Modeling Influential Extrinsic Factors Of Mastitis In Cows: A Case Study

By

MOMINA ZEB

Master of Science in Bioinformatics Fall

2019-MS BI-4-00000320883

Supervisor

Dr. Zamir Hussain

School of Interdisciplinary Engineering and Sciences (SINES)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

June 2022

Declaration

I, *Momina Zeb* declare that this thesis titled "Modeling Influential Extrinsic Factors Of Mastitis In Cows: A Case Study" and the work presented in it are my own and has been generated by me as a result of my own original research.

0.1 Copyright Notice

Dedication

This thesis is dedicated to my parents for always loving and supporting me.But specifically this thesis is dedicated to My father, **'Haji Jahanzeb'** thanks to him for teaching me the core values of hard work and commitment. For supporting me in the whole educational career. Without his support it would be impossible for me to acheive this dream. Daughters are dear to every father but their dreams are dear only those father who are special.My life all great days credit goes to him. I am nothing without him. Words are not enough to explain his efforts which he has done for me.

Acknowledgments

I consider it my atmost obligation to express my gratitude to Allah Almighty, the omnipresent, kind and merciful who gave me the health, thoughts and the opportunity to complete this task. I offer my humble thanks from the core of my heart to the Holy Prophet, Hazrat Muhammad (Peace Be Upon Him) who is forever a torch bearer of guidance and knowledge for humanity as a whole.

In the completion of this work, I was fortunate having the generous advice and encouragement of my supervisor Dr. Zamir Hussain, and my GEC members Dr. Zartasha Mustansar and Dr.Muhammad Tariq Saeed, Faculty of SINES in selecting the research topic, inspiring guidance, sympathetic and unstinted help at every step right from research synopsis to final manuscript writing.

It is my privilege to express deep sense of gratefulness to my kind teacher, Dr. Rehan Zaffar Paracha, Dr. Ishrat Jabeen, for his valuable suggestions and guidance.

I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. I am very much thankful to my friends Saman Ahmed, Maheera , Rida Ayub, Hira Quereshi for their love, understanding, prayers and continuing support to complete this research work. Also I express my thanks to my brothers, Sadiq, Muneeb and Mobeen for their support and valuable prayers. My Special thanks goes to my friend and my best cousin Fawad Zeb for their support not only complete this thesis but for their support in each and every critical situation of my life.

Last but not the least, I would like to thank my cousins Radma Zeb, Rohma Zeb, Amna,Isma, Haseeb, Rizwan, Saad for supporting me spiritually throughout my life.

I thank my fellows and labmates Mirha Malik , Rida ayub chaudry, Hira Qureshi and Aneela in NUST university for the stimulating discussions.Special thanks to my best

friend and my roomate Syeda Aniqa Bukhari for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last two years.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

RF	Random Forest
DT	Decision Tree
LR	Logistic Regression
CMT	California Mastitis Test
SFMT	Surf Feild Mastitis Test
EC	Electrical Conductivity
SCC	Somatic Cell Count
OR	Odd Ratio
ROC	Receiver Operating Characteristic
HPT	Hyper Parameter Tunning
SPSS	Statistical Package for the Social Sciences
ML	Machine Learning

Abstract

Mastitis is an acute disease that mostly occurs in milking cows who experience a red, painful udder with fever.This disease complex is the result of interaction of various factors associated with the host, pathogens and the environment. The current study aims to address the significant features from the available disease dataset using machine learning approaches such as Decision Tree, Logistic Regression and Random Forest.Factors are categorized as external factors in the given dataset. Secondary data is included in this research project which is collected from the Anti Bacter research group of ASAB.Data was gathered from the area of district Rawalpindi in order to address the disease vs. normal cows and to collect the information of each cow by questionnaire survey from farmers and then labeled the data. The questionnaire performa was designed on the basis of 28 extrinsic factors i.e mastitis history of cows, bedding material, housing system, no.of attendees ,management system etc. A total 432 lactating cows data are included in this study.These cows were examined for mastitis by collecting their milk samples.The Surf Field Mastitis test (SFMT) was then used in order to classify the disease vs normal cases.In this study,Chi-square test is used to determine the association between the dependent variables i,e mastitis disease and the independent variables given in the data.Assessment analysis is performed on the predictive models through accuracy, sensitivity, specificity and precision. ROC curve is used for comparative analysis of predictive machine learning models.This study would help to spread awareness among farmers.

Chapter 1

Introduction

Milk, eggs, butter, meat and oils are main sources of nourishment that are enormously important to the good health and adequate nutrition of both the rural and urban populations. Due to socioeconomic issues, the condition of livestock in developing countries is dissimilar to that of developed countries. The majority of livestock is held by small farmers, and mass production is not encouraged because of high transport costs, inadequate infrastructure and other expenses [1].Due to lack of knowledge and resources it is difficult for farmers to maintain hygiene resulting infectious diseases in animals such as tuberculosis, Brucellosis, mastitis etc. Mastitis is a critical and expensive disease in the dairy industry as it is the second most prevalent disease among dairy cows [2].

1.1 Mastitis

Mastitis is the most costly disease in dairy farming today and remains one of the major problems concerning the dairy industry found a mean annual incidence of clinical mastitis between 25 and 30 cases/year per 100 cows.Average economic losses due to mastitis are estimated to be around 150 Euro per cow and year.Early detection of mastitis would reduce milk yield losses . Moreover, early treatment has significantly limited the severity of the disease and, in many cases, prevented the appearance of clinical cases.To sum up, early detection of mastitis is very important not only because of the reduction of the economic impact, but also because of the benefits to the animals welfare [3].Once a cow suffers from mastitis it will never return to its normal milk production.

1.2 Process of Mastitis Infection in Cows:

1. Organisms enter into the udder through teat canal.
2. Migrate up the teat canal and colonize the secretory cells.
3. Colonized organisms produce toxic substances harmful to the milk producing cells.

1.3 Classification of Mastitis:

There are many ways to classify mastitis. Mastitis cases can be divided on the basis of origin into environmental and contagious.

1. Contagious mastitis
2. Environmental mastitis

1.3.1 Contagious Mastitis

Contagious mastitis is due to spread from infected quarter. The most contagious pathogens causing intramammary inflammation are Staphylococcus Aureus, Streptococcus Agalactiae, and Streptococcus Uberis [4].

1.3.2 Environmental Mastitis

Environmental mastitis is caused by environmental pathogens, which are bacterial germs found in the environment. The environmental pathogens causing mastitis in cows are E. coli, Klebsiella (K.) Pneumoniae, Enterobacter Aerogenes, and Streptococcus Uberis [4].

1.4 Symptoms of Mastitis:

The most common symptoms appear in dairy cows during mastitis are following below:

- Inflammation of mammary tissues
- Abnormal milk appears
- Eating disorder
- State of sleepiness [4]

1.5 Factors Associated with Mastitis :

1.5.1 External Factors

The external risk factors which are involved in causing mastitis disease include poor hygiene environment, poor management system, existing trauma, bedding material, floor type, hygiene practice, tick infestation etc [5].

1.5.2 Internal Factors

The intrinsic risk factors which are involved in causing mastitis disease include breed, age, parity number, lactation stage and body condition score [5].

1.6 Types Of Mastitis:

There are two types of mastitis

1. Clinical mastitis
2. Sub-clinical Mastitis

1.6.1 Clinical Mastitis

Clinical mastitis has symptoms such as abnormal milk, udder swelling, elevated temperature, anorexia and lethargy [5].The major pathogens that are involved in clinical mastitis are Escherichia Coli, Staphylococcus Aureus and Streptococcus Uberis [6].

Risk Factors Associated with Clinical Mastitis

Among all the crucial risk factor to analyze their multivariant analysis with respect to mastitis prevalence and reproductive disease in the heifer too. Retained placenta, uterine infections, pyometra, dystocia, as well as twin births were among the reproductive diseases studied. During the prepubertal phase, such abnormalities were found to be linked to clinical mastitis [7].

1.6.2 Sub-Clinical Mastitis

In Sub-clinical mastitis there is no clear sign visible. But the quality of the milk becomes decline[5].The Sub-clinical mastitis is caused by minor pathogens such as Staphlococcus Aureus and Corynebacterium Bovis [6].

Risk Factors Associated with Sub-clinical Mastitis

1. Boosting in concentrate feeding to heifers aged 11–16 months
2. Moving heifers to restricted housing on the day of calving
3. The percentage of cows in the herd who are prone to mastitis
4. The application of restraining measures during milking [7].

1.7 Diagnostic Tests for Mastitis:

Mastitis can be indicated by the use of the following methods.

1. California Mastitis Test (CMT)
2. Surffield Mastitis Test (SMT)
3. Electrical Conductivity (EC)
4. Somatic cell count (SCC)

1.7.1 California Mastitis Test CMT:

This test was done according to the method described by Schalm and Noorlander (9), at cowside, by mixing an equal volume of milk with a 1:1000 dilution of 3% sodium lauryl sulphate and bromocresol. Each quarter's milk sample was placed in 1 clean well of a plastic test paddle, divided into 4 separate wells. As the plate was rotated gently, any color changes or formation of a viscous gel were interpreted. Scores were given within the range 0-3, with 0 for no reaction, 1 for a weak positive, 2 for a distinct positive, and 3 for a strong positive [8].

1.7.2 Surf Field Mastitis Test (SFMT)

The test was carried out by mixing equal amounts of test solution and milk together. 6 teaspoons (about 15 g) of household detergent Surf Excel was dissolved in 1/2 litre of

clean tap water and agitated for about 1 minute to get a 3 percent test solution. The liquid was whirled for one minute, and the emergence of varied degrees of floccules or gel was regarded as a positive SFMT reaction [9].

1.7.3 Electrical Conductivity

EC, which increases during the infection of dairy cows, is also one of the diagnostic method for detection of subclinical mastitis. EC is determined by the concentration of anions and cations. According to Kitchen mastitis increases the EC of milk because of changes in ionic concentrations. As a result of the damage to the udder tissue, concentrations of lactose and K+ decrease, and concentrations of Na+ and Cl increase [8].

1.7.4 Somatic Cell Count (SCC)

The somatic cell count (SCC) can be used to assess the health of an udder. Cows who are healthy or have recovered from mastitis should have an SCC of less than 200,000 cells/mL, whereas cows with counts greater than 400,000 cells/mL should be regarded inframammary infected [4].

1.8 Screening of Mastitis :

The screening of mastitis is normally practicing through by taking milk sample from cow.After doing physical examination of milk if the colour odour and consistency of the milk becomes abnormal then it is futher proceed for microbiological analysis. After performing microbiological investigation on milk sample the bacteria present in it becomes appear on the petri plate. And then antibiotic susceptibility profile test is used for designing medicines according to it.Figure 1.1 shows the procedure of screening the mastitis.

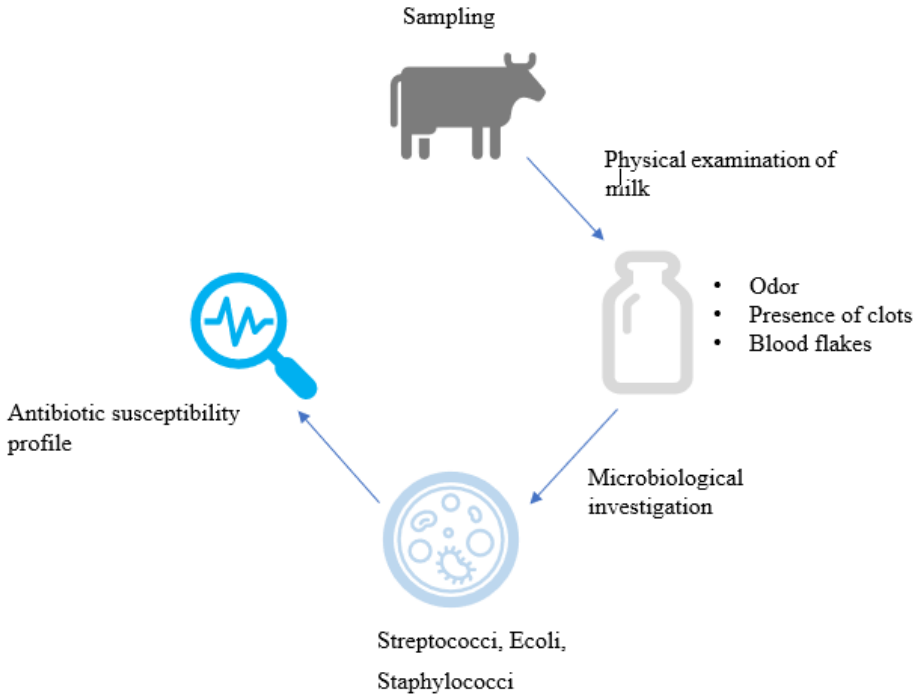

Figure 1.1: Screening Of Mastitis.

1.9 Incidence and Prevalence of Mastitis at International Level:

At a global level, the illness generated almost $35 billion in yearly damages.Mastitis causes an estimated $2 billion in economic losses in the United States each year [10]. According to a survey performed roughly 20 years ago, clinical and subclinical mastitis reduced milk output by 50% and 17.5%, respectively, in India [11].

1.10 Incidence and Prevalence of Mastitis at National Level:

Mastitis is by far the most horrible illness afflicting the dairy sector across the world, but the population of Pakistan is particularly concerning and requires immediate attention for management due to the disease's huge economic costs. Mastitis is prevalent in Pakistan, with a prevalence of 16.72 %. Although it was predicted two decades ago that overall costs caused by clinical mastitis in Punjab province just amounted to Rs.240 million per year, information on current losses due to this illness are not accessible in

Pakistan [10]. Previous research in other parts of Pakistan found that the prevalence of clinical mastitis in buffaloes with cattle was 21.08% and 16.72 %, correspondingly [12].

1.11 Statement of the Problem:

List of significant potential extraneous qualitative factors are heterogenous (varies in literature), and very little is available with respect to the development of predictive modeling to identify the state/class of a unit, i.e. disease/normal.

1.12 Proposed Solution:

To identify another set of significant extraneous factors for local data with the aim to develop predictive models using machine learning techniques. It will help to reduce the prevalence of mastitis in dairy cows at an early stage.

1.13 Objective:

Following is the main objective of this study :
1. Development of predictive models using Logistic Regression, Decision Tree and Random Forest considering a binary dependent variable (state of disease either yes or no) and 28 extrinsic factors provided in Table 1.1 as independent variables.

1.14 Research Gap:

The recent literature review has focused on prevalence of mastitis in dairy cows and identified significant features on the basis of Chi-square analysis.But there is a limited number of extrinsic features are used in previous study held at an international level.Alongwith, there exists a variation between the selection of features extrinsic and intrinsic in the previous literature.The current study aims to identify the comprehensive set of significant features from the available 28 features via machine learning approaches.

1.15 Study Design:

The current project is designed to screen the presence of mastitis in cows via machine learning techniques. For this purpose, the association of extrinsic factors with prevalence of mastitis will be determined by using Chi-square Analysis. For further screening and testing, various machine learning approaches like Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) will be used in order to determine the main extrinsic factors associated with mastitis disease.

1.16 Data Description:

The secondary data of sample size n=432 cows has been collected from the Anti Bacter Research Group of ASAB (Atta-Ur-Rahman School of Applied Biosciences). The respective group collected the data via clinical examination of the udder and teats of each cow by visiting 40 dairy farms of Rawalpindi. The composed data was then screened through Surf Field Mastitis Test (SFMT) to check clinical or subclinical levels. The data was collected on the basis of the following 28 factors shown in 1.1.

Table 1.1: Details of 28 Extrinsic Factors .

S.No	Factors	Categories	Description of Categories
1	Management System	Two	Intensive /Semi intensive
2	Bedding Material	Two	Yes /No
3	Housing	Two	Group /Stall
4	Floor Type	Three	Muddy Concrete Mixed
5	Milking Method	Two	Manual /Machine
6	Milking Routine	Two	Two times /Three times
7	Washing of Udder	Two	Whole udder /Only teats
8	Drying of Udder	Two	Yes /No
9	Position of Udder	Two	Normal /Pendulous
10	Condition of Udder	Three	Atrophy /Normal /Swelling
11	Presence of Ticks	Two	Yes /No
12	Lesions(teat end lesions)	Two	Yes /No
13	Dipping(pre post teat dipping)	Two	Yes /No
14	Standing Position after Milking	Two	Yes /No
15	Last(milking the mastitic cow last)	Two	Yes /No
16	Use of Hormones	Two	Yes /No
17	Use of towel	Two	Yes /No
18	History of Mastitis	Two	Yes /No
19	Udder Hygiene Score	Three	Moderately dirty /Slightly dirty / Very dirty
20	Feed Sharing	Two	Yes /No
21	Manure Removal	Two	Daily /Once a week
22	No. of Attendees	Two	Only one /Two people
23	Location of Farm	Two	Rural /Urban
24	Size of Herd	Two	>10 no. of cows /<10 no. of cows
25	Type of Herd	Two	Mixed /Single type
26	Age	Two	<5 years />5 year
27	Breed	Two	Local /Cross
28	Stage of Lactation	Three	Early /Mid /Late

Chapter 2

Literature Review

In this chapter we will discuss various studies which have been done on the prevalence of mastitis disease at an international and national level. Also, discussed different literatures about intrinsic and extrinsic factors associated with mastitis disease at an international level.

2.1 International Studies

M. Tezera and E. Aman Ali, et al, 2021 performed a cross-sectional study based on California Mastitis Test (CMT) in order to determine the prevalence of mastitis and to identify their intrinsic and extrinsic risk factors in dairy cows in the area of Western Ethiopia.The datasets used in this study consists of total 367 lactating cows which were examined clinically for the detection of clinical and sub clinical mastitis. Based on the CMT results the cow level prevalence of mastitis was 40.3% (n=48), of which 11.99% (n=44) was clinical and 28.34% (n=104) was subclinical mastitis reported respectively. The validation of results was done by using chi square analysis of intrinsic factors and extrinsic factors. Their results showed that intrinsic factors such as breed, stage of lactation and body condition score were statistically differences ($P<0.05$) in the prevalence of mastitis. While extrinsic factors such as hygiene practice and type of floor was significantly associated with the occurrences of mastitis [13].

Nazira Mammadova, İsmail Keskin, et al, (2013) performed a study on an Holstein dairy cattle based on an approach Support Vector Machine (SVM) in order to determine the presence of subclinical and clinical mastitis. The proposed method detected mastitis in a

cross-sectional representative sample of Holstein dairy cattle milked using an automatic milking system. The study used such suspected indicators of mastitis as lactation rank, milk yield, electrical conductivity, average milking duration, and control season as input data. The output variable was somatic cell counts obtained from milk samples collected monthly throughout the 15 months of the control period. Cattle were judged to be healthy or infected based on those somatic cell counts. This study undertook a detailed scrutiny of the SVM methodology, constructing and examining a model which showed 89% sensitivity, 92% specificity, and 50% error in mastitis detection [14].

Another cross-sectional study had been done on bovine mastitis in which they determine the prevalence of bovine mastitis and to assess potential risk factors among lactating cows, both local and crossbreeds in and around the Northeast Algeria.Data was collected in a questionnaire during the farm visit. The sample size of 324 lactating cows was randomly selected (162 for each locality and crossbreeds) managed under extensive, semi-extensive and intensive farming systems. All animals were examined visually for clinical mastitis by clinical and physical examination of the udder and milk, then tested for subclinical mastitis using California Mastitis Test (CMT). Descriptive statistics were performed to summarize the prevalence of mastitis. As a result, 32/324 (9.80%) cows were positive for clinical mastitis and 103/324 (31.79%) cows for sub clinical mastitis were found respectively [9]. Based on the chi-square analysis of risk factors with mastitis it is stated that the prevalence of mastitis was high in late lactation as compared to early and mid-stage lactation. A. Hocine, R. Bouzid, H. Talhi, and D. Khelef, et al, 2021.[5]

A study was conducted on 200 randomly selected farms in each of the Iringa and Tanga regions of Tanzania in order to determine the prevalence and risk factors for subclinical mastitis in dairy cows managed by smallholders. The California mastitis test (CMT) and bacteriological culture of 1500 milk samples taken from 434 clinically normal cows were used in this study to determine the subclinical mastitis. The percentage of cows (and quarters) with subclinical mastitis was 759% (462 %) with a CMT result. While the percentage of cows (and quarters) with subclinical mastitis was $43 \cdot 8$ % ($24 \cdot 3$%) with a culture result. Boran breed, brought-in cow (rather than homebred), peak milk output, and age were all substantially linked with an elevated chance of a CMT-positive quarter. Hand milking with the stripping approach was linked to a considerably decreased occurrence of CMT-positive quarters. CMT-positive cows, as well as brought-in

and older cows, were more likely to be culture positive. Karimuribo ED, Fitzpatrick JL, Swai ES, Bell C et al, 2008 [15].

A research study conducted in Southern Ethiopia from February 2001 to March 2002 by Demelash Biffa, Etana Debela and Fekadu Beyene et al. involving 974 milking cows indicated the prevalence of mastitis and risk factors in the respective dairy cows. Data was gathered on the basis of CMT and clinical inspection of udder. Of the total animals examined, 340 cows were mastitic positive out of which 116 show clinical symptoms while 224 were do not show any symptoms but mastitic positive. Mastitis prevalence were higher on those cows managed under semi-intensive husbandry practice as compared to those cows managed under extensive and intensive environment.Season (rainy), history of mastitis, crossbreeds cows and inadequate sanitation of dairy environment were important factors contributing to high prevalence of mastitis [16].

D.cavero and Kriter et al. (2007) used an SCC (somatic cell count) as well as a neural network to evaluate 478 cows for mastitis prevention and monitoring in an autonomous milking system. For the creation of the Neural Network prediction model, electrical conductivity, milk supply rate, days in milk, and dairy flow rate parameters were employed as input data. The sensitivity, specificity, accuracy, and error rate of the model were used to evaluate it [3].

Karin Östensson, Vo Lam, Ewa Wredle et al. (2013), studied the prevalence of subclinical mastitis etiologic agents at twenty farms. 458 quarters of 115 clinically healthy cows were sampled for milk. The overall prevalence of subclinical mastitis at quarter SCC basis and at cow basis were 63.2% and 88.6%. The most prevalent bacteria species detected was Streptococcus agalactiae. The prevalence of this bacteria is caused by the poor milking hygiene and low awareness of proper measures [17].

2.2 National Studies:

A research study conducted in (2004) by M.Q. Bilal, M.U. Iqbal, G. Muhammad, et al, in the area of Faislabad including peri-urban and urban area in order to identify the factors affecting the prevalence of mastitis in buffaloes. Data was gathered on the basis of questionnaire survey.The questions asked about the number of factors including condition of milk from affected teat, number of animals having swelling /redness of any

teat quarter, floor condition, milking method etc. The results of the conducted study showed that the prevalence of mastitis was higher in peri-urban areas which was 25.12% as compared to rural areas which was 19.74% .The highest incidence was observed during 4 to 6 months after calving both in rural areas (45.08%) and peri-urban (45.76%). Cemented floors are more favourable for mastitis. Animals who was milked by labourer instead of their owner having more chances of mastitis to occur in per-urban areas [10].

Amjad Khan, Muhammad Hassan Mushtaq and Mansur Ud Din Ahmad et al, (2015) reported the results of national survey of clinical mastitis in Khyber Pakhtunkhwa. The total 367 smallholder rural farmers were interviewed and the 606 buffaloes & 611 cattles were examined for one year in the field in order to determine the prevalence of clinical mastitis. Also to check the effect of season on clinical mastitis. As a result of this study, the overall incidence of clinical mastitis was 20.95% in buffloaes and 15.38% in cattles. It was concluded from this study that change in the climate and breed at different altitudes are greatly correlated with the prevalence of clinical mastitis [12].

To investigate the prevalence of sub clinical mastitis in buffalo in the Pothohar region of Pakistan, Asghar Khan1, Aneela Zameer Durrani1 and Arfan Yousaf et al, (2018) conducted a study on milk samples collected from 196 lactating buffaloes.Data was collected on the basis of virtual interviews from owners and farmers. The chemical test which is California Mastitis Test were applied on milk samples which revealed that the overall prevalence of sub-clinical mastitis was 67.3%. Chi-square test was conducted in order to determine the association of health and management factors with disease. On applying multivariable logistic regression several factors such as lactation stage, udder shape, teat dipping, manure removal were considered to be the potential risk factors [18].

Chapter 3

Methodology

The purpose of this study is to identify the significant extrinsic features of Mastitis for the early screening of those cows having mastitis disease through various statistical and machine learning approaches. In this study, for the development of predictive models the following steps had been taken.

1. Data preprocessing
2. Data visualization
3. Feature selection
4. Development of models using machine learning methods
5. Assesment analysis of predictive models

Different softwares such as SPSS, Excel and Anaconda had been used for the analysis of data and the development of models which would be discussed in this chapter.

In this study, secondary data is used for the analysis. Secondary data is categorical in nature which is collected through observation and questionnaires survey of cows. Data is in the form of classes and labels. The categorical data is further split into nominal and ordinal in the context of attributes. Some attributes are nominal in nature while others are ordinal in nature. In our study, age is ordinal in nature while all others 27 factors which are listed in 1.1 in the introduction section are nominal in nature.

This project has four sections. Section 1 deals with Data Preprocessing. Section 2 deals with Feature Selection. Section 3 deals with Model Development. Section 4 deals with Assessment Analysis. Detail of these steps are provided below. A complete workflow of the proposed methodology are shown in 3.1.

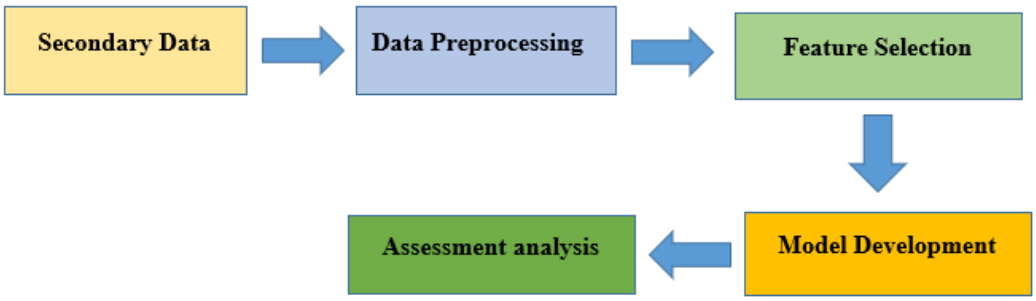

Figure 3.1: Overall workflow Methodology.

3.1 Data Preprocessing

The process of converting raw data into an understandable format is known as data preprocessing. It includes the detection of outliers, estimation of missing values, feature selection etc [19].

In this study data preprocessing includes analyzing completeness of information of each unit i.e. cow with respect to all features and dealing with errors, typos and missing values.

3.1.1 Data Visualization

The process of converting large data sets into charts, graphs, and other graphics is known as data visualization. In this study, Bar charts are used to assess the frequency distribution of categorical data disease vs normal cases.

3.2 Feature Selection

The procedure of acquiring the score for each potential feature and then obtaining the excellent features is known as feature selection.The removal of useless features and the selection of useful features enhances the accuracy and achieving higher performance [20].

Using feature selection, you may better classify and identify the relevance of data contents. In the meanwhile, feature selection has a significant impact on classification results.

In this study feature selection is performed through chi square analysis. The confidence interval taken in order to select the features is 99%. And as a result of this analysis features are selected on the basis of P-value and Chi-square value.

3.2.1 Chi-Square Analysis

Chi-square test is also known as Pearson's chi-square. The standard Chi-square statistics is defined as

$$\chi^2 = \sum\sum (O_i - E_i)^2 / E_i \tag{3.2.1}$$

For feature selection, the most common method used for association of attributes i.e Chi-Square feature statistics (CHI) algorithm is performed [21].Chi-squared is a numerical test that measures deviation from the expected distribution considering the feature event is independent of the class value. This test is used in order to determine the association of two qualitative variables is statistically significant or not [20]. There are five steps to conduct this test

Step 1: Formulation of hypotheses

H_0: There is no association between the extrinsic features and mastitis

H_1: There is an association between extrinsic features and mastitis

Step 2: Specify the expected values for each cell of the table (when the Null Hypothesis is true)

If there is no association between the two variables the expected values in crosstabs specify that what the values of each cell of the table would be.

Step 3: if the data give an evidence against the null hypothesis, compare the observed counts from the sample with the expected counts, assuming H_0 is true.

The observed values are the actual counts computed from the sample.

Step 4: Compute the test statistics

The chi-square statistics compares the observed counts to the expected counts. It is a measure of how far the observed counts are from the expected counts. Where the sum is over all possible values of the categorical variable.

Step 5: Decide if chi-square is statistically significant

The final step of this test is to determine if the value of chi-square is greater enough to reject the null hypothesis.

In this study chi square test is performed in order to identify the association of two qualitative variables is statistically significant or in significant.To highlight the significant and in significant features on the basis of P-value.As a result of this test 21 significant and 7 insignificant features are identified.

3.3 Develoment of Models

In this study, three machine learning models have been applied: Random Forest, Logistic Regression and Decision Tree.

3.3.1 Train-Test Split

Sklearn model selection has a method called train test split that splits data arrays into two subsets: training data and testing data. By the use of this there is no need to divide the dataset manually with this function. Sklearn train test split creates random divisions for the two subsets by default. In our study, the dataset was divided into 80% training and 20% test set shown in figure 3.2

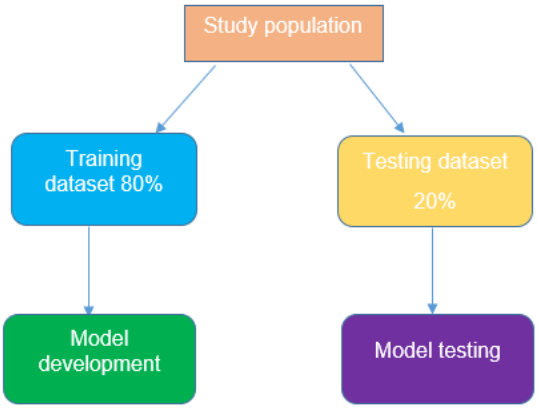

Figure 3.2: Flowchart of dataset splitting for training machine learning models

3.3.2 Logistic Regression

Logistic regression is the commonly used method for binary variables. The model belongs to the generalised linear models family, which explicitly models the relationship between the explanatory and response variables.[22]

3.3.3 Decision Tree

A divide-and-conquer strategy to classification is decision tree analysis [23]. Decision-tree learning is one of the most widely used machine learning algorithm, because it has various attractive features such as no parameters, simplicity, comprehensibility, and being able to handle mixed- type data [19].

In big databases, decision trees can be used to find characteristics and extract patterns that are crucial for discriminating and predictive modelling [23].

3.3.4 Random Forest

Random forest (RF) is a classification and regression method that solves data classification problems using ensemble learning.Decision trees are used in RF to make predictions.During the training phase, a number of decision trees (as defined by the programmer) are built, which are then utilized for class prediction; this is accomplished by taking into account the voted classes of all the individual trees, with the highest vote being considered the output [24].

3.3.5 Confusion Matrix

Confusion matrix is a table which demonstrates the performance of machine learning models. In this table, rows represents the predicted cases by the machine learning models while columns represents the actual cases.

A confusion matrix of size n x n associated with a classifier shows the predicted and actual classification, where n is the number of different classes. Table 3.1 shows a confusion matrix for n = 2, whose entries have the following meanings:

- a is the number of correct negative predictions
- b is the number of incorrect positive predictions

- c is the number of incorrect negative predictions
- d is the number of correct positive predictions[25]

Table 3.1: Two class classification confusion matrix

	Predictive Negative	**Predictive Positive**
Actual Negative	a	b
Actual Positive	c	d

In this study, confusion matrix is used in order to determine the predicted cases of true positive, true negative, false positive and false negative cases identified by model.

3.4 Hyper Parameter Tunning

For a bias-free assessment of a models predictive power, it is necessary to determine the best (hyperparameter) settings for each model. For achieving the better optimal performance of machine learning models there is a need to fine tunned the hyperparameters. Parametric models often do not require tunned parameters for producing optimal results. On the other hand some parametric methods increase their performance through hyperparameters tunning [26]. In this study, parameter tunning is used in order to improve the accuracy of the model.

3.5 Assessment Analysis

In this study, the performance of machine learning methods was evaluated with the following terms.
True Positive (TP) as mastitis cases that are correctly predicted as mastitis.
False Positive (FP) as non-mastitis cases that are incorrectly classified as mastitis.
True Negative (TN) as non-mastitis cases that are correctly predicted as mastitis.
False Negative (FN) as mastitis cases that are incorrectly identified as non-mastitis [27].
Alongwith Receiver Operating Characteristic (ROC) based on sensitivity was also estimated [28].

3.5.1 Stratified 10-fold Cross-Validation

The dataset is divided into 10 parts in 10-fold cross validation; 9 of the 10 parts are used to train the classifier, and the information gained from the training phase is used to validate (or test) the 10th part; this is repeated 10 times, so that each part has been used as both training and testing data at the end of the training and testing phase. This procedure (cross validation) assures that the training data and the test data are distinct. This method is well-known in machine learning for providing a very accurate estimate of a classifiers generalisation error [29]. Figure 3.3 shows the overall flowchart of evaluating the performance.

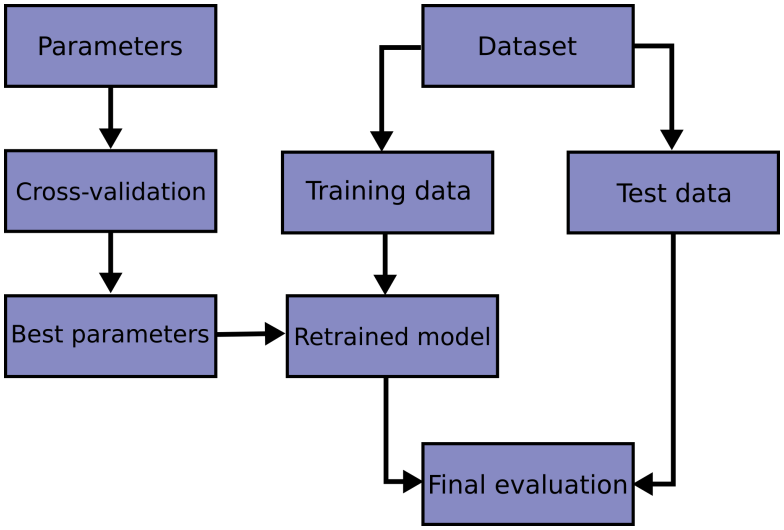

Figure 3.3: Flowchart of evaluating the performance.

Chapter 4

Results

This research project aimed to screen mastitic and non-mastitic cows using machine learning predictive models. For this purpose, the extrinsic features have been used. This section presents the results obtained by performing the proposed approach explained in the previous section. Data preprocessing and feature selection results are explained in this chapter. Different methods of machine learning were used for the development of predictive models for the screening of suspected cows of mastitis. The three machine learning methods used were Logistic regression, Decision trees & Random forest. A comparison of models to determine the best model was also conducted.

4.1 Data Description:

The universe of the research population included Secondary Data taken from Anti Bacter Research Group of ASAB (Atta-Ur-Rahman School of Applied Biosciences). The data collected from the District Rawalpindi (Punjab province), by visiting 40 dairy farms. A total of 432 cows were investigated. The sampling units were cows. The study looked at two categories of determinants: host-related and management-related.

Host-associated determinants included: dairy species (cow), breed, age, stage of lactation, position of udder, udder hygiene score, teat end lesions, condition of udder etc. Similarly, managerial determinants included: condition of floor, bedding material, manure removal, management system, udder washing, number of attendees, milking method, and milking routine etc. All data was gathered using structured questionnaires and a physical inspection of the udder on a pre-designed proforma.

Mastitis was diagnosed based on overt symptoms (clinical mastitis) and the results of the Surf Field Mastitis Test for subclinical mastitis. And the cases and controls are shown in the 4.1.

Table 4.1: Clinical mastitis plus Surf Field Mastitis Test based prevalence of mastitis in cows in the area of Rawalpindi, (Pakistan).

Species	No.of animals examined	Cases	Controls
Cows	432	80	352

4.2 Data Preprocessing

Data preprocessing includes the following two major steps:

1. Analysing completeness of the information with respect to each subject i.e, cow.
2. Dealing typos, errors and missing values.

4.3 Feature Selection for Model Development

The first and the foremost step for the model development is to selection of important features. For this purpose association of attributes is performed through Chi-square analysis.Chi-square test is performed at the 1% confidence interval. Features are selected on the basis of their association of attributes.Features whose P-value is less than 0.01 are considered to be significant while those features whose P-value is greater than 0.01 are considered to be insignificant and then on the basis of significant and insignificant features different subsets of features are selected for model development.

4.3.1 Association of Attributes

Association between the dependent variable i.e disease and the 28 independent features i.e extrinsic features are determined by calculating their chi-square value and P-value. The significance of the available 28 independent features are determined on the basis of P-value.

4.3.2 Significance of Features:

Following are the few steps of association of attributes are given below:

Hypothesis

H_0: There is no association between the extrinsic features and mastitis. (We do not reject our null hypothesis if P>0.01)
H_1: There is an association between the extrinsic features and mastitis. (We reject null hypothesis if P<0.01)

These hypotheses are checked for all the given 28 independent features in the data.
As a result of chi-square analysis twenty-one significant features and seven insignificant features are reported on the basis of P-value.Table 4.30 shows the list of significant and insignificant features.Those features whose P-value is greater than 0.01 are considered to be insignificant. While the features whose P-value is less than 0.01 are considered to be significant.Table 4.2 to Table 4.28 showed the cross tabulation of all the 28 independent features with dependent variable i.e diseasea after performing Chi square test.Cross tabulation table shows the association between the two variables by calculating the observed and the expected counts.If there is a difference between the observed and the expected counts its mean that the association become significant while if there is no or a little difference exist between the two variables this shows that the relationship becomes insignificant.

Table 4.2: Cross tabulation of Mistitis * Age

			Age		Total
			0	1	
Mastitis	0	count	175	177	352
		Expected count	171.1	180.9	352.0
	1	count	35	45	80
		Expected count	38.9	41.1	80.0
Total		count	210	222	432
		Expected Count	210.0	222.0	432.0

Table 4.3: Cross tabulation of Mistitis * Location of Farm

			Location of Farm		Total
			0	1	
Mastitis	0	count	306	46	352
		Expected count	305.6	46.4	352.0
	1	count	69	11	80
		Expected count	69.4	10.6	80.0
Total		count	375	57	432
		Expected Count	375.0	57.0	432.0

Table 4.4: Cross tabulation of Mistitis * Herd Size

			Herd Size		Total
			0	1	
Mastitis	0	count	64	288	352
		Expected count	59.5	292.5	352.0
	1	count	9	71	80
		Expected count	13.5	66.5	80.0
Total		count	73	359	432
		Expected Count	73.0	359.0	432.0

Table 4.5: Cross tabulation of Mistitis * Herd Type

			Herd Type		Total
			1	0	
Mastitis	0	count	221	131	352
		Expected count	240.4	111.6	352.0
	1	count	74	6	80
		Expected count	54.6	25.4	80.0
Total		count	295	137	432
		Expected Count	295.0	137.0	432.0

Table 4.6: Cross tabulation of Mistitis * Breed

			Breed		Total
			1	0	
Mastitis	0	count	191	161	352
		Expected count	199.6	152.4	352.0
	1	count	54	26	80
		Expected count	45.4	34.6	80.0
Total		count	245	187	432
		Expected Count	245.0	187.0	432.0

Table 4.7: Cross tabulation of Mistitis * Management System

			Management System		Total
			1	0	
Mastitis	0	count	246	106	352
		Expected count	254.2	97.8	352.0
	1	count	66	14	80
		Expected count	57.8	22.2	80.0
Total		count	312	120	432
		Expected Count	312.0	120.0	432.0

Table 4.8: Cross tabulation of Mistitis * Milking Method

			Milking Method		Total
			1	0	
Mastitis	0	count	78	274	352
		Expected count	88.8	263.2	352.0
	1	count	31	49	80
		Expected count	20.2	59.8	80.0
Total		count	109	323	432
		Expected Count	109.0	323.0	432.0

Table 4.9: Cross tabulation of Mistitis * Washing of Udder

			Washing of Udder		Total
			1	0	
Mastitis	0	count	170	182	352
		Expected count	197.2	154.8	352.0
	1	count	72	8	80
		Expected count	44.8	35.2	80.0
Total		count	242	190	432
		Expected Count	242.0	190.0	432.0

Table 4.10: Cross tabulation of Mistitis * No. of attendees

			No. of attendees		Total
			1	0	
Mastitis	0	count	232	120	352
		Expected count	242.0	110.0	352.0
	1	count	65	15	80
		Expected count	55.0	25.0	80.0
Total		count	297	135	432
		Expected Count	297.0	135.0	432.0

Table 4.11: Cross tabulation of Mistitis * Manure Removal

			Manure Removal		Total
			0	1	
Mastitis	0	count	163	189	352
		Expected count	149.1	202.9	352.0
	1	count	20	60	80
		Expected count	33.9	46.1	80.0
Total		count	183	249	432
		Expected Count	183.0	249.0	432.0

Table 4.12: Cross tabulation of Mistitis * Feed Sharing

			Feed Sharing		Total
			0	1	
Mastitis	0	count	115	237	352
		Expected count	105.1	246.9	352.0
	1	count	14	66	80
		Expected count	23.9	56.1	80.0
Total		count	129	303	432
		Expected Count	129.0	303.0	432.0

Table 4.13: Cross tabulation of Mistitis * Use of Towel

			Use of Towel		Total
			0	1	
Mastitis	0	count	212	140	352
		Expected count	227.3	124.7	352.0
	1	count	67	13	80
		Expected count	51.7	28.3	80.0
Total		count	279	153	432
		Expected Count	279.0	153.0	432.0

Table 4.14: Cross tabulation of Mistitis * History of Mastitis

			History of Mastitis		Total
			0	1	
Mastitis	0	count	266	86	352
		Expected count	254.2	97.8	352.0
	1	count	46	34	80
		Expected count	57.8	22.2	80.0
Total		count	312	120	432
		Expected Count	312.0	120.0	432.0

Table 4.15: Cross tabulation of Mistitis * Milking Mastitis cow last

			Milking Mastitis cow last		Total
			0	1	
Mastitis	0	count	189	163	352
		Expected count	213.5	138.5	352.0
	1	count	73	7	80
		Expected count	48.5	31.5	80.0
Total		count	262	170	432
		Expected Count	262.0	170.0	432.0

Table 4.16: Cross tabulation of Mistitis * Use of Hormones

			Use of Hormones		Total
			0	1	
Mastitis	0	count	184	168	352
		Expected count	155.6	196.4	352.0
	1	count	7	73	80
		Expected count	35.4	44.6	80.0
Total		count	191	241	432
		Expected Count	191.0	241.0	432.0

Table 4.17: Cross tabulation of Mistitis * Standing Position after Milking

			Standing Position after Milking		Total
			0	1	
Mastitis	0	count	170	182	352
		Expected count	197.2	154.8	352.0
	1	count	72	8	80
		Expected count	44.8	35.2	80.0
Total		count	242	190	432
		Expected Count	242.0	190.0	432.0

Table 4.18: Cross tabulation of Mistitis * Pre/post teat dipping

			Pre/post teat dipping		Total
			0	1	
Mastitis	0	count	189	163	352
		Expected count	213.5	138.5	352.0
	1	count	73	7	80
		Expected count	48.5	31.5	80.0
Total		count	262	170	432
		Expected Count	262.0	170.0	432.0

Table 4.19: Cross tabulation of Mistitis * Teat end Lesions

			Teat end Lesions		Total
			0	1	
Mastitis	0	count	325	27	352
		Expected count	301.5	50.5	352.0
	1	count	45	35	80
		Expected count	68.5	11.5	80.0
Total		count	370	62	432
		Expected Count	370.0	62.0	432.0

Table 4.20: Cross tabulation of Mistitis * Presence of Ticks

			Presence of Ticks		Total
			0	1	
Mastitis	0	count	328	24	352
		Expected count	315.3	36.7	352.0
	1	count	59	21	80
		Expected count	71.7	8.3	80.0
Total		count	387	45	432
		Expected Count	387.0	45.0	432.0

Table 4.21: Cross tabulation of Mistitis * Udder position

			Udder Position		Total
			0	1	
Mastitis	0	count	304	48	352
		Expected count	290.1	61.9	352.0
	1	count	52	28	80
		Expected count	65.9	14.1	80.0
Total		count	356	76	432
		Expected Count	356.0	76.0	432.0

Table 4.22: Cross tabulation of Mistitis * Drying of Udder

			Drying of Udder		Total
			0	1	
Mastitis	0	count	217	135	352
		Expected count	229.8	122.2	352.0
	1	count	65	15	80
		Expected count	52.2	27.8	80.0
Total		count	282	150	432
		Expected Count	282.0	150.0	432.0

Table 4.23: Cross tabulation of Mistitis * Milking Routine

			Milking Routine		Total
			0	1	
Mastitis	0	count	174	178	352
		Expected count	164.6	187.4	352.0
	1	count	28	52	80
		Expected count	37.4	42.6	80.0
Total		count	202	230	432
		Expected Count	202.0	230.0	432.0

Table 4.24: Cross tabulation of Mistitis * Housing

			Housing		Total
			0	1	
Mastitis	0	count	307	45	352
		Expected count	311.3	40.7	352.0
	1	count	75	5	80
		Expected count	70.7	9.3	80.0
Total		count	382	50	432
		Expected Count	382.0	50.0	432.0

Table 4.25: Cross tabulation of Mistitis * Bedding Material

			Bedding Material		Total
			0	1	
Mastitis	0	count	223	129	352
		Expected count	242.8	109.2	352.0
	1	count	75	5	80
		Expected count	55.2	24.8	80.0
Total		count	298	134	432
		Expected Count	298.0	134.0	432.0

Table 4.26: Cross tabulation of Mastitis * Lactation Stage

			Lactation Stage			Total
			0	2	1	
Mastitis	0	count	107	99	146	352
		Expected count	118.1	100.2	133.6	352.0
	1	count	38	24	18	80
		Expected count	26.9	22.8	30.4	80.0
Total		count	145	123	164	432
		Expected count	145.0	123.0	164.0	432.0

Table 4.27: Cross tabulation of Mastitis * Floor Type

			Floor Type			Total
			1	2	0	
Mastitis	0	count	147	131	74	352
		Expected count	133.6	134.4	83.9	352.0
	1	count	17	34	29	80
		Expected count	30.4	30.6	19.1	80.0
Total		count	164	165	103	432
		Expected count	164.0	165.0	103.0	432.0

Table 4.28: Cross tabulation of Mastitis * Udder Condition

			Udder Condition			Total
			2	0	1	
Mastitis	0	count	7	325	20	352
		Expected count	13.0	295.8	43.2	352.0
	1	count	9	38	33	80
		Expected count	3.0	67.2	9.8	80.0
Total		count	16	363	53	432
		Expected count	16.0	363.0	53.0	432.0

Table 4.29: Cross tabulation of Mastitis * Udder Hygeine Score

			Udder Hygeine Score			Total
			1	0	2	
Mastitis	0	count	148	139	65	352
		Expected count	145.0	120.6	86.4	352.0
	1	count	30	9	41	80
		Expected count	33.0	27.4	19.6	80.0
Total		count	178	148	106	432
		Expected count	178.0	148.0	106.0	432.0

Table 4.30: Estimated values of Chi-Square test alongwith P-value and decision to analyze an association between 28 independent features and a binary dependent feature

S.No.	Independent features	Chi-Square value	P-value	Decision
1	Manure Removal	12.12	0.000	reject H_0
2	Towel	15.77	0.000	reject H_0
3	Udder Position	20.52	0.000	reject H_0
4	Presence of Ticks	26.38	0.000	reject H_0
5	Herd Type	26.58	0.000	reject H_0
6	Bedding Material	28.15	0.000	reject H_0
7	Mastitis Cow Last	38.53	0.000	reject H_0
8	Pre /Post dipping	38.53	0.000	reject H_0
9	Udder Hygeine Score	44.05	0.000	reject H_0
10	Washing of Udder	46.02	0.000	reject H_0
11	Standing Position after Milking	46.02	0.000	reject H_0
12	Use of Hormones	50.06	0.000	reject H_0
13	Lesions	69.03	0.000	reject H_0
14	Udder Condition	97.90	0.000	reject H_0
15	History of Mastitis	10.60	0.001	reject H_0
16	Drying of Udder	11.05	0.001	reject H_0
17	Floor Type	14.04	0.001	reject H_0
18	Milking Method	9.51	0.002	reject H_0
19	Lactation Stage	11.95	0.003	reject H_0
20	Feed Sharing	7.16	0.007	reject H_0
21	No.of Attendees	7.14	0.008	reject H_0
22	Milking Routine	5.45	0.02	donot reject H_0
23	Management System	5.17	0.023	donot reject H_0
24	Breed	4.65	0.031	donot reject H_0
25	Housing	2.72	0.099	donot reject H_0
26	Herd Size	2.23	0.135	donot reject H_0
27	Age	0.93	0.335	donot reject H_0
28	Location of Farm	0.03	0.871	donot reject H_0

4.4 Development of ML Models

For the development of predictive model for binary target feature (mastitis non-mastitis), three ML models such as Logistic Regression, Decision Tree and Random Forest have been employed. Four different subsets of features basis on the basis of P-value and chi-square value are considered for model development.The dataset was artbirary randomly divided into 80% training and 20% test set for machine learning model development.The subsets of features selected for model development are listed below.

1. Set of all the 28 available extrinsic features
2. Set of 21 significant features whose P-value < 0.01
3. Set of 14 significant features whose P-value $= 0.000$
4. Set of 5 most significant features whose P-value $= 0.000$ and have high Chi-square value as compared to other significant features

4.4.1 Logistic Regression Model

For the development of LR model, without tunning the parameters subset of 21 significant features , set of all features , set of 14 most significant features and the top most 5 significant features on the basis of P-value and chi-square value are used.

4.4.2 Model Evaluation

Table 4.31: Assessment analysis of Logistic Regression model for different subsets of features

S.No	No. of Features	Accuracy	Sensitivity	Specificity	Precision
1	21	90%	46%	98%	75%
2	28	91%	54%	97%	78%
3	14	88%	38%	97%	71%
4	5	86%	46%	93%	54%

4.4.2.1 Set of 21 Significant Features

After 80% train and 20% test split the LR model assessed that out of 74 normal cases 72 are predict correctly (TP), 2 are those which are incorrectly classified as disease (FN).

And from 13 disease, 7 are wrongly classified as normal (FP) while 6 are predicted correctly (TN), as shown in fig 4.1. The overall model accuracy calculated for 21 significant features is 90%, indicating that 90% of the extrinsic factors are correctly predicted by this model. The sensitivity for this model is 46% indicating 46% subjects are correctly identified. Specificity of the given model indicates that 98% extrinsic factors correctly showed no association with mastitis.The precision for this set of features is 75% shown in Table 4.31.

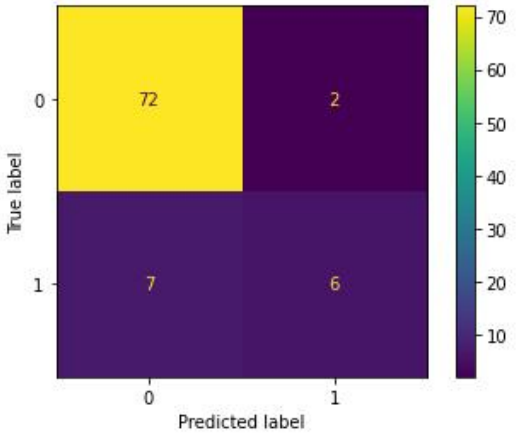

Figure 4.1: Confusion Matrix of LR in classification problem disease vs. normal for 21 significant features.

4.4.2.2 Set of All Features

After 80% train and 20% test split the LR model assessed that out of 74 normal cases 71 are predict correctly (TP), 3 are those which are incorrectly classified as disease (FN).And from 13 disease, 8 are wrongly classified as normal (FP) while 5 are predicted correctly (TN), as shown in fig 4.2.The overall model accuracy calculated for all features is 91%,indicating that 91% of the extrinsic factors are correctly predicted by this model.The sensitivity for this model is 54%.Specificity of the given model indicates that 97 % extrinsic factors correctly showed no association with mastitis.The precision for this set of features is 78% shown in Table 4.31.

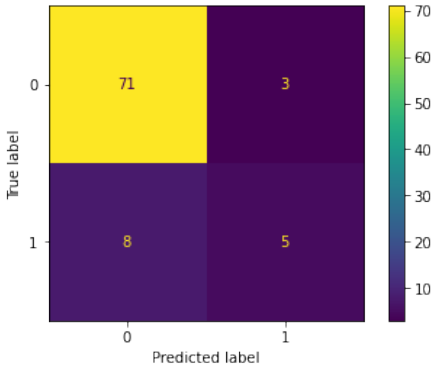

Figure 4.2: Confusion Matrix of LR in classification problem disease vs. normal for all Features

4.4.2.3 Set of 14 most Significant Features

After 80% train and 20% test split the LR model assessed that out of 74 normal cases 72 are predict correctly (TP), 2 are those which are incorrectly classified as disease (FN). And from 13 disease, 8 are wrongly classified as normal (FP) while 5 are predicted correctly (TN), as shown in fig 4.3. The overall model accuracy for the 14 most significant features is 88%, indicating that 88% of the extrinsic factors are correctly predicted by this model. The sensitivity for this model is 38% indicating 38% subjects are correctly identified. Specificity of the given model indicates that 97% extrinsic factors correctly showed no association with mastitis.The precision for this set of features is 71% shown in Table 4.31.

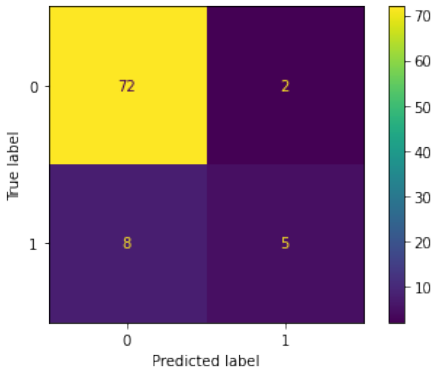

Figure 4.3: Confusion Matrix of LR in classification problem disease vs. normal for 14 significant features

4.4.2.4 Set of 5 most Significant Features

After 80% train and 20% test split the LR model assessed that out of 74 normal cases 69 are predict correctly (TP), 5 are those which are incorrectly classified as disease (FN). And from 13 disease, 7 are wrongly classified as normal (FP) while 6 are predicted correctly (TN), as shown in fig 4.4. The overall model accuracy for the top most 5 significant features is 86%, indicating that 86% of the extrinsic factors are correctly predicted by this model. The sensitivity for this model is 46% indicating 46% subjects are correctly identified. Specificity of the given model indicates that 93% extrinsic factors correctly showed no association with mastitis.The precision for this set of features is 54% shown in Table 4.31.

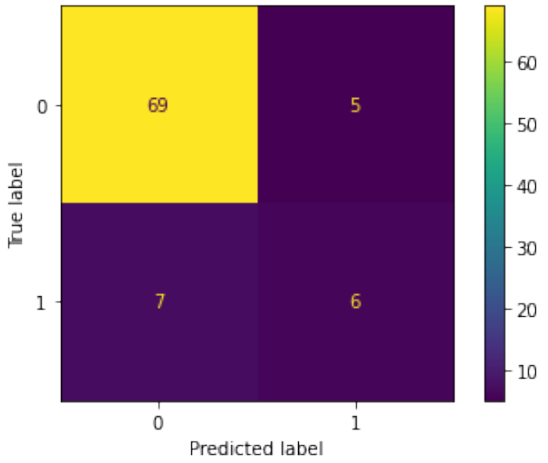

Figure 4.4: Confusion Matrix of LR in classification problem disease vs. normal for 5 Top most significant features

4.5 Decision Tree Model on Different Sets of Features

For the development of DT model, without tunning the parameters subset of 21 significant features , set of all features , set of 14 most significant features and the top most 5 significant features on the basis of P-value and chi-square value are used.

4.5.1 Model Evaluation

Table 4.32: Assessment analysis of DT model for Different Subsets of Features

S.No	No.of Features	Accuracy	Sensitivity	Specificity	Precision
1	21	81.6%	47.3%	91.1%	60%
2	28	79%	47%	88%	53%
3	14	79%	42%	89%	53%
4	5	87%	53%	97%	83%

4.5.1.1 Set of 21 Significant Features

On the diagnol of the matrix is the number of correct classified cases, while other elements of the matrix indicate the number of cases that are incorrectly classified as some of the other classes. Figure 4.5 shows that out of 68 normal cases 62 are predict correctly (TP), 6 are those which are incorrectly classified as disease (FN). And from 19 disease cases, 10 are wrongly classified as normal (FP) while 9 are predicted correctly (TN). Along with the overall accuracy of the DT model for 21 significant features 81.6% indicating 81.6% of the cases are correctly predicted by this model. The sensitivity of this model is 47.3% indicating 47.3% of the mastitis cases are correctly predicted by the DT model without using tunned parameters. The specificity indicates 91.1% of the non-mastitis cases are correctly predicted as non-mastitis by DT. The precision for this model is 60% indicating 60% of the mastitis cases are precisely identified by the DT model as shown in Table 4.32.

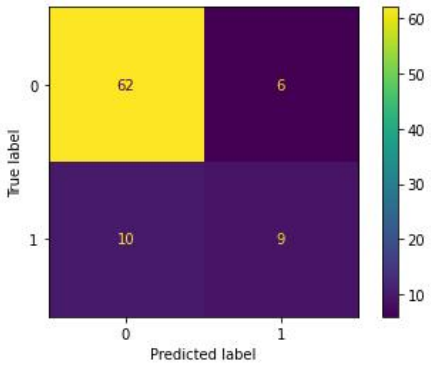

Figure 4.5: Confusion Matrix of DT in classification problem disease vs. normal for 21 features.

4.5.1.2 Set of All Features

After 80% train and 20% test split the DT model assessed that showed in 4.6 out of 68 normal cases 56 are predict correctly (TP), 12 are those which are incorrectly classified as disease (FP). And from 19 disease, 12 are wrongly classified as normal (FN) while 7 are predicted correctly (TN). Along with according to table 4.32, the overall accuracy of the DT model is 79% indicating 79% of the cases are correctly predicted by this model. The sensitivity of this model is 47% indicating 47% of the mastitis cases are correctly predicted by the DT model without using tunned parameters. The specificity indicates 88% of the non-mastitis cases are correctly predicted as non-mastitis by DT. The precision for this model is 53% indicating 53% of the mastitis cases are precisely identified by the DT model shown in Table 4.32.

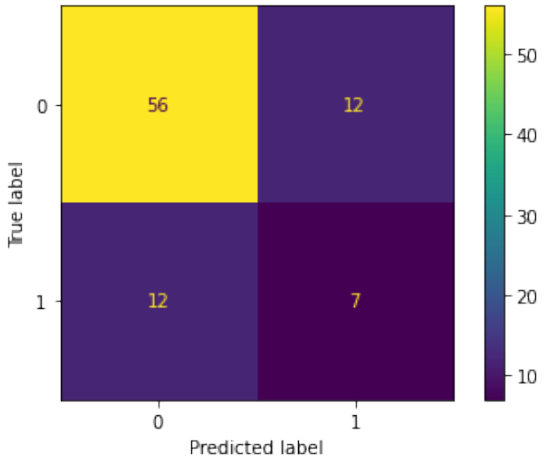

Figure 4.6: Confusion Matrix of DT in classification problem disease vs. normal for all features.

4.5.1.3 Set of 14 most Significant Features

After 80% train and 20% test split the DT model assessed that showed in 4.7 out of 68 normal cases 61 are predict correctly (TP), 7 are those which are incorrectly classified as disease (FP). And from 19 disease, 11 are wrongly classified as normal (FN) while 8 are predicted correctly (TN). Along with according to table 4.32, the overall accuracy of the DT model is 79% indicating 79% of the cases are correctly predicted by this model. The sensitivity of this model is 42% indicating 42% of the mastitis cases are correctly predicted by the DT model without using tunned parameters. The specificity indicates 89% of the non-mastitis cases are correctly predicted as non-mastitis by DT.

The precision for this model is 53% indicating 53% of the mastitis cases are precisely identified by the DT model shown in Table 4.32.

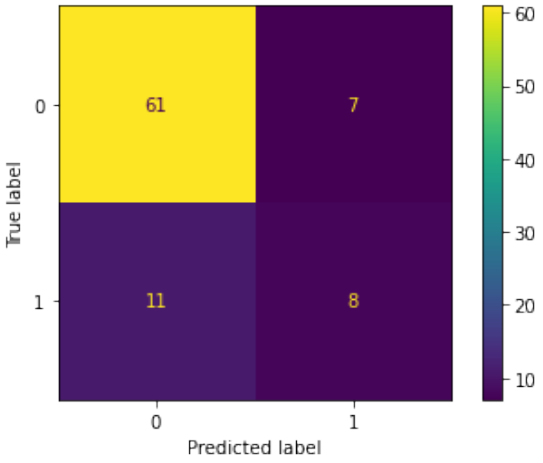

Figure 4.7: Confusion Matrix of DT in classification problem disease vs. normal for 14 significant features.

4.5.1.4 Set of 5 most Significant Features

After 80% train and 20% test split the DT model assessed that showed in 4.8 out of 68 normal cases 66 are predict correctly (TP), 2 are those which are incorrectly classified as disease (FP). And from 19 disease, 9 are wrongly classified as normal (FN) while 10 are predicted correctly (TN). Along with according to table 4.32, the overall accuracy of the DT model is 87% indicating 87% of the cases are correctly predicted by this model. The sensitivity of this model is 53% indicating 53% of the mastitis cases are correctly predicted by the DT model without using tunned parameters. The specificity indicates 97% of the non-mastitis cases are correctly predicted as non-mastitis by DT. The precision for this model is 83% indicating 83% of the mastitis cases are precisely identified by the DT model shown in Table 4.32.

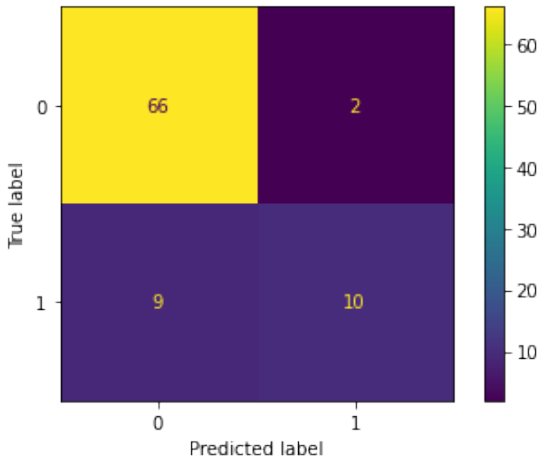

Figure 4.8: Confusion Matrix of DT in classification problem disease vs. normal for 5 top most significant features.

4.6 Random Forest Model for Different Subsets of Features

For the development of RF model, without tunning the parameters subset of 21 significant features , set of all features , set of 14 most significant features and the top most 5 significant features on the basis of P-value and chi-square value are used.

4.6.1 Model Evaluation

Table 4.33: Assessment analysis of RF Model for Different Subsets of Features

S.No	No. of Features	Accuracy	Sensitivity	Specificity	Precision
1	21	87.3%	63.6%	90.7%	50%
2	28	91%	63%	94%	64%
3	14	91%	73 %	93%	62%
4	5	91 %	55%	96%	66 %

4.6.1.1 Set of 21 Significnat Features

After 80% train and 20% test split the RF model assessed that showed in 4.9 out of 76 normal cases 69 are predict correctly (TP), 7 are those which are incorrectly classified as disease (FP). And from 11 disease, 4 are wrongly classified as normal (FN) while 7 are predicted correctly (TN). Along with according to table 4.33, the overall accuracy of the RF model for 21 significant features is 87.3% indicating 87.3% of the cases are correctly predicted by this model. The sensitivity of this model is 63.6% indicating 63.6% of the mastitis cases are correctly predicted by the RF model without using tunned parameters. The specificity indicates 90.7% of the non-mastitis cases are correctly predicted as non-mastitis by RF. The precision for this model is 50% indicating 50% of the mastitis cases are precisely identified by the RF model shown in Table 4.33.

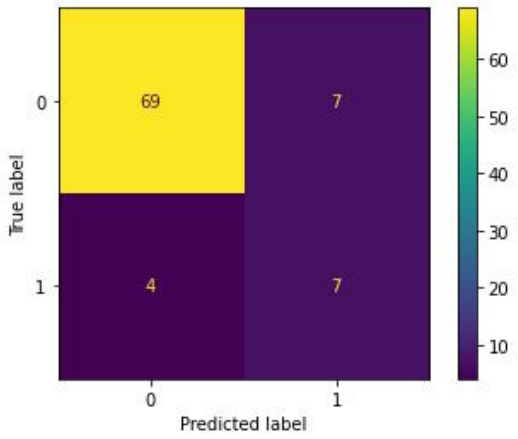

Figure 4.9: Confusion Matrix of RF in classification problem disease vs. normal For 21 Features

4.6.1.2 Set of 14 most Significant Features

After 80% train and 20% test split the RF model assessed that showed in 4.10 out of 76 normal cases 71 are predict correctly (TP), 5 are those which are incorrectly classified as disease (FP). And from 11 disease, 3 are wrongly classified as normal (FN) while 8 are predicted correctly (TN).Along with according to table 4.33, the overall accuracy of the RF model is 91% indicating 91% of the cases are correctly predicted by this model.The sensitivity of this model is 73% indicating 73% of the mastitis cases are correctly predicted by the RF model without using tunned parameters.The specificity indicates 93% of the non-mastitis cases are correctly predicted as non-mastitis by RF.The

precision for this model is 62% indicating 62% of the mastitis cases are precisely identified by the RF model shown in Table 4.33.

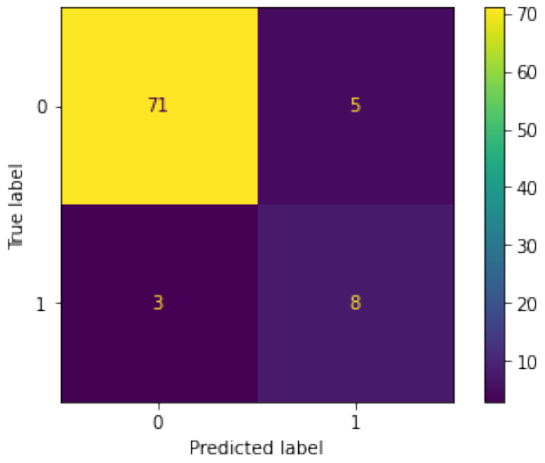

Figure 4.10: Confusion Matrix of RF in classification problem disease vs. normal for 14 significant features.

4.6.1.3 Set of All Features

After 80% train and 20% test split the RF model assessed that showed in 4.11 out of 76 normal cases 73 are predict correctly (TP), 3 are those which are incorrectly classified as disease (FP). And from 11 disease, 6 are wrongly classified as normal (FN) while 5 are predicted correctly (TN). Along with according to table 4.33, the overall accuracy of the RF model for all features is 91% indicating 91% of the cases are correctly predicted by this model. The sensitivity of this model is 63% indicating 63% of the mastitis cases are correctly predicted by the RF model without using tunned parameters. The specificity indicates 94% of the non-mastitis cases are correctly predicted as non-mastitis by RF. The precision for this model is 64% indicating 64% of the mastitis cases are precisely identified by the RF model.

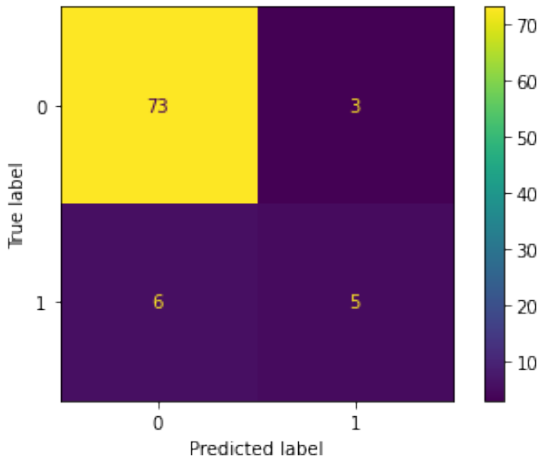

Figure 4.11: Confusion Matrix of RF in classification problem disease vs. normal for all features.

4.6.1.4 Set of 5 most Significant Features

After 80% train and 20% test split the RF model assessed that showed in 4.12 out of 76 normal cases 73 were predict correctly (TP), 3 are those which are incorrectly classified as disease (FP). And from 11 disease, 5 are wrongly classified as normal (FN) while 6 are predicted correctly (TN). Along with according to table 4.33, the overall accuracy of the RF model for top most 5 features is 91% indicating 91% of the cases are correctly predicted by this model. The sensitivity of this model is 55% indicating 55% of the mastitis cases are correctly predicted by the RF model without using tunned parameters. The specificity indicates 96% of the non-mastitis cases are correctly predicted as non-mastitis by RF. The precision for this model is 66% indicating 66% of the mastitis cases are precisely identified by the RF model.

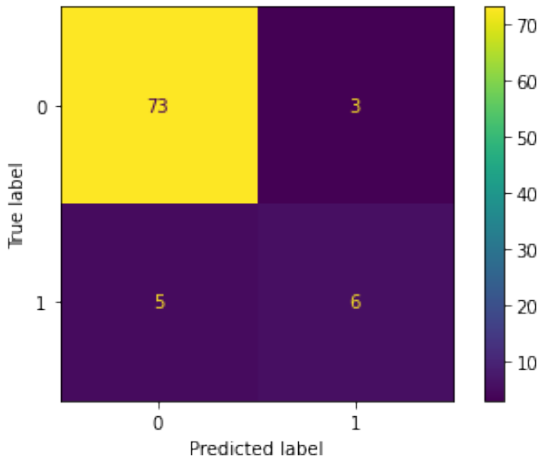

Figure 4.12: Confusion Matrix of RF in classification problem disease vs. normal for top most 5 features.

4.7 Comparative Analysis of the Developed Model:

Comparative analysis of the developed models i.e, LR, DT and RF for different subsets of features have been performed. as shown in table from 4.34 to table 4.37.Features are categorised on the basis of P-value and chi-square value. Comparison of models are performed on the basis of sensitivity measures i.e, how much the model correctly predict the disease cases as diseased. For 21 significant features the best accuracy is obtained by LR model which is 90%. But the best sensitivity is obtained from RF model which is 63.6% shown in table 4.34.

For 14 significant features the best accuracy as well as sensitivity is obtained from RF model which is 91% and 73% shown in table 4.36.

For top most 5 significant features the best accuracy as well as sensitivity is obtained from RF model which is 91% and 55% as compared to others model for this subset of features shown in table 4.37.

For all features the accuracy of LR and Rf model remains same but on the basis of sensitivity RF model gives the better results shown in table 4.35

By comparing all models for different subsets of features four best models on the basis of sensitivity are selected for the selection of final best predictive model shown in table 4.38.Among all the models for different susbets of features the best sensitivity is obtained by RF model for 14 most significant model which is 91% and 73%. The lowest accuracy

observed by DT model for 14 features that is 79% and the lowest sensitivity is obtained by LR model for top most 5 features which is 38%.Hence among all the developed models for different subsets of features RF is selected as the final model.

Table 4.34: Comparison of Developed Models for Set of 21 Features

S.NO	Method	Accuracy	Sensitivity	Specificity	Precision	10-fold cross validation
1	RF	87.3%	63.6%	90.7%	50%	84%
2	DT	81.6%	47.3%	91.1%	60%	80%
3	LR	90%	46%	97%	75%	87%

Table 4.35: Comparison of Developed Models for Set of 28 Features

S.NO	Method	Accuracy	Sensitivity	Specificity	Precision	10-fold cross validation
1	RF	91%	63%	94%	64%	85%
2	DT	79%	47%	88%	53%	81%
3	LR	91%	54%	97%	78%	86%

Table 4.36: Comparison of Developed Models for Set of 14 Features

S.NO	Method	Accuracy	Sensitivity	Specificity	Precision	10-fold cross validation
1	RF	91%	73%	93%	62%	83%
2	DT	79%	42%	89%	53%	81%
3	LR	89%	38%	97%	71%	85%

Table 4.37: Comparison of Developed Models for Set of 5 Features

S.NO	Method	Accuracy	Sensitivity	Specificity	Precision	10-fold cross validation
1	RF	91 %	55%	96%	66 %	85%
2	DT	87 %	53%	97%	83%	85%
3	LR	86%	46%	93%	54%	85 %

Table 4.38: Comparison of All the Selective Best Models

S.NO	Method	Features	Accuracy	Sensitivity	Specificity	Precision	10-fold cross validation
1	RF	21	87.3 %	63.6%	90.7%	50%	84%
2	RF	14	91 %	73%	93%	62%	83%
3	RF	5	91 %	55%	96%	66 %	85%
4	LR	28	91 %	54 %	97 %	78%	86 %

4.8 ROC Curve:

ROC curve is the graphical representation of the true positive and false negative rates predicted by model. ROC curve is the best measurement of sensitivity and specificity for assessing inherent validity of the diagnostic test [28]. ROC curve at 0.5 would considered that there is no discrimination (ability to diagnose the patients with or without disease). On the other hand ROC curve at 0.9 would be considered to be the best showed in fig 4.13. Here ROC curve is used in order to compare the performance of different machine learning models.In this study, ROC curve is used to compare the performance of selective models showed in table 4.38 i.e, RF for all features, RF for 21 features,RF for 14 features and RF for 5 most significant features.As a result of ROC curve we assume that RF model for 14 significant features is considered to be the best model for the screening of mastitis based on the extrinsic factors. The ROC of RF model for 14 significant features lie at 0.83 while ROC of RF for 21 features lie at 0.73 and ROC of RF for all features lie at 0.76 while ROC of RF model for top most 5 significant features lie at 0.75 shown in fig 4.13.

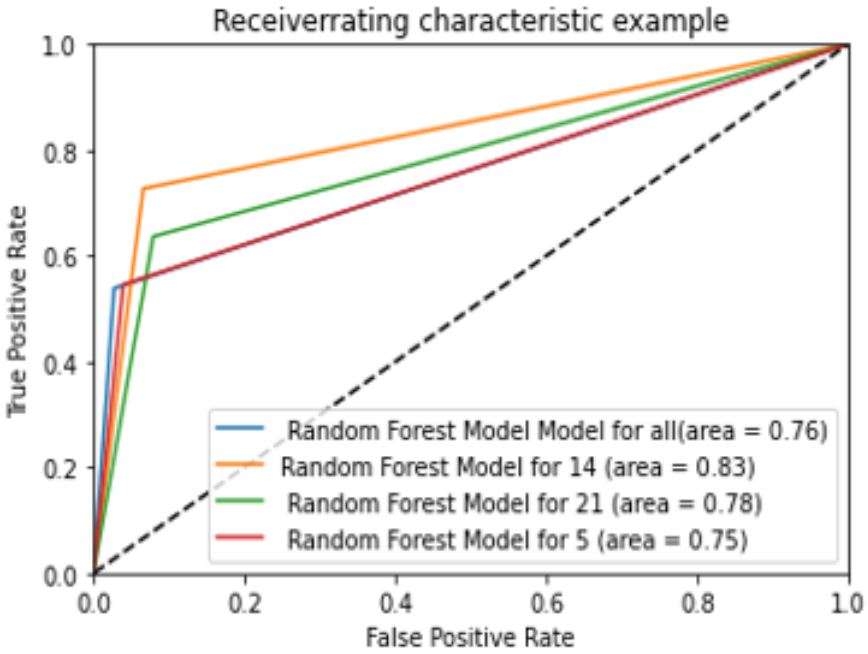

Figure 4.13: ROC curve for all the selective models.

4.9 Hyper Parameter Tunning of RF Regression Model

Hyper parameter tunning of the selective model is performed by using the tunned parameters such as n estimators and criterian = gini in order to improve the accuracy and other assessment measures of the selective model.

4.9.1 Random Forest for 14 most Significant Features

After 80% train and 20% test split and tunned the hyper parameters the RF model assessed that showed in 4.14 out of 76 normal cases 72 were predict correctly (TP), 4 are those which are incorrectly classified as disease (FP). And from 11 disease, 3 are wrongly classified as normal (FN) while 8 are predicted correctly (TN). Along with according to table 4.39, the overall accuracy of the RF model for top most 5 features is 92% indicating 92% of the cases are correctly predicted by this model. The sensitivity of this model is 73% indicating 73% of the mastitis cases are correctly predicted by the RF model without using tunned parameters. The specificity indicates 95% of the non-mastitis cases are correctly predicted as non-mastitis by RF. The precision for this model is 67% indicating 67% of the mastitis cases are precisely identified by the RF model.

Table 4.39: Confusion Matrix of RF in classification problem disease vs. normal for 14 significant features after tunned parameters

Model Name	Features	Accuracy	Sensitivity	Specificity	Precision	10-fold cross validation
RF	14	92 %	73%	95%	67%	83%

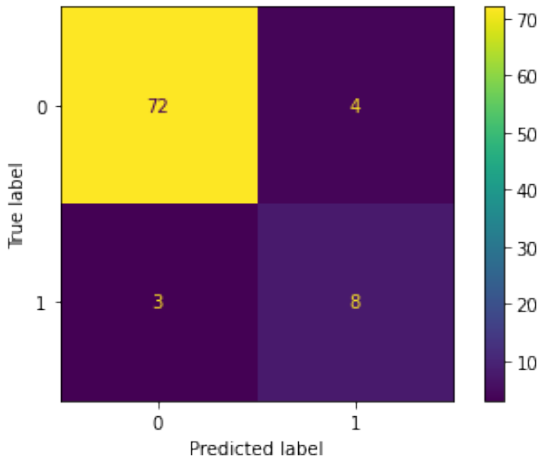

Figure 4.14: Confusion Matrix of RF in classification problem disease vs. normal after tunned parameters.

Chapter 5

Conclusions and Recommendations

Following are the major conclusions of this study are

1. The study investigated the significance of collected 28 features in the development of models with binary target variable using Chi-square association of attributes. Various groups have been tested by introducing certain thresholds with respect the estimated values of chi-square test and corresponding p-values. The combinations include all set of features, then reduced subsets of 21, 14 and 5 features as per their association with the target variable in descending order of magnitude. The identified top 5 features having strong association with target variable are Washing of Udder, Standing position after milking, Use of hormones, Lesions and Udder condition.

2. Three ML methods namely decision tree, logistic regression and random forest regression have been used considering different combinations of features. The assessment analysis reveals that the performance of random forest regression is the best. Moreover, the combination of 14 features as being independent variables are adequate to predict the target class. Accuracy, precision, sensitivity, specificity of the combination of random forest and 14 features is 91%, 73%, 93% and 62%, respectively.

3. For further refinement of the best identified method, i.e. random forest regression, hyper-parameter tunning has been introduced with values of n-estimators as 100, 200 and 500. The performance of model is comparable with respect to the variations in values of n-estimators except a slight increase in assessment measure for n=200.

4. For the validation of developed models, stratified 10-fold cross validation scheme has been used. The assessment measures of this procedure are also in favor of the use of random forest regression with 14 set of features.

Based on the provided details, the study recommends the use of random forest regression with following 14 features to predict category of the target class.

Table 5.1: List of 14 significant features having P-value=0.000

S.No.	Independent features	Chi-Square value	P-value	Interpretation
1	Manure Removal	12.12	0.000	significant
2	Towel	15.77	0.000	significant
3	Udder Position	20.52	0.000	significant
4	Presence of Ticks	26.38	0.000	significant
5	Herd Type	26.58	0.000	significant
6	Bedding Material	28.15	0.000	significant
7	Mastitis Cow Last	38.53	0.000	significant
8	Pre /Post dipping	38.53	0.000	significant
9	Udder Hygeine Score	44.05	0.000	significant
10	Washing of Udder	46.02	0.000	significant
11	Standing Position after Milking	46.02	0.000	significant
12	Use of Hormones	50.06	0.000	significant
13	Lesions	69.03	0.000	significant
14	Udder Condition	97.90	0.000	significant

5.1 Limitations

Certain limitations exist in the current study.

1. There is Multicollinearity Problem exist between attributes.
2. The class imbalance is present in the data.

5.2 Future Recommendation

The future suggestions for this study are:
It could be effcetive and more interesting if we deal with Multicolleanarity problem.Future work concerns deeper analysis of available extrinsic features.Moreover, to try different methods for feature selection other than Chi-square and try different machine learning algorithms for better classification of disease and normal cases.

References

[1] Abdul Rehman, Luan Jingdong, Abbas Ali Chandio, and Imran Hussain. Livestock production and population census in pakistan: Determining their relationship with agricultural gdp using econometric analysis. *Information Processing in Agriculture*, 4(2):168–177, 2017.

[2] Wan-Ting Yang, Chun-Yen Ke, Wen-Tien Wu, Ru-Ping Lee, and Yi-Hsiung Tseng. Effective treatment of bovine mastitis with intramammary infusion of angelica dahurica and rheum officinale extracts. *Evidence-Based Complementary and Alternative Medicine*, 2019, 2019.

[3] D Cavero, K-H Tölle, C Henze, C Buxadé, and J Krieter. Mastitis detection in dairy cows by application of neural networks. *Livestock Science*, 114(2-3):280–286, 2008.

[4] Maros Cobirka, Vladimir Tancin, and Petr Slama. Epidemiology and classification of mastitis. *Animals*, 10(12):2212, 2020.

[5] Abderrazek Hocine, Riad Bouzid, Hamida Talhi, and Djamel Khelef. An epidemiological study of bovine mastitis and associated risk factors in and around eltarf district, northeast algeria. *Veterinarska stanica*, 52(5):0–0, 2021.

[6] Suvi Taponen, Eero Liski, A-M Heikkilä, and S Pyörälä. Factors associated with intramammary infection in dairy cows caused by coagulase-negative staphylococci, staphylococcus aureus, streptococcus uberis, streptococcus dysgalactiae, corynebacterium bovis, or escherichia coli. *Journal of dairy science*, 100(1):493–503, 2017.

[7] LK Fox. Prevalence, incidence and risk factors of heifer mastitis. *Veterinary microbiology*, 134(1-2):82–88, 2009.

[8] GÜVEN KAŞIKÇI, Ömer ÇEtİn, ENVER BARIŞ BİNGÖL, and Mehmet Can GÜndÜz. Relations between electrical conductivity, somatic cell count, california mastitis test and some quality parameters in the diagnosis of subclinical mastitis in dairy cows. *Turkish Journal of Veterinary and Animal Sciences*, 36(1):49–55, 2012.

[9] Ghulam Muhammad, Abeera Naureen, Muhammad Nadeem Asi, Muhammad Saqib, et al. Evaluation of a 3% surf solution (surf field mastitis test) for the diagnosis of subclinical bovine and bubaline mastitis. *Tropical animal health and production*, 42(3):457–464, 2010.

[10] MQ Bilal, MU Iqbal, G Muhammad, M Avais, and MS Sajid. Factors affecting the prevalence of clinical mastitis in buffaloes around faisalabad district (pakistan). *Int. J. Agri. Biol*, 6(1), 2004.

[11] Raveendra Hegde, Shrikrishna Isloor, K Nithin Prabhu, BR Shome, D Rathnamma, VVS Suryanarayana, S Yatiraj, C Renuka Prasad, N Krishnaveni, S Sundareshan, et al. Incidence of subclinical mastitis and prevalence of major mastitis pathogens in organized farms and unorganized sectors. *Indian journal of microbiology*, 53(3): 315–320, 2013.

[12] Amjad Khan, Muhammad Hassan Mushtaq, Mansur Ud Din Ahmad, Mamoona Chaudhry, and Abdul Wali Khan. Prevalence of clinical mastitis in bovines in different climatic conditions in kpk,(pakistan). *Science International*, 27(3), 2015.

[13] Melak Tezera and Endris Aman Ali. Prevalence and associated risk factors of bovine mastitis in dairy cows in and around assosa town, benishangul-gumuz regional state, western ethiopia. *Veterinary Medicine and Science*, 2021.

[14] Nazira Mammadova and Ismail Keskin. Application of the support vector machine to predict subclinical mastitis in dairy cattle. *The Scientific World Journal*, 2013, 2013.

[15] ED Karimuribo, JL Fitzpatrick, ES Swai, Catriona Bell, MJ Bryant, NH Ogden, DM Kambarage, and NP French. Prevalence of subclinical mastitis and associated risk factors in smallholder dairy cows in tanzania. *Veterinary record*, 163(1):16–21, 2008.

[16] Demelash Biffa, Etana Debela, and Fekadu Beyene. Prevalence and risk factors of mastitis in lactating dairy cows in southern ethiopia. *Int. J. Appl. Res. Vet. Med*, 3(3):189–198, 2005.

[17] Karin Östensson, Vo Lam, Natahlie Sjögren, and Ewa Wredle. Prevalence of subclinical mastitis and isolated udder pathogens in dairy cows in southern vietnam. *Tropical animal health and production*, 45(4):979–986, 2013.

[18] Asghar Khan, Aneela Zameer Durrani, Arfan Yousaf, Jawaria Ali Khan, Mamoona Chaudhry, Mumtaz Ali Khan, Amjad Khan, et al. Epidemiology of bovine subclinical mastitis in pothohar region, punjab, pakistan in 2018. *Pakistan Journal of Zoology*, 51(5), 2019.

[19] BM Patil, RC Joshi, and Durga Toshniwal. Association rule for classification of type-2 diabetic patients. In *2010 second international conference on machine learning and computing*, pages 330–334. IEEE, 2010.

[20] Ikram Sumaiya Thaseen and Cherukuri Aswani Kumar. Intrusion detection model using fusion of chi-square feature selection and multi class svm. *Journal of King Saud University-Computer and Information Sciences*, 29(4):462–472, 2017.

[21] Yujia Zhai, Wei Song, Xianjun Liu, Lizhen Liu, and Xinlei Zhao. A chi-square statistics based feature selection method in text classification. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 160–163. IEEE, 2018.

[22] Joshua J Levy and A James O'Malley. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. *BMC medical research methodology*, 20(1):1–15, 2020.

[23] Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6):275–285, 2004.

[24] Zhan Zhang, Hua Han, Xiaoyu Cui, and Yuqiang Fan. Novel application of multimodel ensemble learning for fault diagnosis in refrigeration systems. *Applied Thermal Engineering*, 164:114516, 2020.

[25] Sofia Visa, Brian Ramsay, Anca L Ralescu, and Esther Van Der Knaap. Confusion matrix-based feature selection. *MAICS*, 710:120–127, 2011.

[26] Patrick Schratz, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, and Alexander Brenning. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406:109–120, 2019.

[27] Zeynep TUNÇ KUCUKAKCALİ, İpek BALIKÇI ÇİÇEK, Emek GÜLDOĞAN, and Cemil ÇOLAK. Assessment of associative classification approach for predicting mortality by heart failure. *The Journal of Cognitive Systems*, 5(2):41–45, 2020.

[28] Dieu Tien Bui, Paraskevas Tsangaratos, Viet-Tien Nguyen, Ngo Van Liem, and Phan Trong Trinh. Comparing the prediction performance of a deep learning neural network model with conventional machine learning models in landslide susceptibility assessment. *Catena*, 188:104426, 2020.

[29] Xinchuan Zeng and Tony R Martinez. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1):1–12, 2000.