APPLICATION OF MACHINE LEARNING TECHNIQUES ON THE CONTEXT OF LIVESTOCK ANALYSIS

Md. Shajedul Islam **14101058**

Syed Tahmid Masud **14301004**

Bhuiyan Mustafa Tawheed **14101244**

Supervisor: Mr. Hossain Arif

Department of Computer Science and Engineering
BRAC University

Date of Submission: 16/08/2018

DECLARATION

Thesis Submission to the Department of Computer Science and Engineering, BRAC University, Dhaka, submitted by the authors for the purpose of obtaining the degree of Bachelor of Science in Computer Science and Engineering. We hereby announce that the results of this thesis are entirely based on our research. Resources taken from any research conducted by other researchers are mentioned through reference. This thesis either in whole or in part, was not previously submitted for any degree.

Signature of Supervisor:

_____

Mr. Hossain Arif
Assistant Professor
Department of Computer Science and Engineering
BRAC University

Signature of Authors:

1. _____
MD. Shajedul Islam

2. _____
Syed Tahmid Masud

3. _____
Bhuiyan Mustafa Tawheed

ABSTRACT

Livestock industry has been one of the fastest growing sectors of business and research in Bangladesh over the past decade. This rising, lucrative and profitable industry seems to be attracting a good number of enthusiastic investors to invest their capital, and make a contribution to the country's overall GDP, and recover the deficit in meat production. However, due to lack of suitable resources, reliable data and information, proper knowledge and precise guidance, these investors are lagging behind in generating their expected outcomes. Moreover, these investors and farmers tend to make choices based on their experience only. To come to their assistance in making compatible decisions and provide with a profitable and efficient approach in the investor's business expansion in Bangladesh, this research aims to establish an intelligent prediction methodology through regression analysis, by implementing data mining and supervised machine learning techniques. This research provides some cattle's breed based analysis depending on different related factors, which includes age, current weight of cattle, the environment it is reared on, the diet plan and the geographical region it originated from. The models implemented in this research were, Linear Regression model, Ordinary Least Square Regression model, Polynomial Regression model and Decision Tree learning for attaining this prediction mechanism. For executing the analysis, an unanalyzed data set, having a period of 12 years, were collected from Bangladesh Livestock Research Institute, Savar Dairy Farm, and Meghdubi Agro farm.

Index Terms:

Livestock Analysis; Machine Learning; Regression Models; Decision Trees.

ACKNOWLEDGMENT

Firstly, we would like to thank Almighty Allah for enabling us to put our best efforts into the research and successfully completing it. We would like to express our utmost gratitude and appreciation to our supervisor Mr. Hossain Arif for his attention and time. We would also like to thank him for giving us the opportunity to work on this topic and assisting us throughout the process. In addition to that we would like to give a special thanks to Mr. Kallyan Banik for his support as our research assistant.

Table of Contents

List of Equations

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

k_scoretest_linear – K fold cross validation Linear Score (test)

k_scoretest_linear mean – K fold cross validation Linear Mean Score (test)

k_scoretest_poly mean - K fold cross validation Polynomial Mean Score (test)

k_scoretest_rbf mean - K fold cross validation Radial Basis Function Mean Score (test)

k_scoretest_sigmoid mean - K fold cross validation Sigmoid Mean Score (test)

k_scoretrain_linear mean – K fold cross validation Linear Mean Score (train)

k_scoretrain_poly mean - K fold cross validation Polynomial Mean Score (train)

k_scoretrain_rbf mean - K fold cross validation Radial Basis Function Mean Score (train)

k_scoretrain_sigmoid mean - K fold cross validation Sigmoid Mean Score (train)

k_scoretest_poly - K fold cross validation Polynomial Score (test)

k_scoretest_rbf - K fold cross validation Radial Basis Function Score (test)

k_scoretest_sigmoid - K fold cross validation Sigmoid Score (test)

k_scoretrain_linear – K fold cross validation Linear Score (train)

k_scoretrain_poly - K fold cross validation Polynomial Score (train)

k_scoretrain_rbf - K fold cross validation Radial Basis Function Score (train)

k_scoretrain_sigmoid - K fold cross validation Sigmoid Score (train)

OLS – Ordinary Least Square

SVM – Support Vector Machine

svm_lin –Support Vector Regression Linear Score

svm_poly – Support Vector Regression Polynomial Score

svm_rbf – Support Vector Regression Radial Basis Function Score

svm_sigmoid – Support Vector Regression Sigmoid Score

SVR – Support Vector Regression

Chapter 1
OVERVIEW

1.1 INTRODUCTION

Bangladesh is one of the largest livestock producers in the Southeast Asia. Livestock population in Bangladesh is currently estimated to be 25.7 million cows, 0.83 million buffaloes, 14.8 million goats, 1.9 million sheep, 118.7 million chicken and 34.1 million ducks [1]. In our country, livestock generally includes cows, goats, buffaloes, sheep and poultry. Livestock plays a major role in contributing in the wealth of our country. Leather, Meat and Milk is a big source of revenue for our country, which comes from the livestock industry. It plays an important role in the agriculture production sphere. Statistics shows that around 2.9% of the country's annual GDP is generated through the livestock, and 20% of the families earn their livelihood through rearing livestock [1]. Due to the economic condition of Bangladesh, if given proper guidance, investing in Livestock industry can prove to be quite profitable for the farmers. The first world countries have been successful in bringing a huge advancement in their economic condition through automated agriculture. However, Bangladesh is lagging behind in advancing in increasing the overall GDP significantly through this industry. The prime motive of this research is to automate the animal agriculture sector through the implementation of technological tools and methodologies [4]. Through this research, both the rural area farmers and Agro farm investors will be profitable, as it provides a detailed learning and prediction mechanism of the livestock rearing industry. Furthermore, the new investors, who previously used to take decisions based on their estimations, will gain the most benefit from our research. Farmers with minimum knowledge about a cattle and its growth affecting factors, will be able to learn and get inspiration to take initiate investing in this sector. This research provides a regressive trend analysis of cattle divided by their original breeds. On providing some significantly related factors, the investor will be able to make the choice of the most profitable breed of cattle for rising in their

preferred geographical region. As a result, this technological expansion in livestock sector will contribute in minimizing the protein deficiency of our country, and come to economic assistance. Therefore, if monitored properly, this research can encourage thousands of families to take up the profession of livestock rearing, to bring solvency in their lives. If the importance and profitability of investing in this sector, can be properly pictured, it will in return generate huge revenue for the nation and contribute even more prominently in our GDP.

1.2 MOTIVATION

Bangladesh being a third world country is still on the demographic stage of making its mark in the technological advancement in agriculture. On the other hand, the other business sectors have already laid their hands on automating their system. However, the animal agriculture sector is yet to come to exposure [4]. Automating the agriculture system, is not a trend anymore, in fact is a necessity, in a country like Bangladesh, where over 20% of the families depend on livestock rearing [1]. A good number of young investors plan to start investing in this sector, but eventually back out thinking it needs a lot of practical experience. Thus, this generation is not yet playing a major role in contributing in the Livestock industry.

Another point being is the restriction of import of cattle from India. Over the past few years, India has closed its doors to us for import of beef, which resulted in drastic rise in demands throughout the year. Results show that the price of beef per kg increased up to 38% from 2013 to 2017 [2]. In addition, there have been reports of dead casualties while some businesspersons tried to invade the border to bring cows from India [2]. Bangladesh being a Muslim country has a huge demand for beef throughout the year, especially during the festival of Eid-Ul-Adha. Due to high demand and low supply, the country suffers from a deficiency in production of meat. Consequently, beef, which is a prominent source of protein, is slipping out of the reach of a huge portion of the population. Through this research, we aim to encourage more number

of farmers to invest in this sector and make the journey as simple as possible for them. This will in turn withdraw our reliability on importing beef from outside of our country. If properly monitored, the business will experience significant expansion, and the price of beef will stay stable, if not decrease.

1.3 METHODOLOGY

In the past, many researchers have worked on topics similar to this research for predicting outcomes regarding livestock and Automating the animal agriculture process. Most of the researches incorporated Linear Regression Analysis and Decision Tree algorithm in their work. This research aims to implement the data mining tools and machine-learning models to analyze produce a regression analysis for different breeds of cattle. Our research will take both the test and train datasets, and produce regression analysis for different machine learning algorithms, such as Ordinary least squares model, Linear regression model, Polynomial Regression model in Support Vector Machines (SVM) and Decision tree algorithm. These are supervised machine learning algorithms, which take previous datasets, learn the correlation between different attributes of the model, and use it to predict the future expected outcomes from the given input. The outputs will be displayed in both graphical and tabular forms, where the tabular data will have the numerical properties explained, that the model could derive from the given dataset. These numerical results will portray the characteristic of our dataset. Alongside that, the system will have graphical representation, which will show the complete trend and provide an analysis for the hypothesis among the dependent and independent variables of our dataset. The regression models will generate regression equation based on the analysis of the complete dataset, divided by breeds of the cattle. From this equation, the end user will be able to derive the predicted outcomes of the dependent variable (expected weight of the cattle), by providing the independent variables (age, current weight, food habits and weather conditions) as input variables. Therefore, the system plans to provide a complete insight for a new investor or farmer from the context of the most profitable

breed to purchase based on their geographical region. In addition, plans to provide the exact diet plans to follow based on the age and current weight of the cattle, and thus what weight and growth rate to expect from that cattle in the subsequent years of rearing, through the regression analysis.

1.4 OBJECTIVES

- Increase overall meat production.
- Increase overall nutrient contents and quality of the meat produced, using differentiation approach.
- Reduce the unemployment rate and bring solvency.
- Eliminate dependency on import of meat.

1.5 THESIS CONTRIBUTION

Data collection was the most difficult for us at the very beginning as no one was ready to give his or her confidential data. Especially we want to thank **"Meghdubi Agro"** for their enormous support to share their important information, as they were ready to help youth leaders who wants to enter in this field. Last but not the least Bangladesh Livestock Research Institute also helped us through giving us the chance to visit the biggest government farm of our country **"Savar Dairy Farm"**.

1.6 Hypothesis of the Study

Research hypothesis is a statement, which helps the researcher to draw an assumption on whether his hypothetical assumption is true, or not. Based on the objectives of the study, we are presenting the following hypothesis in the table below.

Factors	Null Hypothesis	Alternative Hypothesis
Origin	$x_{or} = 0$	$x_{or} \neq 0$
Age	$x_{ag} = 0$	$x_{ag} \neq 0$
Current weight	$x_{cw} = 0$	$x_{cw} \neq 0$
Environment	$x_{en} = 0$	$x_{en} \neq 0$
Food Habits	$x_{fh} = 0$	$x_{fh} \neq 0$

Table 1.01 Study Hypothesis

From the collected dataset, we see that all the input parameters are positively correlated with the output parameter.

Chapter 2

LITERATURE REVIEW

2.1 RELATED WORKS

In this chapter, a brief overview on the research works existing and on the ones that have taken place in the past, related to our thesis topic is given. Many researchers, around the world, have taken the concept of automated animal agriculture into account. Advancement of Bangladesh is quite new in this sector. However, many international authors and journalists have placed their systematic investigation in this topic. An official journal of American Society of Animal Science has published a number of papers on cattle production and their relation with genetics, nutrition, feedlots, physiology and other affecting factors [3]. An article on Big Data analytics and Precision animal agriculture uses machine Learning algorithms and Data mining tools to analyze and predict Animal agriculture [4]. Precision animal agriculture plays an important role in uplifting the management, production and sustainability of the livestock industry. However, machine learning and data mining is one of the most effective approach to minimize the challenges in animal sciences, as precision agriculture method in unable to implement complex data generated by fully automated phenotypic platforms [4]. Machine learning is a sector of artificial intelligence for estimation and prediction of algorithms. This article focuses on observing the phenotypes in training data set from the genotypes [4]. Maximum objective function from training dataset signifies the efficiency of a regression model [4] [5]. The regression models used in this research for biological and physical factors of cattle required for automating the livestock analysis were K-means model, principal component analysis, regression models, linear model, random forest and neural networks [4]. The work of Arango and Cundiff in 2004, at the U.S meat animal

research center, implemented covariance functions and random regression models (CF-RRM) for predicting the increase in weight of beef cattle, over a period of eight years [5]. This research model consisted of regressions on weight of a cattle recorded quarterly, season of measurement, and pregnancy-lactation periods. However, when the covariance function was applied for cattle of all ages, it was seen that the older aged cattle disagreed to fit to the model and gave higher variance estimates [5]. The author concluded that supervised machine learning could be a solution to this problem. Another research in 2015 by V. Goldberg focused on the analysis of the growth rate of pasture fed Angus cow's growth curves provide an insight of how the weight of some cattle varies over time [6]. The research produced results using the 3-knot cubic spline function, which is a statistical model, was fitted to model weight change across age. A comparison proved that 3-knot cubic spline function gave better estimated results of a cattle's weight with respect to time, from birth to maturity age, than the "Richard's model" [6]. Moreover, a research in 1983, used polynomial regression to estimate the growth rate curve of animals. Covariance is then performed on the regression coefficients. In case of data with missing variables and observations, this approach provides better results than the Multivariate analysis for growth rate estimation [7]. The work of Karin Meyer also emphasizes on the phenotypic variation of livestock growth behavior using Random regression models of two Australian breeds, Polled Herefords and Wokalups [8]. Here phenotypic random regression model was used to analyze the trend in growth rate in correspondence to weight and environmental factors. The author of the article explained the use of Restricted Maximum Likelihood (REML) model for estimating covariance and error in covariance between regression coefficients [8]. In contrast to the polynomial regression models, the article of R.Schaeffe approaches with Random regression models (RRM) [9]. Genetic evaluation of dairy cattle, survival and fertility data, using test day production records is one of the best application of RRM. This research provides an insight of genetic variability with time, which is another progress on the field of automated livestock rearing [9]. The work of S, Niggol concentrates on a different paradigm of climate change [10]. The author

further says, the large commercial beef cattle rearing farmers of Africa, will face the downside than the small-scale goat, sheep and chicken breeders [10]. Although, there has been quite a number of remarkable work on livestock analysis with big data, the impact of machine learning techniques and data mining and potential in “big data” analysis did not receive adequate appreciated in the animal science community, where this recognition has remained only fragmentary.

2.2 COMPARISON

After reviewing all these past and existing research works on livestock analysis, we were able to find some basic differences between these papers and our model. Each of the past research work mainly focused on a single aspect only. However, in our model we tried incorporating multiple factors that can contribute to cattle’s increase in weight, which is its current weight, age, diet plan, and weather conditions. In addition, we focused on deriving the level of impact each parameter will make on our prediction analysis of the expected weight.

Chapter 3
ALGORITHMS

The following discussion focuses on the techniques used to obtain expected output prediction in this study.

3.0 Supervised data mining

A new way to define the methods are used in the field of statistics, machine learning is called Data Mining [11]. It includes methods from the field of AI, machine learning, statistics that implements pattern recognition in large dataset [12]. The objective of data mining is to train a machine with the help of bulk amount of data into an understandable structure which can be used in different analysis in the basis of the trained data set [12].

From a specified set of predictors (independent variables), target variable (dependent variable) was predicted by supervised data mining [13]. We generate a function that maps inputs to preferred outputs using these set of variables using different models to get the expected output.

3.1 Ordinary least squares

Ordinary least squares (OLS) regression model also known as linear regression model is a regressive model for estimating the unknown variables in a linear regression model. It expresses the linear relationship between an explanatory variable and a dependent variable [14]. OLS selects the parameters of a linear function of the dependent and independent variables by keeping the sum of the

squares of the differences between the observed and predicted values of the dependent variable configured as a straight line from our dataset [14]. If mathematically observed, it can be seen that the smaller the difference of sum of squared distance between each data point in the set and corresponding points on the regression surface, the better the model fits the data.

The general equation for Ordinary least squares regression model can be written as:

$$y = \beta_0 + \beta_1 * x \qquad (1)$$

Where, β_0 stands for a constant or y-intercept.

β_1 stands for the coefficient for the dependent variable x represents the dependent variable and y stands for independent variable.

The coefficients of β is found the minimizing the error of prediction. Error of prediction is the difference between the real dependent value and predicted dependent value [15].

Limitations of OLS:

Collinearity between the x (dependent variables) can make us misinterpret the coefficients β. Solution to the limitation can be using the Principal component regression can be used.

3.2 Multiple Linear regressions

It consists of quite some similarities to the simple linear regression model. However, the basic difference is, it deals with multiple independent variables, that portrays effects on the dependent variable, we are intending to estimate. For that,

we need to take into account multiple coefficients of the independent variables, which contribute to the complex computation of our given variables.

In case of multivariate regression model, it is not efficient to consider all the available independent variables. We intend to select the variables, which contribute most to our dependent parameter, through which we start minimizing the error function [16]. This can be achieved by calculating the correlation function of the all the dependent and independent variables related to our data set. The variables with highest correlation functions indicate on having the most significant effect on our estimated parameter, by minimizing the error function [16]. The independent variables with the lowest correlation factor shows least effect on the result, thus can be completely ignored.

The Multivariate Linear Regression model can be expressed by the following equation [17]:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \cdots + \beta_n * x_n \quad (2)$$

Where,

Correlation factors affecting the independent variable [17].

3.3 Support Vector Machine (SVM):

Support vector machines are supervised learning algorithms that implement classification and regression analysis, by analyzing the dataset [18]. SVM creates hyperplanes between two distinct classes of data categories and selects the one with the maximum margin. SVM works well for classifying data sets with lots of features, also known as the higher-dimensional data.

- Works well for classifying higher-dimensional data (many features) Finds higher-dimensional support vectors across which to divide the data (mathematically, these support vectors define hyperplanes.
- Uses something called the kernel trick to represent data in higher-dimensional spaces to find hyperplanes that might not be apparent in lower dimensions

- The important point is that SVM's employ some advanced mathematical trickery to cluster data, and it can handle data sets with lots of features.
- It is also expensive –the "kernel trick" is the only thing that makes it possible.

Support Vector Machine - Regression (SVR)

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm. The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. For regression in SVM, all the main features that characterize the algorithm must be maintained. Svm regression maps input space data into real numbers [26]. Svm regression is performed for high-dimension feature space mapping from input space features, which is a suitable mechanism for this research. In addition, the svm regression function used relies on the epsilon insensitive (loss function) and reduces the model complexity by minimizing errors [28].

Linear SVR

$$\boldsymbol{y = \sum_{i=1}^{N} * (\alpha_i - \alpha_i^*) * (x_i, x) + b} \quad (3)$$

Non-linear SVR

The kernel functions transform the data into a higher dimensional feature space to make it possible to perform the separation.

$$\boldsymbol{y = \sum_{i=1}^{N} * (\alpha_i - \alpha_i^*) * \{\phi(x_i), \phi(x)\} + b} \quad (4)$$

$$\boldsymbol{y = \sum_{i=1}^{N} * (\alpha_i - \alpha_i^*) * K(x_i, x) + b} \quad (5)$$

Kernel functions [27]

1. Linear splines kernel in one-dimension

It is useful when dealing with large sparse data vectors. It is often used in text categorization. The spline kernel performs well in regression problems. Equation is: [27]

$$\boldsymbol{k(x,y) = 1 + xy + xy\,min(x,y) - \frac{x+y}{2} min(x,y)^2 + \frac{1}{3} min(x,y)^3} \qquad (6)$$

2. Polynomial kernel

It is popular in image processing.
Equation is:

$$\boldsymbol{k(x_i, x_j) = (x_i * x_j + 1)^d} \qquad \textbf{(7)}$$

Where d is the degree of the polynomial. [27]

3. Gaussian radial basis function (RBF)

It is general-purpose kernel; used when there is no prior knowledge about the data. Equation is: [34]

4. Sigmoid kernel

We can use it as proxy for neural networks. Which is given by the equation [34]

$$\boldsymbol{k(x,y) = tanh(\alpha x^T y + c)} \qquad (8)$$

3.4 Decision tree learning

In context of data, mining techniques are concerned with pattern and classification of huge and uncertain data [19]. Decision tree algorithm is a frequently used technique in data mining to classify large amount of data and extract dataset that has similar patterns [19]. Decision tree represents classification model for a data set in the form of a tree structure, breaking down the dataset into smaller subsets. In this research decision tree algorithm builds regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at

the same time an associated decision tree is incrementally developed [20]. The leaf nodes are expected weights and decision nodes has two branches e.g. age difference and current weights, each representing values for the attribute tested. Leaf nodes represent a decision on the numerical target, which is expected value. The topmost decision node in a tree, which is known as root node, was divided into five breeds we worked on. Decision trees can handle both categorical and numerical data.

3.5 Comparison:

- After analyzing multiple algorithms and models for our research, we came to conclusion that Linear Regression Model, Support Vector Machine model and Decision Tree Model (ID3) are the most suitable algorithms that matches with our context and fulfils our needs. We started with analyzing and researching the Time series analysis. While implementing our data set and model, through Time series analysis, we came across multiple limitations. Firstly, to implement Time series analysis, the data set must have at least 30 time lags. However, our data set had only 20 lags, which was insufficient. We tried, but we could not collect data, which was monthly or semiannually lagged. One major reason for not selecting Time series analysis was, we could only predict future outcomes from the Time series of a particular data set. We were not able to implement or work with any data that falls within the Time series executed. Furthermore, Time series analysis could not provide the trend of the growth rate of some cattle, which is a key part of our research. In addition, due to having less number of time lags, we were not able to accurately standardize the data set and check for stationarity of the data set. We also faced difficulty to get the record of consistent time lags, which is a key point for Time series analysis.

- As we are not getting the in between expected weight with the help of Time series analysis we went for the linear regression model which is more effective

for our dataset. Here we are getting the future outcomes as well as the in between expected weight outcomes. In addition, if any user provides us the second years' current weight we can provide them the third or the any future years expected result. Consistent lags of the data set is not required to implement the work process, which is more feasible in comparison to time series analysis.

- Furthermore, we took the help of Decision tree model (ID3). Because, if we have more choices in hand to choose which breed we should go for which aged breed is more profitable and the weight range of particular aged breed, Decision tree model would be more helpful to make the decision on the basis of the alternatives in our hand. From where
  where we get which breed is more profitable in which age and what should be the current weight range to get the expected output.

3.6 K-Fold Cross Validation:

Cross validations are mainly used when there is a prediction model and how effective it performs in practice. We used cross validation to test our model's ability and accuracy to predict new things that were not being estimated [21]. Moreover, we get to know flag problems like overfitting [21] and under-fitting using this validation process whether it will generalize to an independent dataset or not.

In K fold cross validation; we divide the data into K equal subparts at first. Here number of subsets must be at least two and single subpart is retained for the validations process where other k-1 parts are used as the train data set. That is why *k*-fold cross-validation is known as leave one-out cross-validation [22].

Chapter 4

DATA PROCESSING AND RESULT ANALYSIS

Data is the key component for predicting outcomes using regression analysis. By feeding the dataset into the statistical and machine learning models, a comparison is made between the models and a regression analysis for predicting the outcomes, is generated. Therefore, the data sets need to be authentic and complete.

4.1 DATA PROCESSING

The process of collecting the dataset for the research and the standardization form of the raw data is given below.

4.1.1 Data Collection

Firstly, online datasets for livestock of Bangladesh is not available. On searching multiple government websites, we could not find any digital dataset and later discovered they do not keep any record of livestock in digital format. However, we could find some international datasets only. However, we did not consider the international dataset, as we tried to develop a model on the context of Bangladesh completely. Later on, we visited our nearest Department of Livestock Services (DLS), and they referred us to a few government cattle breeding farms. On visiting Savar Dairy farm physically, we could lay our hands on some raw, handwritten data. We had to convert the raw datasets into Microsoft Excel datasheets. We further visited 2 different branches of Meghdubi Agro farm, which is a private cattle farm, for collecting more data required for our research. Meghdubi Agro also could provide us with handwritten datasets only. Both the Agro Farms provided dataset of a cow, having similar attributes. The attributes included, current weight of a cow, its current age, gender, what will be its expected weight about a year later, the diet plan it needs

to follow, the suitable environment it should be kept in and the origin of a cattle. In this model, we worked with five different cattle breeds, Bhutani, Friesian, Local, Brahman and Indian Haryana. The cattle are fed a certain amount of food with respect to their age and current weight. The table showing the average relationship between the diet plan and current weight is given below.

Current Weight of Fattening Bull(kg)	***Daily diet plan(kg)***
100 - 200	**3**
201 – 300	**3.5**
301 – 400	**3.75**
401 – 500	**4.75**
501 – 600	**5.75**
601 – 700	**6.75**
701 – 800	**7.50**
801 ++	**8**

Table 1.02: Diet Plan of cattle

4.1.2 Data Preprocessing

- **Data type conversion**: From our collected dataset, origin of cattle and the weather conditions were categorical data, which needed to be converted into numerical data. For this data type conversion, LabelEncoder method of Sklearn class was used. Due to less variation of the categorical data, we did not require the OneHotEncoder method.

- **Data scaling**: Due to different ranges of data of different parameters, all the data were brought into a standard scale, for better interpretation. After conversion, the

standardized data ranged from -1 to +1. Standardization of the data increased the accuracy of the trend analysis.

- **Feature Selection**: As we could collect from primary sources and had limited attributes, all the input parameters were meaningful and correlated to the output parameter. Therefore, we did not have to run any feature selection algorithm on our dataset.

4.2 RESULT ANALYSIS:

4.2.1 Result analysis based on Ordinary Least Squares Regression Model:

R-square: It is a statistical measure of how close the real data points are fitted onto the regression line. It is known as the coefficient of determination. Its defines how much variation is in the independent variable is explained by the variation in the dependent variable. It ranges from 0 to 1.

A value close to '1' represents most of the data is represented by the model and a value close to 0' represents no relation with the regression line.

Adjusted R-square: It adjusts the statistic value of R-square based on the correlation of an independent variable to the dependent variable. If the correlation is, strong Adjusted R-square value increases and vice versa. Adjusted R-square value must always be less than the actual R-square value.

F-statistic value: The F value represents the ratio of the mean regression sum of squares divided by the mean error sum of squares. Its value will range from zero to an arbitrarily large number [25]. It tests the overall significance of the regression model.

Probability (F): They test the null hypothesis, that all of the regression coefficients are equal to zero [25]. The value of Probability (F) is the probability that the null hypothesis for the full model is true. It is used to determine whether the coefficients are significantly different from zero or not. The closer its value to zero the better.

T-statistic value: It is computed by dividing the estimated value of the parameter by its standard error. This is a statistical measure that the actual significance of the parameter is not zero [25]. The larger the absolute value of t, the less likely that the actual value of the parameter could be zero.

Probability (T): The Probability (t) value is the probability of obtaining the estimated value of the parameter if the actual parameter value is zero [25]. The smaller the value of Probability (t), the more significant the parameter and the less likely that the actual parameter value is zero. The closer the value to zero the significant is the parameter.

Coefficient: Represents the level of impact an independent variable has on the value of the dependent variable.

BHUTANI CATTLE:

Dep. Variable:	y	R-squared:	0.956
Model:	OLS	Adj. R-squared:	0.953
Method:	Least Squares	F-statistic:	319.5
Date:	Sat, 21 Jul 2018	Prob (F-statistic):	1.34e-48
Time:	00:06:11	Log-Likelihood:	-263.16
No. Observations:	79	AIC:	536.3
Df Residuals:	74	BIC:	548.2
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0	0	nan	nan	0	0
x1	-18.2028	1.928	-9.439	0.000	-22.045	-14.360
x2	2.9981	3.865	0.776	0.440	-4.704	10.700
x3	0.0751	0.030	2.533	0.013	0.016	0.134
x4	-4.4649	4.220	-1.058	0.293	-12.873	3.943
x5	9.4150	1.377	6.839	0.000	6.672	12.158

Omnibus:	1.740	Durbin-Watson:	1.878
Prob(Omnibus):	0.419	Jarque-Bera (JB):	1.132
Skew:	0.251	Prob(JB):	0.568
Kurtosis:	3.302	Cond. No.	inf

Table 1.03: Bhutani Train Data

Dep. Variable:	y	R-squared:	0.944
Model:	OLS	Adj. R-squared:	0.926
Method:	Least Squares	F-statistic:	50.70
Date:	Sat, 21 Jul 2018	Prob (F-statistic):	7.24e-09
Time:	00:06:12	Log-Likelihood:	-69.125
No. Observations:	20	AIC:	148.3
Df Residuals:	15	BIC:	153.2
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0	0	nan	nan	0	0
x1	-14.0404	5.417	-2.592	0.020	-25.586	-2.495
x2	6.0306	11.663	0.517	0.613	-18.830	30.891
x3	0.0889	0.153	0.580	0.570	-0.238	0.415
x4	-0.0578	11.479	-0.005	0.996	-24.524	24.409
x5	5.9037	3.535	1.670	0.116	-1.631	13.438

Omnibus:	3.872	Durbin-Watson:	1.785
Prob(Omnibus):	0.144	Jarque-Bera (JB):	2.171
Skew:	-0.779	Prob(JB):	0.338
Kurtosis:	3.421	Cond. No.	inf

Table 1.04: Bhutani Test Data

Train: $y = -18.21 * x_1 - 2.99 * x_2 + 0.075 * x_3 - 4.46 * x_4 - 9.42 * x_5$ (9)

Test: $y = -14.04 * x_1 - 6.03 * x_2 + 0.089 * x_3 - 0.058 * x_4 - 5.90 * x_5$ (10)

R-square: Train Data (0.956) and Test Data (0.944). In both the cases, the dependent variable is well explained by the independent variables, but Train Data is giving better result. The variation between the values are also quite low.

Adjusted R-square: Train Data (0.953) and Test Data (0.926). Both the datasets show very strong correlation with the expected weight, where train data is giving a better output.

F-statistic: Train Data (319.5) and Test Data (50.70). The results show that the regression is more significant using the train data set, in this case.

Probability (F): Both the datasets produce very insignificant P(F) values. This shows that the null hypothesis is almost rejected. Null hypothesis for our research was that there is no relationship between the weather conditions and expected weight of a cattle. In addition, the coefficients are significantly different from zero.

Probability (T): For the train data set, x1, x3 and x5, completely rejects the null hypothesis. For the test data, except for x1, the values comparatively fail to reject the null hypothesis.

Coefficients: For both the train dataset and test dataset, x2, x3 and x5 show positive correlation and x1, x4 shows negative correlation with the expected weight.

FRIESIAN CATTLE:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.892
Model:	OLS	**Adj. R-squared:**	0.890
Method:	Least Squares	**F-statistic:**	349.7
Date:	Sat, 21 Jul 2018	**Prob (F-statistic):**	5.49e-100
Time:	00:06:11	**Log-Likelihood:**	-65.812
No. Observations:	216	**AIC:**	141.6
Df Residuals:	211	**BIC:**	158.5
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0	0	nan	nan	0	0
x1	0.0760	0.031	2.483	0.014	0.016	0.136
x2	-0.4304	0.038	-11.380	0.000	-0.505	-0.356
x3	0.9312	0.149	6.259	0.000	0.638	1.225
x4	-0.0499	0.044	-1.141	0.255	-0.136	0.036
x5	0.3404	0.161	2.113	0.036	0.023	0.658

Omnibus:	1.317	**Durbin-Watson:**	2.300
Prob(Omnibus):	0.518	**Jarque-Bera (JB):**	1.076
Skew:	0.166	**Prob(JB):**	0.584
Kurtosis:	3.097	**Cond. No.**	inf

Table 1.05: Friesian Train Data

Dep. Variable:	y	**R-squared:**	0.862
Model:	OLS	**Adj. R-squared:**	0.848
Method:	Least Squares	**F-statistic:**	61.41
Date:	Sat, 21 Jul 2018	**Prob (F-statistic):**	6.28e-20
Time:	00:06:12	**Log-Likelihood:**	-19.858
No. Observations:	54	**AIC:**	49.72
Df Residuals:	49	**BIC:**	59.66
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0	0	nan	nan	0	0
x1	0.0744	0.066	1.119	0.268	-0.059	0.208
x2	-0.3664	0.085	-4.334	0.000	-0.536	-0.196
x3	0.8096	0.339	2.390	0.021	0.129	1.490
x4	-0.0238	0.109	-0.218	0.828	-0.243	0.195
x5	0.2987	0.406	0.735	0.466	-0.518	1.115

Omnibus:	3.674	**Durbin-Watson:**	1.556
Prob(Omnibus):	0.159	**Jarque-Bera (JB):**	1.917
Skew:	-0.155	**Prob(JB):**	0.383
Kurtosis:	2.130	**Cond. No.**	inf

Table 1.06: Friesian Test Data

Train: $y = 0.07 * x_1 - 0.43 * x_2 + 0.93 * x_3 - 0.04 * x_4 + 0.34 * x_5 (11)$

Test: $y = 0.07 * x_1 - 0.36 * x_2 + 0.81 * x_3 - 0.02 * x_4 - 0.29 * x_5 (12)$

R-square: Train Data (0.892) and Test Data (0.862). In both the cases, the dependent variable is well explained by the independent variables, where Train Data is high. The variation between the values are also quite low.

Adjusted R-square: Train Data (0.890) and Test Data (0.848). Both the datasets show strong correlation with the expected weight, where train data is giving a better output than the test dataset.

F-statistic: Train Data (349.7) and Test Data (61.41). The results show that the regression is more significant using the train data set, in this case.

Probability (F): Both the data sets produce very insignificant values. This shows that the null hypothesis is almost rejected. In addition, the coefficients are significantly

different from zero. However, the train data produces much smaller P (F) value than the test data set.

Probability (T): For the train data set, x1, x2 and x3, completely rejects the null hypothesis. For the test data, except for x2, the values comparatively fail to reject the null hypothesis.

Coefficients: For both the train dataset and test dataset, x1, x3 and x5 show positive correlation and x2, x4 shows negative correlation with the expected weight.

LOCAL CATTLE:

Dep. Variable:	y	**R-squared:**	0.820
Model:	OLS	**Adj. R-squared:**	0.790
Method:	Least Squares	**F-statistic:**	27.34
Date:	Sat, 21 Jul 2018	**Prob (F-statistic):**	2.51e-10
Time:	00:06:11	**Log-Likelihood:**	-90.999
No. Observations:	35	**AIC:**	192.0
Df Residuals:	30	**BIC:**	199.8
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0	0	nan	nan	0	0
x1	-0.8601	1.178	-0.730	0.471	-3.265	1.545
x2	-0.4965	0.437	-1.137	0.265	-1.389	0.396
x3	0.0550	0.027	2.024	0.052	-0.000	0.110
x4	-2.1678	2.294	-0.945	0.352	-6.853	2.518
x5	-3.5121	2.370	-1.482	0.149	-8.352	1.328

Omnibus:	3.476	**Durbin-Watson:**	2.576
Prob(Omnibus):	0.176	**Jarque-Bera (JB):**	3.196
Skew:	0.704	**Prob(JB):**	0.202
Kurtosis:	2.543	**Cond. No.**	inf

Table 1.07: Local Train Data

Dep. Variable:	y	**R-squared:**	0.963
Model:	OLS	**Adj. R-squared:**	0.916
Method:	Least Squares	**F-statistic:**	20.58
Date:	Sat, 21 Jul 2018	**Prob (F-statistic):**	0.00590
Time:	00:06:12	**Log-Likelihood:**	-19.772
No. Observations:	9	**AIC:**	49.54
Df Residuals:	4	**BIC:**	50.53
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0	0	nan	nan	0	0
x1	2.4956	3.412	0.732	0.505	-6.976	11.968
x2	-0.1032	2.795	-0.037	0.972	-7.864	7.658
x3	0.1607	0.060	2.679	0.055	-0.006	0.327
x4	18.6856	9.251	2.020	0.114	-6.998	44.369
x5	-17.2366	6.703	-2.571	0.062	-35.848	1.375

Omnibus:	5.763	**Durbin-Watson:**	1.873
Prob(Omnibus):	0.056	**Jarque-Bera (JB):**	1.968
Skew:	-1.103	**Prob(JB):**	0.374
Kurtosis:	3.614	**Cond. No.**	inf

Table 1.08: Local Test Data

Train: $y = -0.86 * x_1 - 0.50 * x_2 + 0.06 * x_3 - 2.17 * x_4 - 3.51 * x_5$ (13)

Test: $y = 2.50 * x_1 - 0.10 * x_2 + 0.16 * x_3 + 18.69 * x_4 - 17.24 * x_5$ (14)

R-square: Train Data (0.820) and Test Data (0.963). In both the cases, the dependent variable is well explained by the independent variables. However, this time the Test Dataset is giving a better result.

Adjusted R-square: Train Data (0.790) and Test Data (0.916). Both the datasets show strong correlation with the expected weight, where test data is giving a significantly better output than the train dataset.

F-statistic: Train Data (27.34) and Test Data (20.58). The results show that the regression is more significant using the train data set, in this case. However, both the values are not impressive in proving the significance of the regression model.

Probability (F): Both the datasets produce very insignificant values. This shows that the null hypothesis is almost rejected. And also, the coefficients are significantly different from zero.

However, the train data produces much smaller P(F) value than the test data set.

Probability(T): For the train data set, except for x3, the values are high enough for not being able to reject the null hypothesis. For the test data, the values comparatively fail to reject the null hypothesis.

Conclusion: For the train dataset only x3 shows positive correlation and for the test dataset, parameters x1, x3 and x4 show positive correlation and the rest seem to affect the expected weight inversely

BRAHMAN CATTLE:

Dep. Variable:	y	R-squared:	0.930
Model:	OLS	Adj. R-squared:	0.913
Method:	Least Squares	F-statistic:	56.44
Date:	Sat, 21 Jul 2018	Prob (F-statistic):	1.36e-09
Time:	00:06:11	Log-Likelihood:	-59.341
No. Observations:	21	AIC:	126.7
Df Residuals:	17	BIC:	130.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0	0	nan	nan	0	0
x1	1.795e-13	1.36e-13	1.321	0.204	-1.07e-13	4.66e-13
x2	-2.2840	0.877	-2.604	0.019	-4.134	-0.434
x3	0.0905	0.036	2.489	0.023	0.014	0.167
x4	8.9729	2.884	3.111	0.006	2.887	15.059
x5	-6.3798	3.759	-1.697	0.108	-14.311	1.551

Omnibus:	0.303	Durbin-Watson:	1.582
Prob(Omnibus):	0.859	Jarque-Bera (JB):	0.016
Skew:	-0.063	Prob(JB):	0.992
Kurtosis:	2.949	Cond. No.	inf

Table 1.09: Brahman Train Data

Dep. Variable:	y	R-squared:	0.992
Model:	OLS	Adj. R-squared:	0.977
Method:	Least Squares	F-statistic:	64.34
Date:	Sat, 21 Jul 2018	Prob (F-statistic):	0.0154
Time:	00:06:12	Log-Likelihood:	-9.8186
No. Observations:	6	AIC:	27.64
Df Residuals:	2	BIC:	26.80
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0	0	nan	nan	0	0
x1	-4.471e-11	1.32e-11	-3.378	0.078	-1.02e-10	1.22e-11
x2	5.0776	2.189	2.320	0.146	-4.340	14.495
x3	0.5556	0.140	3.978	0.058	-0.045	1.156
x4	-52.3364	16.847	-3.107	0.090	-124.822	20.150
x5	-51.3392	13.558	-3.787	0.063	-109.673	6.995

Omnibus:	nan	Durbin-Watson:	2.390
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.143
Skew:	-0.106	Prob(JB):	0.931
Kurtosis:	2.275	Cond. No.	inf

Table 1.10: Brahman Test Data

Train: $y = 1.79 * 10^{-13} * x_1 - 2.2 * x_2 + 0.09 * x_3 + 8.97 * x_4 - 6.38 * x_5 (15)$

Test: $y = -4.47 * 10^{-11} * x_1 + 5.07 * x_2 + 0.55 * x_3 - 52.33 * x_4 - 51.33 * x_5 (16)$

R-square: Train Data (0.930) and Test Data (0.992). In both the cases, the dependent variable is very well explained by the independent variables. However, this time the Test Dataset is giving an almost accurate result.

Adjusted R-square: Train Data (0.913) and Test Data (0.977). Both the datasets show strong correlation with the expected weight, where test data is giving a significantly better output than the train dataset.

F-statistic: Train Data (56.44) and Test Data (64.34). The results show that the regression is more significant using the test data set, in this case. However, both the values are not impressive in proving the significance of the regression model.

Probability (F): Both the data sets produce very insignificant values. This shows that the null hypothesis is almost rejected. And also, the coefficients are significantly

different from zero. However, the train data produces much smaller P(F) value than the test data set, in case of Brahman breed.

Probability (T): For the train data set, except for x2 and x4, the values are high enough for not being able to reject the null hypothesis. For the test data, the values comparatively fail to reject the null hypothesis.

Coefficients: For the train dataset, in case of Brahman breed, x1, x3 and x4, shows positive correlation and for the test dataset, x2 and x3 show positive correlation and the rest seem to affect the expected weight inversely.

INDIAN HARYANA CATTLE:

Dep. Variable:	y	**R-squared:**	0.934
Model:	OLS	**Adj. R-squared:**	0.920
Method:	Least Squares	**F-statistic:**	68.13
Date:	Sat, 21 Jul 2018	**Prob (F-statistic):**	2.18e-13
Time:	00:06:11	**Log-Likelihood:**	-76.073
No. Observations:	29	**AIC:**	162.1
Df Residuals:	24	**BIC:**	169.0
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0	0	nan	nan	0	0
x1	-6.5100	1.431	-4.548	0.000	-9.464	-3.556
x2	-11.5906	2.671	-4.340	0.000	-17.102	-6.079
x3	0.0453	0.030	1.514	0.143	-0.016	0.107
x4	6.5227	2.827	2.308	0.030	0.689	12.356
x5	2.4204	3.147	0.769	0.449	-4.074	8.915

Omnibus:	1.021	**Durbin-Watson:**	1.912
Prob(Omnibus):	0.600	**Jarque-Bera (JB):**	0.878
Skew:	-0.165	**Prob(JB):**	0.645
Kurtosis:	2.214	**Cond. No.**	inf

Table 1.11: Indian Train Data

Dep. Variable:	y	**R-squared:**	0.979
Model:	OLS	**Adj. R-squared:**	0.945
Method:	Least Squares	**F-statistic:**	28.25
Date:	Sat, 21 Jul 2018	**Prob (F-statistic):**	0.00999
Time:	00:06:12	**Log-Likelihood:**	-17.038
No. Observations:	8	**AIC:**	44.08
Df Residuals:	3	**BIC:**	44.47
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0	0	nan	nan	0	0
x1	-5.2934	5.458	-0.970	0.404	-22.662	12.075
x2	-0.9015	4.897	-0.184	0.866	-16.487	14.684
x3	-0.1671	0.182	-0.918	0.426	-0.746	0.412
x4	-0.4295	5.170	-0.083	0.939	-16.883	16.024
x5	19.1740	17.189	1.115	0.346	-35.530	73.879

Omnibus:	0.253	**Durbin-Watson:**	1.134
Prob(Omnibus):	0.881	**Jarque-Bera (JB):**	0.196
Skew:	-0.246	**Prob(JB):**	0.907
Kurtosis:	2.411	**Cond. No.**	inf

Table 1.12: Indian Test Data

Train: $y = -6.51 * x_1 - 11.59 * x_2 + 0.05 * x_3 + 6.52 * x_4 - 2.42 * x_5$(17)

Test: $y = -5.29 * x_1 - 0.90 * x_2 - 0.17 * x_3 - 0.43 * x_4 - 19.17 * x_5$ (18)

R-square: Train Data (0.934) and Test Data (0.979). In both the cases, the dependent variable (expected weight) is very well explained by the independent variables. However, this time the Test Dataset is giving a better result, in case of Indian Haryana Breed.

Adjusted R-square: Train Data (0.920) and Test Data (0.945). Both the datasets show strong correlation with the expected weight, where test data is giving a significantly better output than the train dataset.

F-statistic: Train Data (68.13) and Test Data (28.25). The results show that the regression is more significant using the train data set, in this case of Indian breed. However, both the values are not impressive in proving the significance of the regression model.

Probability (F): Both the data sets produce very insignificant values. This shows that the null hypothesis is almost rejected. In addition, the coefficients are significantly different from zero. However, the train data produces much smaller P (F) value than the test data set, in case of Brahman breed.

Probability (T): For the train data set, x1 and x2 completely rejects the null hypothesis, as the values are 0.0000. However, for the test data, the values comparatively fail to reject the null hypothesis.

Coefficients: For the train dataset, in case of Indian Haryana breed, x3, x4 and x5, shows positive correlation and for the test dataset only x5 show positive correlation and the rest seem to affect the expected weight inversely.

4.2.2 Individual attribute regression analysis of Friesian breed

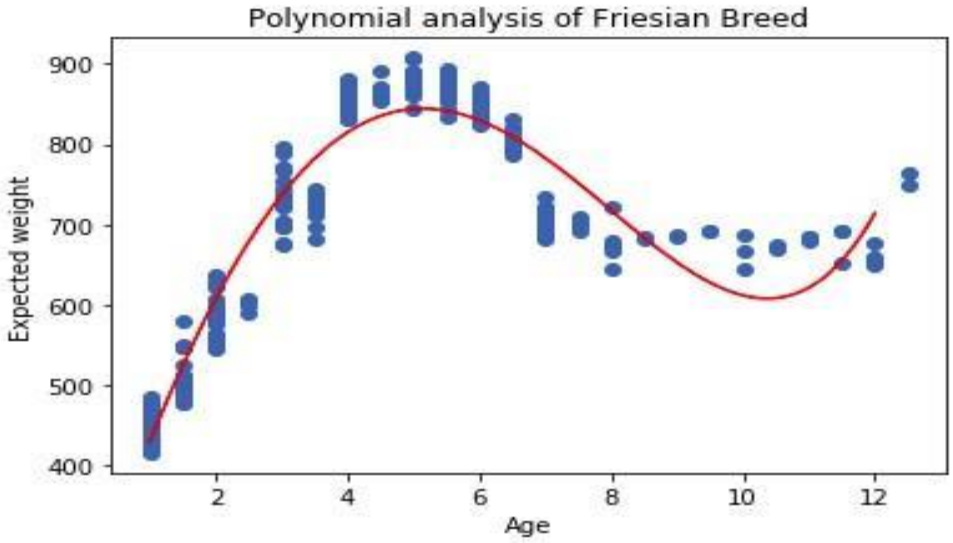


Fig-1: Regression Analysis of Age vs. Expected Weight of Friesian Breed

From the graph it can be seen that till 5 years of age, the expected weight of a Friesian cattle increases proportionately. From 6 to 10 years of age the expected weight decreases and again from 11 to 12 years there is an upward trend.

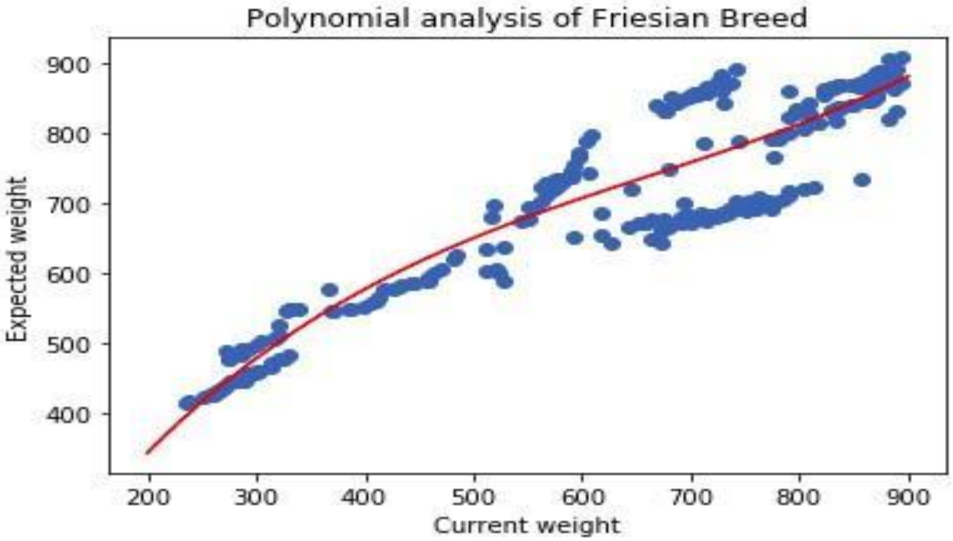


Fig-2: Regression Analysis of Current weight vs. Expected Weight of Friesian Breed

This graph shows that with increase in current weight of a Friesian cattle, the expected weight also increases.

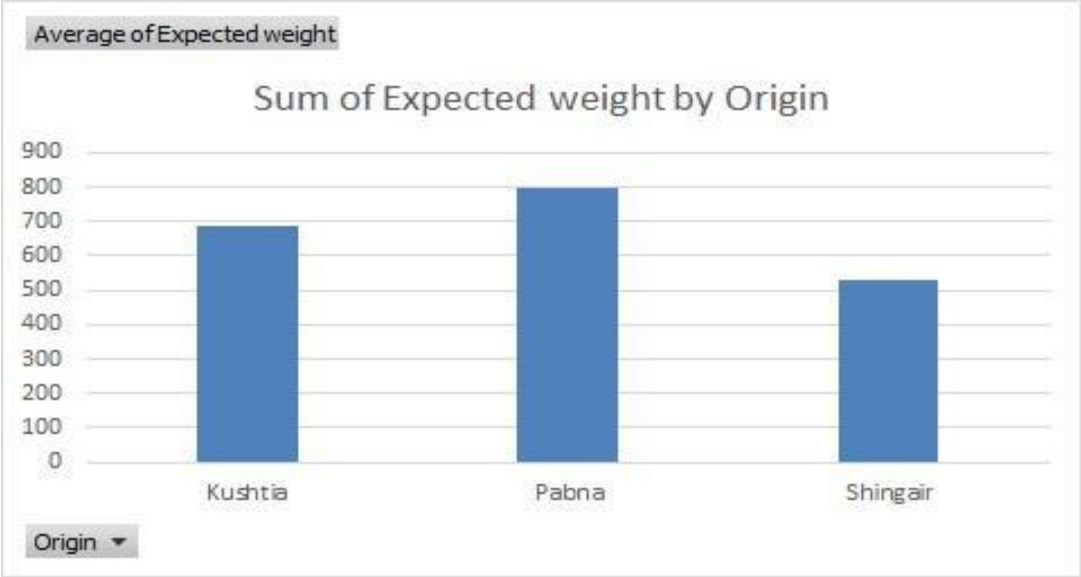


Fig-3: Regression Analysis of Origin vs. Expected Weight of Friesian Breed

The graph shows that Pabna region is best for rearing Friesian cattle compared to Kushtia and Shingair region.

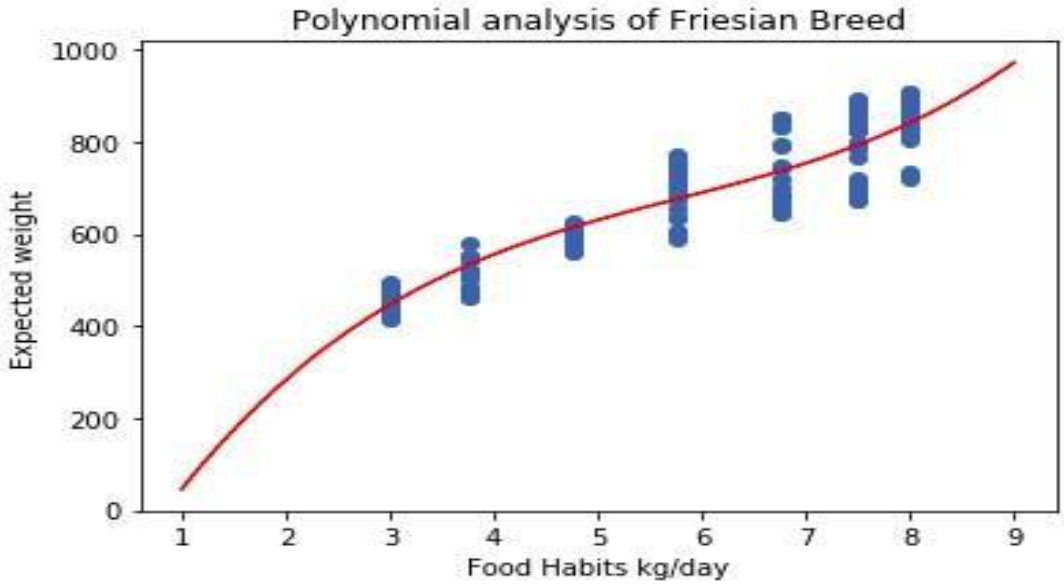


Fig-4: Regression Analysis of Food Habits vs. Expected Weight of Friesian Breed

With the increase in amount of food provided to a cattle the expected weight increases.

4.2.3 Result analysis based on Support Vector Machines (SVM)

Support Vector Regression (SVR)

BREED	*Bhutani*	*Friesian*	*Local*	*Brahman*	*Indian-Haryana*
Linear Regression Score	0.6865	0.4036	0.2855	0.5721	0.0745
svm_lin	0.1049	0.4857	-0.8538	0.4191	0.0094
k_scoreTrain_linear mean	0.1713	0.4857	-0.8310	0.1485	-0.0669
k_scoreTest_linear mean	-0.9836	0.4602	-4.7962	-0.0146	-0.0908
svm_poly	-0.2645	0.1708	-0.1794	0.4696	-0.0476
k_scoreTrain_poly mean	-0.1651	0.2609	-1.5114	0.3140	-0.1974
k_scoreTest_poly mean	-1.2729	0.1454	-1.5866	-0.2853	-0.1224
svm_rbf	0.5777	0.5218	-0.4294	0.4629	-0.3044
k_scoreTrain_rbf mean	0.1714	0.4857	-0.8310	0.1485	-0.0669
k_scoreTest_rbf mean	-0.3614	0.5181	-6.3857	-0.7273	-0.1467
svm_sigmoid	-7.1099	-30.0922	-0.8399	0.4056	0.0112
k_scoreTrain_sigmoid mean	-2.8733	-18.8017	-0.5980	0.0293	0.2433
k_scoreTest_sigmoid mean	-2.3269	-0.2530	-6.8875	-0.4004	-0.0790

Table: 1.13 SVR models

Result Interpretation:

Bhutani breed: Linear Regression gives the best interpretation for the coefficient of determination that is 68.65%. Whereas, the SVM for Sigmoid model gives the poorest fit to our given dataset.

Friesian Breed: Support Vector Machine for Radial Basis Function produces the best score of 52.18%. This explains that 52.18% of the times the model will predict accurate results for the required newly given data. In this case, SVM for sigmoid gives the lowest score of regression analysis.

Local Breed: Linear Regression model gives the best score that is 28% approximately, among the other regression models. However, for this breed, the highest coefficient of determination is comparatively lower to other breeds. In this case, SVM for sigmoid model represents a highly poorly fitted negative correlation.

Brahman Breed: Out of the all the breeds, Brahman breed shows the highest positive correlation for all the models. However, linear regression model produces the best score for the coefficient of determination of 57.2%, among the five models, which represents that 57.2% of the dependent variables are explained by our given independent variables. SVM for sigmoid, in this case also, produces the lowest score.

Indian-Haryana breed: produces the worst fitted scores among the other breeds. Although having a low score, linear regression gives a comparatively better output, out of the five other models. However, in this case SVM for Radial Basis Function produces the worst fit with our given dataset.

Comparison: Comparing the selected algorithms, it is found that Radial Basis Function (RBF) model and linear model of the SVM regression, proves to be the most preferable algorithms for this proposed model, as these algorithms generate the highest accuracy in the regression score. However, different models fit different breeds more accurately.

4.2.4 Result Analysis Based on Decision Trees:

FIG-5: LOCAL:

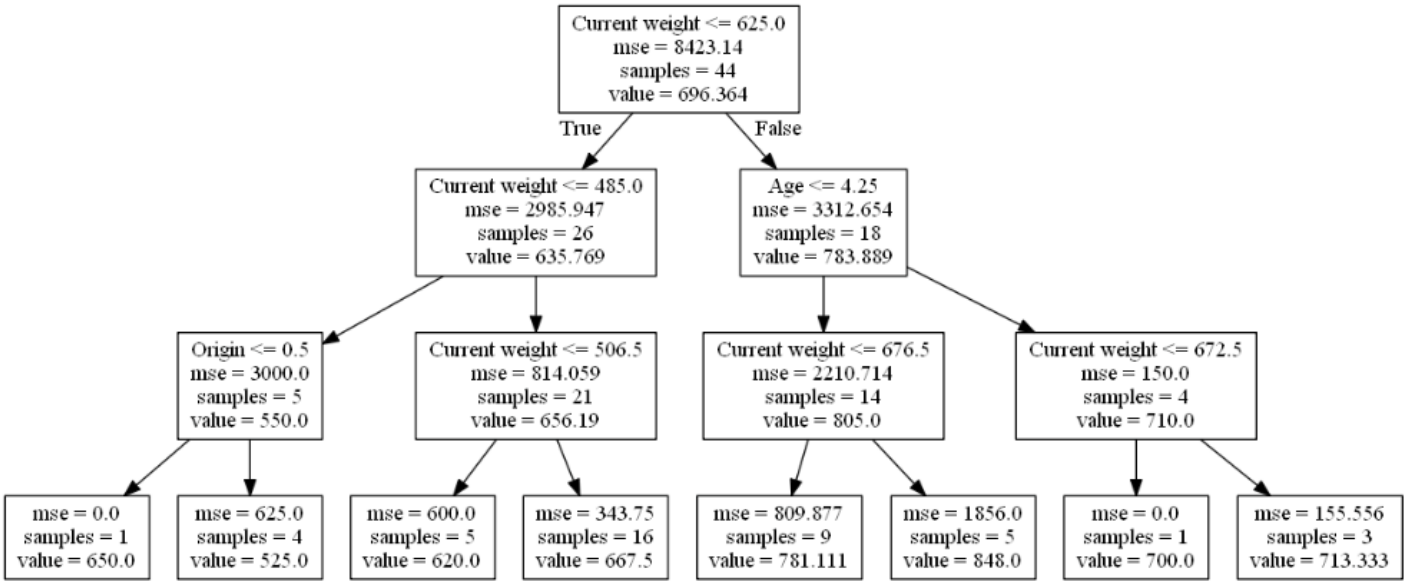


FIG-6: BHUTANI:

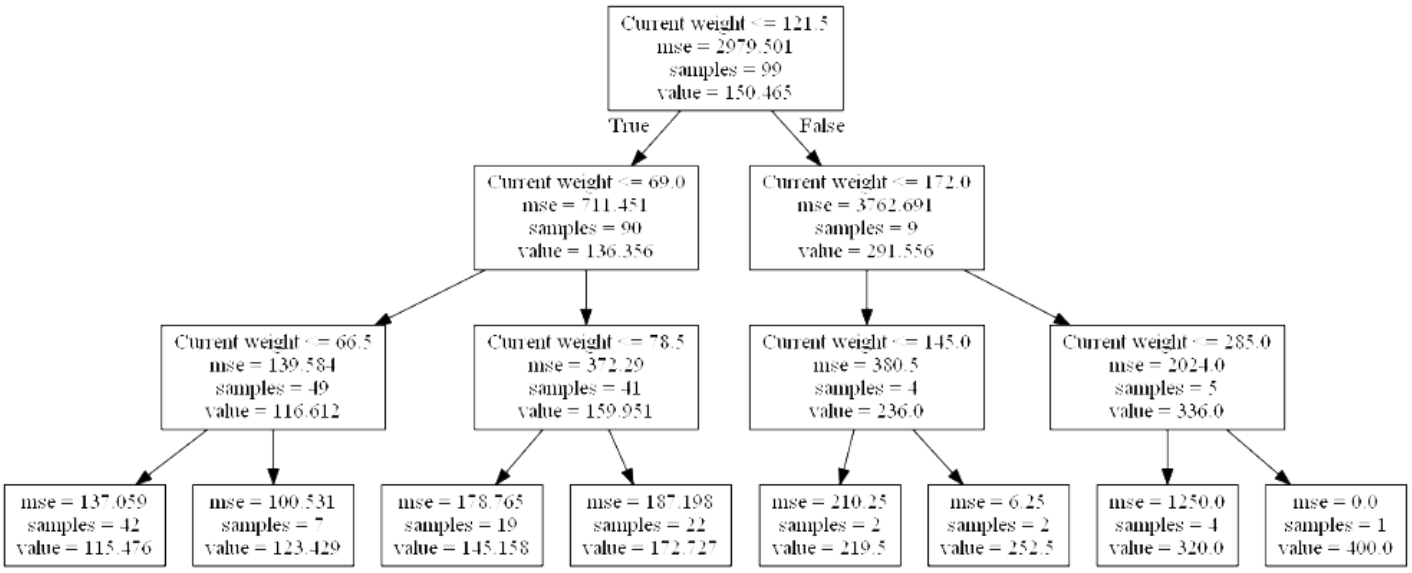


FIG-7: FRIESIAN:

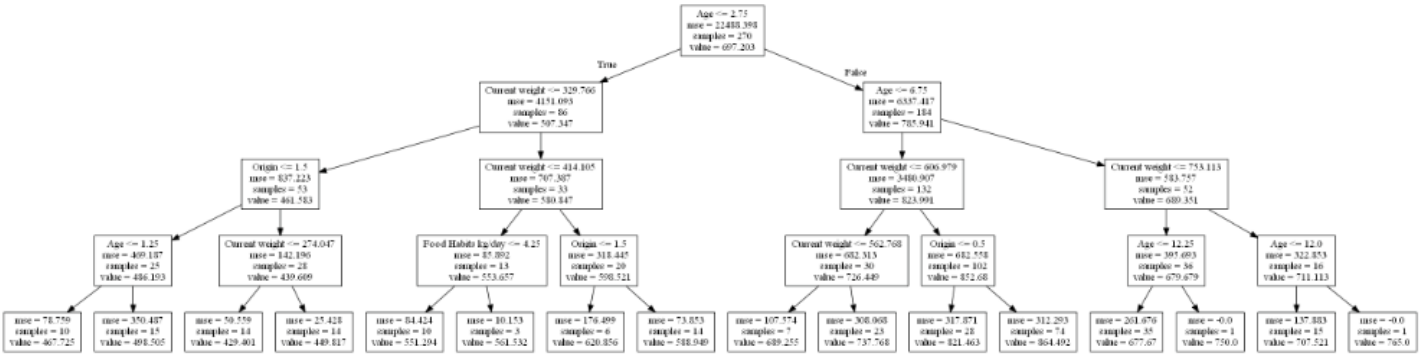


FIG-8: BRAHMA:

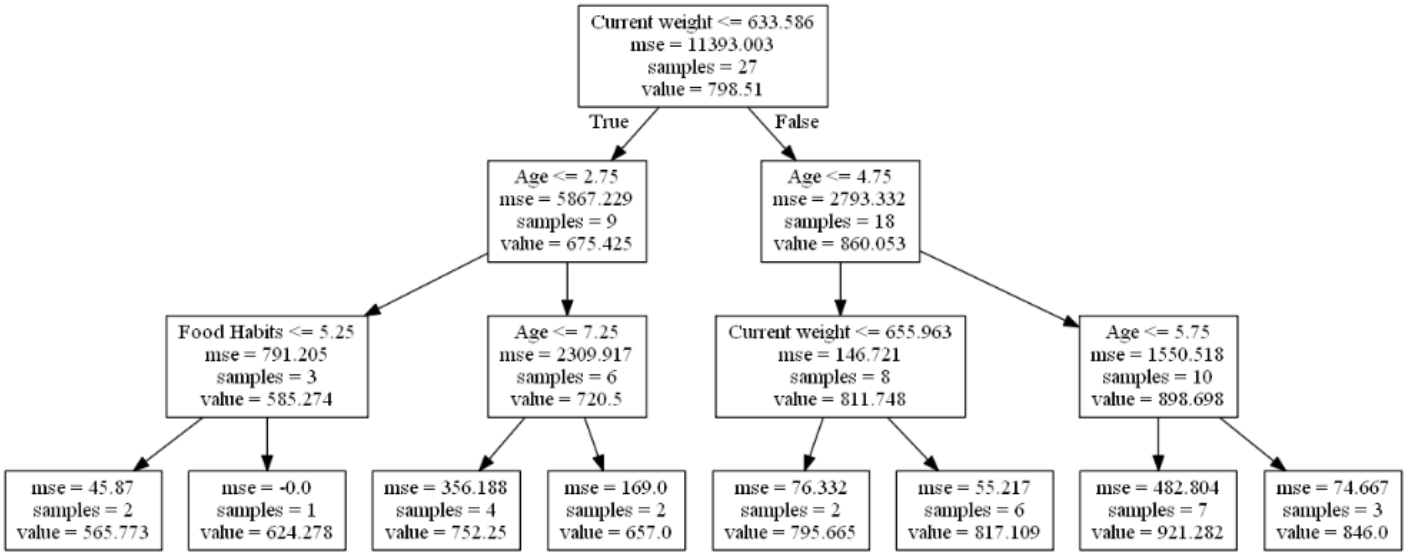
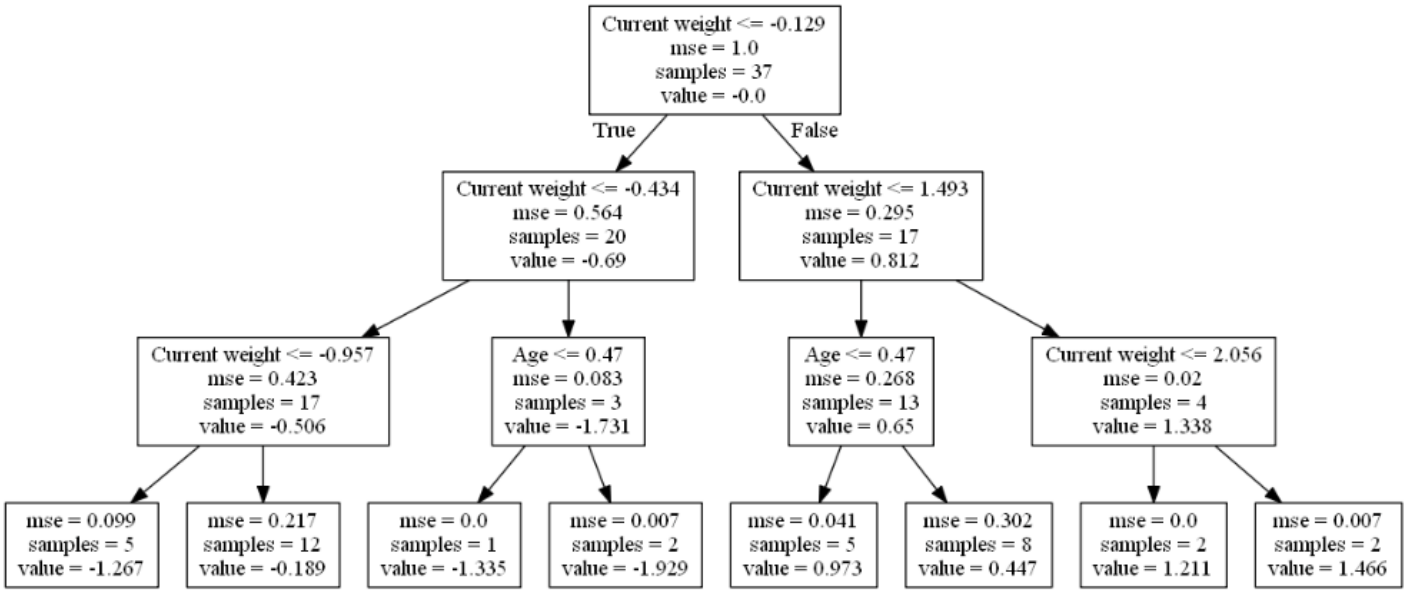


FIG-9: INDIAN HARYANA:



Chapter 5
TOPIC ANALYSIS

This chapter analyzes our research topic in terms of its feasibility according to the market situation. Also the requirements of the information, how much importance they bear, screenshots of the findings and what function they are significant, features of the analysis, limitations while working on the topic– all will be discussed in this chapter.

5.1 FEASIBILITY ANALYSIS

We did feasibility analysis of our project in four different sectors:

5.1.1 Topic Feasibility:

In this world of digitalization, every sector around us is making their advancement through technical automation. However, the livestock industry did not experience much exposure yet [23]. Therefore, this research will be in huge demand to the end users, as it is considering all the factors for analyzing big data by implementing machine-learning techniques. The findings of this research will be a learning guide for the young investors aiming to start bull fattening for meat production. Moreover, through the revolution of this sector we can also develop our leather sector more effectively without depending on other countries. There are very few livestock analysis systems based on this information from where the new investors and farmers would be able to clarify all their doubts and get a step closer in achieving their goals. Moreover, the farmers will gain the complete idea of the most profitable breed, based on their region, feedlot, and climate conditions [23]. The investors with minimum knowledge will be able to get optimum outputs by following this research, adds to the feasibility and demand of this topic. Moreover, this research will be a great business tool for the investors of both public and private sectors in Bangladesh.

5.1.2 Technological Feasibility:

Our system provides a technically sound and user-friendly platform, through which even the end users with minimum technical knowledge can learn and generate their desired outcomes. In Bangladesh, there is not yet any digital database, which records updated data of the livestock. As a result, the investors have to collect the hand written or analogue information from their nearest district livestock ministry office, which is quite a troublesome process. However, by only accessing our system, the investors will not only get all the digital information, but also a complete insight of trend analysis from that huge data set. The end user only needs to give input of the independent variables he/she has, and our system will find a match with that input of the other independent factors, and provide the desired output. The system will consist of all the latest features and a smooth graphical interface with both English and Bengali language support, which will give a good experience to our system users. The outputs of our trend analysis is generated by implementing the latest machine learning models and data mining tools, which will keep the end users updated with the ongoing trends around the world.

5.1.3 Economical Feasibility:

Analyzing the total cost of a hypothesis with the total revenue it can generate is the concept of economic feasibility [24]. The output of this research will give a trend analysis based on the given inputs by the user. These results will be displayed in a graphical and tabular format. People coming from all sectors of Bangladesh will have an easy access to this graphical analysis. Currently, we are not focusing for any business initiation with this project. However, our main motive at the moment is to develop this livestock industry of our country, which needs immediate digitalization. The main labor we had to incur during this research was while collecting data. However, we did not have to incur any financial major cost in this process. Basically, all the necessary data for our findings was collected from Bangladesh Livestock

Research Institute (BLRI), Department of Livestock services (DLS), Savar government dairy farm and Meghdubi Agro farm. Consequently, we do not require generating significant revenue from our project now. However, in future this model will become a very useful business tool for becoming beneficial from meat production in our country. Thus, this research can be said to be economically feasible from all perspectives.

5.1.4 Market Feasibility:

In this world of massive competition, the inclination towards livestock industry in increasing every day. Therefore, the youth investors or entrepreneurs and the farmers who are looking for the information to initiate their business in this sector, can be considered our prime target customers. Moreover, the existing Agro farm owners will also consider our system to be a beneficial tool for their business expansion, as it contains huge number of data sets and is technically updated. As this concept is still quite fresh in our country, our research and its outcomes will surely be successful in marking its place in the automated solutions' market and will act as an inspiration for our target customer base.

5.2 LIMITATIONS

While progressing through our research we came across multiple limitations that caused a hindrance for us in achieving the optimum outcomes.

Data collection: While carrying out the data collection process, we could only gather handwritten analogue data instead of having any access to online database record of livestock. As a result, we had to physically visit the private and government agro farms for collecting the available, handwritten data sets and which were later converted by us into digitized data sheets. Moreover, the analogue datasets collected did not have any proper orientation and were in scattered forms, which further added to our

difficulties. However, we were successful in collecting different breeds of cattle and their effective factors, which is the key portion of our research.

Short time lag: The agro farms did not keep the weight record on daily basis, they were able to provide us with the annual weight records only. This shortened the number of time lags of the dataset, that we required in order to carry out Time series analysis on the. We had only 20 lags in our dataset, where at least 30 lags were the least amount needed to implement the Time series analysis model. Furthermore, we would have been unable to predict the accurate growth rate and expected outcomes, with this short time lagged dataset.

Decentralized data sets: On visiting the Department of Livestock Services (DLS), we could not find any central database. They referred us to go to remote areas' government farms to collect the actual valid analog data sets. We faced troubles when we had to travel many places to collect data. On visiting some of the district-based offices, we discovered that the records they keep are quite scattered and unorganized.

Confidentiality of dataset: As the data sets are the key components for the business in livestock industry, very less agro farm owners were willing to share their confidential information related to the actual diet plan of a cattle. However, the diet plan for divided by specific breeds of cattle was later collected from the government sectors.

Limited Resource: As this research is still quite fresh, so we could find very limited research works related to this topic. If, we could study more number of past-related works, it would have benefit to the accuracy of our trend analysis.

Chapter 6

CONCLUSION AND FUTURE WORKS

This chapter draws a conclusion on the work that we have done till date and gives a picture of the future possibilities of our contribution in other sectors of the livestock industry as well. It describes the functionalities, which, upon addition, will convert the prototype developed in this research to an operational application.

6.1 CONCLUSION

In a developing country like Bangladesh, livestock industry makes up for a significant amount of world's livestock resources. Both the national economy as well as socio-economic growth of country is backed by the livestock sector. Through this research, a complete guideline is provided to the enthusiast investors in livestock sector and a technological, economic and market feasibility of this business sector is introduced to them as well. In addition, they will be able to make the best choice of cattle, based on their geographical regions and other significant features mentioned in this research. Livestock productions have also poverty alleviation benefits. The increasing demand for livestock products continues to be a key opportunity for poverty reduction and economic growth. In the future, livestock production will also help to enrich the leather business that will be very helpful for developing countries for making some valuable job sectors and for achieving foreign earnings by exporting leather products and meat as well. Moreover, if properly nourished, the livestock industry will not only be able to fulfil our national meat deficiency need, but also we will be able to earn by exporting the beef produced. This approach of livestock analysis through machine learning and regression analysis, will add a completely new dimension in automating the livestock agriculture.

6.2 FUTURE WORK

- **Increase data set:** In future, we would love to work on more breeds growth rate when we get the required information. Moreover, in the future we can approach our concept by implementing Time series analysis, by integrating all the data sets, provided we get sufficiently lagged datasets for our analysis.
- **More Accuracy:** In near future if we get the chance we will try to cover the prediction based on geographical regions of our country, which will be more economically feasible. Farmers will be able to easily get access from the regional data for almost all the breeds.
- **Deploying user interface:** As we are currently running program code to deliver the trend analysis, we aim to portray the research findings to the end users as an application or a website database system, in the near future.
- **Covering more sectors:** in the End, we would love to work on leather sector and dairy sector as both of these sectors have a closely related with our research findings.
- **Disease detection:** Some of the virus diseases can prove to be fatal for cattle breeding. If we can detect the probabilities of particular diseases, in due time, the farmers can take suitable steps for preventing them.

REFERRENCES

[1] Banglapedia.com, 'National Encyclopedia for Bangladesh', 2015. [Online]. Available: http://en.banglapedia.org/index.php?title=Livestock. [Accessed: 12 February 2015].

[2] R. Nair and R. Paul. "India's push to save its cows starves Bangladesh of beef", *Reuters*, p.5, July 3, 2015.

[3] S. Johnson, *Journal of ANIMAL SCIENCE*, vol.96, July 2018. [Online]. Available: https://academic.oup.com/jas . [Accessed July 21, 2018].

[4] G. Morota, FF. Silva, M. Koyama, "BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture, *Journal of ANIMAL SCIENCE*, vol. 96, pp. 1540-1550, April 2018.

[5] J. A. Arango, L. V. Cundiff, L. D. Van Vleck, "Covariance functions and random regression models for cow weight in beef cattle", *Journal of ANIMAL SCIENCE*, vol. 82, pp. 54-67, Jan. 2004.

[6] V. Goldberg, O. Ravagnolo, "Description of the growth curve for Angus pasture-fed cows under extensive systems", *Journal of ANIMAL SCIENCE*, vol. 93, pp. 4285-4290, Sept. 2015.

[7] O. Brian Allen J. H. Burton J. D. Holt, "Analysis of Repeated Measurements from Animal Experiments using Polynomial Regression", *Journal of ANIMAL SCIENCE*, vol. 57, pp. 765770, Sept. 1983.

[8] K. Meyer, "Random regressions to model phenotypic variation in monthly weights of Australian beef cows", *American Journal of Kidney Diseases*, vol. 65, pp. 19-38, July 2000.

[9] J. Anim, "Application of random regression models in animal breeding", *Livestock Production Science*, vol. 16, pp. 335-348, Jan. 2004.

[10] S. Niggol, Robert, “Measuring Impacts and Adaptations to Climate Change: A Structural Ricardian Model of African Livestock Management”, *Agricultural Economics*, vol. 38, pp. 151-165, Jan. 2008.

[11] D. J. Hand, "Data Mining: Statistics and More?" *The American Statistician*, vol. 52, no. 2, pp. 112–112, 1998.

[12] S. I. G. K. D. D. Blog, “Data Mining Curriculum: A Proposal,” *SIGKDD*.[Online].Available:http://www.kdd.org/curriculum/index.html. [Accessed: 21-Jul-2018].

[13] S. Ray and Business Analytics and Intelligence, “Essentials of Machine Learning Algorithms (with Python and R Codes),” Analytics Vidhya, 11-Mar-2018. [Online]. Available:http://www.analyticsvidhya.com/blog/2015/08/common-machine learningalgorithms/. [Accessed: 21-Jul-2018].

[14] “Ordinary least squares,” Wikipedia, 16-Jul-2018. [Online].Available:https://en.wikipedia.org/wiki/Ordinary_least_squares. [Accessed: 21-Jul-2018].

[15] “Ordinary Least Squares regression (OLS),” *Xlstat, Your data analysis solution*.[Online].Available:https://www.xlstat.com/en/solutions/features/ordinary-least-squares-regressionols. [Accessed: 21-Jul-2018].

[16] “Multivariate linear regression Tutorials & Notes | Machine Learning | HackerEarth,” *Innovation Management |HackerEarth Blog*. [Online]. Available:Https://www.hackerearth.com/practice/machine-learning/linear-

[17] http://www.public.iastate.edu/~maitra/stat501/lectures/MultivariateRegression.pdf.(2018).

[18] “Support Vector Machine,” Hierarchical Clustering. [Online]. Available:http://www.saedsayad.com/support_vector_machine.htm. [Accessed: 21-Jul-2018].

[19] Entropy: M. Somvanshi and P. Chavan, & quot;A review of machine learning techniques using decision tree and support vector machine," 2016 International Conference on Computing Communication Control and automation (ICCUBEA), Pune, India.

[20] J. R. Quinlal, "Induction of decision trees", *Machine Learning*, vol. 1, p.p 81-106, Jan. 1986.

[21] A. Bronshtein, "Train/Test Split and Cross Validation in Python – Towards Data Science," Towards Data Science, 17-May-2017. [Online]. Available:https://towardsdatascience.com/train-test-split-and-cross-validation-in-python80b61beca4b6. [Accessed: 21-Jul-2018].

[22] Hastie and Tibrishani, "Cross Validation and Bootstrap", vol.1, p.p. 14-19, Feb. 2009.

[23] B. Hochner, "A model for a comparative analysis of the evolution of Learning and Memory Mechanisms", *Biological Discovery in woods hole*, vol. 210, p.p. 308-317.

[24] "What is economic feasibility? Definition and meaning," BusinessDictionary.com. [Online]. Available: http://www.businessdictionary.com/definition/economicfeasibility.html. [Accessed: 21-Jul-2018].

[25] "Understanding the Results of an Analysis," Logistic Growth Curve -- AIDS Infections. [Online]. Available: http://www.nlreg.com/results.htm. [Accessed: 21-Jul-2018].

[26] "Support Vector Regression," Hierarchical Clustering. [Online]. Available:http://www.saedsayad.com/support_vector_machine_reg.htm. [Accessed: 21-Jul-2018].

[27] "Kernel Functions-Introduction to SVM Kernel & Examples," Data Flair, 12-Aug-2017. [Online]. Available: https://data-flair.training/blogs/svm-kernel-functions/. [Accessed: 21 Jul-2018].

[28] Kernel Svm.tripod.com. (2018). Support Vector Machine Regression. [Online] Available at: http://kernelsvm.tripod.com/ [Accessed 28 Jul. 2018].