Chapter 3: Methodology

3.1 Diabetes

3.1.1 Analysis of the diabetes dataset

3.1.1.1 Sources of data on diabetes

The Pima Indians Diabetes Dataset used in this paper is downloaded from Kaggle (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database), the data came from the UCI Machine Learning Repository. The Pima Indians in the Phoenix, Arizona neighbourhood have a high prevalence of diabetes and since 1965 the population has been used by the National Institute of Diabetes to study digestive and kidney diseases. Every community resident over the age of 5 is required to undergo a standard and comprehensive screening test every two years, which includes a glucose tolerance test for oral glucose. Diabetes is diagnosed according to WHO diagnostic guidelines, i.e., if the blood glucose concentration is greater than 200 m g/dL after two hours of glucose administration or if the community health service hospital finds that the glucose concentration in a routine medical examination is at least 200 mg/l (Bennett et al. ,1971).

3.1.1.2 Description of diabetes dataset

The dataset included a total of 768 women who were Pima Indians living near Phoenix, Arizona, and all were over the age of 21 (Smith et al. ,1988). There are nine attributes in this dataset, including eight feature attributes and a binary category label. The first five samples of this dataset are shown in Table 4.1.

Table 4.1: The first five rows of the dataset that were not processed.

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DPF	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

A detailed description of each attribute is shown in Table 4.2.

Table 4.2: Attribute Introduction

Attribute Number	Attribute Name	Detailed Description
1	Pregnancies	The number of pregnancies per woman
2	Glucose	Blood glucose concentration 2 hours after oral glucose tolerance test
3	Blood Pressure	Diastolic blood pressure (mm HG)
4	Skin Thickness	Triceps skinfold thickness(mm)
5	Insulin	Two hours of serum insulin (μ U/ml)
6	BMI	Body mass index (Weight Kg/(Height m)2）
7	DPF	Diabetes Pedigree Function
8	Age	Age of the sample
9	Outcome	Classification results of diabetes

The diabetes pedigree function uses family history of diabetes to derive an individual's risk value for diabetes (Smith et al. ,1988). The eight feature attributes of each sample are numerical, and the outcome attribute is nominal. Among the outcome attribute of the dataset, 1 represents diabetic patients and 2 represents non-diabetic patients. Simple statistic was performed on the outcome attribute in the dataset. As shown in Table 4.3, among all the samples, there were 268 patients with diabetes, accounting for 34.89%, and 500 patients without diabetes, accounting for 65.11%.

Table 4.3: Number of people with and without diabetes

outcome variable	Number of samples	The percentage
0	268	34.89%
1	500	65.11%

3.1.2 Data Processing

In a data set, there are usually problems with the data, such as noisy data that can cause classification errors, incorrect data that needs to be identified, duplicate data that may have been entered, and so on, so a bad data set can have a significant impact on the classification results.

For these problems, if the original data is used directly for modelling, the results are often inaccurate and may not be meaningful. Firstly, for incorrect or abnormal data (deviations from the expected value), it is common practice to delete the incorrect or abnormal data to ensure consistency within the data; secondly, for incorrect data, i.e., the input data does not match the valid values of the fields, it can be deleted or replaced by the mean or median of the fields; Thirdly, duplicate values can also be removed for repeated input data to leave separate valid data; Third, for duplicate input data, duplicate values can also be removed to leave separate valid data; fourth, in terms of missing data, it may be found that within the attributes, there is a certain amount of data missing or there are null values, so in the case of judging the validity of other characteristics of this sample, deleting the sample values may cause a reduction in the sample size, and in order to make this part of the information valid, it is common to use a numerical value to replace it, such as the mean or median. Fifth, for high-dimensional data, that is, data with multiple attributes with redundant information, the data needs to be dimensioned down.

3.1.3 Dataset Analysis

3.1.3.1 Missing value detection

Missing value checking of data is about checking whether there are missing values in the various attributes of the data and how these missing values are dealt with subsequently.

There are three ways to deal with missing values: removing them, filling them manually and filling them automatically. Commonly used methods for filling missing values are mean interpolation, high-dimensional mapping, multiple interpolation, and so on. In mean interpolation, if the null value is numeric, the missing attribute value is filled in based on the mean of the values of that attribute for all other objects. In high-dimensional mapping, it retains all the information of the original data without adding any additional information, but the disadvantage is that it is much more computationally intensive and only works well when the sample size is very large. In multiple interpolation, m reasonable interpolation values are generated for each missing value, so that m sets of complete data are obtained after interpolation, and each set is analysed to fuse the results.

The detection of missing values for the Pima diabetes dataset in this paper is shown in Figure 4.1.

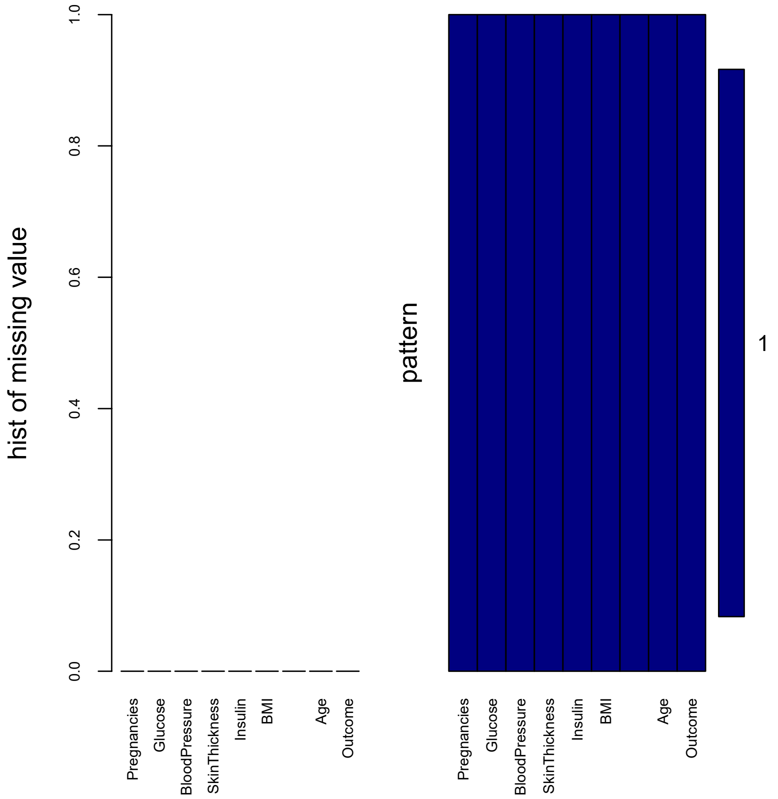


Figure 4.1: Detection of missing value

Figure 4.1 shows that there are no null values in the Pima dataset, as can be seen from the detection of null values in the dataset. The missing values in the dataset with zero values are also analysed in this paper. In the Pima diabetes dataset, the pregnancy attribute is first analysed as the number of pregnancies in women aged 21 years or older, and it is observed that there is more data available in this attribute, so it is necessary to fill in the missing values. In the case of plasma glucose concentration, there are 5 samples with a value of 0. To ensure the integrity of the data and based on the reasonableness of the value of glucose concentration, the value of this attribute is replaced by the mean value. In terms of diastolic blood pressure, some of the missing values are found, but as a human being, diastolic blood pressure is bound to exist, so the average value is considered for subsequent filling. Triceps skinfold thickness is a standard used to detect body obesity. Many values of this attribute are 0. Too many missing values may cause inaccurate classification results, so this attribute value is replaced by the mean value. In terms of 2-hour serum insulin, for patients with type I diabetes, the main reason is that there is no insulin in the body, and they mainly rely on external insulin for treatment, so insulin is completely impossible to exist in the body. However, the patients in the Pima Indian diabetes dataset are all type II diabetic patients. It is possible for patients with type II diabetes to have insulin content in their bodies. It can be judged that the insulin value of 0 in this diabetes dataset is the missing data according to the above standards. Therefore, the treatment method for the missing value of insulin content is to use the average value instead. For BMI (body mass index), everyone should have, should not appear 0 value, so there is a missing value should be dealt with. In general, the missing values in this attribute can be either deleted or replaced by the mean or median. However, deleting the missing data may lose part of the data and lead to inaccurate prediction results of the model. Therefore, filling method is considered to replace the missing values in the BMI attribute.

In summary, in this paper, the values of Glucose, BloodPressure, SkinThickness, Insulin, and BMI that are zero in 5 features are replaced by the mean value.

3.1.3.2 Outlier Detection

Before detecting outliers, the paper performs a statistical analysis of the sample data so that a description of the data in the Pima dataset can be made. In the data description, the squared deviation(sd), median, mean, max, min, points in the upper quartile (Q_1) and points in the lower quartile (Q_3) of each characteristic in the data set are included. Table 4.4 presents a descriptive analysis of the characteristics of the initial diabetes data.

Table 4.4: Number of people with and without diabetes

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DPF	Age
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.08	21.00
1st Q	1.00	99.00	62.00	0.00	0.00	27.30	0.24	24.00
Median	3.00	117.00	72.00	23.00	30.50	32.00	0.37	29.00
Mean	3.84	120.90	69.11	20.54	79.80	31.99	0.47	33.24
3rd Q	6.00	140.20	80.00	32.00	127.20	36.60	0.63	41.00
Max	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00
Sd	3.37	31.97	19.36	15.95	115.24	7.88	0.33	11.76

Using the descriptive analysis table in Table 4.4 for the raw data we can see that the mean value of insulin is 79.79, while the maximum and minimum values are 864 and 0. The maximum value of 864 is much larger than the mean, and therefore it is probably an outlier. At the same time, the normal value of the BMI attribute is between 18.5 and 24, while the maximum value of 67 in the BMI feature is much higher than the normal standard and not in accordance with medical standards, so it may be an outlier. The descriptive analysis table does not allow for an accurate analysis of the outliers for the attributes such as blood pressure, age and number of pregnancies, so further analysis is required in order to determine if there are any abnormal values.

In order to further analyse the data for outliers, a box plot of the eight characteristic attributes is drawn using the plot function. In the box plot, the quartile distance IQR is considered to be $Q_3 - Q_1$. The outliers in the box plot are identified by setting upper and lower limits on the box plot, and if they are above the upper limit or below the lower limit, they are considered as outliers. In this paper, the upper limit of the box line diagram is set to $Q_1 + IQR$ or and the lower limit of the box line diagram is set to $Q_3 - IQR$. For clarity of graphical visualisation, the features are grouped according to the maximum values of the eight features and the box line diagram for each feature is obtained as shown in Figure 4.2.

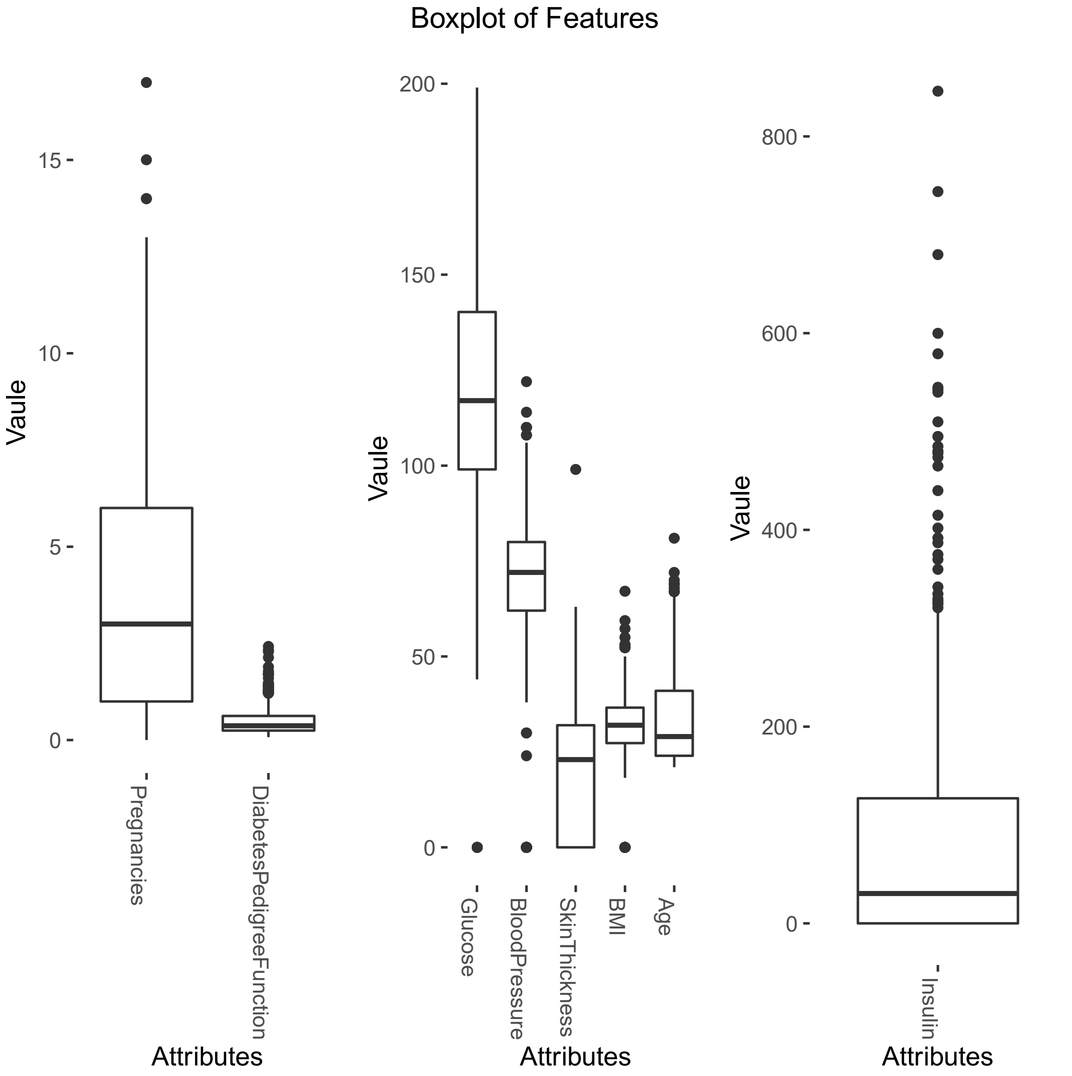


Figure 4.2: Boxplot of Features

Based on the above theory of outlier detection, and to simplify the calculations, only finite numbers with two decimal places are used in this paper, so Table 4.5 shows the quadrature differences for each feature.

Table 4.5: Quadrature Differences for Each Feature

	Q_1	Q_3	IQR	Q_1-1.5IQR	Q_3+1.5IQR
Pregnancies	1.00	6.00	5.00	-6.50	13.50
Glucose	99.00	130.25	31.25	37.12	202.12
Blood Pressure	62.00	80.00	18.00	35.00	107.00
Skin Thickness	0.00	32.00	32.00	38.00	80.00

Insulin	0.00	127.25	127.25	-190.87	318.12
BMI	27.30	36.60	9.30	13.35	50.55
DPF	0.23	0.63	0.39	0.33	1.21
Age	23.00	31.00	17.00	-1.50	66.5

For the results table, which is based on the quadrature technique of the box plot, the observations, particularly for the number of pregnancies, blood pressure and age, indicate that there are values outside the range $Q_1 + IQR$ and $Q_3 - IQR$. For these three features, the data are acceptable under the current circumstances, so that the problematic values of 15, 176, 444 and 581 are found. In summary, the above four outliers are removed. After deletion, 764 sample data remain in this paper.

3.1.3.3 correlation

The correlation coefficient table shows the degree of relationship between each attribute and can provide some basis for dimensionality reduction of the data. The 8 attributes in this study are not easy to use in the diagnosis of diabetes, so it is important to choose as few features as possible. This is of great practical importance. The correlation between the characteristics in the diabetes dataset is shown in the figure.

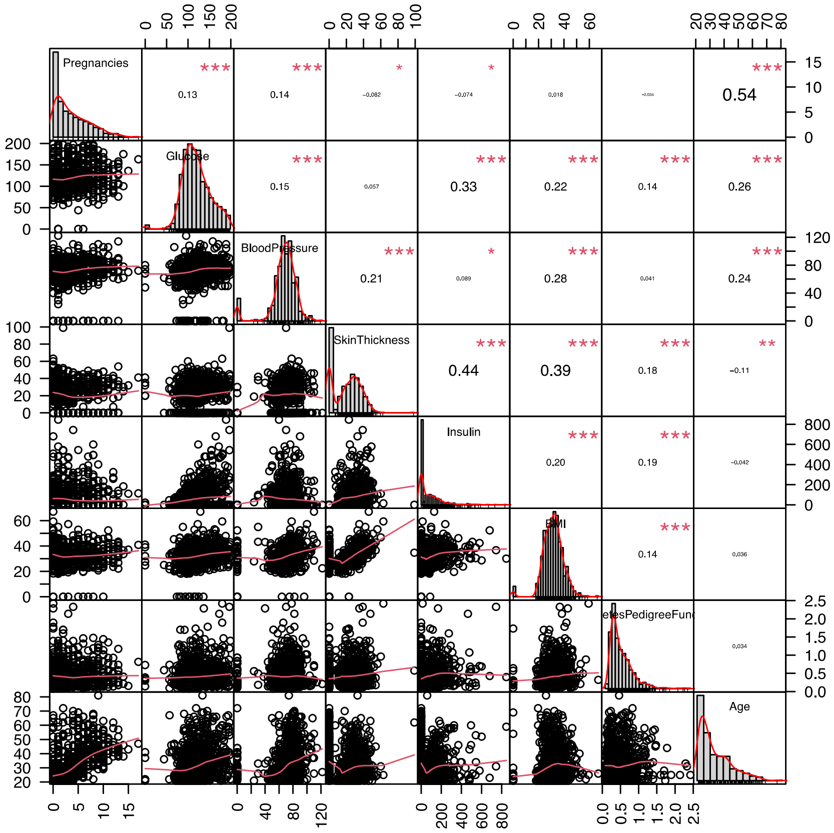


Figure 4.3: Correlation between Features

In Figure 4.3, the numerical values express the correlation between the two features, while the number of stars indicates the strength of the correlation between the two features. As shown in the figure, among the characteristics of pregnancies, the characteristics that are strongly correlated with it are Glucose, blood pressure and age, with Age being the most strongly correlated with it. Within the Glucose feature, the features that are strongly correlated are Blood Pressure, Insulin, BMI, DiabetesPedigreeFunction, and age. Within the feature SkinThickness, the strongly correlated features were Insulin, BMI, DiabetesPedigreeFunction, and age. within the feature Insulin, the strongly correlated features were BMI and DiabetesPedigreeFunction. within the feature BMI, the strongly correlated features were DiabetesPedigreeFunction.

The above graphs only observe the correlation between the features in the diabetes dataset in addition to the outcomes. In order to verify the correlation between the features and the outcomes, a correlation table between the features containing the outcomes was created using the ggpairs function.

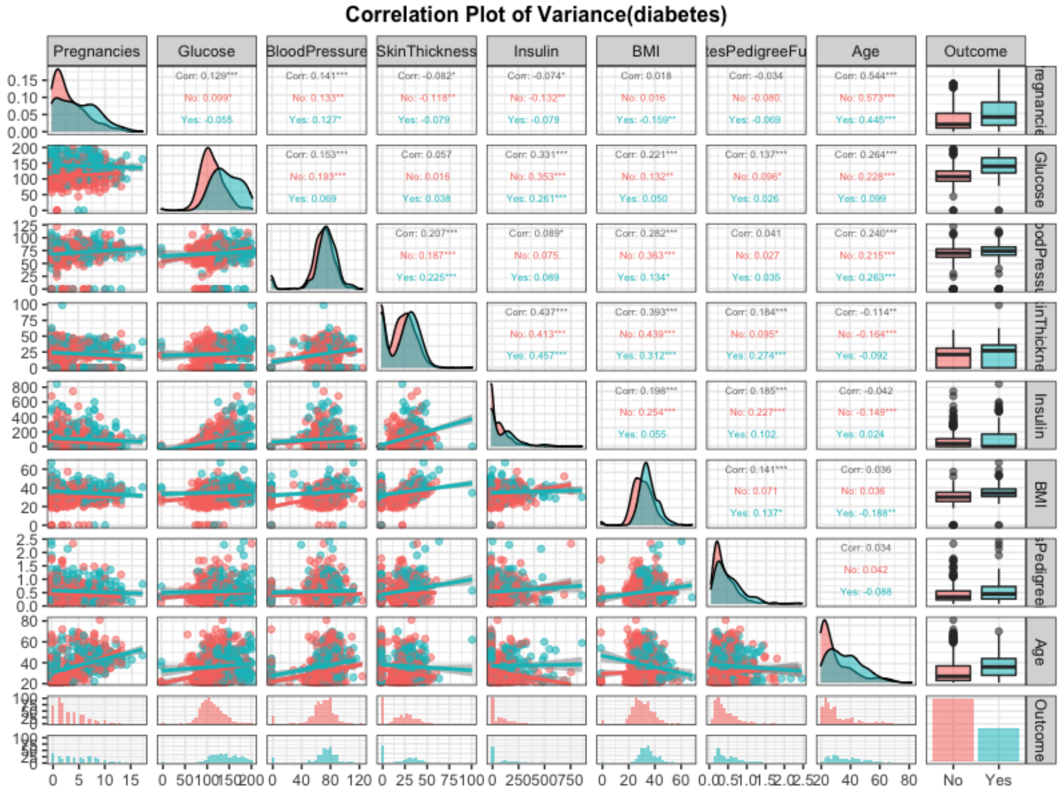


Figure 4.4: Correlation between All Features

As shown in Figure 4.4, the correlations between the features containing the outcome features in the diabetes dataset are shown, along with the correlations between each feature and having diabetes and not having diabetes, respectively.

3.1.4 Feature Extraction

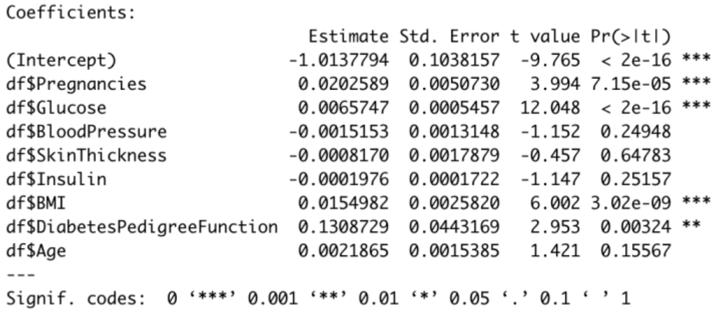
The multivariate linear stepwise regression modelling process in this research uses the step function in RStudio software (with the default parameter backward, which stands for backward stepwise regression), which is based on minimising the AIC value and gradually eliminates all explanatory variables that reduce the AIC value of the equation.

3.1.4.1 Stepwise Analysis

As a backward stepwise regression is used, all variables are introduced into the equation at the beginning and the regression equation in the current case is shown below:

$$Outcome = preg + glu + bp + st + insu + bmi + pf + age \qquad (4.1)$$

In the above equation, $preg$ stands for pregnancies, glu stands for Glucose, bp stands for Blood Pressure, st stands for SkinThickness, $insu$ stands for Insulin and pf stands for DiabetesPedigreeFunction. By setting up the regression equation as above, the results of the multiple linear regression are shown in the figure below.

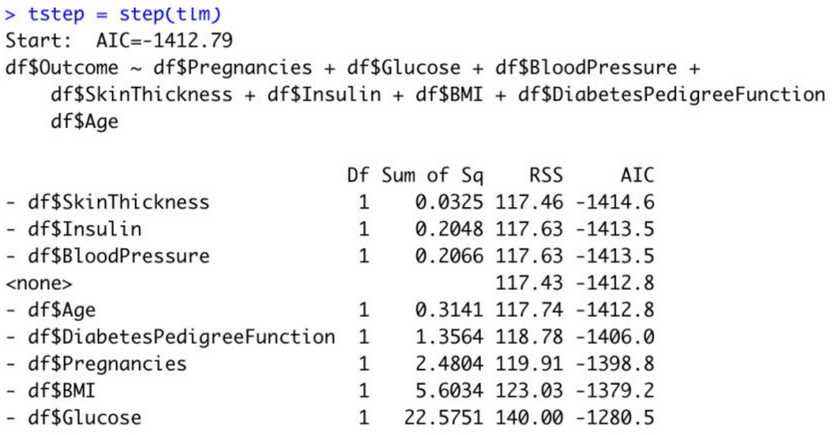


```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -1.0137794  0.1038157  -9.765  < 2e-16 ***
df$Pregnancies           0.0202589  0.0050730   3.994 7.15e-05 ***
df$Glucose               0.0065747  0.0005457  12.048  < 2e-16 ***
df$BloodPressure        -0.0015153  0.0013148  -1.152  0.24948    
df$SkinThickness        -0.0008170  0.0017879  -0.457  0.64783    
df$Insulin              -0.0001976  0.0001722  -1.147  0.25157    
df$BMI                   0.0154982  0.0025820   6.002 3.02e-09 ***
df$DiabetesPedigreeFunction  0.1308729  0.0443169   2.953  0.00324 ** 
df$Age                   0.0021865  0.0015385   1.421  0.15567    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4.5: Step1 of Feature Extraction

By looking at Figure 4.5， only Pregnancies, Glucose, BMI and DIabetesPedigreeFunction passed the significance test in the regression equation set out in Equation 4. 1.

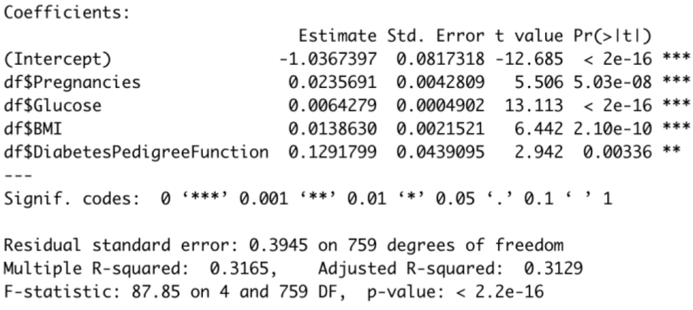
The results of the backward stepwise regression analysis of the above multiple linear regression results are shown below.



```
> tstep = step(tlm)
Start:  AIC=-1412.79
df$Outcome ~ df$Pregnancies + df$Glucose + df$BloodPressure + 
    df$SkinThickness + df$Insulin + df$BMI + df$DiabetesPedigreeFunction
    df$Age

                              Df Sum of Sq    RSS     AIC
- df$SkinThickness             1    0.0325 117.46 -1414.6
- df$Insulin                   1    0.2048 117.63 -1413.5
- df$BloodPressure             1    0.2066 117.63 -1413.5
<none>                                     117.43 -1412.8
- df$Age                       1    0.3141 117.74 -1412.8
- df$DiabetesPedigreeFunction  1    1.3564 118.78 -1406.0
- df$Pregnancies               1    2.4804 119.91 -1398.8
- df$BMI                       1    5.6034 123.03 -1379.2
- df$Glucose                   1   22.5751 140.00 -1280.5
```

Figure 4.6: Start of Stepwise Analysis

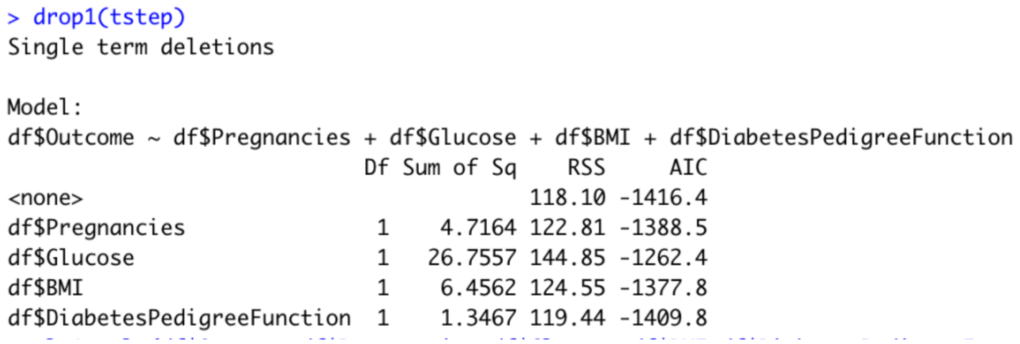
At the beginning of the back stepwise regression, the value of AIC can be reduced to -1412.8 by removing the three features SkinThickness, Insulin and blood pressure. When the stepwise analysis is terminated, the results are shown in the figure below.



```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -1.0367397  0.0817318 -12.685  < 2e-16 ***
df$Pregnancies           0.0235691  0.0042809   5.506 5.03e-08 ***
df$Glucose               0.0064279  0.0004902  13.113  < 2e-16 ***
df$BMI                   0.0138630  0.0021521   6.442 2.10e-10 ***
df$DiabetesPedigreeFunction  0.1291799  0.0439095   2.942  0.00336 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3945 on 759 degrees of freedom
Multiple R-squared:  0.3165,	Adjusted R-squared:  0.3129 
F-statistic: 87.85 on 4 and 759 DF,  p-value: < 2.2e-16
```

Figure 4.7: End of Stepwise Analysis

According to Figure 4.7, by sequentially removing features, only four features, Pregnancies, Glucose, BMI and DiabetesPedigreeFunction, remained in the regression equation at the end of the back stepwise regression analysis, and the significance level of each feature as a regression coefficient improved and all passed the significance test. In an attempt to optimise the stepwise regression analysis, an attempt was made to reduce one of the characteristic coefficients using the method of Drop1, the results of which are shown below.



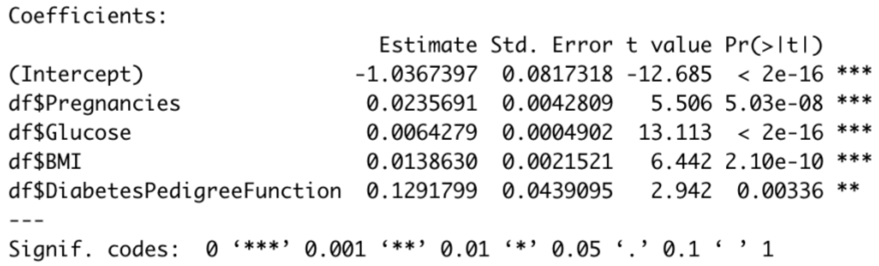
```
> drop1(tstep)
Single term deletions

Model:
df$Outcome ~ df$Pregnancies + df$Glucose + df$BMI + df$DiabetesPedigreeFunction
                           Df Sum of Sq    RSS     AIC
<none>                                  118.10 -1416.4
df$Pregnancies              1    4.7164 122.81 -1388.5
df$Glucose                  1   26.7557 144.85 -1262.4
df$BMI                      1    6.4562 124.55 -1377.8
df$DiabetesPedigreeFunction 1    1.3467 119.44 -1409.8
```

Figure 4.8: Optimization of Stepwise Analysis

After optimisation, the regression equation with the four regression coefficients mentioned above is found to be the optimal regression equation. The new regression equation is as follows.

$$Outcome = preg + glu + bmi + pf \tag{4.2}$$

Further multiple regression analysis is carried out on it and the results are shown below.



```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -1.0367397  0.0817318 -12.685  < 2e-16 ***
df$Pregnancies              0.0235691  0.0042809   5.506 5.03e-08 ***
df$Glucose                  0.0064279  0.0004902  13.113  < 2e-16 ***
df$BMI                      0.0138630  0.0021521   6.442 2.10e-10 ***
df$DiabetesPedigreeFunction 0.1291799  0.0439095   2.942  0.00336 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4.9: Result of Multiple Regression Analysis

By conducting a new multiple linear regression on the optimised regression equation, the results of the analysis are shown above and all four coefficients in the regression equation passed the significance test, hence the optimised regression equation is shown below.

$$Outcome = -1.03674 + 0.02357preg + 0.00643glu + 0.01386bmi + 0.12918pf \tag{4.3}$$

In summary, after screening and testing, the final four variables that will be used for model building are Pregnancies, Glucose, BMI and DiabetesPedigreeFunction.

3.1.5 Build training and test sets

When applying a model to data analysis, the performance of the model directly determines the accuracy and effectiveness of the application. We can use the data from the training set to train the model, and then use the error on the test set as the generalisation error of the final model in response to realistic scenarios. Therefore, the training set is the subset used to train the model and the test set is the subset used to test the trained model. In this paper, we use R to train the prediction model by using Stratified Sampling on 70% of the data as the training set and the remaining 30% as the test set. In summary, the number of samples in the training set in this paper is 516, while the number of samples in the test set is 248.

3.2 Model

In this paper, four models were implemented, namely KNN, SVM, RF and ANFIS, and the performance of each of the four models is optimised to enable the models to better classify diabetes.

3.2.1 KNN

Since the size of K is one of the main factors affecting KNN models, the size of K needs to be considered for the construction and optimisation of KNN models. Regarding the selection of K-values, the 10-fold cross-validation method is used in this paper. Under 10-fold cross-validation, the data is divided randomly, and the model is trained with a training set by varying the number of parameters, changing the values of the parameters, etc. The model with the lowest measurement error, i.e. the highest accuracy, is found from these models.

In this paper, the KNN classification model is implemented using the R language and a call to the knn method in Caret package. Figure 3.10 shows the accuracy of the KN N model for different K values, where the horizontal axis represents the number of nearest neighbours, i.e. the K value, and the vertical axis represents the accuracy of the test set for a specific value of K. In the parameter setting of the KNN model, the range of k values is [5,23].

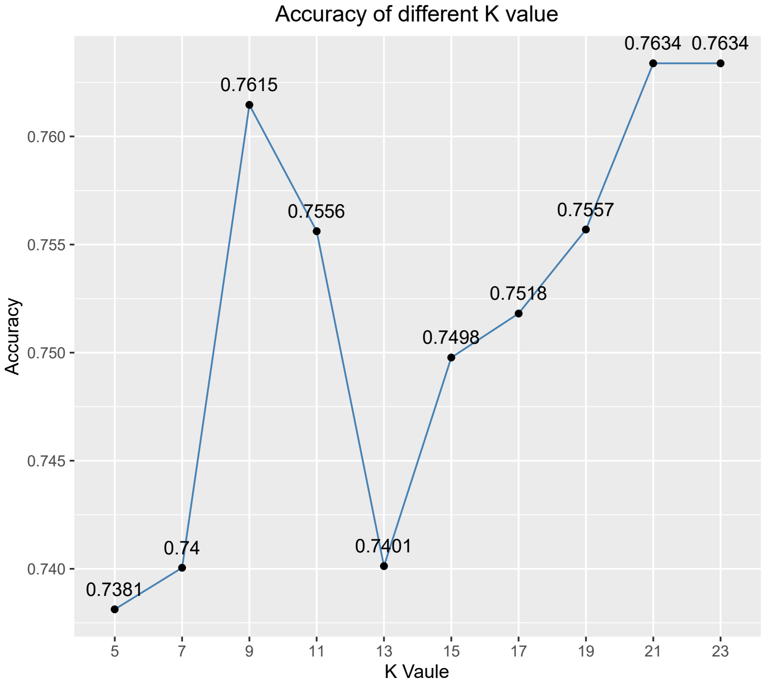


Figure 4.10: Accuracy of K value

From the above graph we can see that the accuracy rate of the training set tends to increase as the value of k increases, but when the value of k is greater than 9, the accuracy rate decreases to 74.01% at a value of k of 13. However, when the value of k was greater than 13, the accuracy rate tended to increase again, and reached a maximum of 76.34% when the value of k was greater than 21 and remained unchanged. By observing the accuracy of the KNN model at various values of k, we can conclude that the accuracy of the KNN model is higher at k = 21, so we have chosen k = 21 for the modelling. The confusion matrix of the KNN model is shown in Table 3.6.

Table 3.6: KNN Confusion Matrix

Reference Prediction	0	1
0	142	40
1	21	45

The confusion matrix of the classification results for KNN modelling with K = 21 shows that the number of samples with actual class 0 and predicted class 0 is 142; the number of samples with actual class 1 and predicted class 0 is 40; the number of samples with actual class 0 and predicted class 1 is 21; and the number of samples with actual class 1 and predicted class 1 is 45. In summary, the KNN model has an accuracy of 0.754 on diabetes when K = 21, and its ROC curve and AUC values are shown below.

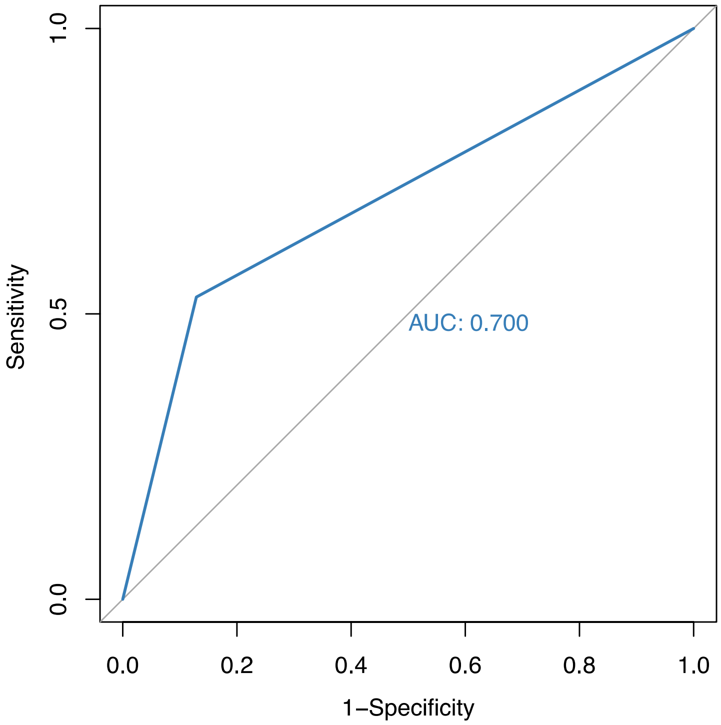


Figure 4.11: ROC of KNN

As shown in Figure 4.11, its vertical coordinate indicates the sensitivity of the model, which means the true positive rate, and the horizontal coordinate indicates the False positive rate of the model, which can be expressed using 1-specificity. As shown in the figure, the Sensitivity of the KNN model is 0.8712, its Specificity is 0.5294 and its F1-Score is 0.8232, and the AUC value of the model is 0.7, which means that it can effectively classify diabetes.

3.2.2 SVM

The two most important parameters of the SVM are the penalty factor C, which controls the stringency of the SVM classification criteria by varying its size. When C tends to be infinitely large, it means that there can be no bias in classification stringency; when C tends to be very small, it means that there can be a high tolerance for error. At the same time, the choice of kernel function is one of the most important factors for SVM. If a suitable kernel function is not chosen, the feature space will be poorly selected and the ideal model will not be obtained. The kernel functions commonly used are the Sigmoid kernel function, RBF, and the linear kernel.

A commonly used method for tuning model parameters is the grid search algorithm, which iterates through a number of given parameter combinations and then brings them into the model for training, eventually finding the optimal model from them. In this paper, we use the Caret package in the R language and call the SVM method in it for 10 cross-validation. The validation shows that the SVM model with Radial Basis Function Kernel works best on diabetes, so on this basis, parameter optimisation of the penalty factor C and sigma values is performed. In parametric optimisation, the penalty factor C is searched in (0, 1),where the interval is 0.1, while sigma is searched in (0,1.2), where the interval is 0.2. Based on the above settings, the variation in accuracy under the search range is shown in Figure 4.12.

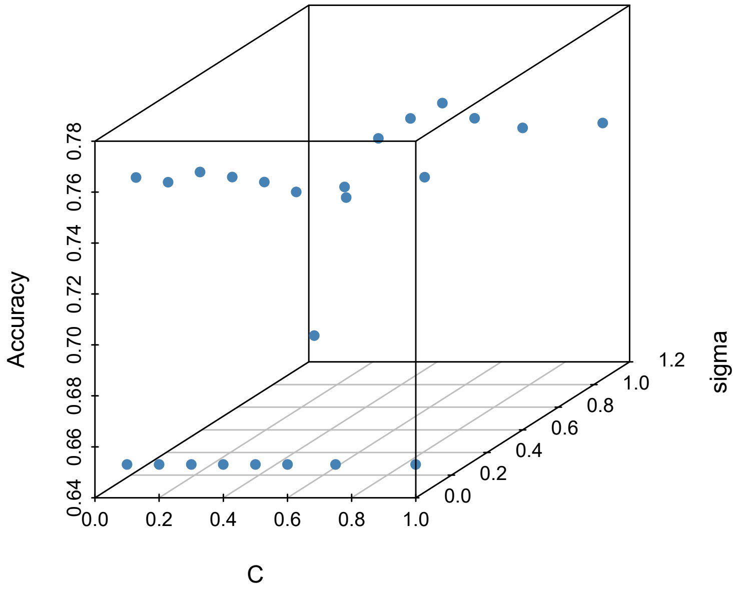


Figure 4.12: Tunning of SVM

In this paper, the process of tunning in SVM is constructed in the R language using the scatterplot3d package. In Figure 4.12, its z-axis indicates the classification accuracy of the SVM in diabetes with different parameters, while the x-axis indicates different values of the penalty factor C, and the y-axis indicates different values under sigma. As shown above, in the SVM model of Radial Basis Function Kernel, the highest accuracy rate of the SVM model is achieved when the penalty factor C is 0.2 and the sigma is 0.05, which is 0.7711 on the diabetic training set.

Therefore, after tunning, the kernel in this paper is set to Radial Basis Function Kernel with a penalty factor C of 0.2 and a sigma value of 0.05 when constructing the SVM model. The confusion matrix of the SVM model is shown in Table 3.7.

Table 3.7: SVM Confusion Matrix

Reference Prediction	0	1
0	142	40

1	21	45

The confusion matrix of the classification results of the SVM model with the above parameter settings shows that the sample size of actual class 0 and predicted class 0 is 142; the sample size of actual class 1 and predicted class 0 is 40; the sample size of actual class 0 and predicted class 1 is 21; and the sample size of actual class 1 and predicted class 1 is 45. In summary, after the optimization of the parameters of the SVM, the accuracy of the SVM model in the testing test of diabetes is 0.754, and its ROC curve and AUC values are shown below.

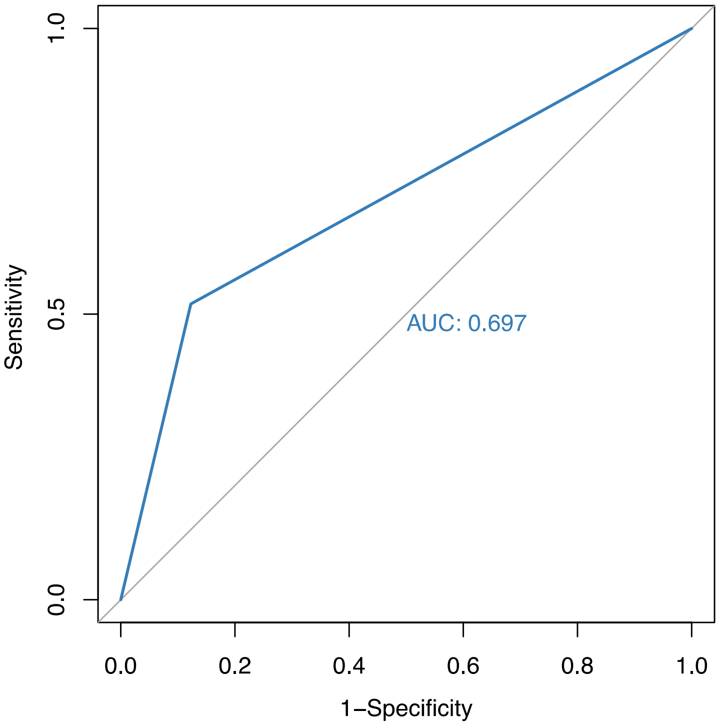


Figure 4.13: ROC of SVM

As shown in Figure 4.13, its vertical coordinates indicate the true positive rate of the model, and the horizontal coordinates indicate the false positive rate of the model. As shown, the SVM model has a sensitivity of 0.8773, a specificity of 0.5176, an F1-Score of 0.8242 and an AUC value of 0.697 for the model, which indicates its effectiveness in classifying on the diabetes dataset.

3.2.3 RF

Random forests are ensemble learning algorithms that use decision trees as base classifiers, which allow rapid notational parallelization of computation and the detection of influence relationships between features. In practical problems, the number of trees is an important parameter of a random forest, and the number of decision trees in a random forest cannot be infinite.

In tuning the RF, this paper uses the Caret package in R and calls the rf method, in which a 10-fold cross-validation is set using the train_fit method. In this method, the parameters of the RF are tuned mainly by the method of the parameter mtry in it. After the 10-fold cross-validation, the number of decision trees in the RF model in this paper was chosen to be 500, the number of nodes is set to 2, while the maximum number of nodes is set to 10. Meanwhile, the value of the parameter Mtry is tuned, and the random search method is used in this paper, and the search range of Mtry is set to [1,10]. The accuracy of the RF model with different mtry values on the training set is shown below.

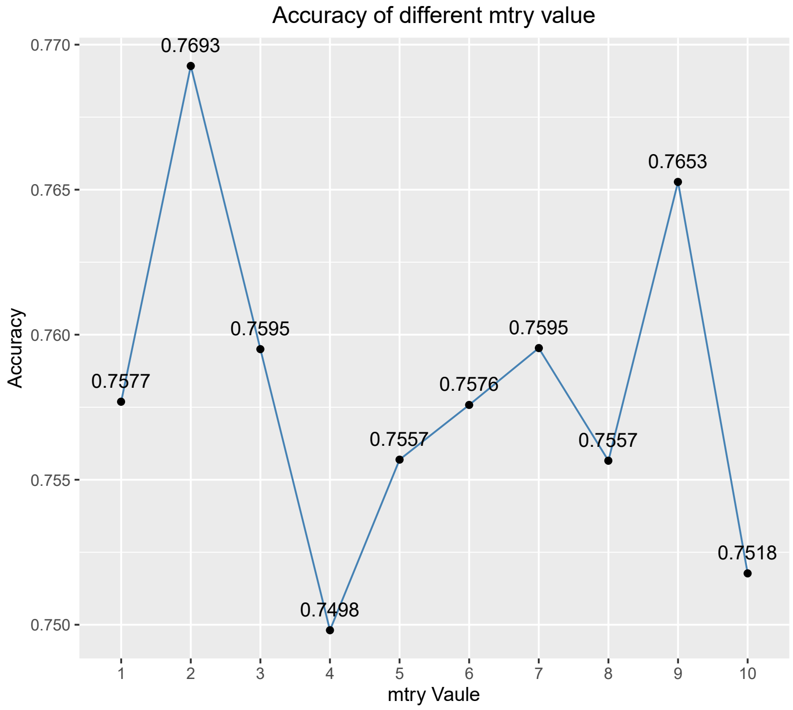


Figure 4.14: Tunning of RF

As shown in Figure 4.14, when the value of mtry is 2, the accuracy of the model of RF reaches the highest at 0.7693. when the value of mtry is 4, the accuracy of the model of RF is the lowest at 0.7498. After that, as the value of mtry increases, the accuracy of the model of RF also increases. In summary, it can be seen that the accuracy of the RF model on diabetes is higher when mtry = 2. Therefore, in this paper, when modelling RF, the value of mtry is set to 2 and the number of decision trees is set to 500, the number of nodes is set to 2, while the maximum number of nodes is set to 10. Based on the above parameter settings of the optimised RF model, the confusion matrix of the RF model on the diabetes test set is shown in Table 3.8.

Table 3.8: RF Confusion Matrix

Reference Prediction	0	1
0	134	34
1	30	50

With the above parameter settings, the confusion matrix of the RF model classification results shows that the sample size of actual class 0 and predicted class 0 is 134; the sample size of actual class 1 and predicted class 0 is 34; the sample size of actual class 0 and predicted class 1 is 30; and the sample size of actual class 1 and predicted class 1 is 50. In summary, after optimisation of the RF parameters, the accuracy of the RF model in the diabetes test set is 0.7419, and its ROC curve and AUC values are shown below.

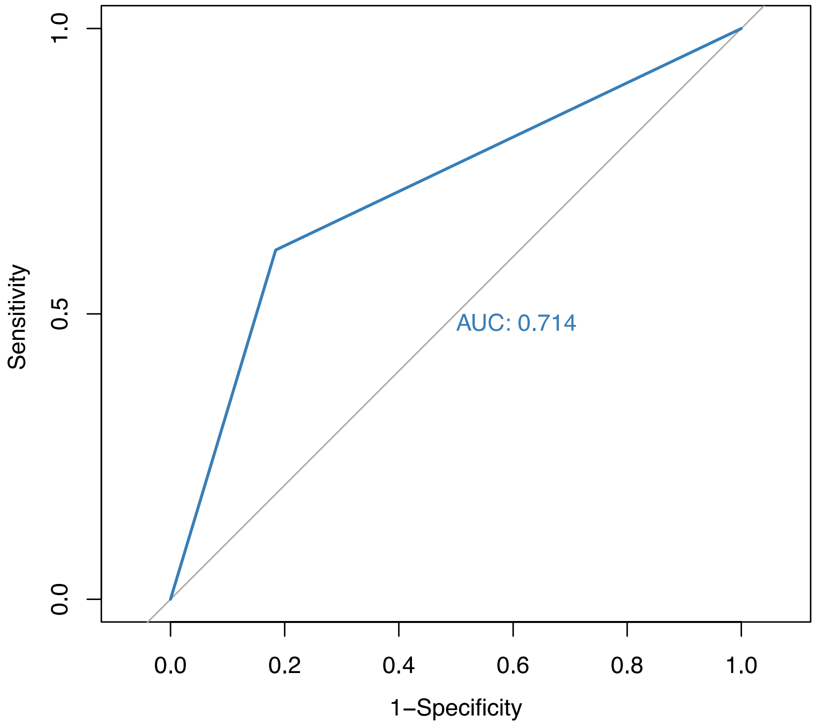


Figure 4.15: ROC of RF

As shown in Figure 4.15, its vertical coordinates indicate the true positive rate of the model, and the horizontal coordinates indicate the false positive rate of the model. As shown, the sensitivity of the RF model was 0.8221, the specificity was 0.6000, the F1-score is 0.8097 and the AUC value of the model is 0.714, which indicates that its classification on the diabetes dataset is the best among the three benchmark models.

3.2.4 ANFIS

In the ANFIS model, the ANFIS model converts attribute clarity values into fuzzy values using a subordination function, and this subordination value is fed into a rule-based neural network model linked to the inference model. Finally, the results are output via the neural network. However, due to the rule-based system, it takes more time to train the model and it is also important to generate the correct set of rules to predict the output.

3.2.4.1 Configuration

In the construction of the model for ANFIS, this paper uses the frbs package in the R language and calls the frbs.learn method for model training. In the method of frbs. learn, which is used in this paper, the ANFIS method is used for model learning. In the ANFIS method of frbs. learn, the parameters of the model are set to include the type of FIS model, the operators of tnorm and snorm, the method of defuzzification, the type of membership function, the number of verbal variables, the maximum number of iterations, the step size, and so on. The parameter settings for the ANFIS method in this paper are shown in Table 3.9.

Table 3.9: The parameter settings for the ANFIS Model

Type of FIS model	Type of mf	Type of tnorm	Type of snorm	Type of defuzzy	Num of labels	Maxfof iteration	Step Size
TSK	GAUSSIAN	MIN	MAX	NULL	5	10	0.01

As shown in Table 3.9, in the ANFIS model constructed in this paper, the type of FIS model chosen is TSK, the type of membership function is Gaussian, the operator of tnorm is MIN, and the operator of snorm is max. since ANFIS does not require defuzzification, this parameter is set to NULL. where the number of labels is 5, so the input variables of the model have 5 linguistic terms, while the maximum number of iterations in the ANFIS model is set to 10 and the step size is 0.01. The input to the model has four features, namely Pregnancies, Glucose, BMI, and DiabetesPedigreeFunction. The membership functions for the above input variables are shown below.

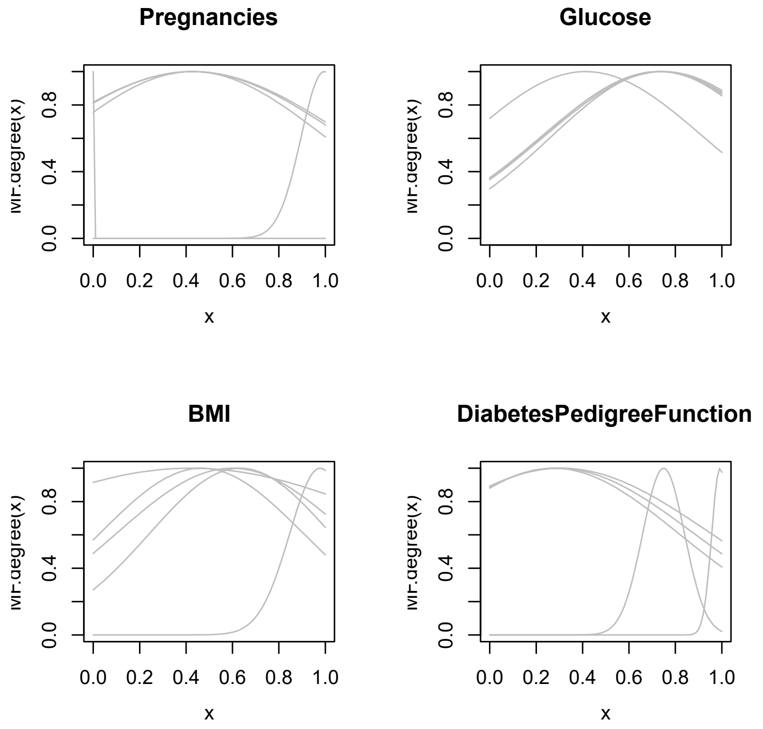
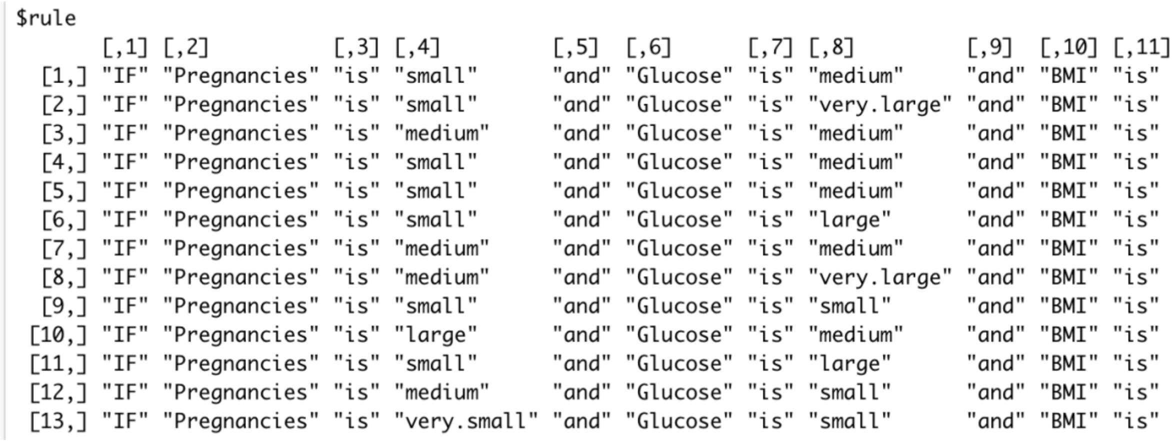


Figure 4.16: MFs of ANFIS

By fuzzing the input variables, some of the rules in the rule base generated in ANFIS are shown below, and all of its rules can be found in the appendix.



```
$rule
      [,1] [,2]          [,3] [,4]         [,5]  [,6]      [,7] [,8]         [,9]  [,10] [,11]
 [1,] "IF" "Pregnancies" "is" "small"      "and" "Glucose" "is" "medium"     "and" "BMI" "is"
 [2,] "IF" "Pregnancies" "is" "small"      "and" "Glucose" "is" "very.large" "and" "BMI" "is"
 [3,] "IF" "Pregnancies" "is" "medium"     "and" "Glucose" "is" "medium"     "and" "BMI" "is"
 [4,] "IF" "Pregnancies" "is" "small"      "and" "Glucose" "is" "medium"     "and" "BMI" "is"
 [5,] "IF" "Pregnancies" "is" "small"      "and" "Glucose" "is" "medium"     "and" "BMI" "is"
 [6,] "IF" "Pregnancies" "is" "small"      "and" "Glucose" "is" "large"      "and" "BMI" "is"
 [7,] "IF" "Pregnancies" "is" "medium"     "and" "Glucose" "is" "medium"     "and" "BMI" "is"
 [8,] "IF" "Pregnancies" "is" "medium"     "and" "Glucose" "is" "very.large" "and" "BMI" "is"
 [9,] "IF" "Pregnancies" "is" "small"      "and" "Glucose" "is" "small"      "and" "BMI" "is"
[10,] "IF" "Pregnancies" "is" "large"      "and" "Glucose" "is" "medium"     "and" "BMI" "is"
[11,] "IF" "Pregnancies" "is" "small"      "and" "Glucose" "is" "large"      "and" "BMI" "is"
[12,] "IF" "Pregnancies" "is" "medium"     "and" "Glucose" "is" "small"      "and" "BMI" "is"
[13,] "IF" "Pregnancies" "is" "very.small" "and" "Glucose" "is" "small"      "and" "BMI" "is"
```

Figure 4.17: Rules of ANFIS

3.2.4.2 Training and Testing

When training the ANFIS model, this paper first uses the Scale method to normalise the input features in the training and test sets, thus making the ANFIS model ready to use the data for classification. In the frbs package used in this paper, the main task of the ANFIS method is to handle the regression task, so this paper assigns separate values to diabetes when classifying the results of the diabetes dataset according to whether the patient has the disease or not. The range of values assigned to those who do not have the disease is [0.6,1], while the range of values assigned to those who have heart disease is [0,1,0.5].When the trained model is used to classify the test set, the model output is firstly the predicted value, which is a regression value, so the predicted value is converted to a classification value by setting an appropriate threshold. Through 10-fold cross-validation, 0.94 is chosen as the threshold for distinguishing whether a patient has diabetes in this paper. Table 3.10 shows the confusion matrix for the ANFIS model on the diabetes test set.

Table 3.10: ANFIS Confusion Matrix

Reference Prediction	0	1
0	151	44
1	12	41

With the above parameter settings, Table 3.10 shows the confusion matrix for the classification results of the ANFIS model on the diabetes test set with a sample size of 151 for actual class 0 and predicted class 0; sample size of 44 for actual class 1 and predicted class 0; sample size of 12 for actual class 0 and predicted class 1; and sample size of 41 for actual class 1 and predicted class 1. In summary, the accuracy of the ANIFS model in the diabetes test set is 0.7742 and its ROC curve and AUC values are shown below.

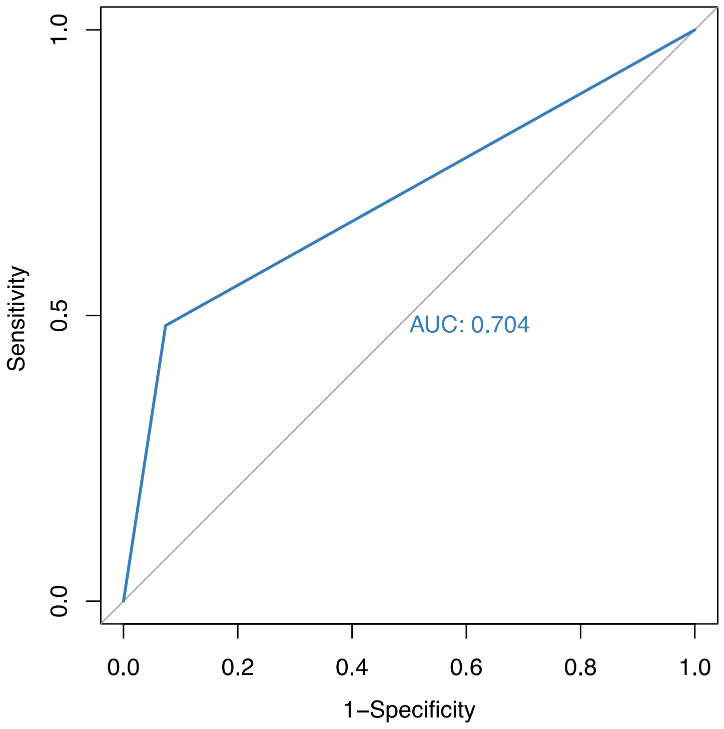


Figure 4.18: ROC of ANFIS

As shown in Figure 4.18, the vertical coordinates indicate the true positive rate of the model and the horizontal coordinates indicate the false positive rate of the model. As illustrated, the ANFIS model had a sensitivity of 0.9264, a specificity of 0.4824, an F1 score of 0.8436 and an AUC value for the model of 0.704. In summary, the ANFIS model had the highest accuracy on the test set in diabetes.

Chapter 4: Result

In the methodology, the paper describes the implementation methods and processes for each model, and completes the construction of the models and the classification on the test set. In order to validate the performance of each model, the paper is analysed in terms of accuracy, consistency, ROC curves and AUC values.

Firstly, the accuracy of the classification of each model on the diabetes test set is compared by comparing the correctness of the models. The difference in accuracy between the models is shown below.

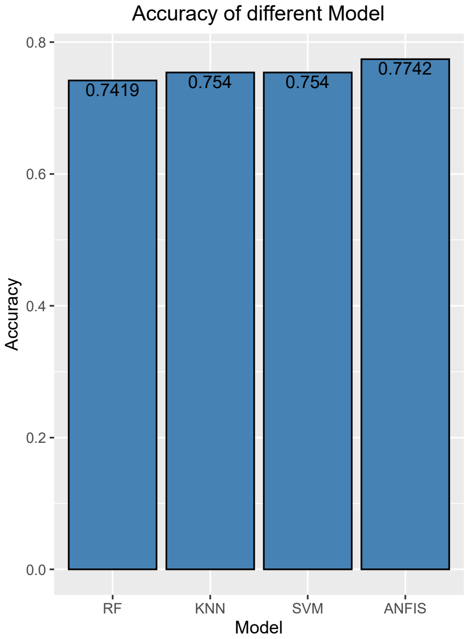


Figure 4.1: Accuracy of Models

In Figure 4.1, the vertical axis represents the accuracy of each model and the horizontal coordinate represents the name of the model. As shown, the ANFIS model had the highest accuracy among all models in the diabetes test set, with 77.42%, while in contrast, the RF model had the lowest classification accuracy, at 74. 19%. Meanwhile, the SVM model and the KNN model are the same in terms of classification accuracy, both at 75.4%.

Secondly, the consistency of the models is compared in this paper. As the 10-fold cross-validation method is used in this paper to cross-validate the models, the model will give different cross-validation results when classified on different test sets, thus the results of the model need to be tested for consistency. The consistency of the model can be described by the kappa value. In a kappa test, a kappa value below 0.4 is usually considered to be a relatively unreliable model, while a value above 0.4 is considered to be a relatively reliable model. The kappa values under each model are shown below.

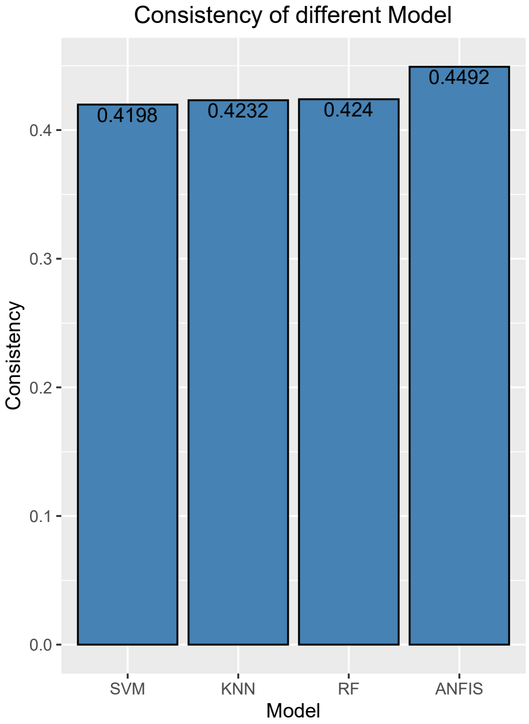


Figure 4.2: Consistency of Models

As shown in Figure 4.2, the kappa values for the consistency of each model are indicated on the vertical axis, and the names of the models are indicated on the horizontal axis. As shown in Figure 4.2, the kappa values for the consistency of each model are indicated on the vertical axis, and the names of the models are indicated on the horizontal axis. As shown above, the ANFIS model has the highest kappa value among the four models, which reaches 0.4492, which means that the model has the best consistency and the model is the most reliable among the four models. In contrast, the SVM model has the lowest kappa value among the four models with a kappa value of 0.4198. At the same time, the kappa values show that the consistency of KNN and RF is almost the same, which are 0.4232 and 0.424 respectively. In summary, the four classification models constructed in this paper are relatively stable on the diabetes dataset.

Thirdly, the models are also evaluated in this paper by means of ROC curves and AUC values. In the ROC curve, it is possible to combine the specificity and sensitivity of the model. At the same time, although the ROC curve can reflect the performance of the model in these two aspects, it sometimes does not intuitively reflect the performance advantage of the model in these two aspects, so this paper also evaluates the performance of the model by its AUC value. The ROC curves and AUC values for the four models constructed in this paper on the diabetes test set are shown in Figure 4.3 and Figure 4.4.

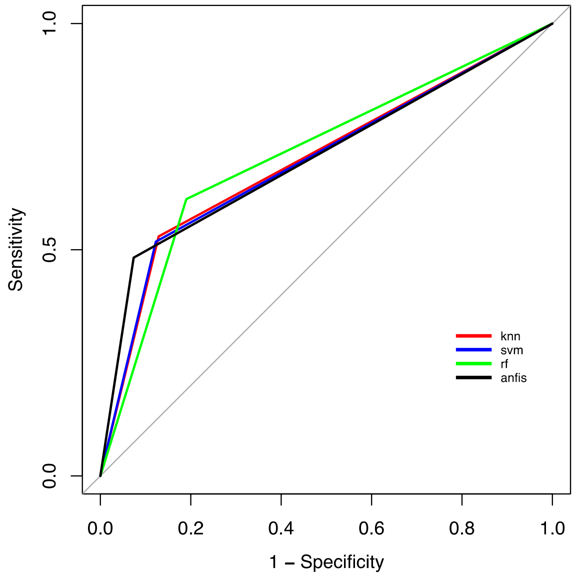


Figure 4.3: ROC Curve of Models

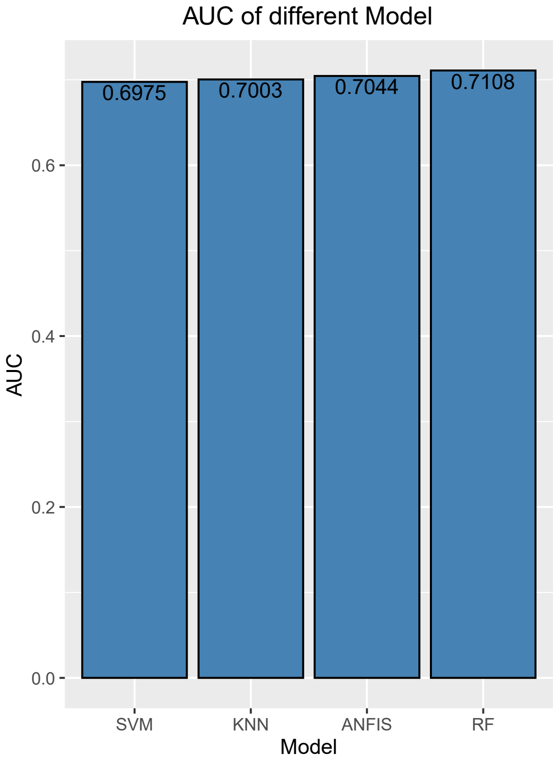


Figure 4.4: AUC Value of Models

According to Figures 3.4 and 3.5, the AUC values of the four models used in this paper are all basically above 0.7. According to the definition of AUC value, it can be seen that when the AUC value is greater than 0.5, it indicates that the model is better than the random classification model. Therefore, it can be seen from the AUC values that all four models constructed in this paper outperform the stochastic model. Among the four models, the RF model has the highest

AUC of 0.7108, followed by the ANIFS model with 0.7044. The AUC values of the KNN and SVM models are both around 0.7, 0.6975 and 0.7003 respectively.

Chapter 5: Conclusion

As society continues to modernize, people can get a better life, but at the same time there are many negative effects such as unhealthy diets, the ageing process of the population, mental stress, lack of physical activity and obesity. Among the many problems, the high incidence of diabetes amplifies the impact of the disease on human beings. Severe diabetes brings with it a variety of complications that can be unbearably devastating not only for the family of each patient, but also for every country and the global community in terms of the enormous medical burden.

The ANFIS model used in this paper is described in the literature review. The 5-layer structure of ANFIS is explained in principle and synthesised in the context of its current diabetes scenario, respectively. The three machine learning methods used, such as KNN, SVM and RF models, are also explained and analysed. As this paper also requires data processing and feature processing, the AIC method used for feature extraction is presented, which includes an explanation of the principles of the stepwise analysis method. Finally, in this section the principles of the various metrics used in the evaluation of the model are analysed, including the accuracy of the model, the F1-Score, and so on.

In the methodology section of this paper, the sources and datasets of the diabetes dataset are presented and analysed. The individual features of the diabetes dataset are analysed in detail in this paper. Secondly, missing values and outliers in the diabetes dataset are examined in this paper using missing value plots and box line plots respectively, and outliers are removed. At the same time, values that do not meet the criteria for each feature in the diabetes dataset are replaced by the mean value. Correlation plots and ggpairs were also used in this paper for correlation and association between features in the diabetes dataset. Finally, in the data processing part, this paper used stepwise analysis and combined with AIC to gradually complete the feature extraction, and finally four features were used in this paper, namely Pregnancies, Glucose, BMI and DiabetesPedigreeFunction. Also, the division of the data set in diabetes was done according to the ratio of 70% and 30%. In the methodology, the four models from the literature review are also constructed, while the parameters of the models are optimized so that the performance of each model is optimized, and the performance performance such as the confusion matrix and ROC curve of the models are analysed. The ANFIS model has the highest accuracy on the diabetes test set, with 77.42%.

Finally, in the results section of the paper, the four models are analysed in terms of accuracy, consistency and ROC curves and AUC values. In terms of accuracy and consistency, the ANFIS model performed the best on the diabetes dataset. Also, the performance of the ANFIS model on both ROC curves and AUC values was in line with the range of excellent performance. In summary, the ANFIS model proposed in this paper is valuable on the diabetes dataset.