but it is not easy to catch. The chances of catching TB from someone you live or work with are much higher than from a stranger. Most people with active TB who have received appropriate treatment for at least 2 weeks are no longer contagious. Since antibiotics began to be used to fight TB, some strains have become resistant to drugs. Multidrug-resistant TB (MDR-TB) arises when an antibiotic fails to kill all of the bacteria, with the surviving bacteria developing resistance to that antibiotic and often others at the same time.

MDR-TB is treatable and curable only with the use of very specific anti-TB drugs, which are often limited or not readily available. In 2012, around 450,000 people developed MDR-TB.

2.1 Treatments for tuberculosis

The majority of TB cases can be cured when the right medication is available and administered correctly. The precise type and length of antibiotic treatment depend on a person's age, overall health, potential resistance to drugs, whether the TB is latent or active, and the location of infection (i.e., the lungs, brain, kidneys). People with latent TB may need just one kind of TB antibiotics**,** whereas people with active TB (particularly MDR-TB) will often require a prescription of multiple drugs. Antibiotics are usually required to be taken for a relatively long time. The standard length of time for a course of TB antibiotics is about 6 months.

TB medication can be toxic to the liver, and although side effects are uncommon, when they do occur, they can be quite serious. Potential side effects should be reported to a doctor and include:

- Dark urine
- Fever
- Jaundice
- Loss of appetite
- Nausea and vomiting

It is important for any course of treatment to be completed fully, even if the TB symptoms have gone away. Any bacteria that have survived the treatment could become resistant to the medication that has been prescribed and could lead to developing MDR-TB in the future. Directly observed therapy (DOT) may be recommended. This involves a healthcare worker administering the TB medication to ensure that the course of treatment is completed.

2.3 Prevention of tuberculosis

If you have active TB, a face mask can help lower the risk of the disease spreading to other people. A few general measures can be taken to prevent the spread of active TB. Avoiding other people by not going to school or work, or sleeping in the same room as someone, will help to minimize the risk of germs from reaching anyone else. Wearing a mask, covering the mouth, and ventilating rooms can also limit the spread of bacteria.

2. 4 Risk factors

People with compromised immune systems are most at risk of developing active tuberculosis. For instance, HIV suppresses the immune system, making it harder for the body to control TB bacteria. People who are infected with both HIV and TB are around 20-30 percent more likely to develop active TB than those who do not have. Tobacco use has also been found to increase the risk of developing active TB. About 8 percent of TB cases worldwide are related to smoking.

People with the following conditions have an increased risk:

- diabetes
- certain cancers
- malnutrition
- kidney disease

Also, people who undergo cancer therapy anyone either young or old, and people who abuse drugs are more at risk. Travel to certain countries where TB is more common increases the level of risk, too.

The following countries have the highest TB rates, globally:

- Africa - particularly West African and sub-Saharan Africa
- Afghanistan
- Southeast Asia: including Pakistan, India, Bangladesh, and Indonesia
- China
- Russia
- South America
- Western Pacific region - including the Philippines, Cambodia, and Vietnam

2.5.1 Complications

If left untreated, TB can be fatal. Although it mostly affects the lungs, it can also spread through the blood, causing complications, such as:

- Meningitis: swelling of the membranes that cover the brain.
- Spinal pain.
- Joint damage.
- Damage to the liver or kidneys.
- Heart disorders: this is rarer.

2.6 Outlook

Fortunately, with proper treatment, the vast majority of cases of tuberculosis are curable. Cases of TB have decreased in the United States since 1993, but the disease remains a concern. Without proper treatment, up to two-thirds of people ill with tuberculosis will die.

2.7 Symptoms of tuberculosis

While latent TB is symptomless, the symptoms of active TB include the following:

- Coughing, sometimes with mucus or blood
- Chills
- Fatigue
- Fever
- Loss of weight
- Loss of appetite
- Night sweats

Tuberculosis usually affects the lungs, but can also affect other parts of the body. When TB occurs outside of the lungs, the symptoms vary accordingly. Without treatment, TB can spread to other parts of the body through the bloodstream:

- TB infecting the bones can lead to spinal pain and joint destruction
- TB infecting the brain can cause meningitis
- TB infecting the liver and kidneys can impair their waste filtration functions and lead to blood in the urine
- TB infecting the heart can impair the heart's ability to pump blood, resulting in a condition called cardiac tamponade that can be fatal

2.8 TB vaccination

In some countries, BCG injections are given to children to vaccinate them against tuberculosis. It is not recommended for general use in the U.S. because it is not effective in adults, and it can adversely influence the results of skin testing diagnoses.

The most important thing to do is to finish entire courses of medication when they are prescribed. MDR-TB bacteria are far deadlier than regular TB bacteria. Some cases of MDR-TB require extensive courses of chemotherapy, which can be expensive and cause severe adverse drug reactions in patients.

III. Bayesian Classification Model

Extensive review of literature on related areas in Tuberculosis is prediction was carried carried out and interview was conducted with the health officers in order to identify required prediction variables for tuberculosis predictive model was developed using supervised machine learning techniques (Bayesian and Naïve Bayes classifier was used). The validation of the model was done by comparing the historical data collected from tuberculosis and leprosy center (TBL) in Ado Ekiti, Ekiti State with predicted values. The performance of this model was tested and we found out that it can predict patients with tuberculosis with an accuracy, based on selected dependable variables.

Bayesian approaches are powerful tools for decision and reasoning under uncertain conditions. Bayesian classifiers employ probabilistic concept representations, and range from the Naïve Bayes (NB) to Bayesian networks (Domingos and Pazzi, 1997). Bayesian reasoning is based on the assumption that the relation between attributes can be represented as a probability distribution (Maimon and Rokach, 2005). Naïve Bayes, used in this study, is the simplest and most straight forward Bayesian learning method based on strong independence assumption. Here, the problem examined is supervised in nature, and then the objective is to find the conditional distribution of the target attribute given the input attribute.

3.1 Bayesian Network

Bayesian networks are one of the most widely used graphical models to represent and handle uncertain information (Amor *et al*, 2004). A Bayesian network is a graphical model for probabilistic relationships among a set of variables. It specifies joint conditional probability distributions. A Bayesian network is defined by two components – a directed acyclic graph (DAG) and a set of conditional probability tables (figure 3b). Each node in the DAG represents events or variables. The variables may be discrete or continuous-valued. Figure 3a is a simple Bayesian network adapted from Russel *et al* (1995) for six variables

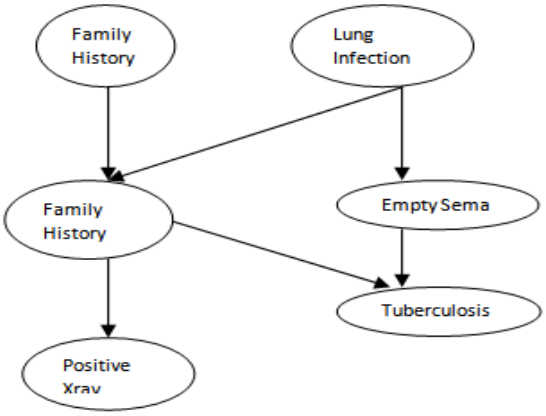

Figure 3.1a: A Simple Bayesian Network

	FH,TB/H	FH,~TB/H	~FH,TB/H	~FH,~TB/H
TB	0.8	0.5	0.7	0.1
~TB	0.2	0.5	0.3	0.9

Figure 3.1b: A prior probability for a simple Bayesian Network of Figure 31a

The arrows in figure 3.1(a) called a representation of causal knowledge. For instance, having lung infection is influenced by a person's family history of tuberculosis infection, as well as whether or not the person is having tuberculosis infection or HIV. The variable Positive X-Ray is independent of whether the patient has a family history of tuberculosis or is a latent and the active tuberculosis infection given that we know that the patient has lung.

The variable Positive X-Ray is independent of whether the patient has a family history of Tuberculosis or HIV given that we know that the patient has Lung Infection Also, the arcs show that the variable Tuberculosis is conditionally independent of Empty sema, given its parents, Family History and active or passive tuberculosis. Figure 3.1(b) shows the conditional probability for each possible combination of values of its parents. For instance,

P(Tuberculosis = yes/FamilyHistory = yes, Lung Infection = yes) = 0.8

P(Tuberculosis = no/FamilyHistory = no, Lung Infection = no) = 0.9

Given a data tuple $X = (x_1 \dots x_n)$ described by the attribute $X = (y \dots y_n)$, the joint probability distribution is $P(x_1 \dots x_2) = \prod_{i=1}^{n} P(x_1 \,|Parents(Y_i)), where\ P(x_1 \dots x_n)$ is the probability of a particular combination of values of X, and the values for P(*x*i/Parents(Yi)) correspond to the entries in the conditional probability table for Yi.

Applying Bayesian network to tuberculosis infection detection going by the popular work of Axelsson (1999) on intrusion detection where Bayesian rule of conditional probability was used to point out the implications of the base-rate fallacy for intrusion detection. He observed that the tests or models that identify malicious events very accurately may raise many false alarms because a-priori probability of an attack in the input data stream is usually very low. Kruegel *et al* (2007) uses Bayesian networks to improve the aggregation of different model outputs and integration of additional information into the decision process. It was observed that accuracy of event classification process improved significantly.

In this paper, a simple Bayesian network known as naïve Bayes or naïve Bayesian is used. A naïve Bayesian network is a very simple network with two layers and assumes complete independence between the information nodes. These limitations result in a directed acylic graph (DAG) with only one root node (called parent), representing the unobserved node, and several children, corresponding to observe nodes and no other casual relationship between nodes.

3.2 Bayes Theorem

Let X be a data tuple with a set of n attributes. Recall the general formula for conditional probability. P (AB) is called the joint probability and P (B) is called the prior probability of B.

$$P(A\backslash B) = \frac{P(AB)}{P(B)}$$

The conditional probability formula (4.1) in reverse produces

$$P(B\backslash A) = \frac{P(AB)}{P(A)}$$

Which gives

$$P(AB) = P(B\backslash A)P(A)$$

substituting equation 4.3 into 4.1 to get the Bayes' theorem

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

3.3 Naïve Bayesian Classifications

A Bayesian classifier is a probabilistic model of what is happening in data, which estimates the class for new data item. Naïve Bayesian has been successfully applied in solving various problems (Christopher *et al.,* 2003, Amor *et al*., (2004)). According to Zaki (2003), a naïve Bayesian Classification works as follows:

(i). Given a training set of tuples and their associated class labels. The training data has nattributes, which may be categorical or numeric, $\{A_1, A_2, \ldots . A_3\}$ and each tuple is represented by an n-dimensional attribute vector $X = (x_1, x_2, \ldots . x_n)$, depicting n measurements made on the tuple from n-attributes.

(ii). Suppose that there are k classes, $C_1, C_2, \ldots ., C_k$ and each data point belongs to one of the k classes. The goal is to develop a Bayes classifier$M(X) \rightarrow C_i$, where X can be any point, not necessarily a member of the training data. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class C_1 if and only if $P(C_i\backslash X) > P\big(C_j\backslash X\big) for\ 1 \leq j \leq k, j \neq i.$

Thus, we maximize conditional probability $P(C_i|X)$ over class C_i. The class C_i for which$P(C_i|X)$ is maximized is called the maximum posteriori hypothesis. By Bayes theorem in equation 4.4

$$P(C_i\backslash X) = \frac{P(X\backslash C_i)P(C_i)}{P(X)}$$

(iii). Since $P(X)$ is uniform across all classes Ci, task here amounts to maximizing $P(X\backslash C_i)P(C_i)$ over all C_i. $P(C_i)$is estimated by looking at the given dataset. Let N be the number of entries in the entire dataset and n_i the number of examples with class C_i. Then $P(C_i) = \frac{n_i}{N}$.

And $P(X\backslash C_i)$ will be computed from the training dataset, which means finding the probability distribution over a n-dimensional space: $P(X = (x_1, x_2, \ldots \ldots, x_n)/C_i)$.

(iv). The naïve assumption of class conditional probability is made to reduce computation in evaluating $P(X\backslash C_i)$. This presumes that the values of attributes are conditionally independent of one another, given the class label of the tuple.

Thus, $P(X|C_i) = \prod_{k-1}^{n} P(x_k|C_i)$

$= P(x_1\backslash C_i)xP(x_2\backslash C_i)x \ldots \ldots xP(x_n\backslash C_i)$

The probabilities $P(x_1\backslash C_i), P(x_2\backslash C_i) \ldots . . P(x_n\backslash C_i)$ can easily be estimated from the training tuples. Here x_k refers to the value of attribute A_k for tuple X. In other to compute $P(x_1\backslash C_i)$ for each attribute, we consider whether the attribute is categorical or continuous valued.

(a) Categorical attributes. If Ak is categorical, then $P(x_1\backslash C_i)$ is the number of tuples of class C_i in N having the value x_k for A_k.

$$P(x_k\backslash C_i) = \frac{Number\ of\ Cases\ with\ C_i\ that\ have\ A_k = x_k}{N}$$

(b) **Numerical attributes.** If A_k is continuous-valued, then there is need for some assumptions about the distribution of A_k for C_i. General assumption is that continuous-valued attributes have a Gaussian or normal distribution with a mean μ and standard deviation σ defined by

$$f(x_k|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Picking out all samples with a given class C_i for attribute A_k. Then compute μC_i and σC_ionly for that class. This gives

$$f(x_k|\mu C_i, \sigma C_i) = \frac{1}{\sqrt{2\pi\sigma\, C_i}} e^{\frac{(x-\mu C_i)^2}{2\sigma {C_i}^2}}$$

So

$$P(x_k \backslash C_i) = f(x_k | \mu C_i, \sigma C_i)$$

Other forms of normalization could as well be used such as Fuzzy logic, kernel density estimates among others to handle continuous-valued attributes.

Laplace Adjustment: This is used to avoid computing probability values of zero. Let $n_i(x_k)$ denote the number of examples in the training dataset that have class C_i and have $A_k = x_k$. Consider the problem where some $n_i(x_k) = 0$ (for some dimension). Then the entire probability $P(x_k \backslash C_i) = 0$. The solution is to initialize $n_i(x_k) =$ to 0.00001 or 1, then add the observed counts.

3.4 Predicting a class label using naïve Bayesian

The training data in Table 3.1 will be used as illustration. The data tuples are described by the class label attribute (category) values are normal and infection. Let C1 correspond to category = normal and C2 to category = infection.

For Tuberculosis infection predictive model, the variables used are Chest pain, Not having any appetite, Weakness, Weight loss, Chills, Very pale skin, Listless eyes. They are represented as X (input value) and the output class by C – Diagnosed and not Diagnosis. Consider the training data provided and the classification of the following data, X containing the values of each attributes for Tuberculosis Infectious patients identified

X=(Weight loss(X_1)=”value”, Listless eyes(X_2)=”value”,chest pain(X_3)=”value”,weakness (X_4)=”value”)

a. First determine the probability of the output class being YES

$$P(YES|X_i) = (P(X_1|YES) * P(X_2|YES) * P(X_3|YES) * P(X_4|YES)) \cdot P(YES)$$

b. Second, determine the probability of the output class being NO

$$P(NO|X_i) = (P(X_1|NO) * P(X_2|NO) * P(X_3|NO) * P(X_4|NO)) \cdot P(NO)$$

c. Determine the maximum class probability

$$Diagnosed_{class} = \boldsymbol{MAXIMUM}[P(YES|X_i), P(NO|X_i)]$$

Each probability $P(Attribute, X_i|Class, C_i)$ is calculated using the formula in equation (1) below.

$$P(X_i|C_j) = \frac{P(X_i \cap C_i)}{P(C_i)}$$

P(X|Ci) = P(x_1|Ci)* P(x_2|Ci)* P(x_3|Ci)*…….*P(x_k|Ci)

The classification of each training data was performed via the implementation of the Naïve Bayes' classification algorithm which calculates the probability and manipulates them into the Necessary results. A typical demonstration of how this is achieved is shown as follows:

For Tuberculosis infection predictive model, the variables used are: Not Having Appetite (NHA), Chills (CHI), Fever (F), Difficult Breathing (DB), Feeling very Tired (FT), Weight Loss (WL), Listless Eyes (LE), Chest Pain(CP) and Very Pale Skin (VP), Sweating (SW) and they are represented as X (input value) and the output class by C.

X= (x_1=''Weight Loss status = value'', x_2= ''Chest Pain = value'', x_3= ''Listless Eyes=value'', x_4= ''Very Pale Skin = value'')

There is need to maximize $P(X|C_i)P(C_i) for\ I = 1,2$. The prior probability of each class $P(i)$ can be computed based on the training data.

P(Category = normal) = 4/6 = 0.667

P(Category = infection) = 2/6 = 0.333

To compute $P(X|C_i)$ for i = 1,2, the following conditional probabilities are computed

P(TB Infection = Weight loss | Category = normal) = 0.5

P (TB Infection = Weight loss | Category = intrusion) = 0.997

P (TB Infection = Listless Eyes | Category = normal) = 0.5

P (TB Infection = Listless Eyes| Category = intrusion) = 0.003

P (TB Infection = Chest Pain | Category = normal) = 0.250

P (TB Infection = Chest Pain| Category = intrusion) = 0.499

P (TB Infection = Weakness | Category = normal) = 0.250

P (TB Infection = Weakness| Category = intrusion) = 0.499

P (TB Infection = Fever | Category = normal) = 0.499

P (TB Infection = Fever | Category = intrusion) = 0.003

P (TB Infection = Sweating | Category = normal) = 0.50

P (TB Infection = Sweating| Category = intrusion) = 0.003

P (TB Infection = Difficult Breathing| Category = normal) = 0.50
P (TB Infection = Difficult Breathing | Category = intrusion) = 0.997
It should be noted here that Laplace adjustment are used in the computation of conditional probabilities and $n_i(x_k)$ was set to 0.0001.
Using the above probabilities, we obtain

$$P(X|Category = normal) = P(TInfection = Weight\ Loss|Category = normal)x$$
$$P(Infection = Chest\ Pain\ |Category = normal)xP(Infection = Difficult\ Breathing = normal = 0.5 * 0.499 * 0.5.$$

Similarly,

$$P(X|Category = Tinfection) = 0.997 * 0.003 * 0.003 = 8.773 * 10^{-6}$$

To find the class C_i, that maximizes$P(X|C_i)P(C_i)$, we compute

$$P(X|Category = normal)P(Category = normal) = 0.125 * 0.667 = 0.083$$
$$P(X|Category = Tinfection)P(X|Category = Tinfection) = 8.773 * 10^{-6} * 0.333 = 2.821 * 10^{-6}$$

Therefore, the naïve Bayesian Classifier predicts Category = normal for Tuberculosis infectious patient with data X.

Table 3.1: The accuracy obtained by changing the number of instances in the testing data

Training Dataset	Testing Dataset	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy (%)
450	200	185	15	92.00
450	300	285	15	95.00
450	350	333	18	95.10
450	400	380	20	95.00

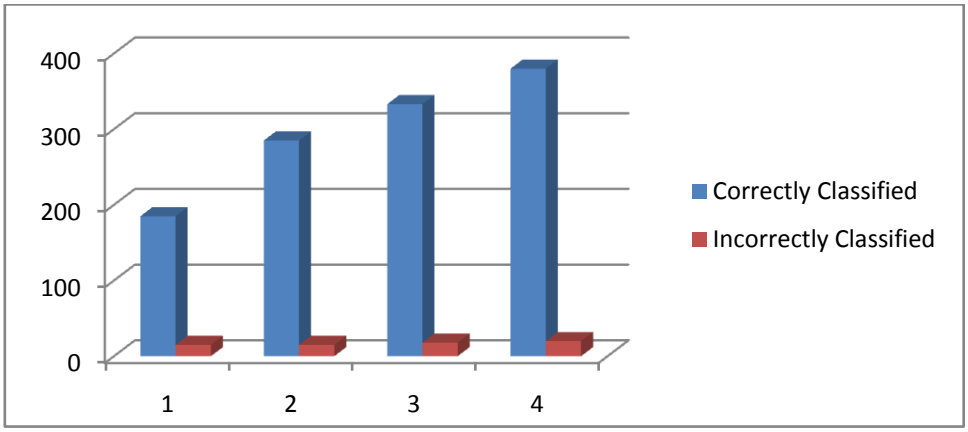

Figure 3.2: column chart results

IV. Discussions

From the result of the analysis made on the dataset using Naïve Bayes' classification in developing the predictive model of Tuberculosis prediction for patients; the following were observed from the table 3.2.
Out of 699 data collected from the TBL (Tuberculosis and Leprosy centre) in Ado Ekiti , after using Naïve Bayes' Classifier to train the data the model developed using 10 fold cross validation, it was discovered that Naïve Baye,' prediction model made the following results. 200 records were selected at random for testing, 185 were correctly classified and 15% were incorrectly classified with 92% accuracy. At the second instance 300 were tested, 279 were correctly classified while 21 were incorrectly classified with accuracy of 95% . at the third instance, 350 records were tested, 333 instances were correctly classified while 18 instances were incorrectly classified with accuracy of 95%. At the fourth instances 400 were tested, 380 instances were correctly classified while 20 instances were incorrectly classified with accuracy of 95%. This Naiye Bayes' shows a very good prediction with minimum of 92% of accuracy. This shows that the Naïve Bayes' Models has a low rate of bias, an average of 0.01(1% of results)

Also, figure 2 shows the graph of the results of the classification the Naïve Bayes' prediction model blue identify yes while red identify NO.

Table 3.3 Validation of the Model with Actual Data

Chest Pain	Weakness	Listless Eyes	Cough	Prediction	Naïve Bayes'
Yes	Yes	Yes	No	Yes	Yes
No	No	No	No	No	No
No	Yes	Yes	No	No	No
Yes	Yes	Yes	No	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	No	Yes	Yes	Yes
No	Yes	Yes	No	Yes	Yes

VALIDATION OF THE MODEL

Validation is the task of demonstrating the model is a reasonable representation of the actual system. It reproduces system behaviour with enough fidelity to satisfy analysis objectives. In table 3.3, the validation of the model with actual values collected from the TBL centre and it was discovered that the values from the Naïve Baye's is almost the same as the actual values using the input variables (Chest Pain, fever, sweat and listess Eyes) from the results of the analysis made on the dataset using Naïve Bayes' classification in developing the predictive model of TB for patients living with the disease. The following were discovered: that out of 430 dataset which had Yes prediction, 423 were classified as Yes and 7 were classified as No and of the 269 dataset which had a prediction of NO, 264 were classified as YES and 5 were classified as NO the model also produced an accuracy of 98% in prediction. This was also representing with a confusion matrix in Table 3.4.

Table 3.4: Confusion Matrix of Naïve Bayes' Classification

	YES	NO
YES	423	7
NO	264	5

V. Conclusion

Mortality is a factor that can be associated with the well being of a population and taken as one of the development indicators of health and socioeconomic status in any country. TB epidemic has a devastated many individuals, families and communities. Therefore, in order to reduce mortality which is one of the millennium goals, there is need to have effective and efficient model that can be used to predict patients living with TB. Early prediction of patients needs for treatment will equally reduce the rate of mortality and morbidity among individuals. It will also help individuals, NGO and the government to make adjustments in taking care of the affected individuals, the Naiye Bayes' predictive model serves as an effective model from the analysis above and will recommend for use to individuals with TB.

VI. Acknowledgement

I acknowledge the contribution of AvH, BMBF and AGNES for the Award of AGNES Junior Researcher Grant, I am indeed grateful. Thank you very much.

References

[1]. Amor, N.B., Beferhat, S. and Elouedi, Z.(2004) Naïve Bayes vs Decision Trees in Intrusion
[2]. Detection Systems, ACM Symposium on Applied Computing, pp. 420 – 424
[3]. Benko, A & Wilson, B. (2003) online decision support gives plans an edge. Managed healthcare executive, Vol. 13 No. 5, p.20
[4]. Christopher, K., Darren, M., William, R.and Fredik, V. (2003) "Bayesian Event classification for intrusion detection", Proceedings of the 19th Annual Computer Security Applications Conference (ACSAC'03), 2003
[5]. Rupali .R. Patil (2014) Heart Disease Prediction system using Naïve Bayes' and Jlinck Mercer Smoothing Kumar V. et al 2007. Robbins basic pathology (8th ed.). Saunders Elsevier. pp. 516–522. ISBN 978-1-4160-2973-1.
[6]. Konstantinos A (2010). "Testing for tuberculosis". Australian Prescriber 33 (1): 12–18. http://www.australianprescriber.com/magazine/33/1/12/18/.
[7]. Peter G. Gibson (ed) (2005). Evidence-based respiratory medicine. Oxford: Blackwell. p. 321. ISBN 978-0-7279-1605-1.
[8]. World Health Organization (2009). "The Stop TB Strategy, case reports, treatment outcomes and estimates of TB burden". Global tuberculosis control: epidemiology, strategy, financing. pp. 187–300. ISBN 978-92-4-156380-2. Retrieved 14 November 2009.
[9]. World Health Organization. November 2010. Retrieved 26 July 2011.
[10]. 10. World Health Organization (2009). "Epidemiology" (PDF). Global tuberculosis control: epidemiology, strategy, financing. pp. 6–33. ISBN 978-92-4-156380-2. Retrieved 12 November 2009.
[11]. World Health Organization (2011). "The sixteenth global report on tuberculosis" (PDF).
[12]. Lawn, SD; Zumla, AI (2011). "Tuberculosis". Lancet 378 (9785): 57–72. doi:10.1016/S0140- 6736(10)62173-3. PMID 21420
[13]. Thomas M. Daniel. (2006). The history of tuberculosis. *Respiratory Medicine*. Volume 100, Issue 11, Pages 1862–1870. Retrieved from http://www.resmedjournal.com/article/S0954-6111(06)00401-X/abstract
[14]. Domingos, P., and Pazzani, M. (1997). On the Optimality of the Naïve Bayes Classifier under zero-one Loss, Machine Learning, 29:2, 103-130.
[15]. Lawn, SD; Zumla, AI (2011). "Tuberculosis". Lancet 378 (9785): 57–72. doi:10.1016/S0140- 6736(10)62173-3. PMID 21420

[16]. Pulmonary tuberculosis. (n.d.). Retrieved from http://bestpractice.bmj.com/best- practice/monograph/165/diagnosis/step-by-step.html
[17]. Zaki, M. (2003) Data Mining Class Notes
[18]. Rupali .R. patil (2014): Heart Disease Prediction System Using Naïve Bayes' and Jelinek mercer smoothing.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Bukola Badeji – Ajisafe "Bayesian Classification Model in Predicting Tuberculosis Infection." IOSR Journal of Computer Engineering (IOSR-JCE) 20.4 (2018): 06-16.