
Model Selection in Diabetes Prediction

Master Thesis



Srividhya Hariharasudhan

Universität Trier, FB IV

Supervisor: Dr.Charlotte Articus

Submission: January 10, 2022

Contents

1	Introduction	2
2	Model Variable Selection	3
2.1	what is variable selection	3
2.2	Aspects of Variable Selection	4
2.3	Strategies	5
3	Variable selection methods	6
3.1	AIC	8
3.2	Properties of AIC	10
3.3	BIC	11
3.4	p-value	15
4	Application	15
4.1	Dataset Description	15
5	Data analysis	17
5.1	Model Variable selection in PIMA Dataset	21
	References	21

1 Introduction

In the recent years , the hospitals and many medical institutions are maintaining the patients medical health record electronically which improves the quality of the patient care and increases the efficiency of accessing the patient health easily and computer based analysis are more developed than the traditional manual analysis which enabled to be the great improvement for the monitoring the patient and giving suggestions easily and conveniently [6]. Diabetes Mellitus – DM which is also known as diabetes and it is one of the metabolic disorders with the raised blood glucose levels. The consumption of carbohydrates is converted into the type of sugar called the glucose and it will be released to the bloodstream [14]. Diabetes is considered to be the cause of many long term diseases like stroke, cardiovascular disease, heart attack,kidney failure. Overall around 122 million people are affected by diabetes and resulted in the increasing number of patients every day by day and also the deaths are increased everyday [9].Diabetes is closely related to the one of the two mechanisms : Insufficient production of the insulin that is the insulin produced is not used by the body efficiently and the cells in the body are very much sensitive to the action of the insulin [13].

Diabetes is classified into different types such as Type I diabetes, Type II diabetes , gestational diabetes.

1. Type I diabetes: This type of diabetes usually occurs in the young children, teens and you g adults whose age is less than 30 years of age [9].Usually this type of diabetes occurs when there is no insulin secretion in the body or presence of very less amount of insulin can cause diabetes which is mainly caused due to the less amount of insulin present in the body and it needs insulin to be injected into the body. Patients affected with this type of diabetes are also called as the insulin Diabetes Dependent patients. [6]. The symptoms of the Type I diabetes are thirst,constant hunger,weight loss,vision changes and fatigue [9]. According to the International diabetes Federation [7] there were around 4,97,100 people diagnosed with Type I diabetes worldwide in 2013 and number of newly diagnosed cases per year was 78900. The person with the less intake of protein are less resistant to the Type I diabetes , the more disease come into exist if the person have the intake of high amount of protein to their diet [15].

2. Type II Diabetes: This type of type I diabetes occurs in the adults over 45 years of age which are often associated with the persons having obesity, hypertension,asteriosclerosis, and other diseases [9]. Type II diabetes is mainly caused due to the resistance less in insulin in the body in other words the insulin in their body is not sufficient and if the patients body mass index is greater than 25 then there is greater risk of this type II diabetes. This diabetes is also called as the Non-Insulin dependent diabetes [6].Diet also plays a major role where the intake of carbohydrate should be replaced by the high-dietary fiber , the food rich in fiber is considered to be more effective against the type II diabetes and

it also have lower cholesterol levels. And the people with the family history of diabetes should be avoided with the intake of fatty foods and the vegetarians are associated with the lower risk of diabetes, cardio-vascular diseases and cancer [15]. The Type II diabetes can be easily avoided by maintaining the healthy diet, proper body weight and avoiding the use of Tobacco [13].

3. Gestational diabetes: The pregnant women are more affected by this gestational diabetes and it can be cured after the pregnancy period that is after the birth of the child. If the proper preventive measures and medications are not followed there are many chances of converting into the type II diabetes [6]. They are detected to have the abnormal glucose tolerance ranging from mild to high. Proper screening and diagnosis are needed to the people affected by gestational diabetes and also the relevant family history is considered to be the another cause of this type of diabetes. [3]. Mostly this type of diabetes is noted in the third tri-semester of pregnancy in most of the women with the genetic effect of the beta-cell function and this type of diabetes is also called as the MODY or the maturity onset diabetes [2].

The Medical health care system are loaded with lot of information. The predictions can be done using the data more precisely. This paper mainly focus on the PIMA dataset by National institute of diabetes and Digestive and Kidney diseases. The primary key area is to find whether the patient shows the sign of diabetes or not according to the World Health organization . The prediction model plays a important role in selecting the suitable variables for the model and proper variable selection methods are applied inorder to predict the most efficient variables for the inclusion in the model which is considered to be most important. [4]. In this paper we will discuss more deeply on the model selection strategies , the importance of the model selection and their advantages and their disadvantages and after finding the suitable variable for the prediction , the detailed description of the use of the PIMA dataset and how the PIMA dataset is used in the model selection, then proceeded by carrying out the model validation using the logisitic regression process with the inclusion of the efficient variable form the previous step and splitting the data into training and test. As the result the model is resulted with the best performance accuracy on both the training and the test data with predicting the patients having the diabetes or not.

2 Model Variable Selection

2.1 what is variable selection

The variable selection of purpose of finding out the best set of variables among all the variables by neglecting the variables which are not relevant or redundant and to be included in the model and it is considered to be the special case of

model selection.[4]. The variable selection methods provide solution to most of the important problems in statistics and there exist many variable selection methods to be used in the model for predictions , so the best performing models needs to be selected [11]. In simple words it can be stated that eliminating the insignificant variables and including the most significant variables in the model. The better way of making the predictions and also improvement on the performance of the model can be done using the variable selection and removal of many irrelevant variables, noisy variable from the dataset will also enhance the performance. Trying on all the possible combinations of the variables and selection of the best suited one is the finest way of performing the variable selection . For many computational reasons the variable selection are more relevant [1]. So appropriate variable are selected in order to avoid the noise and the outliers in the final model.

2.2 Aspects of Variable Selection

The large collection of complex data plays a vital role in the health care that solves many public health and the precision medical approaches with their huge source of data which tends to grow rapidly. [4]. There are many significant advantages of using the variable selection the first thing is that it increases the model performance by including the most relevant variables in the model , it contributes to predictive accuracy . It makes the model to be less complex where the storage and the processing time significantly reduces the cost involved in the overall process. The analysis is more informative and understandable and more knowledge can be gained because of inclusion of the significant variables in the model. There are cases where the huge set of variables contains only fewer relevant variables having the relevant information which is required for the model prediction in such cases it is important to consider only the important variables to carry out the task more efficiently [12]. There are usually several reasons in selecting the variables , so the large set of variables are not selected because the information about the variables may not be available sometimes and it is very expensive to collect and the simple models are preferred when compared to the complex models and also focus on the less computational time and its complexity . The importance in working with the variable selection is that important variables should not be excluded when performing the simple model [4].

There are usually more benefits in considering the variable selection that is it helps in the data visualization and more understanding of the data , it reduces the time consumption and increases the dimensionality in improving the model performance . The good predictor is build based on the selecting the good set of variables , mostly the variables should be redundant [5]. There is no rule in selection of variables to be included in the given model for the prediction however it depends on the certain parameters from the given dataset , sometimes there may be the situation where very few number of variables might

be needed or if more variables are needed for the given model this may leads to the over fitting . The relation between the variables and outcome exists if there are more variables in a model in respect to the number of observation or the dataset. The analysing of the data plays a vital role in using the actual variables in the final prediction model . The variable selection can be stated as the determination of the set of variables for the final model , for the making the model complete, the variable must be related to the outcome and the increase in the complexity and the decrease in the precision can be done by eliminating the insignificant variables and including the fewer set of variables and it also creates the balance between the fit and the simplicity [4].For the prediction model the candidate variables should be considered and also the variables are determined by analysing the data where it is used for two important reasons one is, it is helpful in determining all the variables related to the outcome and the another reason is fewer variables are selected for the model by eliminating the irrelevant variables , where it makes the model more accurate and more complete [4].

2.3 Strategies

The choosing of candidate variables has to be done at first in order to restrict the potential variables, where the candidate variables can be identified by the performance of the systematic reviews and by analyses . Based on the knowledge about the subject the candidate variables for the specific topic can be selected before the actual study is started . Depending on the performance with the outcome the candidate variables for the specific topic are demonstrated . A problem may occur in the applications of the model if the variables with the large number of missing values are included which may affect the persons who are working with due to the lack of estimation. In order to build the good prediction model usually 5-20 candidate variables are sufficient. [4]. From the set of variables for instance $x_1, x_2, x_3, \dots, x_{10}$ we will find to use the best set combination of the variables for example (x_1, x_4, x_7, x_{10}) which has increased information content and this also form the another form of the original variables . It is said that the it limits in building the best model because the variables without any structure gives more power to the model for prediction. A variable selection is considered to be good if it transforms the individual variable and the modelling and analysis provide the model with the good results and facilities the interpretation of the variables effect in the analysis. [11].

And also it is important to check the certain factors in the variable selection like over fitting , outliers, redundancy . if there are two or more similar variables , it is better to locate the variable closely . If the simple model is to be generated then the it is necessary to remove the redundant variables and sometimes removing the redundant variables causes the issue and therefore it is better to understand and interpret it easily and carefully . There may be the situation where the removal of the redundant variables also makes the performance low and probably will not improve the model.[1]. The variable selection has been

performed in order to select the best set of variables for the prediction and also it improves the compatibility of the data, less time for training and test , and it makes it easier to interpret for the users or the researchers, simplification of the models. The best set of variables for the model provides the reliable characterization of the sources for the scientific interpretation and the selected model should not be too sensitive to the sample size. Consistently selected models gives sufficiently many data samples and the predictive performance is considered to be the best and in case the complex models are selected then the predictions are unreasonable, unreliable and misleading , the data are different on which the selection are made.

It is always important to check whether the under the certain assumptions the variable selection works and provide the better solution and they also highlights which variables seems to be important and which seems to be unimportant, so with the understanding of the data, the analyst makes the clear decisions on how to implement the results of the variable selection. So using of the proper variable selection is the iterative process in order to provide the best results [1].The large number of variables sometimes becomes the time consuming to carryon the normal variable selection process and the small number of variables helps in identifying the good set of variables that are taken in our statistical model, the data on the performance illustrate more on the variable selection process. By removing the irrelevant data it allows the model to only focus on the important features of the data and it also reduces the computation time involved to fit the model. The model is considered to be more interpretable if it contains the smaller number of variables and overall the variable selection are being able to predict the values more accurately.

3 Variable selection methods

The selection of the variables to be further included in the model can be done once the candidate variables are identified from the list of all the available variable from the dataset. There are many ways in selecting the variable for the model. Once there contain many candidate variables and there are lot of uncertainty , in regard to which should be included in the final model then the formal variable selection are done . [4]. The variable selection methods plays an important role in providing the solution to the important problems in statistics and there exist many variable selection methods . The different subsets are usually produced by the different set of elements and in that different subsets is common with the certain number of variables and they are small in size and all the size of the subsets are different [11].The variable selection is considered to be the most important research areas in the field of statistics and it is difficult to apply for the high dimensional data and they are very flexible enough to work with many complex data. The variable selection methods usually includes forward selection,backward selection and stepwise selection And the stepwise

selection is usually the combination of both the forward and the backward selection and these methods work well in the many regression models .(Modern variable selection).

The backward elimination is considered to be the simplest when compared to the other variable selection methods. It diagnosis all the variables and it is included in the model and then it proceed by excluding the variables one by one from the full model till finding for the significant contribution to the outcome. The variables are considered to be less significant variables if the variables have the smaller test statistic which is less than the cut-off value are to be excluded . [4]. This is also the iterative approach and for the every iteration the predictor is removed from the model which produces the larger change in the information criterion value and if there is no predictor then there will be change in the information criterion value from that of the previous iteration and it terminates from the previous iteration [8]. The backward selection covers all the variables and at the end of each and every step it will try to eliminate the variables which is least important and which does not meet the criteria. The forward selection evaluates each variables and repeat the steps and finally selects the variables which improves the criteria the most are added to the outcome of the model(Modern variable selection techniques). The forward selection is the opposite of the backward elimination method .At each step the variables are excluded from the model and it is continously tested for the inclusion in the model and incase if the excluded variables are included in the model , depending on the calculation of the p value and the cut-off value it is selected and added to the model that is most significant variables are added to the model. The process continues until there are no variables remaining as significant and if the variables it added to the model it remains in the model. The advantage of the forward selection is that the process begins with the smaller models and it has very high intercorrelations among the independent variables but the backward elimination , forward selection also has the drawbacks. There is the possibility of making the existing variable insignificant by including the new variable in the model and the backward elimination and the forward elimination need to be balanced and the balance can be achieved in the stepwise selection [4]. The stepwise selection is the popular technique and it is used everywhere and it is mostly used in the medical field and this is said to be the combination of the forward and the backward selection procedure . It can either begin with the forward selection in which the variables are added to the model and after adding the model it need to be checked whether the variables included are statistically significant or not and this step continues until all the variables in the model are significant. if the stepwise selection starts with the backward elimination the variables are excluded from the model and it is added later when they appear to be significant and the less significant variables are excluded from the model as a whole for the stepwise selection the backward elimination is given more preference and in this full model is considered .

3.1 AIC

The Akaike information criterion is developed by the Japanese statistician . AIC predicts which model is likely to be expressed by the data from the sample. Usually in the statistics, AIC which is Akaike information criterion is the often used in the many model selection and usually after calculation by comparing the AIC values with the other possible models it is better to chose the one that best fits the data and the AIC usually lowers the number of variables in the model and p-values becomes the less which decreases the complexity of the model.(comparson of the variable selection methods for the clinical predictive modelling). AIC usually determines the value of the model using the maximum likelihood estimate and the number of the parameters which is considered to be the independent variables in the model and the formula for the AIC is

$$AIC = 2K - 2\ln(L) \quad (1)$$

Where k is the number of the independent variables and the L is considered to be the log estimate . AIC is need to be calculated for each model and if the model performance has AIC more than 2 units lower than the another then it is considered to be significantly better than that model and calculation of the AIC is very easy if the value of the log likelihood of your model is known. The performance of the AIC statistic are affected by the sample size in this case the AIC performs poorly if it contains too many parameters relative to the size of the sample (AIC and model selection). The AIC is used to estimate whether the prediction error is high or low with respect to the data outside the sample , for the given amount of data the quality or the statistical models can be easily accessed and AIC also plays a important role in the machine learning in the area of inference. How fast and precise the process tends to take place, the more precisely the model is likely to fit the available data and prediction or the detection rate is also more reliable.AIC is based on the information theory, sometimes there might be amount of information lost due to the process that represent the data which are used in the model, such information are lost due to model which is used to represent the process and the lost information can be easily estimated using the AIC , if the model loses very less amount of information then the quality of the model will be high . AIC has two important aspects for the model, one it represents the simplicity of the model and other aspect is the fit of the model depending on the amount of data available . there are certain risks which tend to occur which is if the model estimated too precisely to the amount of the data then there might be the chance of the over fitting and on the other hand under fitting which is the quality will be poor which can lead to the problem of inaccuracies. AIC usually does not care about the absolute quality of the model there will not be any issue if the candidate models does not match well and in order to determine the absolute quality of the model it is better to perform the validation including the residual values and testing can be using the model predictions. If you have number of candidate models for the given set of data , the preferred model would be the one with the lowest AIC value and AIC values the accuracy fit of the model with respect to the number

of the parameters and discourages the over fitting although the number of the parameters leads to the higher accuracy of the fit.

The AIC predicts the accuracy of the new data during the performance of the model there are important factors about considering the AIC which is all the process uses the same dataset and likelihood function can be estimated if the sample size is large enough and there is proper estimation of the parameters if the number of data n is sufficiently large it follows the multivariate normal distribution and the differences in the AIC allows for the ranking of the candidate models and if the difference in the AIC difference of the model is large then there is the less probability that it is the best model .(model selection and model averaging in phylogenetics advantages of akaike information criterion and Bayesian approach over likelihood ratio tests). The process of combining the estimation with the structural and the dimensional determination into the single procedure highlights the idea of the Akaike information criterion. The complexity of the model is increased and the model becomes more capable of adapting to the characteristics of the data and selection of the most fitted model maximizes the empirical likelihood .

The AIC has many advantages , that is applicable to all the range of the modelling frameworks and the criterion does not require the assumption that one of the candidate models is true or correct model and also it is used to compare the non-tested models . AIC is used to compare the models based on the different probability distributions , while the criterion values are computed , from the good fit term no constants should be removed as the result the minimum AIC value is identified as the optimal fitted model and the models with the similar values should also receive the same ranking in assessing the criterion preferences. Akaike information criterion is used in many statistical models compares of both the theoretical and the empirical aspects and involved in specifying the significance level in testing the models and there is no restriction in the comparison between the models and the AIC can be calculated once the maximum likelihood estimators of the parameters are calculated , among the several models the AIC is chosen to be the best fitting model and minimum value of the AIC is chosen to be the best fitting model and emphasizes the goodness of the model. The most important aspect of the AIC is to identify the true model and the true model does not always mean that it is the best fitting model and it is mostly in approximate to the true model and the model can be estimated more accurately with the larger sample size parameters and best fitting model can change the function of the sample size .(Akaike information criterion (AIC)- Introduction.

The computation of the AIC takes place between the goodness of the fit and the model variability where it is easier to rank the different models in order to select the best performing models where the highest complexity of the model can be found by incorporating the best performing model in the model selection and the AIC model statistics stated that the full models containing all the

possible values are not well defined . Models having the very less parameters are considered to be under-fit which subjects to the loss of information in the model and model having too many parameters are considered to be over-fitted and lack of the precision and therefore it involves the relation between the bias and the variance and also deals with depending on estimation of the number of parameters , the level of the model accuracy can be reached to optimum level . The AIC models can be build using different methods which deals with the model selection uncertainty considered to be the greatest strength and usually the AIC is the sum of the bias and the uncertainty which are called as the penalty terms and always the inclusion of the parameters always increases the likelihood score and the over parameterized model is not included and the over-fitted parameters are estimated to have the higher AIC values. (An information theory approach to hypothesis testing in criminology research).

Use of AIC in different scenarios: According to this (model selection and akaike information criteria :An example from the wine rating and the prices) demonstrates first time in the wine research how well the AIC is used in comparison with the different models and describes the relationship between the ratings and the prices of the wines. AIC plays an important role in order to identify the new comings in the wine ratings and how the variables are influenced in the determining the ratings of the wine which includes the price,what type of wine used and from which region the wine is produced also being the part of the wine determination . The selection of the model is important as the under-fitting model does not give the variability in the outcome variable and the over-fitting models has lost the generality . The AIC is used mainly in different fields like biological science, environmental, pharmacological sciences and many other where they are significantly used in improving the model development and in many areas of research.

3.2 Properties of AIC

According to (Properties of the Akaike information criterion) there are general properties of the AIC for both the discrete and the continuous models

Theorem 1: States that the from some distribution $x = (x_1, x_2, x_3 \dots x_n)$ is considered to be the random sample with distribution function $f(x, \theta)$, the number of parameters in θ is k and $S = S(x_1, x_2 \dots x_n)$

Is said to be sufficient statistic then the $AIC(x) \neq AIC(S)$ where the $AIC(x)$ and the $AIC(S)$ are the AIC in the random sample x and in statistic S respectively thus the AIC violates both the likelihood principle and the principle of the invariance . Proof: The factorization theorem the joint distribution function of $x = (x_1, x_2 \dots x_n)$ is

$$F(x; \theta) = k(x).g(S(x); \theta) = A(S(x)).g(S(x); \theta) \quad (2)$$

Moreover, the Maximum likelihood estimate for theta is a function of S(x) hence,

$$AIC(x) - AIC(S) = -2[\log f(x; \theta) - \log g^s(S(x); \theta)] \quad (3)$$

$$= -2\log \left(\frac{k(x)}{A(S(x))} \right) \quad (4)$$

which is not zero.

Theorem 2 : AIC satisfies the compositivity and the convexity properties i.e. if $f(x) = \alpha f_1(x) + (1 - \alpha)f_2(x)$ where $0 < \alpha < 1$ and α is known then
1) $AIC(f) = \psi AIC(f_1, AIC(f_2), \alpha)$ for some function ψ and
2) $AIC(f) \leq \alpha AIC(f_1) + (1 - \alpha)AIC(f_2)$.

Proof: The results follow the fact that the logarithmic function is concave.

Theorem 3 : AIC is not strongly additive and it is not weakly additive

Proof: It is known that if h and q are two functions on the same domain and each has a maximum value on its domain then

$$\max_w h(w) + q(w) \neq \max_w h(w) + \max_w q(w). \quad (5)$$

AIC is based on the combined sample of size $n_1 + n_2$ is not the sum of the two values of AIC obtained from the O_{n_1} and O_{n_2} respectively. The AIC is not weakly additive is that it is based on the sample from the bivariate distribution of (X,Y) and is not the sum of AIC based on the sample from the marginal distribution of X and AIC is based on the sample from the conditioned distribution of Y—X. The above mentioned properties gives the theoretical interpretation and the information criterion is defined where the AIC is almost applied everywhere to solve the statistical problems and leads to AIC.

3.3 BIC

BIC is the Bayesian information criterion and it is the another form of the variable selection method and it is similar to the AIC , different parameters are followed for the number of variables to be included in the model . There are both the similarities and the differences exists between the AIC and the BIC , the similarity is considered to be the balancing criteria between both the simplicity and the model fit . In general, BIC is calculated for each of the candidate models from the given dataset and value corresponding to the minimum BIC is chosen. The best model is selected to maximise the predictive accuracy and represents the true relation and it is consistent. Ultimately the goal is to select the best model from among the candidate models that represents the true model. (Variable selection strategies and its importance in the clinical prediction modelling). The Bayesian information criterion is derived as

$$BIC = 2\ln(L) + K \log(n)$$

It is stated that if the true model is present in the set of the models the BIC selects the true model with the probability 1 and n always tends to infinity and it really concentrates on having the true model in the set of the candidate models. BIC has the possibility in increasing the likelihood by adding the parameters by adding so it may result in the overfitting but the BIC reduces this problem by introducing the penalty term equivalent for the number of parameters in the model. The penalty term is larger in BIC than in the AIC and most closely related to the AIC. The value of the BIC increases with respect to the number of the explanatory variables and value with the lower BIC implies either the fewer explanatory variables ,better fit or the both the criteria's . When the numerical value of the dependent variables are identical for the all the estimates then the BIC compares the estimated model. The BIC are used in the time series identification and the linear regression and most likely applied to the set of maximum likelihood based models and also in many different applications. The BIC has the property of consistency which indicates the candidate modes to be specified correctly and the BIC always choose the models that are more parasimonious than the AIC in reference with small to the moderate sample size settings and the BIC model converges with the probability one. (The Bayesian information criterion: background, derivation, and applications). BIC provides the large-sample estimator of a transformation of the Bayesian posterior probability associated with the approximating model and by choosing the best candidate model which is corresponding to the minimum value of the BIC, it is better to select the candidate model corresponding to the highest posterior probability.

According to this paper [10] the the Bayesian information criterion is derived using a data y described using the model selected from the set of the candidate models $M_{k_1}, M_{k_2} \dots M_{k_L}$. The candidate models M_k , for k belongs to $(k_1, \dots k_L)$. The derivatives of the likelihood function $L(\theta_k|y)$ upto order two exist with respect to θ_k , and are continuous and suitably bounded for all $\theta_k \in \Theta(k)$. Let $\pi(k), k \in k_1, \dots, k_L$, denote a discrete prior over the models $M_{k_1}, M_{k_2}, \dots, M_{k_L}$. Let $g(\theta_k|k)$ denote a prior on θ_k given the model M_k . Applying Bayes Theorem, the joint posterior of M_k and θ_k can be written as

$$h((k, \theta_k) | y) = \frac{\pi(k)g(\theta_k | k) L(\theta_k | y)}{m(y)}$$

where $m(y)$ denotes the marginal distribution of y . A Bayesian model selection rule aims to choose the model which is a posteriori most probable. The posterior probability for M_k is

$$P(k | y) = m(y)^{-1} \pi(k) \int_{\Theta(k)} L(\theta_k | y) g(\theta_k | k) d\theta_k$$

Now consider minimizing $2 \ln P(k|y)$ as opposed to maximizing $P(k|y)$. We have

$$-2 \ln P(k | y) = 2 \ln \{m(y)\} - 2 \ln \{\pi(k)\} - 2 \ln \left\{ \int L(\theta_k | y) g(\theta_k | k) d\theta_k \right\}.$$

The term involving $m(y)$ is constant with respect to k ; thus for the purpose of model selection, this term can be discarded. We define

$$S(k | y) = -2 \ln \{\pi(k)\} - 2 \ln \left\{ \int L(\theta_k | y) g(\theta_k | k) d\theta_k \right\}.$$

Consider the integral that appears in Eq. (6). In order to obtain an approximation to the integrant, We have second order Taylor series expansion of the log-likelihood about $\hat{\theta}_k$. we have

$$\begin{aligned} \ln L(\theta_k | y) &\approx \ln L(\hat{\theta}_k | y) + (\theta_k - \hat{\theta}_k)' \frac{\partial \ln L(\hat{\theta}_k | y)}{\partial \theta_k} \\ &\quad + \frac{1}{2} (\theta_k - \hat{\theta}_k)' \left[\frac{\partial^2 \ln L(\hat{\theta}_k | y)}{\partial \theta_k \partial \theta_k'} \right] (\theta_k - \hat{\theta}_k) \\ &= \ln L(\hat{\theta}_k | y) - \frac{1}{2} (\theta_k - \hat{\theta}_k)' \left[n \bar{\mathcal{I}}(\hat{\theta}_k, y) \right] \\ &\quad \times (\theta_k - \hat{\theta}_k), \end{aligned}$$

where

$$\bar{\mathcal{I}}(\hat{\theta}_k, y) = -\frac{1}{n} \frac{\partial^2 \ln L(\hat{\theta}_k | Y_n)}{\partial \theta_k \partial \theta_k'}$$

is the average observed Fisher information matrix.

$$\begin{aligned} L(\theta_k | y) &\approx L(\hat{\theta}_k | y) \\ &\times \exp \left\{ -\frac{1}{2} (\theta_k - \hat{\theta}_k)' \left[n \bar{\mathcal{I}}(\hat{\theta}_k, y) \right] (\theta_k - \hat{\theta}_k) \right\}. \end{aligned}$$

We therefore have the following approximation for the integral in Eq. (2):

$$\begin{aligned} &\int L(\theta_k | y) g(\theta_k | k) d\theta_k \\ &\approx L(\hat{\theta}_k | y) \int \exp \left\{ -\frac{1}{2} (\theta_k - \hat{\theta}_k)' \left[n \bar{\mathcal{I}}(\hat{\theta}_k, \right. \right. \end{aligned}$$

$$y \left(\theta_k - \hat{\theta}_k \right) \times g \left(\theta_k \mid k \right) d\theta_k \quad (6)$$

The Taylor series approximation in Eq. (3) holds when θ_k is close to $\hat{\theta}_k$. Thus, the approximation in Eq. (4) should be valid for large n . In this instance, the likelihood $L(\theta_k \mid y)$ should dominate the prior $g(\theta_k \mid k)$ within a small neighborhood of $\hat{\theta}_k$. Outside of this neighborhood, $L(\theta_k \mid y)$ and the exponential term should be small enough to force the corresponding integrands in Eq. (4) near zero. Therefore, it is defensible to simplify the justification by using the noninformative prior $g(\theta_k \mid k) = 1$. In this case, we can evaluate the second integral in Eq. (4) as

$$\begin{aligned} & \int \exp \left\{ -\frac{1}{2} \left(\theta_k - \hat{\theta}_k \right)' \left[n \bar{\mathcal{I}} \left(\hat{\theta}_k, y \right) \right] \left(\theta_k - \hat{\theta}_k \right) \right\} g \left(\theta_k \mid k \right) d\theta_k \\ &= (2\pi)^{(k/2)} \left| n \bar{\mathcal{I}} \left(\hat{\theta}_k, y \right) \right|^{-1/2} \end{aligned}$$

This leads to an approximation of the first integral in Eq. (4) as

$$\begin{aligned} & \int L \left(\theta_k \mid y \right) g \left(\theta_k \mid k \right) d\theta_k \\ & \approx L \left(\hat{\theta}_k \mid y \right) (2\pi)^{(k/2)} \left| n \bar{\mathcal{I}} \left(\hat{\theta}_k, y \right) \right|^{-1/2} \\ & = L \left(\hat{\theta}_k \mid y \right) \left(\frac{2\pi}{n} \right)^{(k/2)} \left| \bar{\mathcal{I}} \left(\hat{\theta}_k, y \right) \right|^{-1/2} \end{aligned}$$

The former can be viewed as a variation on the Laplace method of approximation. The approximation in Eq. (5) is valid so long as $g(\theta_k \mid k)$ is noninformative or 'flat' over the neighborhood of $\hat{\theta}_k$ where $L(\theta_k \mid y)$ is dominant (see Cavanaugh and Neath⁶ for the formalities), although the choice of $g(\theta_k \mid k) = 1$ makes the derivation more tractable. We can now approximate $S(k \mid y)$ in Eq. (2) as

$$\begin{aligned} S(k \mid y) & \approx -2 \ln \{ \pi(k) \} \\ & -2 \ln \left[L \left(\hat{\theta}_k \mid y \right) \left(\frac{2\pi}{n} \right)^{(k/2)} \left| \bar{\mathcal{I}} \left(\hat{\theta}_k, y \right) \right|^{-1/2} \right] \\ & = -2 \ln \{ \pi(k) \} - 2 \ln L \left(\hat{\theta}_k \mid y \right) + k \left\{ \ln \left(\frac{n}{2\pi} \right) \right\} \\ & \quad + \ln \left| \bar{\mathcal{I}} \left(\hat{\theta}_k, y \right) \right| \end{aligned}$$

Ignoring the terms in Eq. (6) that are bounded as the sample size grows to infinity, we obtain

$$S(k | y) \approx -2 \ln L(\hat{\theta}_k | y) + k \ln n.$$

With this motivation, the BIC is defined in Eq. (1) as an asymptotic approximation to $-2 \ln P(k | y)$, a transformation of the Bayesian posterior probability of model M_k .

The important property of BIC is consistency and it can measure the efficiency of the parameterized model in terms of predicting the data and it is widely used in the model identification in the time series and in the linear regression and applied quite to the maximum likelihood based models and reduces to maximum likelihood selection because the number of parameters is equal for the models of the interest. BiC is used to compare the estimated models where the dependent variables are identical for all the estimates.

3.4 p-value

The significance level of the pvalue is 0.05 or 0.10 and depending on the p-value the variable to be included or excluded in the model is decided and it is suggested to be 1. The higher significant value need to be considered and included in the model as it relevant to the outcome and excluding the less significant variables may lead to the practical reasoning [4].

4 Application

4.1 Dataset Description

The dataset used for the study is PIMA Indian dataset by National institute of diabetes and digestion. The dataset consist of 768 rows and 9 columns having 268 positive for diabetes and 500 negative for diabetes and there are 8 variables taken as the indicators of the dataset and the variable outcome stated that whether or not the person has the diabetes by showing the result value as 0 for NO and 1 for YES. The objective of the dataset is to diagnostically predict whether the patient has the diabetes based on the certain diagnostic measurements included in the dataset. The variables that are involved in the prediction of the diabetes are as follows the patients has had the Pregnancies, BMI, insulin level, age, Blood pressure, Skin thickness, Glucose, Diabetes Pedigree Function the label outcome.

PIMA Indian Dataset		
S.No	Attribute	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration by oral glucose tolerance test
3	Blood Pressure	Diastolic blood pressure(mmHg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2 Hour serum Insulin(mm U/ml)
6	BMI	Body Mass Index(BMI)
7	DiabetesPedigreeFunction	Diabetes Pedigree function
8	Age	Age(in years)
9	Outcome	class variable(0 or 1)

The pregnancies is calculated depending on the number of times pregnant. The data required for the glucose is done by using the oral Glucose tolerance test result which checks how your body has the sugar from the blood into the tissues . This test is used in the diagnosis of diabetes.It predicts how the sugar present in the blood like muscle and the fat.There are certain important factors to be considered for the test the most popular glucose tolerance test is the oral glucose tolerance test and the factors includes before starting the test the sample of the blood will be taken . It is necessary to drink a water which contains the glucose which is approximately 75 grams . After consuming the water for about 30 to 60 minutes after the blood will be taken for the diagnosis this is how the oral glucose tolerance test is carried out.

Types	Normal persons	Criteria for Diagnosing Diabetes	Criteria for Diagnosing Impaired Glucose Tolerance(IGT)
Fasting	< 110 mg/dl	> 126 mg/dl	110 to 126 mg/dl
1 hr (after glucose administration)	< 160 mg/dl	Not Prescribed	Not Prescribed
2 hr(after glucose administration)	< 140 mg/dl	> 200 mg/dl	140 to 199 mg/dl

The BloodPressure is diagnosed and measured based on the diastolic Blood pressure values in terms of (mm Hg). The value is recorded during the time where the pressure in the arteries and also when the heart rests between the

beats and filled with blood and gets the oxygen . It is said to be normal when the heart bear is lower than 80 and for hypertension the rate is around 80-89 ad the hypertension crisis is said to have 120 or more . Through this value the diastolic pressure values are measured and the people with high blood pressure is tend to be more likely having the Diabetes. The Skinfold thickness is predicted based in the total amount of the body fat and the triceps skinfold is calculated based on the upper arm muscle circumference and amount of thickness is gives the information about the amount of fat present in the body. The value of calculated muscle mass gives the information about the protein present. The value of the triceps skinfolds varies and for example for adults the value of the triceps skinfolds are 2.5mm for men and 18.0mm for women.If the value range is above 15% percent it is considered to be either borderline or presence of fat. And the value over 20mm for men and 30mm for women is said to be about 85% and can be considered. Insulin is a type of hormone taken from the bloodstream from the cells and it is helpful in the blood sugar levels. and if it is greater than 150 mu U/ml it is related to the insulin therapy where is the present in the people having type 1 diabetes and the type2 diabetes and it is mainly targeted to keep the blood sugar level in the fixed range. The BMI is the body mass index is the way of finding where the patient is either over weight or the under weight and if the patient is having the greater body mass index, it is said as excess body and moreover the athletes are considered to be having over weight and their result of the skin fold test shows the presence of the normal amount of the adipose tissue and is usually done by using the patients weight in terms of (kg) and the square of their height in terms of (meters). BMI is calculated based on

$$\text{weight(kg)}_{\text{height(m)}^2}$$

Category	BMI (kg/m2)
Underweight	< 18.5
Normal weight	18.5-24.9
Overweight	25.0-29.9
Obesity(class I)	30.0-34.9
Obesity(class II)	35.0-39.9
Obesity(class III)	> 40.0

The data related to the Diabetes Pedigree function represents the history among the family relatives and the genetic relationship of those relatives to the patients. The influence of the genetic might have the risk of the diabetes mellitus.The age of the patients is calculated in years and the outcome represents the class 1 indicates the person having the diabetes and class 0 indicates the person not having diabetes.

5 Data analysis

The diabetes dataset is taken from the PIMA dataset and the data preprocessing and data cleaning methodologies are used for cleaning the dataset. The

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree-Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	Diabetes
1	85	66	29	0	26.6	0.351	31	Not Diabetes
8	183	64	0	0	23.3	0.672	32	Diabetes
1	89	66	23	94	28.1	0.167	21	Not Diabetes
0	137	40	35	168	43.1	2.288	33	Diabetes

Table 1: Record of PIMA Dataset

PIMA Indians Diabetes contains the patients details with the outcome of Diabetes and the Non-diabetes . The prediction of the Diabetes is predicted using the patients information from the dataset. The diabetes mellitus contains 768 records and the dataset contains such as number of times pregnant, plasma glucose concentration, diastolic blood pressure, Triceps skinfold thickness, 2-hour serum insulin, Body mass index, Diabetes pedigree function, Age in years, Class variable ie that describing diabetes or not as presented in the table 1 [14]

The missing and the null values are checked in the dataset where it results that there is no missing values and the values from the dataset and further proceed with the duplication of the values in the dataset, since there is no duplicate values to be founded in the dataset such that the new dataset contains the unique 768 rows which has the same number of records with our original dataset. Both contains 768 rows and thus conclude again having no duplicated values in the dataset. The next step proceed with the testing of the outliers with each of the variable.

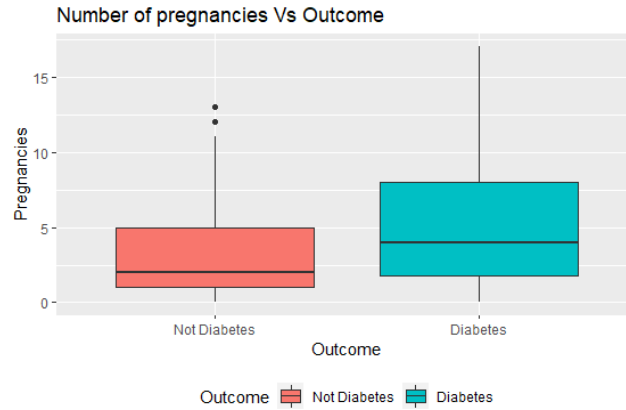


Figure 1: Pregnancies Vs Outcome

The figure 1 shows the outlier detection for the number of pregnancies com-

pared with the outcome results that the women in the non-diabetic group have fewer pregnancies compared to those who are in the diabetic group and the distribution of the pregnant women in non-diabetic group is skewed to right.

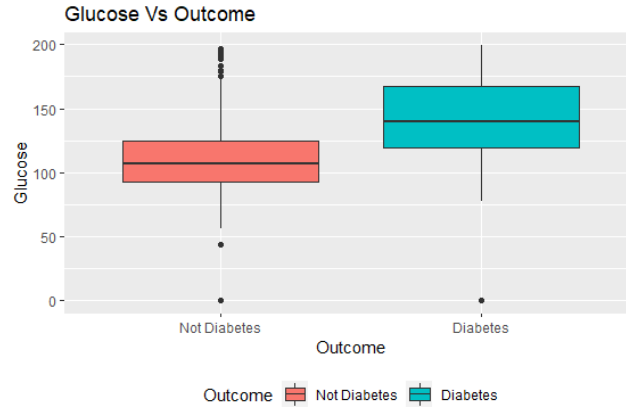


Figure 2: Glucose Vs Outcome

The figure 2 depicts the outlier detection for the glucose rate with the outcome shows that the diabetic women appear to have higher glucose concentrations and more outliers are detected in the non-diabetic women.

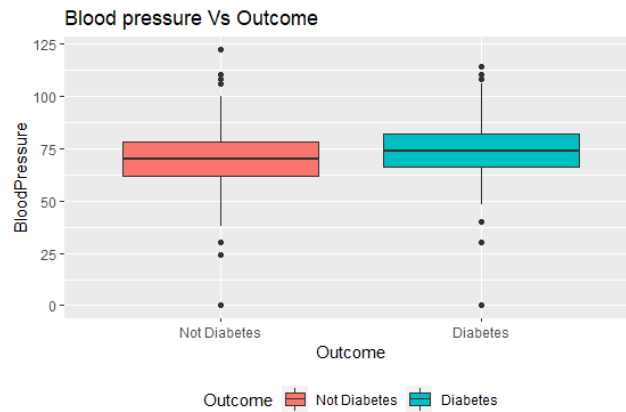


Figure 3: Blood Pressure Vs Outcome

The figure 3 gives the outlier detection for the blood pressure shows that the two groups that is the blood pressure and the outcome have the similar blood pressure measurement and there are zero values in both the non-diabetic and the diabetic group.

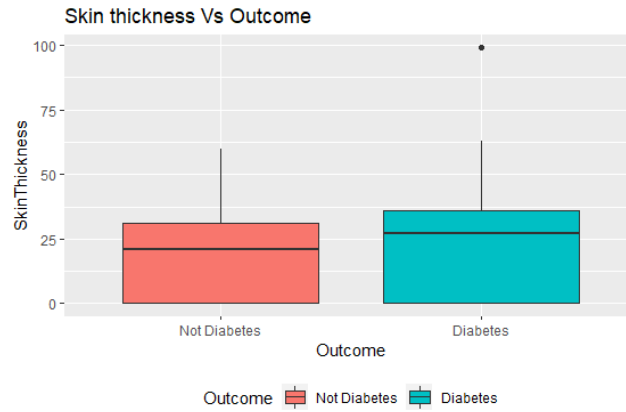


Figure 4: Skin Thickness Vs Outcome

The figure 4 depicts the outliers detection of skin thickness with respect to the outcome shows that there are very fewer number of outliers in the diabetic group and found to have more zero values in both the diabetic and the non-diabetic group.

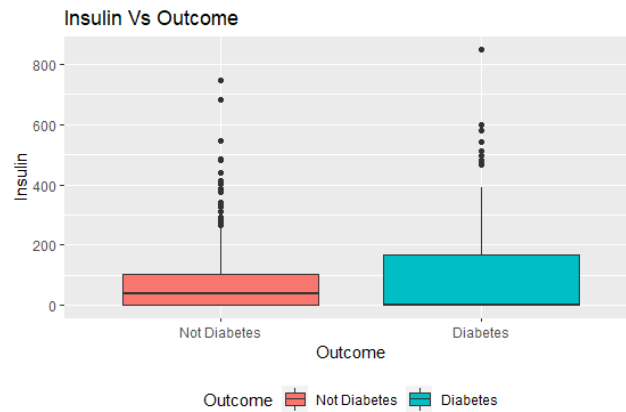


Figure 5: Insulin Vs Outcome

The figure 5 shows the detection of the insulin with respect to the outcome and there are many outliers from both the groups especially women with diabetes are heavily skewed to the right and there detects the zero values in the both the groups .The BMI the diabetic women have slightly higher BMI than other group . The pedigree distribution function in both the groups have the outliers and have positive skew. The average age of the women in the diabetic group seems older than the women in the non-diabetic group. At last there are invalid values to be found in some of the columns and that is the

dataset is found to be incomplete so in order to make the dataset more relevant and reasonable, the rows contained the zero values in Blood Pressure, Glucose, Skin fold Thickness, Insulin and the BMI variables are replaced with the median value which is the midpoint of the frequency distribution of the observed values, the value of the mean is not taken into account because it does not reflect the reality in the observation. The next process of the explanatory data analysis proceed with the checking of the correlations between the variables.

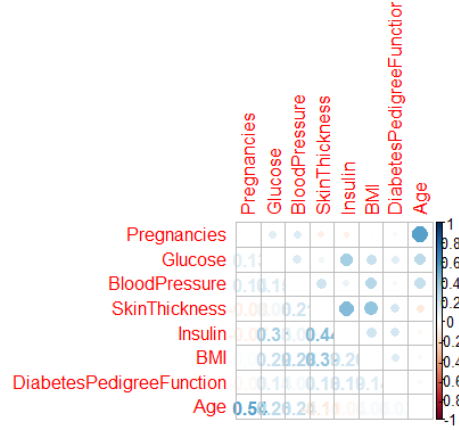


Figure 6: Coorelation matrix

In the figure 6 it is observed that the Skinthickness might have correlation with the BMI. The coefficient in the correlation plot shows that the 0.57 between BMI and Skin Thickness which means there is moderate positive relationship. It also shows correlation coefficient of 0.54 between age and Pregnancies and 0.49 between Insulin and Glucose . These coefficients measures the strength and the direction between the two variables. However these coefficients from the scatter plot are not strong enough to assure that there are significant relationship among the covariates so further analysis can be done without dropping the columns. Also the distribution can be visualized from the ggplot function in each regressor. Blood pressure and BMI seems to follow the normal distribution. Pregnancies, Age, Insulin and Diabetes Pedigree function are skewed to the right.

5.1 Model Variable selection in PIMA Dataset

The variables which highly predict the cause of diabetes from the PIMA dataset that influences the outcome. To find out which of these variables are important for predicting the relationship between the chance of diabetes and the outcome we predict the variables using the several variable selection such as p-value, AIC, BIC. The prediction of the model is done by how each variable performs followed by the combination of the variables. At first the model selection is performed by taking the combination of the dependent variables with respect to the independent variables, as a result of this performance the

References

- [1] Charlotte Møller Andersen and Rasmus Bro. Variable selection in regression—a tutorial. *Journal of chemometrics*, 24(11-12):728–737, 2010.
- [2] Habtamu W Baynes. Classification, pathophysiology, diagnosis and management of diabetes mellitus. *J diabetes metab*, 6(5):1–9, 2015.
- [3] Thomas A Buchanan, Anny H Xiang, et al. Gestational diabetes mellitus. *The Journal of clinical investigation*, 115(3):485–491, 2005.
- [4] Mohammad Ziaul Islam Chowdhury and Tanvir C Turin. Variable selection strategies and its importance in clinical prediction modelling. *Family medicine and community health*, 8(1), 2020.
- [5] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.
- [6] Musavir Hassan, Muheet Ahmad Butt, and Majid Zaman Baba. Logistic regression versus neural networks: the best accuracy in prediction of diabetes disease. *Asi. J. of Comp. Sci. and Tech*, 6:33–42, 2017.
- [7] Akram T Kharroubi and Hisham M Darwish. Diabetes mellitus: The epidemic of the century. *World journal of diabetes*, 6(6):850, 2015.
- [8] Charles Lindsey and Simon Sheather. Variable selection in linear regression. *The Stata Journal*, 10(4):650–669, 2010.
- [9] Md Maniruzzaman, Md Jahanur Rahman, Benojir Ahammed, and Md Menhazul Abedin. Classification and prediction of diabetes disease using machine learning paradigm. *Health information science and systems*, 8(1):1–14, 2020.
- [10] Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.
- [11] Bruce Ratner. Variable selection methods in regression: Ignorable problem, outing notable solution. *Journal of Targeting, Measurement and Analysis for Marketing*, 18(1):65–75, 2010.
- [12] Sailee Rumao. *Exploration of Variable Importance and Variable selection techniques in presence of correlated variables*. Rochester Institute of Technology, 2019.
- [13] Suyash Srivastava, Lokesh Sharma, Vijeta Sharma, Ajai Kumar, and Hemant Darbari. Prediction of diabetes using artificial neural network approach. In *Engineering Vibration, Communication and Information Processing*, pages 679–687. Springer, 2019.

- [14] Prasannavenkatesan Theerthagiri, J Vidya, et al. Diagnosis and classification of the diabetes using machine learning algorithms. 2021.
- [15] Jaakko Tuomilehto and Eva Wolf. Primary prevention of diabetes mellitus. *Diabetes care*, 10(2):238–248, 1987.