Sudan University of Science and Technology

College of Graduate Studies

Classification of Diabetic Patients using Computational Intelligent Techniques

تصنيف مرضى السكرى باستخدام التقنيات الحسابية الذكية

Submitted for the degree of Doctor of Philosophy
in Computer Science

By

Ahlam Ali Sharif Elhussein

Supervisor

Mohamed Elhafiz, Dr

Co-Supervisor

Talat Wahabi, Dr

March, 2018

Dedication

This thesis is dedicated to

My parents,

My husband,

My children,

For their endless love, support and encouragement.

Acknowledgements

After an intensive work, today is the day: writing this note of thanks is the finishing touch on my thesis. It has been a fruitful period and it has an excellent effect on me. So I would like to thank all those who have helped me specially my colleagues at Sudan University and I will never forget their willing to help whenever I needed that help.
My great appreciation to my supervisor who has done wonderful things to enable me to finish my thesis. Thank to you Dr Mohammed Elhafiz. Then a lot of thanks for my co-supervisor Dr Talat Wahabi for his valuable guidance and following his guides enable me to finish this thesis successfully.
My great thanks to my family who were even with me ,till I get my doctorate. My thanks and love to my husband who is with me all the time giving advice and help. Also I would like to thank my colleagues at the office especially Istaz Ali Elsheikh, Mogahid , Tahani and all the team. Finally, there are my friends. We were not only able to support each other by deliberating over our problems and findings, but also happily by talking about things other than just our papers.
Thank you very much, everyone.

Abstract

Diabetes Mellitus is one of the fatal diseases growing at a rapid rate in developing countries. This rate is also critical in the developed countries, Diabetes Mellitus being one of the major contributors to the mortality rate. Detection and diagnosis of Diabetes at an early stage is the need of the day. It is required that a classifier is be designed so as to work efficient, convenient and most importantly, accurate.

Artificial Intelligence and Soft Computing Techniques mimic a great deal of human ideologies and are encouraged to involve in human related fields of application. These systems most fittingly find a place in the medical diagnosis. As much as there was a need for exact classification with accuracy, it should be understood that detection of a diabetic situation is highly beneficial to the community. The propose number of research methods expected for detection of the diabetic conditions so as to provide a sound warning before they had happened.

The experimental result done using Pima Indian dataset which can even be retrieved from UCI Machine Learning Repositorys web site.

In this research Genetic Programming Toolbox For Multigene Symbolic Regression (GPTIPS), used to build a mathematical model for predict the diabetes class. After that simplified the model by selecting the weighted features that affected on the prediction model. The Neural Network, Fuzzy logic and Genetic Programming are used to check the accuracy when using the new features.

The conclusion of that three features can be used to predict the class. The mathematical model become simple and convenient. As a feature work improving the performance by using the optimization methods like Grey Wolf Optimization (GWO) and Particle Swarm Optimization (PSO).

المستخلص

يعتبر مرض السكرى واحد من الامراض القاتلة التي تنمو بمعدل هائل في الدول النامية، وهذا المعدل يعتبر حرجا أيضا في الدول المتقدمة، فالاكتشاف المبكر لمرض السكرى وتشخيصه هو ما نحتاجه اليوم. وهذا يتطلب تصميم مصنف قليل التكلفة يكون ملائما ودقيقا.

ان الذكاء الاصطناعي وتقنيات برمجيات الحاسوب تحاكى الى حد بعيد الأيدولوجيات البشرية وقد وجدت التشجيع لتشترك في مجالات التطبيق المتعلقة بالبشر. وتجد هذه الأنظمة أكثر الأماكن مناسبة لها في التشخيص الطبي. وبقدر ما كانت هنالك حاجة لتصنيف صحيح مع الدقة فانه يتوجب علينا اكتشاف حالة من حالات مرض السكرى يكون ذو فائدة عالية للمجتمع.

من المتوقع اقتراح عدد من أساليب البحث لاكتشاف حالات مرض السكرى لكي نوفر تحذير مبكر للمرض. ان أساليب التعلم الالية تساعدنا كثيرا في تشخيص مرض السكرى وتظهر مستوى معقول من الكفاءة، ولكن هذه المعلومات كثيرة جدا في الطبيعة ولدرجة مزعجة مما يؤثر سلبا على عملية ملاحظة المعرفة والانماط المفيدة. وقد جذبت تقنيات التعلم الالية انتباها كبيرا الى الباحثين لتحويل هذه البيانات الى معرفة مفيدة.

نتائج التجربة أجريت باستخدام بيانات مجموعة متجانسة تسكن المنطقة حول أمريكا، لكنها تحظى بشعبية لكونها المجموعة الأكثر إصابة بالنوع الثاني من داء السكري. تم الحصول على البيانات من موقع مستودع تعلم الآلة.

في هذا البحث تم استخدام أدوات البرمجة الجينية والتي تستخدم لبناء نموذج رياضي للتنبؤ بتصنيف المرض. وبعد ذلك العمل على تبسيط ذلك النموذج عن طريق اختيار الميزات التي اثرت عليه. وتم استخدام الشبكة العصبية والمنطق الضبابي والبرمجة الجينية للتحقق من الدقة عند استخدام الميزات الجديدة.

الخلاصة يمكن استخدام ثلاثة ميزات فقط لبناء نموذج التنبؤ لمرض السكرى ويمكن تحسين الأداء باستخدام أساليب التحسين مثل طريقة الذئب الرمادى او طريقة سرب الجسيمات.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

DM	Diabetes Mellitus
GA	Genetic Algorithm
NN	Neural Network
GP	Genetic Programming
GPTIPS	genetic programming & symbolic regression for MATLAB
NIDDK	National Institute of Diabetes and Digestive and Kidney Diseases
WHO	World Health Organization
SPR	Statistical Pattern Recognition
ANN	Articial Neural Network
IDDM	Insulin Dependent Diabetes Mellitus
NGSP	National Glycohemoglobin Standardization Program
DCCT	Diabetes Control and Complications Trial
MLP	Multi layer Percepton
LDA	Linear Discriminant Analysis
FNN	Fuzzy Neural Networks
KNN	K Nearest Neighbor
SVM	Support Vector Machine
NBTree	Naive Bayess Tree
REPTree	Reduced Error Pruning Tree
LM	Levenberg Marquardt

List of Publications

1. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. (IJARAI) International Journal of Advanced Research in Artificial Intelligence,Vol.3, No.10, 2014

2. Comparative Study of Diabetic Patient Data Using Classification Algorithm in Weka, International Journal of Science and Research (IJSR), Volume 6 Issue 10, October 2017, 141 - 146

Chapter One

Introduction

1.1 Introduction/Overview

Medical diagnosis tasks can be very effectively supported by intelligent computational techniques that, after a proper training on the problem implied, they are able to make a reliable suggestion about a particular instance of the problem. So, a doctor who will use the system will have the option to take into consideration another opinion that concerns the healthiness of a person. In this way, computer aided diagnosis problems are essentially Pattern Recognition tasks since they are trying to classify an instance of a problem to one out of all possible classes for it. For example, a patient can be diabetic or non-diabetic, which, for a diabetes diagnosis problem, means healthy. If it is possible to describe a persons healthiness that concerns diabetes by a set of data, then a medical diagnosis system will have to decide if that person is diabetic or not based only on these data provided. So, the system performs classification of data. However, every classification procedure inevitably suffers by errors that reduce the reliability of the classification. Though, in medical applications, it is critical to have high classification performance because diagnosis suggestions need to be accurate. So, it is crucial to search for techniques that will decrease the error rate (Tsirogiannis et al., 2004).

Diabetes Mellitus : It is simply caused by the failure of the body to produce the right amount of insulin to stabilize the amount of sugar in the body

Patil et al. (2010). Most patients who suffer this type of body failure are recommended to take insulin injection Tresp et al. (1999). This is called diabetes type I, where diabetes type II suffer from their body rejection to insulin. This type of patient is recommended to undergo certain health meal program as well as performing exercises to lose weight, plus taking oral medication. But heart diseases are likely to strike these patients in the long run Zecchin et al. (2011).

Gestational Diabetes which occur temporarily during Pregnancy is called as Gestational Diabetes which occur during to the hormonal changes and usually begins in the fifth or sixth month of pregnancy (between the 24th and 28th weeks) Sc, Phil, and D (Sc et al.). Gestational diabetes usually resolves once the baby is born. However, 25-50 percent of women with gestational diabetes will eventually develop diabetes later in life, especially in those who used to take insulin during pregnancy and those who are overweight after their delivery. Diagnosing diabetes diseases require well trained and expert technicians to compare the case before facing them with typical cases of diabetes. The number of such technicians should also be great so as to come to robust clear cut decision to every patient at risk.

General diabetes statistics : Due to the wide spread of type II in America, a survey has been conducted by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in collaboration with the American Diabetes Association Estrada et al. (2010), the result is that 17.9 million have been diagnosed while 5.7 million are unaware that they are infected by the disease. Statistically 23.6 million people in America have been diagnosed type II diabetes positive. This form 90 to 95 percent. Those affected include:

- US women aged 20 years and older form 11.5 million. which represent 10.2% of women in USA.

- US men aged 20 years and older form 12 million. which represent 11.2% of men in USA.

- people under 20 years form 186,300.

- Adults over 60 years form 12.2 million.

- African Americans aged 20 years and older form 3.7 million (14.7 percent of all African Americans age 20 years and older).

- Hispanic/Latino Americans form 2.5 million (9.5 percent of all Hispanic/Latino Americans).

- Caucasian Americans aged 20 years and older form 14.9 million (9.8 percent of all Caucasian Americans age 20 years and older).

Table 3.1 showed the Gestational diabetes in the Middle East and Northern Africa (WHO, 2016).

Table 1.1: Gestational diabetes in Middle East and Northern Africa

Country	**Extrapolated incidence**	**Population Estimated used**
Sudan	19.430	$39,148,162^2$
Iran	33.503	$67.503.205^2$
Iraq	12.594	$25.374.691^2$
Jordan	2.784	$5.611.202^2$
Kuwait	1.120	$2.257.549^2$
Lebanon	1.874	$3.777.218^2$
Saudi Arabia	12.803	$25.795.938^2$
Syria	8.942	$18.016.874^2$
UAE	1.252	$2.523.915^2$
Yemen	9.938	$20.024.867^2$
Egypt	37.778	$76.117.421^2$
Libya	2.795	$5.631.585^2$

Data classification can be defined as assigning a class label to a data instance based upon knowledge gained from previously seen class-labeled data. Various classification algorithms have been proposed and the frequency of their usage depends upon many things including their simplicity, comprehensibility and accuracy Alaydie et al. (2012). Techniques like decision trees are simple and comprehensive but applicable to small data sets only; only one tree can be evolved for one set of training samples and they suffer from lack of robustness in presence of noise or missing values. Statistical techniques like bayesian nets, neural networks and support vector machines are complex and results of classification decision are not comprehensible. Figure 1.1 showed an example of animal classification.

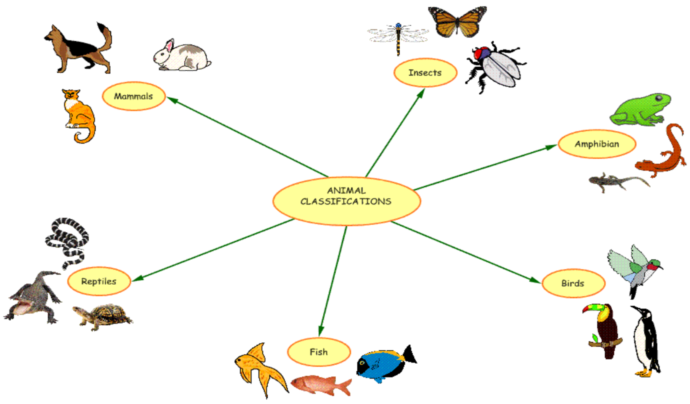


Figure 1.1: Example of animal classification

Consequently, most of the research efforts on Diabetics diagnosis and classification .Our goal in this dissertation is to find a methodology for detection of the diabetic conditions so as to provide a sound warning a head before

1.2 Problem Statement

A physician has to analyze lot of factors before diagnosing the diabetes which makes physicians job very difficult. Normally physicians make their decisions by comparing the test results of current patient with some previous patients who also had similar conditions. This depends not only on the physicians knowledge but depends strongly on the experience of the physician as well. This is not an easy job as the physician has to consider many factors while making a decision Muhammad Waqar Aslam (2010). Also there will be demand for a large number of physicians when everybody at risk will need to be tested. As physicians need to have a look at previous results while making their decision, they may need a tool for listing all the previous decisions made on the patients who have similar conditions. So a classifier system is needed which can classify that list according to the decisions made by experts. There is no doubt that the most important factors in diagnosis are the data taken from the patient and experts opinion on that data but the use of different intelligence techniques and classifiers also help a lot. That is why the use of classifier systems in medical diagnosis are increasing. Although the prediction and classification of diabetes problems is a hot research area that has many work, some of the problems associated with classification are still there and need to resolve like performance optimization, reduce complexity, reduce error etc.

1.3 Research Objectives

The main objective of this work is to define a mapping from the original representation space into a new space where the classes are more easily separable. This will reduce the classifier complexity, increasing in most cases classifier accuracy. There are sub objectives listed below:

- Preprocessing the data.

- Developing a prediction model to diagnose diabetic patients.

- Build a rule base system for diabetic disease.
- Select the best attributes that affect on the performance of prediction and classification of diabetic patients.
- Evaluate the model.

1.4 Research Methodology

Hospital information systems have been frustrated by problems that include congestion, long wait time, and delayed patient care over decades. To solve these problems, Computation Intelligence techniques have been used in medical research for many years and are known to be effective. Therefore, this study examines building a hybrid Computation Intelligence methodology, combining medical domain knowledge and associative classification rules. Real world emergency data will be collected from a hospital and the methodology will evaluated by comparing it with other techniques. The methodology is expected to help physicians to make rapid and accurate diagnosis of diabetes diseases. Table 4.1 show the scope of dataset and machine learning methods that used in this research. In this work four paradigms of Computation Intelligence (CI) used listed bellow:

1. Neural Network (NN).
2. Fuzzy System (FS).
3. Evolutionary computation (Genetic Programming).

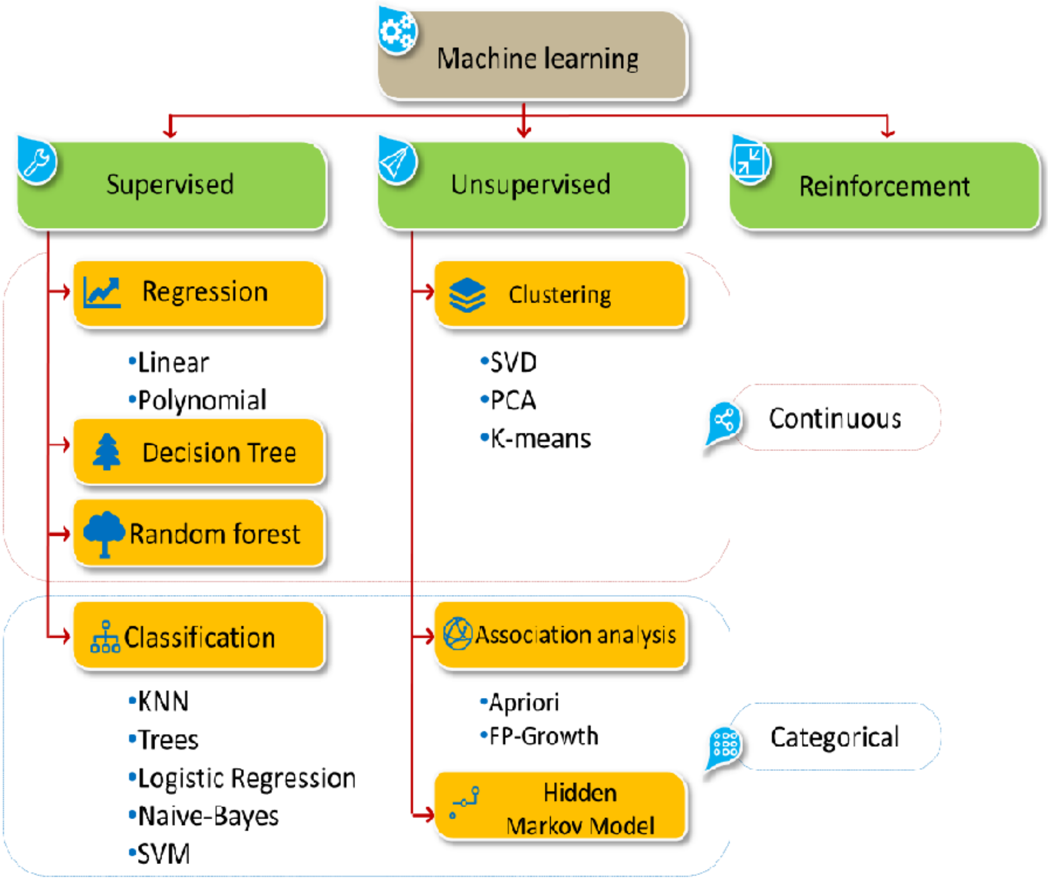


Figure 1.2: Types of Machine Learning

Table 1.2: Dataset and Machine Learning Methods

Medical Datasets	Brest Cancer Liver Disorder Heart Disease Diabetes	Diabetes	Pima Indian Dataset	Group with type II diabetes
Machine Learning	Supervised Unsupervised Reinforcement	Supervised	Classification	Neural Network Fuzzy logic Genetic Programming

1.5 Research Questions

The main question of this study is what is the best model that can be used to predict and classify diabetes disease? But there are other questions which can be listed as follows:

- How can a classifier be designed that is cost efficient, convenient and most importantly, accurate?
- What are the best features that affect in the accuracy of classification?
- How to measure the performance and accuracy of the classifier?

1.6 Research Hypothesis/Philosophy

The goal of this dissertation is the development of diabetes prediction and classification framework that improves the prediction of diabetes performance. In order to accomplish this objective, the work in this dissertation is divided into five tasks.

The first task: focuses on developing a prediction model using Neural Network. Neural networks are developed as an attempt to realize simplified mathematical models of brain-like systems. The main advantage of NN is its capability to learn from examples instead of requiring an algorithmic way Rahamneh et al. (2010). Data recorded by the patient may be missing or invalid, causing incomplete or wrong features. However, correctly recorded patient data should correspond to the change in blood glucose levels, improving the accuracy of blood glucose prediction models. This motivates the inference of missing data values that were recorded by the patient.The proposed system will use data preprocessing techniques that were carried out on patient data prior to classification and prediction. The proposed system will use a Neural Networks for design of a predictor for detection of diabetes. Neural Networks are popular for their dynamic nature in terms of learning, which are preferably chosen for medical diagnosis Allam et al. (2011).

The second task: is using Fuzzy logic to build fuzzy rule base. Fuzzy logic represents an exciting technology with a wide scope for potential applications. Fuzzy logic provides an outstandingly simple way to draw conclusions from vague, uncertain or rough information. Fuzzy logic approach is particularly a preferable tool for dealing with problems with uncertainties and imprecise information Rahamneh et al. (2010). In spite of having such a great quality, the neural networks cannot predict with a remarkable accuracy. The fuzzy systems, though not as dynamic as neural networks, can work accurately owing the fact that they have rule bases that mimic human thinking Sapna and Tamilarasi (2009).

The Third task: is to use Genetic Programming (GP) to build a mathematical model for the classification. Genetic Programming based classifier can assist in the diagnosis of diabetes disease. GP showed quite good classification accuracies. Authors believe that the Genetic Programming Toolbox for the Identification of Process Systems (GPTIPS) provides free complementary alternatives to current data analysis techniques and has a wide domain of application. This is because the GPTIPS transforms linear combinations out of non-linear ones Hinchliffe et al. (1996a). Then the transformed structure is forced to a smart regression of the input-output process and thus constructs a standard model that allows evolution of accurate and compact mathematical method even when there is a large number of input data. This regression limited the probability of mistakes and errors in the transformed model. Thus giving it peculiarity.

The last task is to select the best attributes which affect on the performance of prediction and classification model. In a decision-theoretic or statistical approach to pattern recognition, the classification or description of data is based on the set of data features used. Therefore, feature selection and extraction are crucial in optimizing performance, and strongly affect classifier design. Defining appropriate features often requires interaction with experts in the application area. In practice, there is much noise and redundancy in most high dimensionality, complex patterns. Therefore, it is sometimes difficult ,even for experts, to determine a minimum or optimum feature set. The curse of dimensionality becomes an annoying phenomenon in both statistical pattern recognition (SPR) and Artificial Neural Network (ANN) techniques. Researchers have discovered that many learning procedures lack the property of scaling, these procedures either fail or produce unsatisfactory results when applied to problems of larger size Jain and Mao (1997). To address this scaling problem, we have developed two approaches, one based on genetic algorithms (GA's). The basic operation of these approaches utilizes a feedback linkage between feature evaluation and classification. That is, we carry out feature extraction (with dimensionality reduction) and classifier design simultaneously, through learning model and evolution. The objective of these approaches is to find a reduced subset among the original N features such that useful class discriminatory information is included and redundant class information and/or noise is excluded.

1.7 Thesis Contribution

- A model is constructed to solve the diabetics classification and diagnosis.
- Simulation for model that used.
- The best attributes that affect on the performance of prediction and classification of diabetic patients is selected.
- Comparing results with the benchmark machine learning techniques.
- Publishing articles related to this field in academic journals.

1.8 Thesis Structure

This thesis is built in five chapters to cover the objectives of the study and details of methodologies followed to achieve the goals. These five chapters are organized as follows:

Chapter 1 Introduction : by now the reader has learned about this chapter, which presented a general discussion about this research by giving a brief background about the topic. Then how the problems were stated, and the research objectives were shown, which is followed by scope and significance of the study. Finally the research contribution of this thesis is summarized.

Chapter 2 Previous Studies : this chapter introduce the previous studies in this research area.

Chapter 3 Theoretical Background : this chapter presents a comprehensive review of all areas related to this research. The chapter starts by briefing diabetes and complications associated with diabetes mellitus and how is diagnosed, then important prediction algorithms are shown. After that various applications, techniques and researches of classification are presented. The chapter summarizes the literature in a simple way.

Chapter 4 Research Methodology : this chapter describes the methodology used and the steps followed to achieve the objectives of this research. A methodology is generally a guideline for solving a research problem. It contains the generic framework of the research and the steps required to carry out the research systematically.

Chapter 5 Experimental Results : this chapter gives the research details by showing the algorithms used to build the models and select the best features. Also the chapter presents the initial comparison between algorithms, the details of each classification model. Also it shows the evaluation of each model and a comparison of each model result with the actual values using graphs. the chapter also presents a detailed discussion and interpretation of the obtained results.

Chapter 6 Conclusion and future work : this chapter presents the research conclusion by highlighting the research contributions and the findings of its work. The chapter also presents suggestions and recommendations for future study.

Chapter Two

Previous Studies

A prediction for the diabetes is highly needed. Not only this, but also a prediction that is extremely automated and with less human interference. A diabetic prediction should meet the following specification; efficient modelling, applicability and accuracy and be trusted. It should be compatible with various diagnostic techniques.

Many prediction techniques are used, but the Multi-layer Percepton(MLP) is the most common Silva et al. (2008a); Elkamel et al. (2001); Selvaraj et al. (2010). ANN consists of fully connected layers. In the training phase of the prediction, the learning algorithm examines the inputs. While during the testing phase, it examines the outputs and the other unexamined parts during the training phase.

Anthropometrical Body surface scanning data was used to construct a classification modell for diabetes type II in Su et al. (2006). The model applies four data mining approaches. This model is meant to select and point out the appropriate and necessary decision tree for classifying diabetic diseases. It incorporates Artificial Neural Network, Decision Tree, Logistic Regression and Rough sets. In Tennis (2002) authors used the classification tree for the classification and regression with a binary target. It introduces ten attributes including age, sex, emergency department visits, comorbidity index, dyslipidemia, hypertension, cardiovascular disease, and retinopathy and end-stage renal disease. The cascade learning system which is based on generalized discriminant analysis was introduced by Polat et al. (2008). It has also linked the system with the least square support vector machine in order to perform the classification of diabetes diseases. This uses the classification accuracy, k-fold cross-validation method and confusion matrix.

A method to discover key attributes affecting diabetic diseases was introduced in Huang et al. (2007). The method is called feature selection method. Then it introduced the three classification complementary techniques including Naive Bayes and C4.5. In Çalisir and Dogantekin (2011) authors developed and upgraded the Linear Discriminant Analysis (LDA) and integrated it into the automatic diagnosis system. All these models functions primarily in the area of classification. But this method is proved to be accurate and well performed.

The fuzzy approaches have recently become the well-known approaches for improving classification models. Fuzzy Neural Networks (FNNs) and artificial neural networks have been recently integrated hybrid classification model that helps well in diagnosing and classifying the state of the diabetic diseases. This model was presented by Kahramanli and Allahverdi (2008). Multi-objective genetic programming approach is proposed by Mugambi and Hunter (2003) to develop Pareto optimal decision trees in diabetes classification. In Aslam et al. (2013), GP was used to generate new features by making combinations of the existing diabetes features.

Authors in Pradhan et al. (2012) propose a multi-class genetic programming (GP) based classifier design that will help the medical practitioner to confirm his/her diagnosis towards pre-diabetic, diabetic and non-diabetic patients.

Kumari and Chitra (2013) uses Support Vector Machine (SVM), a machine learning method as the classifier for diagnosis of diabetes. The machine learning method focus on classifying diabetes disease from high dimensional medical dataset. The experimental results obtained show that support vector machine can be successfully used for diagnosing diabetes disease.

Table 2.1: The Previous Studies

Author	Classifier	Dataset	Number of Features	Accuracy
Zahed Soltani [2016]	A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II	Pima Indian dataset	8	%89.56
Ramalingaswamy Cheruku [2016]	Diabetes Classification using Radial Basis Function Network by Combining Cluster Validity Index and BAT Optimization with Novel Fitness Function	Pima Indian dataset	8	%70.00
Mehmet Recep BOZKURT [2014]	Comparison of different methods for determining diabetes	Pima Indian dataset	8	%76.00
V. Anuja Kumari [2013]	Uses Support Vector Machine (SVM)	Pima Indian dataset	8	%78.00
Mehdi Khashei [2012]	A hybrid binary classification model is proposed for diabetes type II classification	Pima Indian dataset	8	%82.4
Muhammad Waqar Aslam [2010]	Detection of Diabetes Using Genetic Programming with comparative partner selection (CPS)	Pima Indian dataset	8	%76.5
This Study	Classification of Diabetes using Genetic Programming	Pima Indian dataset	8 4 3	%77.4 %77.0 %77.0

Chapter Three

Theoretical Background

The theoretical background is divided into three sections. The first section reviews the definition and classification of diabetes. The second section looks deeply in classification approaches. The third section is about the feature selection.

3.1 The Definition and Classification of Diabetes

Diabetes is the condition in which the body does not properly process food for use as energy. Most of the food we eat is turned into glucose, or sugar, for our bodies to use it for energy. The pancreas, an organ that lies near the stomach, makes a hormone called insulin to help glucose get into the cells of our bodies. When you have diabetes, your body either doesn't make enough insulin or can't use its own insulin as well as it should be. This causes sugars to build up in your blood. This is why many people refer to diabetes as sugar. Diabetes can cause serious health complications including heart disease, blindness, kidney failure, and lower-extremity amputations. Diabetes is consider the seventh leading cause of death .

3.1.1 Type 1 Diabetes

Type 1 diabetes is an autoimmune condition. It's caused by the body attacking its own pancreas with antibodies. In people with type 1 diabetes, the damaged pancreas doesn’t make insulin. This type of diabetes may be caused by a genetic predisposition. It could also be the result of faulty beta cells in the pancreas that normally produce insulin.Type 1 diabetes, previously called insulin-dependent diabetes mellitus (IDDM) or juvenile-onset diabetes, may account for 5 percent to 10 percent of all diagnosed cases of diabetes. A number of medical risks are associated with type 1 diabetes. Many of them stem from damage to the tiny blood vessels in your eyes (called diabetic retinopathy), nerves (diabetic neuropathy), and kidneys (diabetic nephropathy). Even more serious is the increased risk of heart disease and stroke. Treatment for type 1 diabetes involves taking insulin, which needs to be injected through the skin into the fatty tissue below. The methods of injecting insulin include:

i) Syringes.

ii) Insulin pens that use pre-filled cartridges and a fine needle.

iii) Jet injectors that use high pressure air to send a spray of insulin through the skin.

iv) Insulin pumps that dispense insulin through flexible tubing to a catheter under the skin of the abdomen.

3.1.2 Type 2 Diabetes

Type 2 diabetes is called adult-onset diabetes, but with the epidemic of obese and overweight kids, more teenagers are now developing type 2 diabetes. Type 2 diabetes was also called non-insulin-dependent diabetes. Type 2 diabetes is often a milder than diabetes type 1. Nevertheless, type 2 diabetes can still cause major health complications, particularly in the smallest blood vessels of the body that nourish the kidneys, nerves, and eyes. Type 2 diabetes also increases the risk of heart disease and strokes. With Type 2 diabetes, the pancreas usually produces some insulin. But the amount produced is not enough for the body needs, or the body cells are resistant to it. Insulin resistance, or lack of sensitivity to insulin, happens primarily in fat, liver, and muscle cells. People who are obese – more than 20 percent over their ideal body weight for their height – are at particularly high risk of developing type 2 diabetes and its related medical problems. Obese people have insulin resistance. With insulin resistance, the pancreas has to work hard to produce more insulin. But even then, there is not enough insulin to keep sugars normal. There is no cure for diabetes. Type 2 diabetes can, however, be controlled with weight management, nutrition, and exercises. Unfortunately, type 2 diabetes tends to progress, and diabetes medications are often needed. An A1C test is a blood test that estimates average glucose levels in your blood over the previous three months. Periodic A1C testing may be advised to see how well diet, exercise, and medications are working to control blood sugar and prevent organ damage. The A1C test is typically done a few times a year. Type 2 diabetes may account for about 90 percent to 95 percent of all diagnosed cases of diabetes. Risk factors for Type 2 diabetes include older age, obesity, family history of diabetes, prior history of gestational diabetes, impaired glucose tolerance, physical inactivity, and race/ethnicity. African Americans, Hispanic/Latino Americans, American Indians, and some Asian Americans and Pacific Islanders are at particularly high risk for type 2 diabetes.

3.1.3 Gestational diabetes

Diabetes that's triggered by pregnancy is called gestational diabetes (pregnancy, to some degree, leads to insulin resistance). It is often diagnosed in middle or late pregnancy. Because high blood sugar levels in a mother are circulated through the placenta to the baby, gestational diabetes must be controlled to protect the baby's growth and development.

According to the National Institutes of Health, the reported rate of gestational diabetes is between 2 percent to 10 percent of pregnancies. Gestational diabetes usually resolves itself after pregnancy. Having gestational diabetes does, however, put mothers at risk for developing type 2 diabetes later in life. Up to 10 percent of women with gestational diabetes develop type 2 diabetes. It can occur anywhere from a few weeks after delivery up to month or a year ahead.

With gestational diabetes, risks to the unborn baby are even greater than risks to the mother. Risks to the baby include abnormal weight gain before birth, breathing problems at birth, and higher obesity and diabetes risk later in life. Risks to the mother include needing a caesarean section due to an overly large baby, as well as damage to heart, kidney, nerves, and eye. Gestational diabetes occur more frequently in African Americans, Hispanic/Latino Americans, American Indians, and people with a family history of diabetes than in other groups. Obesity is also associated with higher risk. In some studies, nearly 40 percent of women with a history of gestational diabetes develop diabetes in the future. Other specific types of diabetes result from specific genetic syndromes, surgery, drugs, malnutrition, infections, and other illnesses. Such types of diabetes may account for 1 percent to 2 percent of all diagnosed cases of diabetes.

3.1.4 Complications associated with diabetes mellitus

i) Diabetic retinopathy is a leading cause of blindness and visual disability. Diabetes mellitus is associated with damage to the small blood vessels in the retina, resulting in loss of vision. Findings, consistent from study to study, make it possible to suggest that, after 15 years of diabetes, approximately 2 percent of people become blind, while about 10 percent develop severe visual handicap. Loss of vision due to certain types of glaucoma and cataract may also be more common in people with diabetes than in those without the disease.

ii) Good metabolic control can delay the onset and progression of diabetic retinopathy. Loss of vision and blindness in persons with diabetes can be prevented by early detection and treatment of vision-threatening retinopathy: regular eye examinations and timely intervention with laser treatment, or through surgery in cases of advanced retinopathy. There is evidence that, even in developed countries, a large proportion of those in need are not receiving such care due to lack of public and professional awareness, as well as an absence of treatment facilities. In developing countries, in many of which diabetes are now common, such care is inaccessible to the majority of the population.

iii) Diabetes is among the leading causes of kidney failure, but its frequency varies between populations and is also related to the severity and duration of the disease. Several measures to slow down the progress of renal damage have been identified. They include control of high blood glucose, control of high blood pressure, intervention with medication in the early stage of kidney damage, and restriction of dietary protein. Screening and early detection of diabetic kidney disease are an important means of prevention.

iv) Heart disease accounts for approximately 50 percent of all deaths among people with diabetes in industrialized countries. Risk factors for heart disease in people with diabetes include smoking, high blood pressure, high serum cholesterol and obesity. Diabetes negates the protection from heart disease which is difficult for women without diabetes experience. Recognition and management of these conditions may delay or prevent heart disease in people with diabetes.

v) Diabetic neuropathy is probably the most common complication of diabetes. Studies suggest that up to 50 percent of people with diabetes are affected to some degree. Major risk factors of this condition are the level and duration of elevated blood glucose. Neuropathy can lead to sensory loss and damage to the limbs. It is also a major cause of impotence in diabetic men.

vi) Diabetic foot disease, due to changes in blood vessels and nerves, often leads to ulceration and subsequent limb amputation. It is one of the most costly complications of diabetes, especially in communities with inadequate footwear. It results from both vascular and neurological disease processes. Diabetes is the most common cause of non-traumatic amputation of the lower limb, which may be prevented by regular inspection and good care of the foot.

3.1.5 How is diabetes diagnosed?

Diabetes is diagnosed with fasting sugar blood tests or with A1c blood tests, also known as glycated hemoglobin tests. A fasting blood sugar test is performed after you have had nothing to eat or drink for at least eight hours.

Normal fasting blood sugar is less than 100 mg/dl (5.6 mmol/l). You do not have to be fasting for an A1c blood test.

Diabetes is diagnosed by one of the following:

(i) Your blood sugar level is equal to or greater than 126 mg/dl (7 mmol/l).

(ii) You have two random blood sugar tests over 200 mg/dl (11.1 mmol/l) with symptoms.

(iv) You have an oral glucose tolerance test with results over 200 mg/dl (11.1 mmol/l).

(v) Your A1c test is greater than 6.5 percent on two separate days.

An A1c test should be performed in a laboratory using a method that is certified by the National Glycohemoglobin Standardization Program (NGSP) and standardized to the Diabetes Control and Complications Trial (DCCT) assay.

- Fasting Glucose Test Normal: Less than 100 Pre-diabetes: 100-125 Diabetes: 126 or higher
- Random (anytime) Glucose Test Normal: Less than 140 Pre-diabetes: 140-199 Diabetes: 200 or higher
- A1c Test Normal: Less than 5.7 percent Pre-diabetes: 5.7 - 6.4 percent Diabetes: 6.5 percent or higher

3.2 Classification Approaches

In this section, the basic concepts and modelling approaches for classification are briefly reviewed.

3.2.1 Multi-Layer Perceptrons (MLPs)

Artificial neural networks (ANNs) are computer systems developed to mimic the operations of the human brain by mathematically modelling its neurophysiological structure. Artificial neural networks have been shown to be effective at approximating complex nonlinear functions Zhang (2001). For classification tasks, these functions represent the shape of the partition between classes. In artificial neural networks, computational units called neurons replace the nerve cells and the strengths of the interconnections are represented by weights, in which the learned information is stored. This unique arrangement can acquire some of the neurological processing ability of the biological brain such as learning and drawing conclusions from experience. Artificial neural networks combine the flexibility of the boundary shape found in K-nearest neighbor with the efficiency and low storage requirements of discriminant functions. Like the K-nearest neighbor, artificial neural networks are data driven; there are no assumed model characteristics or distributions, as is the case with discriminant analysis.

Multi-layer perceptrons (MLPs) are one of the most important and widely used forms of artificial neural networks for modelling, forecasting, and classification Silva et al. (2008b). These models are characterized by the network of three layers of simple processing units connected by acyclic links. Data enters the network through the input layer, moves through hidden layer, and exits through the output layer. Each hidden layer and output layer node collects data from the nodes above it (either the input layer or hidden layer) and applies an activation function. Activation functions can take several forms. The type of activation function is indicated by the situation of the neuron within the network. In the majority of cases input layer neurons do not have an activation function, as their role is to transfer the inputs to the hidden layer. In practice, simple network structure that has a small number of hidden nodes often works well in out-of-sample forecasting. This may be due to the overfitting effect typically found in the neural network modelling process. An over-fitted model has a good fit to the sample used for model building but has poor generalizability to data out of the sample. There exist many different approaches such as the pruning algorithm, the polynomial time algorithm, the canonical decomposition technique, and the network information criterion for finding the optimal architecture of an artificial neural network. These approaches can be generally categorized as follows Khashei and Bijari (2010):

i) Empirical or statistical methods that are used to study the effect of internal parameters and choose appropriate values for them based on the performance of model. The most systematic and general of these methods utilizes the principles from Taguchis design of experiments.

ii) Hybrid methods such as fuzzy inference where the artificial neural network can be interpreted as an adaptive fuzzy system or it can operate on fuzzy instead of real numbers.

iii) Constructive and/or pruning algorithms that, respectively, add and/or remove neurons from an initial architecture using a previously specified criterion to indicate how artificial neural network performance is affected by the changes.

iv) Evolutionary strategies that search over topology space by varying the number of hidden layers and hidden neurons through application of genetic operators and evaluation of the different architectures according to an objective function Benardos and Vosniakos (2007).

Although many different approaches exist in order to find the optimal architecture of an artificial neural network, these methods are usually quite complex in nature and are difficult to implement. Furthermore, none of these methods can guarantee the optimal solution for all real forecasting problems. To date, there is no simple clearcut method for determination of these parameters and the usual procedure is to test numerous networks with varying numbers of hidden units, estimate generalization error for each and select the network with the lowest generalization error. Once a network structure is specified, the network is ready for training a process of parameter estimation. The parameters are estimated such that the cost function of neural network is minimized.

3.2.2 K-Nearest Neighbor (KNN)

The K-nearest neighbor (KNN) model is a well known supervised learning algorithm for pattern recognition that first introduced by Fix and Hodges in 1951, and is still one of the most popular nonparametric models for classification problemsTsypin and Röder (2011). K-nearest neighbor assumes that observations, which are close together, are likely to have the same classification. The probability that a point x belongs to a class can be estimated by the proportion of training points in a specified neighborhood of x that belong to that class.

The point may either be classified by majority vote or by a similarity degree sum of the specified number (k) of nearest points. In majority voting, the number of points in the neighborhood belonging to each class is counted, and the class to which the highest proportion of points belongs is the most likely classification of x. The similarity degree sum calculates a similarity score for each class based on the K-nearest points and classifies x into the class with the highest similarity score. Due to its lower sensitivity to outliers, majority voting is more commonly used than the similarity degree sum Chaovalitwongse et al. (2007). In order to determine which points belong in the neighborhood, the distances from x to all points in the training set must be calculated. Any distance function that specifies which of two points is closer to the sample point could be employed (Fix and Hodges 1951). In general the following steps are performed for the K-nearest neighbor model:

i) Chosen of k value.

ii) Distance calculation.

iii) Distance sort in ascending order.

iv) Finding k class values.

v) Finding dominant class.

One challenge to use the K-nearest neighbour is to determine the optimal size of k, which acts as a smoothing parameter. A small k will not be sufficient to accurately estimate the population proportions around the test point. A larger k will result in less variance in probability estimates but the risk of introducing more bias. K should be large enough to minimize the probability of a non-Bayes decision, but small enough that the points included give an accurate estimate of the true class. Authors in Enas and Choi (1986) found that the optimal value of k depends upon the sample size and covariance structures in each population, as well as the proportions for each population in the total sample. This model presents several advantages:

- Its mathematical simplicity, which does not prevent it from achieving classification results as good as (or even better than) other more complex pattern recognition techniques.
- It is free from statistical assumptions, such as the normal distribution of the variables.
- Its effectiveness does not depend on the space distribution of the classes.

In additional, when the boundaries between classes cannot be described as hyper-linear or hyper-conic, K-nearest neighbour performs better than the

(LDA) and (QDA) functions. Author in Enas and Choi (1986) found that the linear discriminant performs slightly better than K-nearest neighbour when population covariance matrices are equal, a condition that suggests a linear boundary. As the differences in the covariance matrices increases, K-nearest neighbor performs increasingly better than the linear discriminant function. However, despite of the all advantages cited for the K-nearest neighbour models, they also have some disadvantages. K-nearest neighbour model cannot work well if large differences are present in the number of samples in each class. K-nearest neighbor provides poor information about the structure of the classes and of the relative importance of each variable in the classification. Furthermore, it does not allow a graphical representation of the results, and in the case of large number of samples, the computation can become excessively slow. In addition, K-nearest neighbour model much higher memory and processing requirements than other methods. All prototypes in the training set must be stored in memory and used to calculate the Euclidean distance from every test sample. The computational complexity grows exponentially as the number of prototypes increases.

3.2.3 Support Vector Machines (SVMs)

Support vector machines (SVMs) are a new pattern recognition tool theoretically founded on Vapniks statistical learning theory Vapnik (1998). Support vector machines, originally designed for binary classification, employs supervised learning to find the optimal separating hyper-plane between the two groups of data. Having found such a plane, support vector machines can then predict the classification of an unlabeled example by asking on which side of the separating plane the example lays. Support vector machine acts as a linear classifier in a high dimensional feature space originated by a projection of the original input space, the resulting classifier is in general non-linear in the input space and it achieves good generalization performances by maximizing the margin between the two classes.

Support vector machines differ from discriminant analysis in two significant ways. First, the feature space of a classification problem is not assumed to be linearly separable. Rather, a nonlinear mapping function (also called a kernel function) is used to represent the data in higher dimensions where the boundary between classes is assumed to be linear Murtagh and Farid (2001). Second, the boundary is represented by support vector machines instead of a single boundary. Support vectors run through the sample patterns which are the most difficult to classify, thus the sample patterns that are closest to the actual boundary. Over-fitting is prevented by specifying a maximum margin that separates the hyper plane from the classes Andrew (2000).

3.2.4 Decision tree

Decision tree is a supervised approach to classify a large number of datasets that make up the structure of the rules are simple, clear and easy to understand Munandar and Winarko (2015). The decision tree is used to examine the data and form a rules in a tree that will be used for forecasting needs. There are many types of algorithms in the decision tree that can be utilized for a variety of needs Iyer et al. (2015), some of them and used in this study are J48 which is the development of C4.5, Naive Bayess Tree (NBTree) and Reduced Error Pruning Tree (REPTree).

3.2.5 Feed Forward Back propagation Neural Network

Neural systems are prescient model that have capacity to learn, examine, sort out the information and foresee test outcomes in like manner. Among a few sorts of neural systems, Feed forward neural system is generally utilized in therapeutic conclusion applications and others. These systems are prepared by a situated of examples called preparing set, whose result is as of now known. In our study, Multilayer Perceptron Feed forward back spread Neural Network prepared with Levenberg Marquardt (LM) calculation is requisitioned arrangement. LM preparing calculation does not get stuck in nearby minima and produces a superior expense capacitybeyli (2009).

Feed Forward NN comprises of info, shrouded and a yield layer, and the information works in forward course, and the lapse is back spread to upgrade the weights at each age to lessen blundersParashar et al. (2014).

Extensive studies by many researchers have demonstrated higher performance and accuracy in predicting clinical outcomes of diabetes diagnosis using neural network strategies (Table 3.1).

3.2.6 Genetic Programming

GP works on a population of individuals, each of which represents a potential solution to a problem. GP was introduced by J. Koza in 1992 at Stanford. A flow chart for GP evolutionary process is shown in Figure 3.1. In order to solve a problem, it is necessary to specify the following Koza (1992):

- **The terminal set:** A set of input variables or constants.
- **The function set:** A set of domain specific functions used in conjunction with the terminal set to construct potential solutions to a given

Table 3.1: Artificial intelligence approaches for early diabetes detection

Author	Algorithm	Accuracy
Kayaer, Yildirim	MLP	77.08
Kayaer, Yildirim	RBF	68.23
Kordos et al.	k-nearest-neighbor	77
Barakat et al.	SVM	94
Ster, Dobnikar	k-NN	71.9
Ster, Dobnikar	CART	72.8
Ster, Dobnikar	MLP	75.2
Ster, Dobnikar	LVQ	75.8
Ster, Dobnikar	LDA	77.5
Polat et al	GDA and LS-SVM	78.21
Polat, Gunes	PCA-ANFIS	89.47
Dogantekin et al	LDA-ANFIS	84.61
Ubeyli	MLPNN	91.53
Ubeyli	MME	99.17

problem. For symbolic regression this could consist of a set of basic mathematical functions, while Boolean and conditional operators could be included for classification problems.

- **The fitness function:** Fitness is a numeric value assigned to each member of a population to provide a measure of the appropriateness of a solution to the problem in question.

- **The termination criterion:** This is generally a predefined number of generations or an error tolerance on the fitness.

In order to further illustrate the coding procedure and the genetic operators used for GP, a symbolic regression example will be used. Consider the problem of predicting the numeric value of an output variable, y, from two input variables a and b. One possible symbolic representation for y in terms of a and b would be: $y = \frac{a-b}{3}$.

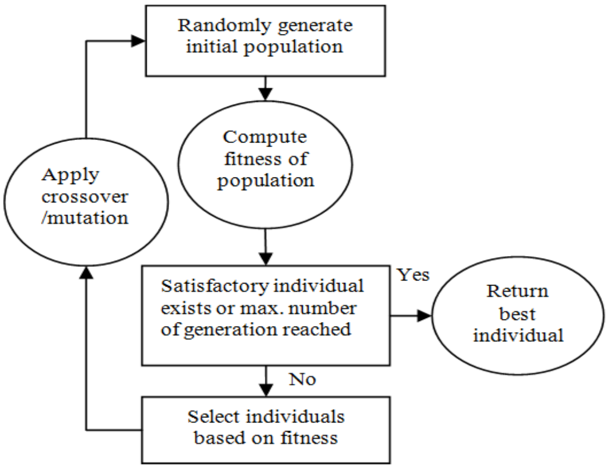

Randomly generate initial population
Compute fitness of population
Apply crossover /mutation
Satisfactory individual exists or max. number of generation reached
Yes
Return best individual
No
Select individuals based on fitness

Figure 3.1: Flow chart of the GP algorithm Sheta et al. (2014)

Figure 3.2 demonstrates how this expression may be represented as a tree structure. With this tree representation, the genetic operators of crossover and mutation must be posed in a fashion that allows the syntax of resulting expressions to be preserved. Figure 3.3 shows a valid crossover operation where the two parent expressions are given in Equations 3.1 and 3.2. The two offspring are given in Equation 3.3 and 3.4. Parent 1 (y^1) and Parent 2 (y^2) are presented in Equations 3.1 and 3.2. The developed offspring 1 (y^3) and offspring 2 (y^4) are presented in Equation 3.3 and 3.4.

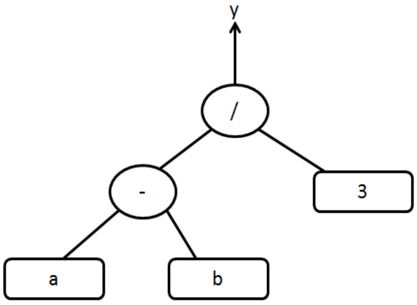


Figure 3.2: Representation of a numeric expression using tree structure

$$y^1 = \frac{a-b}{3} \tag{3.1}$$

$$y^2 = (c-b) \times (a+c) \tag{3.2}$$

$$y^3 = \frac{a-b}{a+c} \tag{3.3}$$

$$y^4 = (c - b) \times 3 \tag{3.4}$$

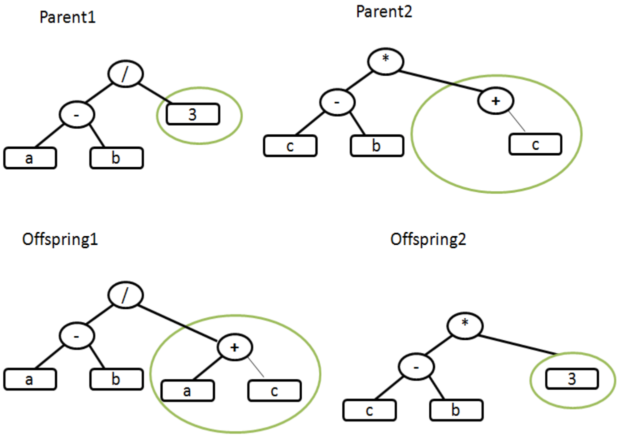


Figure 3.3: A typical crossover operation

3.2.7 Multigene Symbolic Regression GP

Typically, symbolic regression is performed by using GP to evolve a population of trees, each of which encodes a mathematical equation that predicts $n \times 1$ vector of outputs y using a corresponding $n \times m$ matrix of inputs X where N is the number of observations of the response variable and M is the number of input (predictor) variables Koza (1992). In contrast, in Multigene symbolic regression each symbolic model (and each member of the GP population) is a weighted linear combination of the outputs from a number of GP trees, where each tree may be considered to be a gene Koza (1991). For example, the Multigene model shown in Figure 3.4 predicts an output variable using input variables x_1, x_2, x_3. This model structure contains non-linear terms (e.g. the hyperbolic tangent) but is linear in the parameters with respect to the coefficients $\alpha_0, \alpha_1, \alpha_2$.

In practice, the user specifies the maximum number of genes G_{max} and the maximum tree depth D_{max} therefore an exert can control the model complexity. In particular, we have found that enforcing stringent tree depth restrictions (i.e. maximum depths of 4 or 5 nodes) often allows the evolution of relatively compact models that are linear combinations of each model, the linear coefficients are estimated from the training data using ordinary least squares techniques.

Hence, Multigene GP combines the power of classical linear regression with the ability to capture non-linear behavior without the need to pre-specify the structure of the non-linear model. In Hinchliffe et al. (1996b) it was shown that Multigene symbolic regression can be more accurate and computationally efficient than the standard GP approach for symbolic regression.

Here, the first parent individual contains the genes (G_1 G_2 G_3) and the second contains the genes (G_4 G_5 G_6 G_7) where G_{max} equals to 5. Two randomly selected crossover points are created for each individual. The genes enclosed by the crossover points are denoted by [ ].

(G_1 [G_2] G_3) (G_4 [G_5 G_6 G_7])

The genes enclosed by the crossover points are then exchanged resulting in two new individuals as follows:

(G_1 G_5 G_6 G_7 G_3) (G_4 G_2)

Two point high level crossover allows the acquisition of new genes for both individuals but also allows genes to be removed. If an exchange of genes results in an individual containing more genes than G_{max} then genes are randomly selected and deleted until the individual contains G_{max} genes.

The user can set the relative probabilities of each of these recombination processes. These processes are grouped into categories called events. The user can then specify the probability of crossover events, direct reproduction events and mutation events. These must sum to one. The user can also specify the probabilities of event subtypes, e.g. the probability of a two point high level crossover taking place once a crossover event has been selected, or the probability of a sub tree mutation once a mutation event has been selected.

An example of Multigene model is shown in Figure 3.4. The presented model can be introduced mathematically as given in Equation 3.5. GPTIPS Matlab Toolbox provides default values for each of these probabilities so the user does not need to explicitly set them Searson et al. (2010).

$$\alpha_{\mathbf{0}} + \alpha_{\mathbf{1}}(0.41x_1 + tanh(x_2x_3)) + \alpha_{\mathbf{2}}(0.45x_3 + \sqrt{x_2}) \tag{3.5}$$

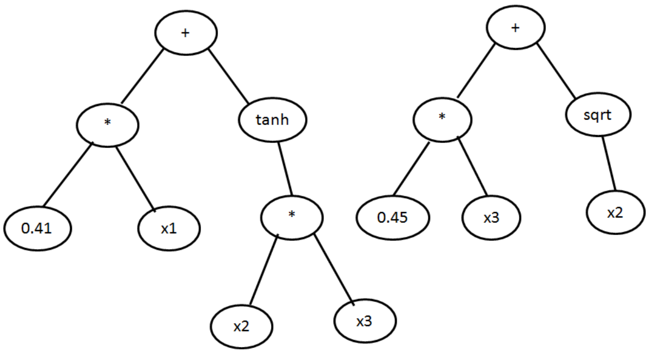


Figure 3.4: Example of a Multigene Symbolic Regression Model

3.3 Feature Selection

The feature selection algorithm removes the irrelevant and redundant features from the original dataset to improve the classification accuracy. The feature selections also reduce the dimensionality of the dataset increase the learning accuracy, improving result comprehensibility. The feature selection avoid over fitting of data. The feature selection also known as attributes selection which is used for best partitioning the data into individual class. The feature selection method also includes the selection of subsets, evaluation of subset and evaluation of selected feature. The two search algorithms forward selection and backward eliminations are used to select and eliminate the appropriate feature. The feature selection is a three step process namely search, evaluate and stop. Different kinds of feature selection algorithms have been proposed. The feature selection techniques are categorized into three Filter method, Wrapper method, and Embedded method. Every feature selection algorithm uses any one of the three feature selection techniques Kumar and (2014). According to the class label present or not the feature selection can be further classified into two categories. Supervised and unsupervised feature selections. In the supervised method the relevance between the feature and the class is evaluated by calculating the correlation between the class and the feature. The relevance is evaluated by checking some property of the data in an unsupervised method. In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for four reasons:

i) simplification of models to make them easier to interpret by researchers.

ii) shorter training times.

iii) to avoid the curse of dimensionality.

iv) enhanced generalization by reducing overfitting (formally, reduction of variance).

3.3.1 General Approach for Feature Selection

There are three general approaches for feature selection figure 3.5 shows the Feature Selection for Classification.

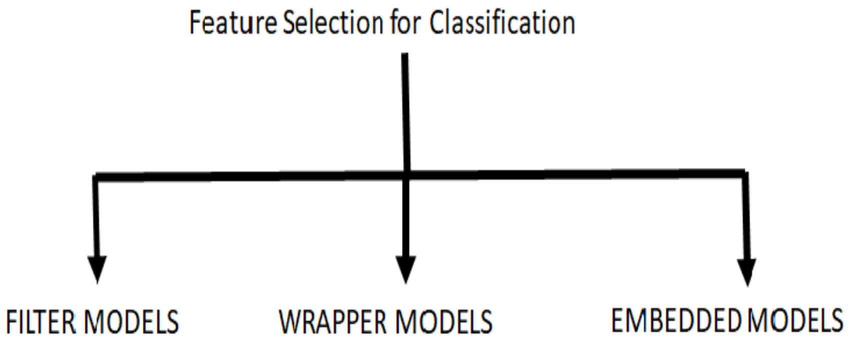


Figure 3.5: General Approach for Feature Selection

3.3.2 Filter Method

The filter approach incorporates an independent measure for evaluating features subsets without involving a learning algorithm. This approach is efficient and fast to compute (computationally efficient). However, filter methods can miss features that are not useful by themselves but can be very useful when combined with others. The graphical representation of the filter model is shown in Figure 3.6

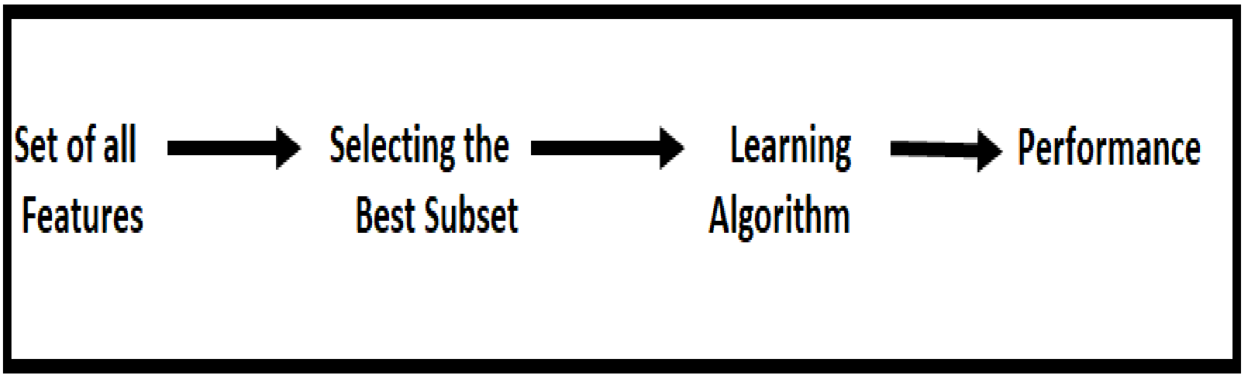


Figure 3.6: Filter Method for feature selection

3.3.3 Wrapper Method

Wrapper methods evaluate subsets of variables which allows, unlike filter approaches, to detect the possible interactions between variables. In wrapper methods, we try to use a subset of features and train a model using them. Based on the inferences that we draw from the previous model, we decide to add or remove features from your subset. The problem is essentially reduced to a search problem. These methods are usually computationally very expensive. Some common examples of wrapper methods are forward feature selection, backward feature elimination, recursive feature elimination, etc. Figure 3.7 shows the Wrapper Method for Feature selection.

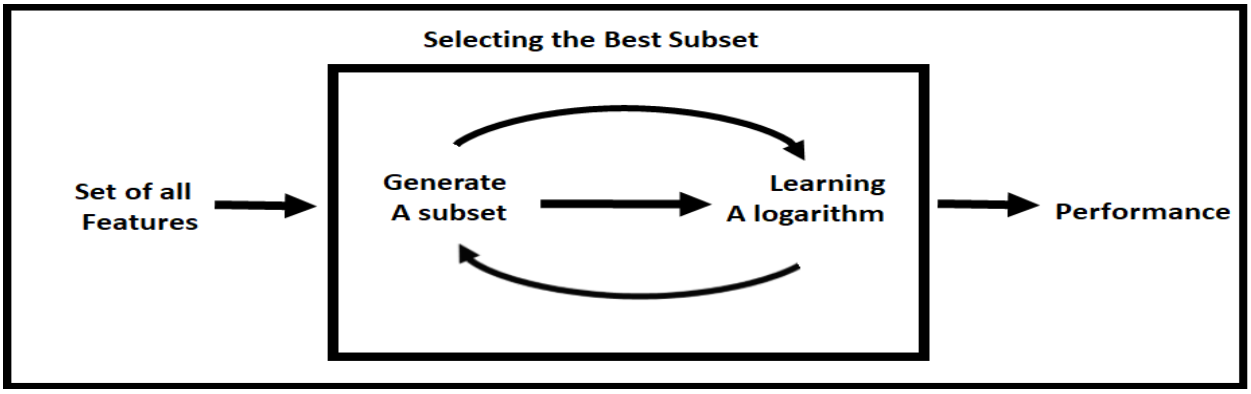


Figure 3.7: Wrapper Method for Feature selection

Forward Selection: Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

Backward Elimination: In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.

Recursive Feature elimination: It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.

The two main disadvantages of these methods are :

i) The increasing overfitting risk when the number of observations is insufficient.

ii) The significant computation time when the number of variables is large.

3.3.4 Embedded Methods

Embedded methods combine the qualities of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods. Figure 3.8 shows Embedded method for Feature selection.

Some of the most popular examples of these methods are LASSO and RIDGE regression which have inbuilt penalization functions to reduce overfitting.

i) Lasso regression performs L1 regularization which adds penalty equivalent to absolute value of the magnitude of coefficients.

ii) Ridge regression performs L2 regularization which adds penalty equivalent to square of the magnitude of coefficients.

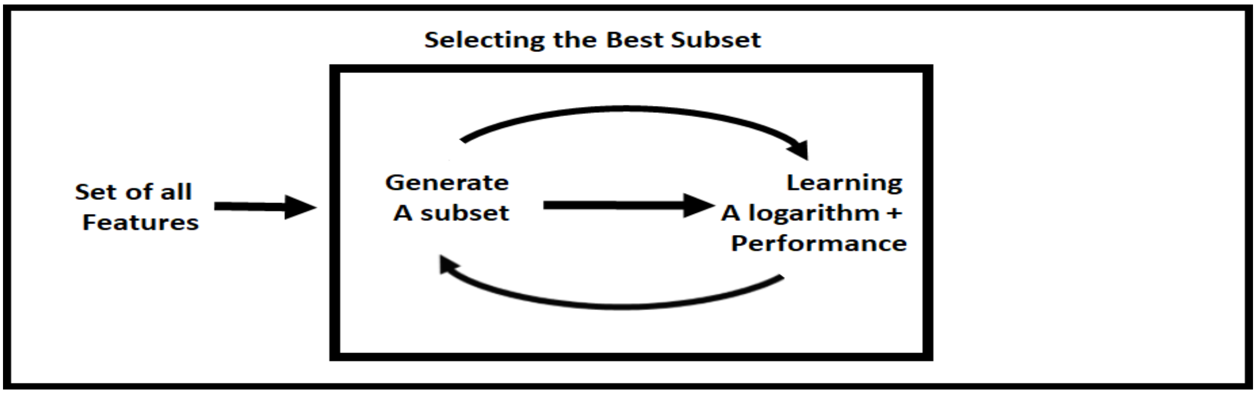


Figure 3.8: Embedded method for Feature selection

Chapter Four

Research Methodology

4.1 Introduction

This chapter presents the methodology followed in this research. It shows the steps followed to achieve the research objectives.The goal of this dissertation is the development of diabetes prediction and classification framework that improves the prediction of diabetes performance. Another objective is to select the weighted features from the mathematical model produced by Genetic programming (GP) model. The first section is about the dataset that used in this research. The second section shows the research framework. The third section is about the evaluation criteria.

4.2 Data Set Description

The kind of the data set use in this research is concise in following table 4.1.Pima Indian is a homogeneous group that inhabits the area around American, but they are popular for being the most infected group with type II diabetes. Pima Indians diabetes data can even be retrieved from UCI Machine Learning Repository's web siteAslam and Nandi (2010). So, they are subject of intense studies in type II diabetes. The detailed descriptions of the data set are available at UCI repository. This data set was diagnosis of diabetes of Pima Indians. Based on personal data, such as age, number of times pregnant, and the results of medical examinations, e.g., blood pressure, body mass index, result of glucose tolerance test, etc., it is tried to decide whether a Pima Indian individual was diabetes positive or not. Pima Indian Diabetes Data (PIDD) set is publicly available from the machine learning database at UCI table 4.2 show sample of data. All patients represented in this data set are females with at least 21 years old of Pima Indian heritage living near Phoenix, Arizona. The problem posed here is to predict whether a person would test positive given a number of physiological measurements and medical test results. The data set in the UCI repository contains 768 observations and 9 variables with no missing values reported. However, as some researchers point out, there are a number of impossible values, such as 0 body mass index and 0 plasma glucose. Furthermore, one attribute (2 hour serum insulin) contains almost 50 percent impossible values. To keep the sample size reasonably large, this attribute is removed from analysis. There are 236 observations that have at least one impossible value of glucose, blood pressure, triceps skin thickness, and body mass index. There are nine variables, including the binary response variable, in this data set; all other attributes are numeric valued. The attributes are given below:

1) Number of times pregnant .

2) Plasma glucose concentration a 2 hours in an oral glucose tolerance test.

3) Diastolic blood pressure (mm Hg).

4) Triceps skin fold thickness (mm) .

5) 2 Hour serum insulin (mu U/ml) .

6) Body mass index (weight in kg/(height in m)2).

7) Diabetes pedigree function.

8) Age (years) .

9) Class variable (0 or 1) .

Table 4.1: Inputs and Output for Diabetic prediction model

Inputs	The number of times pregnant	x_1
	The results of an oral glucose tolerance test	x_2
	Diastolic blood pressure (mm/Hg)	x_3
	E Triceps skin fold thickness (mm)	x_4
	2 h serum insulin (micro U/ml)	x_5
	Body mass index	x_6
	Diabetes pedigree function	x_7
	Age (year)	x_8
Output	Predicted class	y

Table 4.2: Sample of Pima Indian dataset

No.	preg	plas	pres	skin	insu	mass	pedi	age	class
1	6	148	72	35	0	33.6	0.627	50	tested_positive
2	1	85	66	29	0	26.6	0.351	31	tested_negative
3	8	183	64	0	0	23.3	0.672	32	tested_positive
4	1	89	66	23	94	28.1	0.167	21	tested_negative
5	0	137	40	35	168	43.1	2.288	33	tested_positive
6	5	116	74	0	0	25.6	0.201	30	tested_negative
7	3	78	50	32	88	31	0.248	26	tested_positive
8	10	115	0	0	0	35.3	0.134	29	tested_negative
9	2	197	70	45	543	30.5	0.158	53	tested_positive
10	8	125	96	0	0	0	0.232	54	tested_positive

4.3 Research Framework

The framework should be well prepared and organized to describe the exact steps that followed, research phases, experiments, and results evaluation methods. Figure 4.1 illustrate the framework for this research.

In order to accomplish this objective, the work in this dissertation is divided into five tasks.

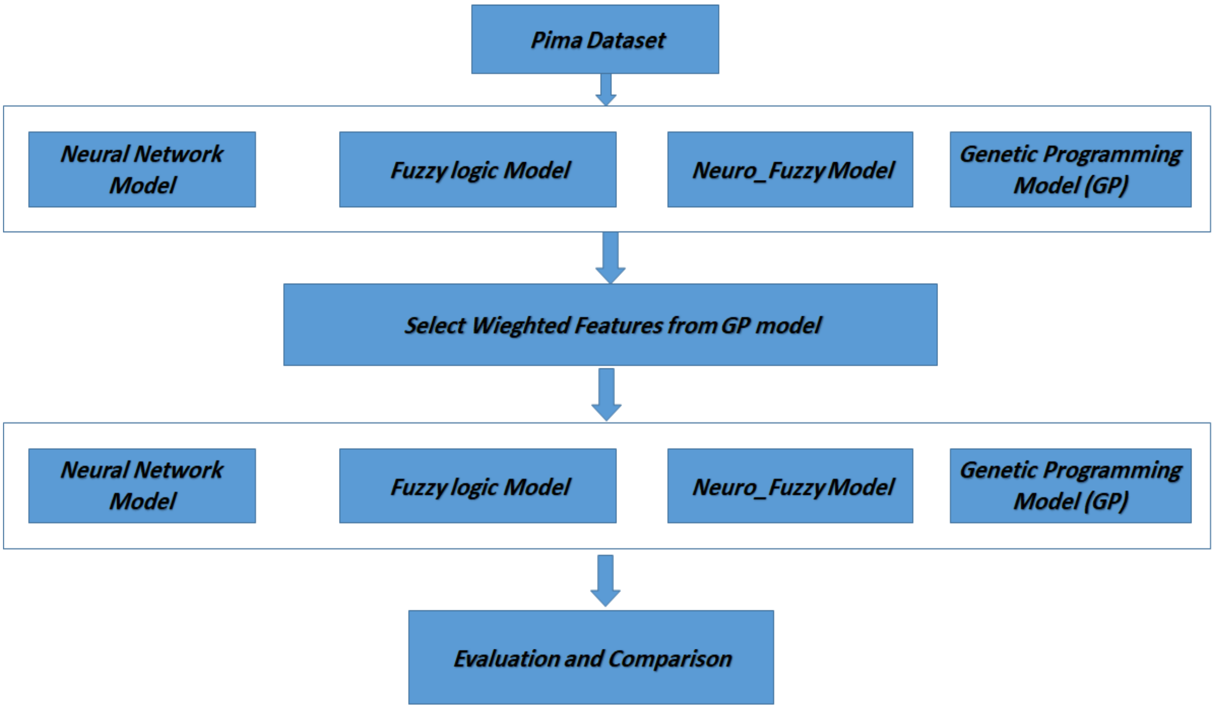

Pima Dataset
Neural Network Model
Fuzzy logic Model
Neuro_Fuzzy Model
Genetic Programming Model (GP)
Select Wieghted Features from GP model
Neural Network Model
Fuzzy logic Model
Neuro_Fuzzy Model
Genetic Programming Model (GP)
Evaluation and Comparison

Figure 4.1: Research Framework

1. Neural Network model.

2. Fuzzy model.

3. Genetic Programming model.

4. Feature Extraction Model.

4.3.1 Neural Network model

The first task focuses on developing a prediction model using Neural Network as shown in 4.2. Neural networks are useful when we have training pattern set Sheta et al. (2009). We do not need any knowledge of the modeling of the problem. A trained neural network is a black-box that represents knowledge in its distributed structure. However, any prior knowledge of the problem cannot be incorporated into the learning process Zitar and Al-Jabali

(2005). It is difficult for human beings to understand the internal logic of the system. Nevertheless, by extracting rules from neural networks, users can understand what neural networks have learned and how neural networks predict. Learning system is fed with the input data and generates output, which is then compared with the target to compute the error signal by arbitrator.The error is sent to the learning system for further training until the minimum value of error is generated.

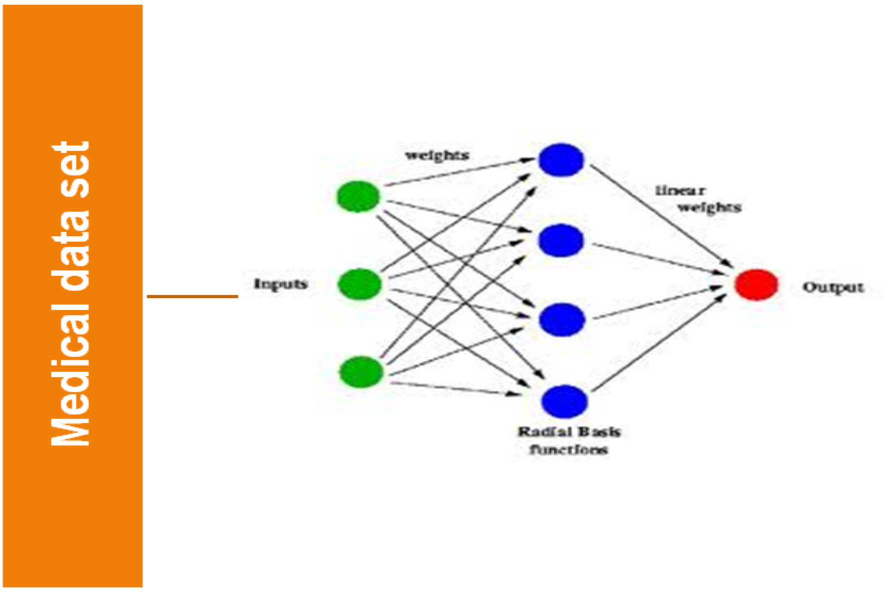


Figure 4.2: Neural Network model

4.3.2 Fuzzy model

The second task is using Fuzzy logic to build fuzzy rule base figure 4.3. During the last few years observations show an escalating use of fuzzy set theory in the field of medical diagnostics. Obviously, fuzzy set theory responds effectively to the non-statistical uncertainty, which is circumscribed in problems of the medical domain Lekkas and Mikhailov (2010). Fuzzy systems usually generate human interpretable rules which take the form of IF-THEN statements. Such rules which correspond to the knowledge granules of the diagnostic systems can be comprehended by human operators.

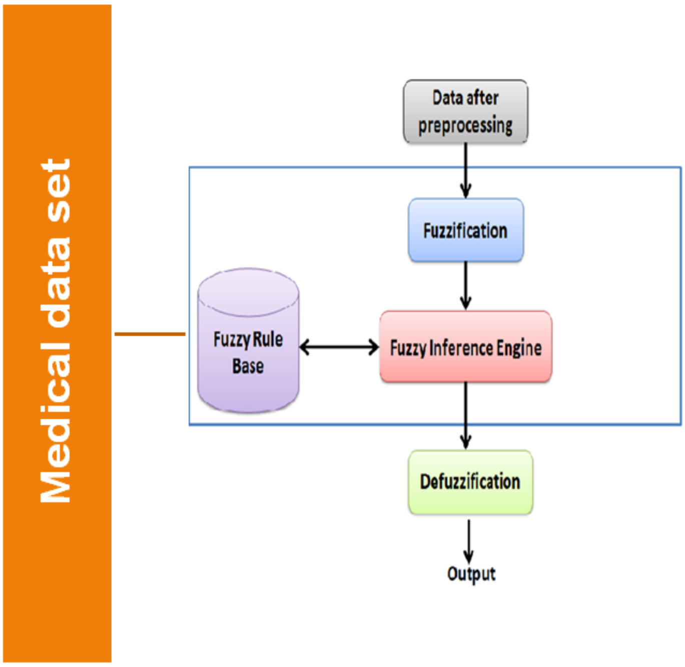


Figure 4.3: Fuzzy model

4.3.3 Genetic Programming model

The Third task is to use Genetic Programming (GP) to build a mathematical model for the classification figure 4.4. When the task is building an empirical mathematical model of data acquired from a process or system, the GP is often known as symbolic regression Al-Afeef et al. (2010). Unlike traditional regression analysis (in which the user must specify the structure of the model), GP automatically evolves both the structure and the parameters of the mathematical model Hinchliffe and Willis (2003). Symbolic regression has had both successful academic and industrial applications.

4.3.4 Feature Extraction Model

The last task is to select the best attributes that affect on the performance of prediction and classification model. It is possible to use regression methods to solve classification tasks. In order to apply the continuous prediction

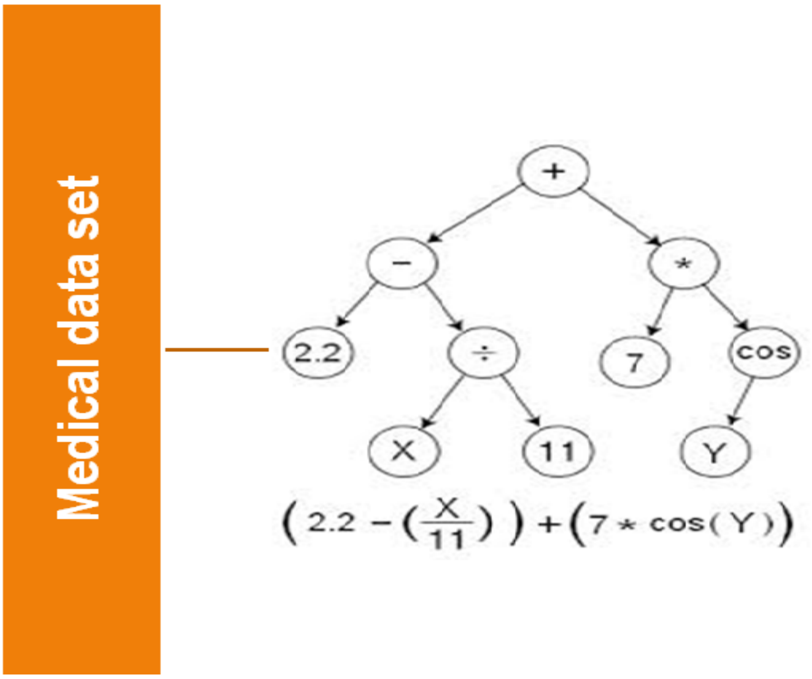


Figure 4.4: Genetic Programming model

technique of regression models to discrete classification problems, an approximation of the conditional class probability function can be considered. During classification, the class whose model yields the greatest approximated probability value is chosen as the predicted class.

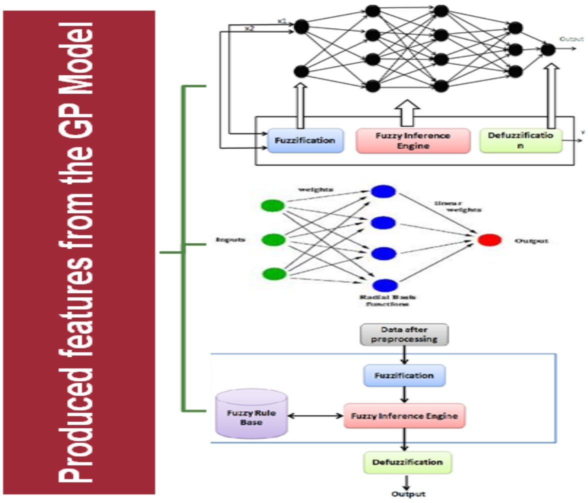


Figure 4.5: Feature Extraction Model

4.3.5 Evaluation Criteria

The Evaluation Criteria we used to evaluate the model is:

- Sensitivity (Sens):

$$Sens = \frac{TP}{TP + FN} \tag{4.1}$$

- Specificity (Spec):

$$Spec = \frac{TN}{FP + TN} \tag{4.2}$$

- Positive Predicted Value (PPV):

$$PPV = \frac{TP}{TP + FP} \tag{4.3}$$

- Negative Predicted Value (NPV):

$$NPV = \frac{TN}{FN + TN} \tag{4.4}$$

- Accuracy (Acc):

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.5}$$

Given that:

- True Positive (TP): Sick people correctly diagnosed as sick.
- False Positive (FP): Healthy people incorrectly identified as sick.
- True Negative (TN): Healthy people correctly identified as healthy.
- False Negative (FN): Sick people incorrectly identified as healthy.

4.4 Genetic Programming Toolbox For Multi-gene Symbolic Regression (GPTIPS)

GPTIPS is a free, open source MATLAB based software platform for symbolic data mining (SDM). It uses a multigene variant of the biologically inspired machine learning method of genetic programming (MGGP) as the engine that drives the automatic model discovery process. Symbolic data mining is the process of extracting hidden, meaningful relationships from data in the form of symbolic equations. In contrast to other data-mining methods, the structural transparency of the generated predictive equations can give new insights into the physical systems or processes that generated the data. Furthermore, this transparency makes the models very easy to deploy outside of MATLAB.

The rationale behind GPTIPS is to reduce the technical barriers to using, understanding, visualising and deploying GP based symbolic models of data, whilst at the same time remaining highly customisable and delivering robust numerical performance for power users Searson (2015).

GPTIPS is a predominantly command line driven open source toolbox that requires only a basic working knowledge of MATLAB. A run is configured by a simple configuration M file and there are a number of command line functions to facilitate postrun analyses of the results. Whilst not an exhaustive list, GPTIPS currently contains the following configurable GP features: tournament selection and plain lexicographic tournament selection Luke and Panait (2002) , elitism, three different tree building methods (full, grow and ramped half and half) and six different mutation operators:

(1) subtree mutation

(2) mutation of constants using an additive Gaussian perturbation

(3) substitution of a randomly selected input node with another randomly selected input node

(4) set a randomly selected constant to zero

(5) substitute a randomly selected constant with another randomly generated constant

(6) set a randomly selected constant to one.

In addition, GPTIPS can, without modification in the majority of cases, use nearly any built in MATLAB function as part of the function set for a run. The user can also write bespoke function node M files and fitness functions. In addition, GPTIPS has a number of features that are specifically aimed at the creation, analysis and simplification of multigene symbolic regression models. These include:

(1) use of a holdout validation set during training to mitigate the effects of overfitting.

(2) graphical display of the results of symbolic regression for any multigene model in the final population.

(3) mathematical simplification of any model.

(4) conversion to LaTex format of any model.

(5) conversion to PNG (portable network graphics) file of the simplified equation of any model.

(6) conversion of any model to standalone M file for use outside GPTIPS.

(7) graphical display of the statistical significance of each gene in a model.

(8) functions to reduce the complexity of any model using gene knockouts to explore the trade off of model accuracy against complexity.

(9) graphical population browser to explore the trade off surface of complexity/accuracy.

(10) graphical input frequency analysis of individual models or of a user specified fraction of the population to facilitate the identification of input variables that are relevant to the output.

The Symbolic Math toolbox (a commercial toolbox available from the vendors of MATLAB) is required for the majority of the post run simplification and model conversion features and the Statistics Toolbox is required for the display of gene statistical significance. The core functionality of GPTIPS and the ability to evolve multigene models does not, however, require any specific toolboxes. In the following section, some of these features will be demonstrated using a real world modelling example.

Chapter Five

Result and Discussion

5.1 Introduction

The research target could be divided into two parts: the first is to build a mathematical model using genetic programming then filter the model parameters to select the best features that affected on the performance, and the second is to build a prediction model that uses the selected features to predict the diabetes class. To achieve these goals, three different models were used (GP,ANN,Fuzzy logic). This chapter starts by describing the datasets, then basic statistical analysis about these datasets is shown. After that the model evaluation criteria is describe.

5.2 Models Inputs and Output

Pima Indian is a homogeneous group that inhabits the area around American, but they are popular for being the most infected group with type II diabetes. Pima Indians diabetes data can even be retrieved from UCI Machine Learning Repository's web site Aslam and Nandi (2010). So, they are subject of intense studies in type II diabetes. The data consist of eight input variables and one output (0,1). The GP mathematical model has the inputs and output presented in Table 5.1. Table 5.2 shows statistical analysis for mean and standard deviation in Pima Indians Diabetes Dataset. We used 500 samples as a training set and 100 samples as a testing set. The data set was normalized according to Equation 5.1.

$$x^{new} = \frac{x^{old} - x_{min}}{x_{max} - x_{min}} \tag{5.1}$$

x_{max} and x_{min} are the maximum and minimum values of the array x, respectively. x^{new} is the newly computed value based on the value of x^{old}.

Table 5.1: Inputs and Output for Diabetic Prediction Model

Inputs	The number of times pregnant	x_1
	The results of an oral glucose tolerance test	x_2
	Diastolic blood pressure (mm/Hg)	x_3
	E-Triceps skin fold thickness (mm)	x_4
	2-h serum insulin (micro U/ml)	x_5
	Body mass index	x_6
	Diabetes pedigree function	x_7
	Age (year)	x_8
Output	Predicted class	y

5.2.1 Evaluation Criteria

The Evaluation Criteria we used to evaluate the model is:

Table 5.2: Statistical Analysis For Mean And Standard Deviation In Pima Indians Diabetes Dataset

No of Feature	Feature Name	Mean	Standard Deviation
1	Number of times pregnant	3.8	3.4
2	Plasma glucose concentration	120.9	32.0
3	Diastolic blood pressure	69.1	19.4
4	Triceps skin fold thickness	20.5	16.0
5	2 -Hour serum i insulin	79.8	115.2
6	Body mass index	32.0	7.9
7	Diabetes pedigree function	0.5	0.3
8	Age	33.2	11.8

- Sensitivity (Sens):

$$Sens = \frac{TP}{TP + FN} \tag{5.2}$$

- Specificity (Spec):

$$Spec = \frac{TN}{FP + TN} \tag{5.3}$$

- Positive Predicted Value (PPV):

$$PPV = \frac{TP}{TP + FP} \tag{5.4}$$

- Negative Predicted Value (NPV):

$$NPV = \frac{TN}{FN + TN} \tag{5.5}$$

- Accuracy (Acc):

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.6}$$

Given that:

- True Positive (TP): Sick people correctly diagnosed as sick.
- False Positive (FP): Healthy people incorrectly identified as sick.
- True Negative (TN): Healthy people correctly identified as healthy.
- False Negative (FN): Sick people incorrectly identified as healthy.

5.3 Models of Classification Diabetes using Full Features

5.3.1 Genetic Programming model

Experimental Setup In this research, a GPTIPS toolbox Searson et al. (2010) adopted to develop our results. In GPTIPS, the initial population is constructed by creating individuals that contain randomly generated GP trees with between 1 and G_{max} genes. During the run, genes are acquired and deleted using a tree crossover operator called two point high level crossover. This allows the exchange of genes between individuals and it is used in addition to the standard GP recombination operators.

Some parameters have to be defined by the user at the beginning of the evolutionary process. They include: population size, probability of crossover, mutation probability and the type of the selection mechanism. User has also to setup the maximum number of genes G_{max} where a model is allowed to have. The maximum tree depth D_{max} allows us to change the complexity of the evolved models. Restricting the tree depth helps evolving simple model but it may also reduce the performance of the evolved model.

A prior knowledge on the problem domain helps in designing a function set which could speed up the evolutionary process for model development. The adopted function set to develop the GP model is given as:

$$F = \{+, -, \times\}$$

Table 5.3: GP Tuning Parameters

Population size	100
Number of generation	500
Selection mechanism	Tournament
Max. tree depth	7
Probability of Crossover	0.85
Probability of Mutation	0.1
Max. No. of genes allowed in an individual	6

Crossover was performed with the two-point high-level crossover operator. Once the two parent individuals have been selected, two gene crossover points are selected within each parent. Then the genes enclosed by the crossover points are swapped between parents to form two new offspring.

Developed Mathematical GP Model The data set described earlier was loaded then the Multigene GP was applied using GPTIPS Tool. The parameters of the algorithm were tuned as listed in Table 5.3. In Figure 5.1, we show the convergence of GP over 500 generations. It can be clearly seen that the final model is a simple and compact mathematical model which is easy to evaluate. The performance measurements for the model was computed and summarized in Table 5.5. The best generated diabetic prediction Multigene GP model is given in Table 5.4.

Table 5.4: A GP model with Inputs: $x_1, \ldots, x_8$

$$\begin{aligned} y = \quad & 0.5546\,x_2 - 0.2773\,x_3 + 0.2773\,{x_1}^2\,(3.852\,x_3 + 3.852\,x_7)\,(x_5 + x_7 + x_1\,x_5) - 0.2291 \\ + \quad & 0.3613\,x_6\,(2\,x_6 + 2\,x_7 + x_3\,(x_8 - 1.94\,x_5 + x_8\,(x_3 + x_5)) + x_5\,x_8) \\ - \quad & 3.689\,x_6\,x_8\,(x_2 - x_8)\,(x_2 + 2\,x_7 + x_5\,x_8)\,(2\,x_7\,x_8 - x_7 + (x_6 + x_8)\,(x_2 + 2\,x_7) + x_7\,(x_3 + x_5 + x_8) - 2.253) \end{aligned}$$

Table 5.5: Performance of the GP model with $x_1, \ldots, x_8$ as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.8899	0.9176
Specificity	0.5714	0.5698
Accuracy	0.7740	0.8060
Positive Predicted Value	0.7839	0.8186
Negative Predicted Value	0.7482	0.7656

5.3.2 Neural Network model

Neural networks are non linear modelling of intelligent computational techniques which in recent years as advances in computing and information processing tools obtained an important and advances position in science, and the results have been favourable. Feed forward neural networks, are useful type of artificial neural networks, because feed forward neural network with a hidden layer, suitable activation function in the hidden layer and the

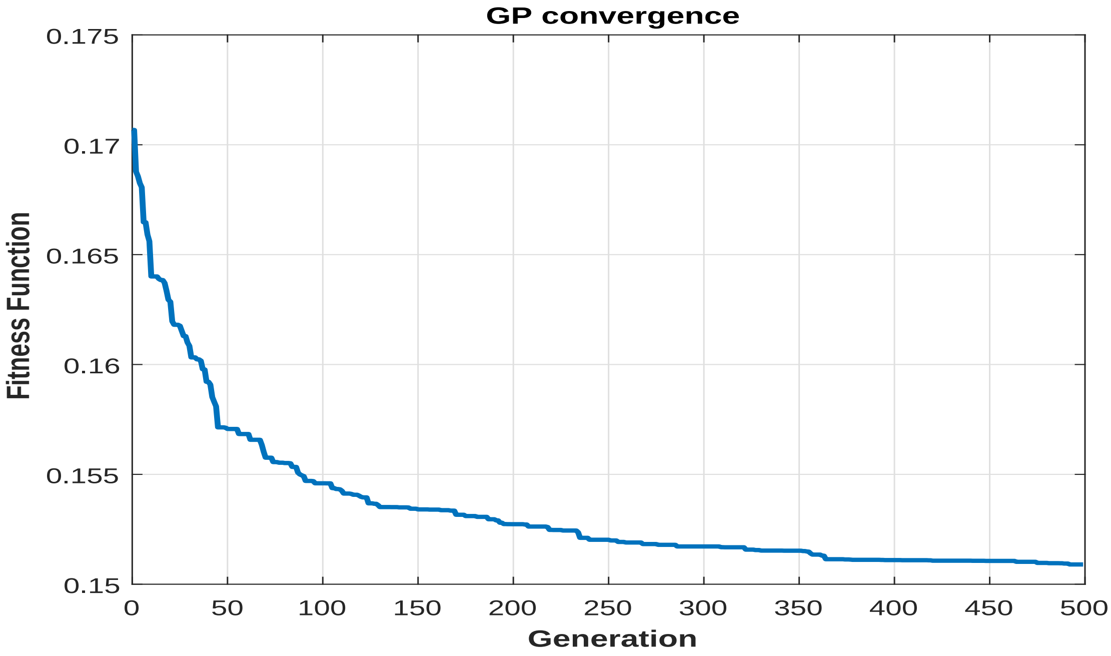


Figure 5.1: Convergence of the GP evolutionary process

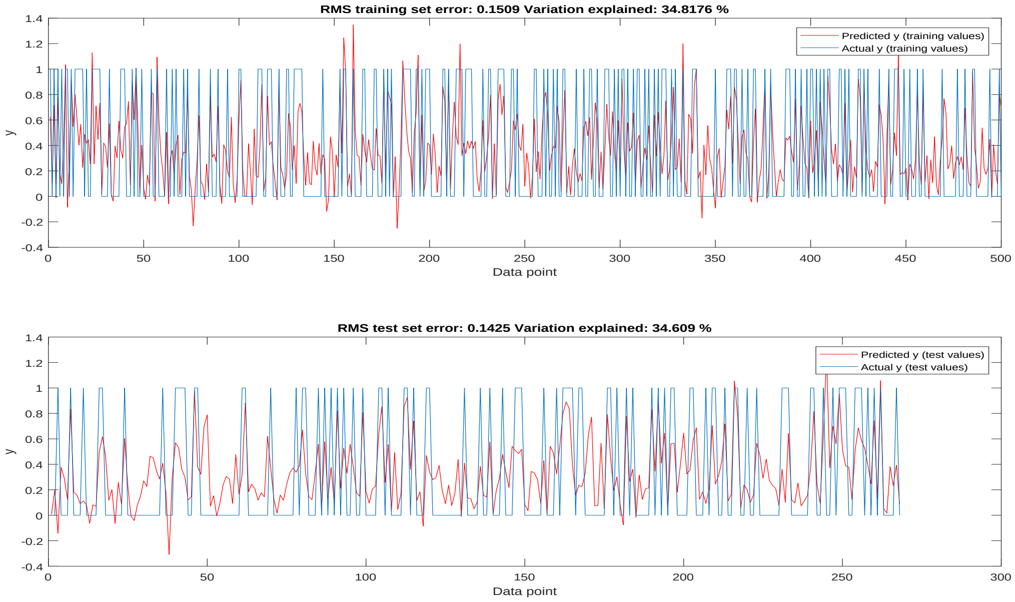


Figure 5.2: Measure class and Predicted Class in Training and Testing of GP Model

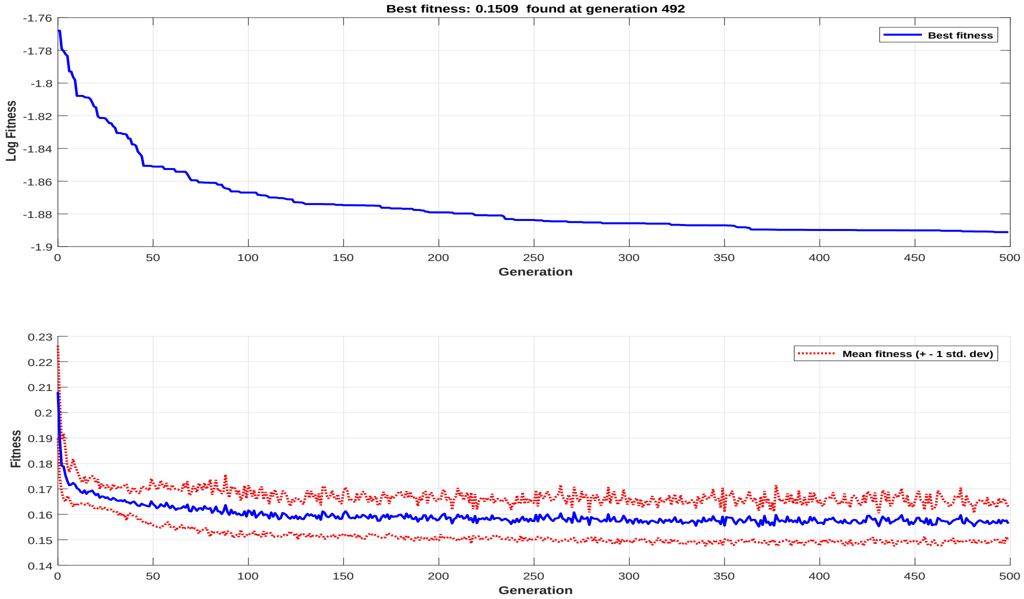


Figure 5.3: Best Fitness of the GP Model

enough hidden layer neurons are able to approximate any function with arbitrary precision. For this reason, in the following is presented a structure of feed forward neural network modelling to prediction diabetes problem. In general, the artificial neural networks are three types of neuronal layers as each layer is as follows:

- Input layer: Get the raw data that has been fed to the network.
- Hidden layers: the function of this layers is determined by inputs and weight and the relationship between them and the hidden layers. Weights between input and hidden units determine when a hidden unit to be activated.
- Output layer: output unit function depending on activity and weight of the hidden unit and the connection between hidden units and output. Created Neural network in this article has 1-15-10 structure that has shown in Figure 5.43.

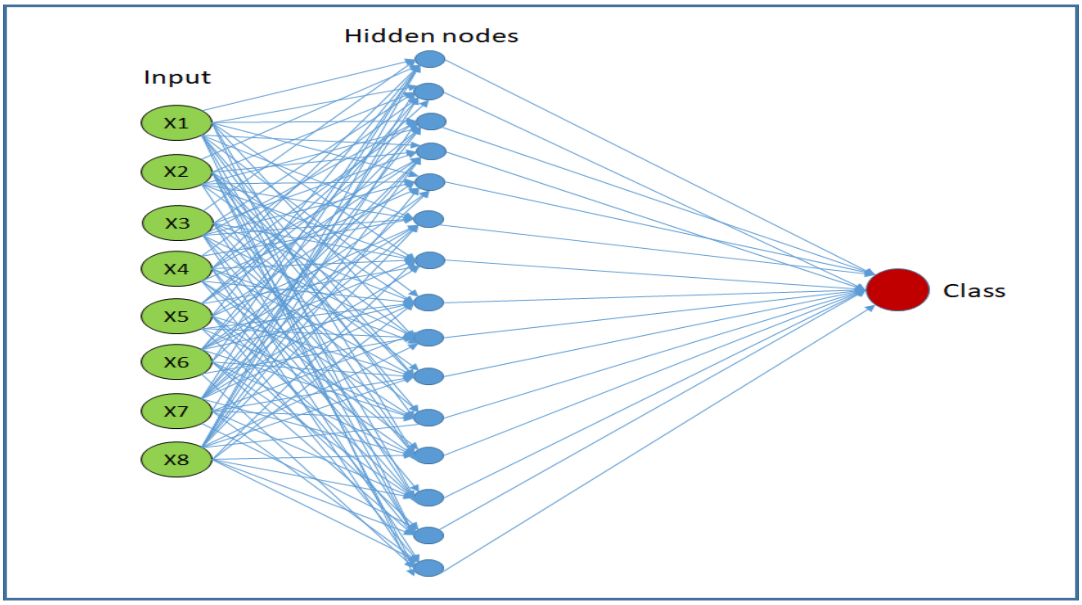


Figure 5.4: Artificial Neural Network Structure

Experimental Setup In this model feed forward neural network is used. Some parameters are define at the beginning like: number of epochs , training set ,testing set and number of hidden nodes. Table 5.6 shows the parameters of the neural network. Table 5.7 shows the performance measurement of neural network model. In Figure 5.5, we show the convergence of Neural Network over 500 number of epochs. Figure 5.6 and figure 5.7 describe the measure class and predicted class in training and testing respectively. Figure 5.6 shows the training state of the model. In figure 5.9 we see the performance of the model.

Table 5.6: Neural Network Parameters

max epoch number	500
Number of hidden nodes	15
Training	Levenberg Marquardt
Number of neurons in hidden layer	10
Training Set	500
Testing Set	268
Number of features	8

Figure 5.10 show regression graph of data for training process, testing, validation and total data. The horizontal axis is the output target and vertical

Table 5.7: Performance of the NN model with $x_1, \ldots, x_8$ as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.9117	0.8833
Specificity	0.6685	0.5244
Accuracy	0.8233	0.7710
Positive Predicted Value	0.8281	0.8030
Negative Predicted Value	0.8121	0.6719

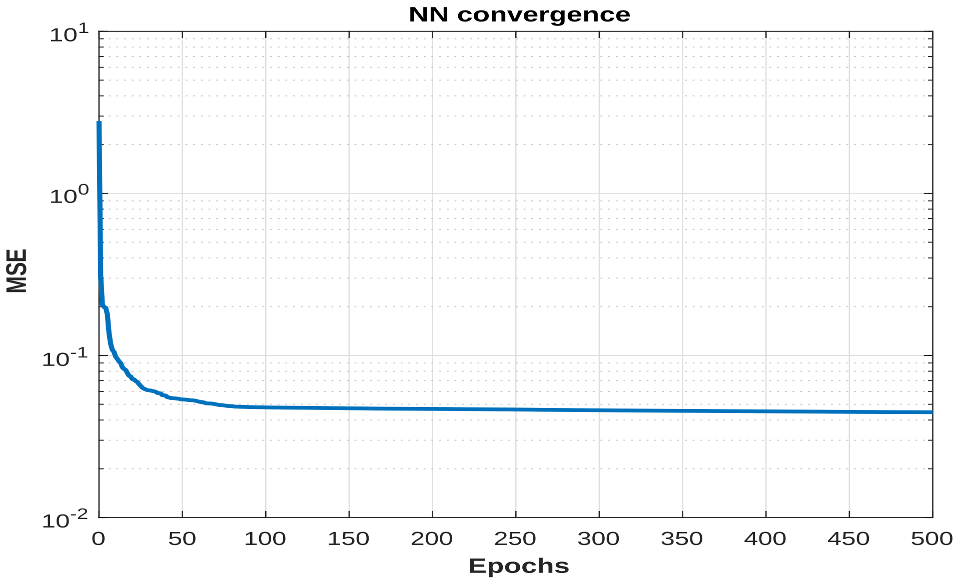


Figure 5.5: Convergence of the Artificial Neural Network Process

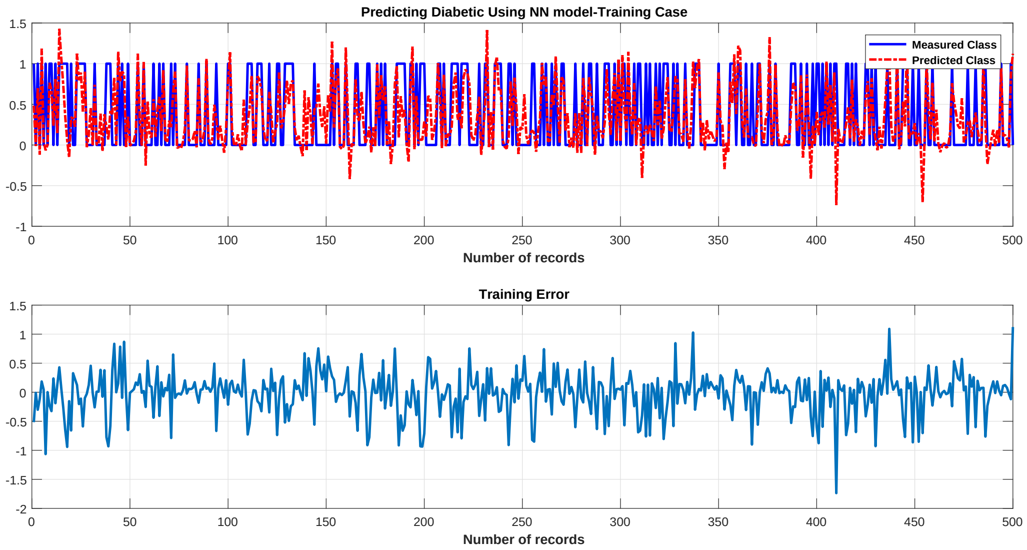


Figure 5.6: Measure class and Predicted Class in Training

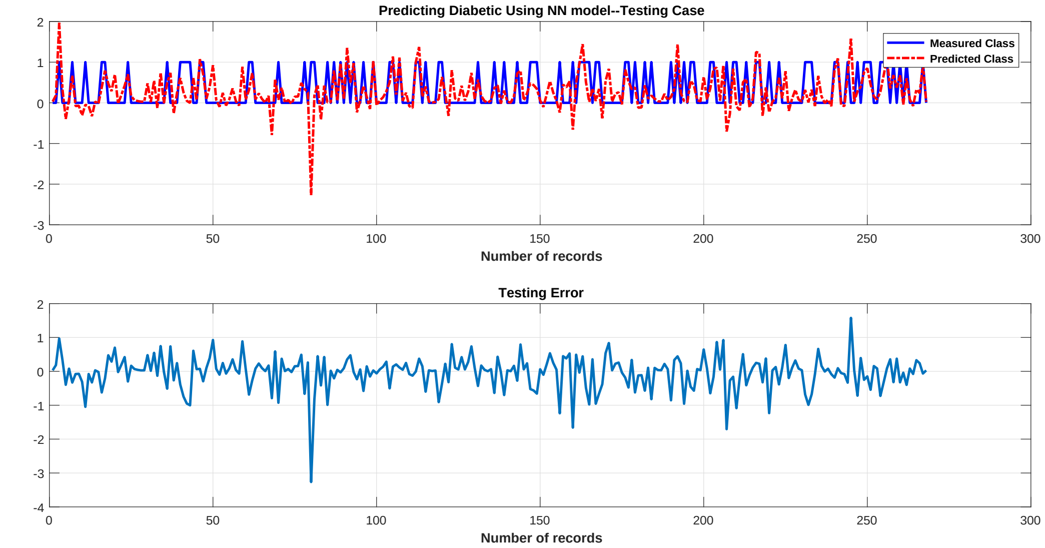


Figure 5.7: Measure class and Predicted Class in Testing

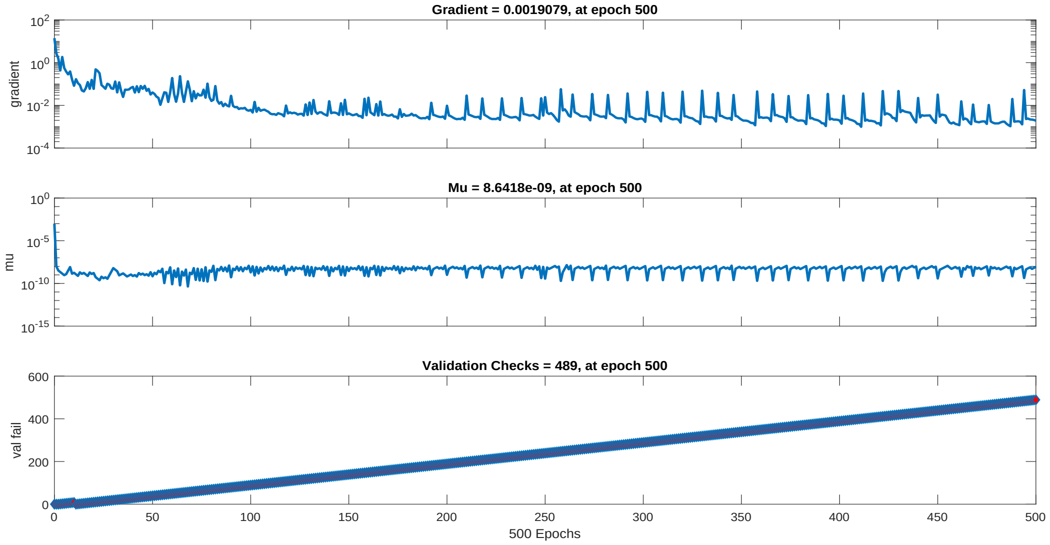


Figure 5.8: NN Model Training State

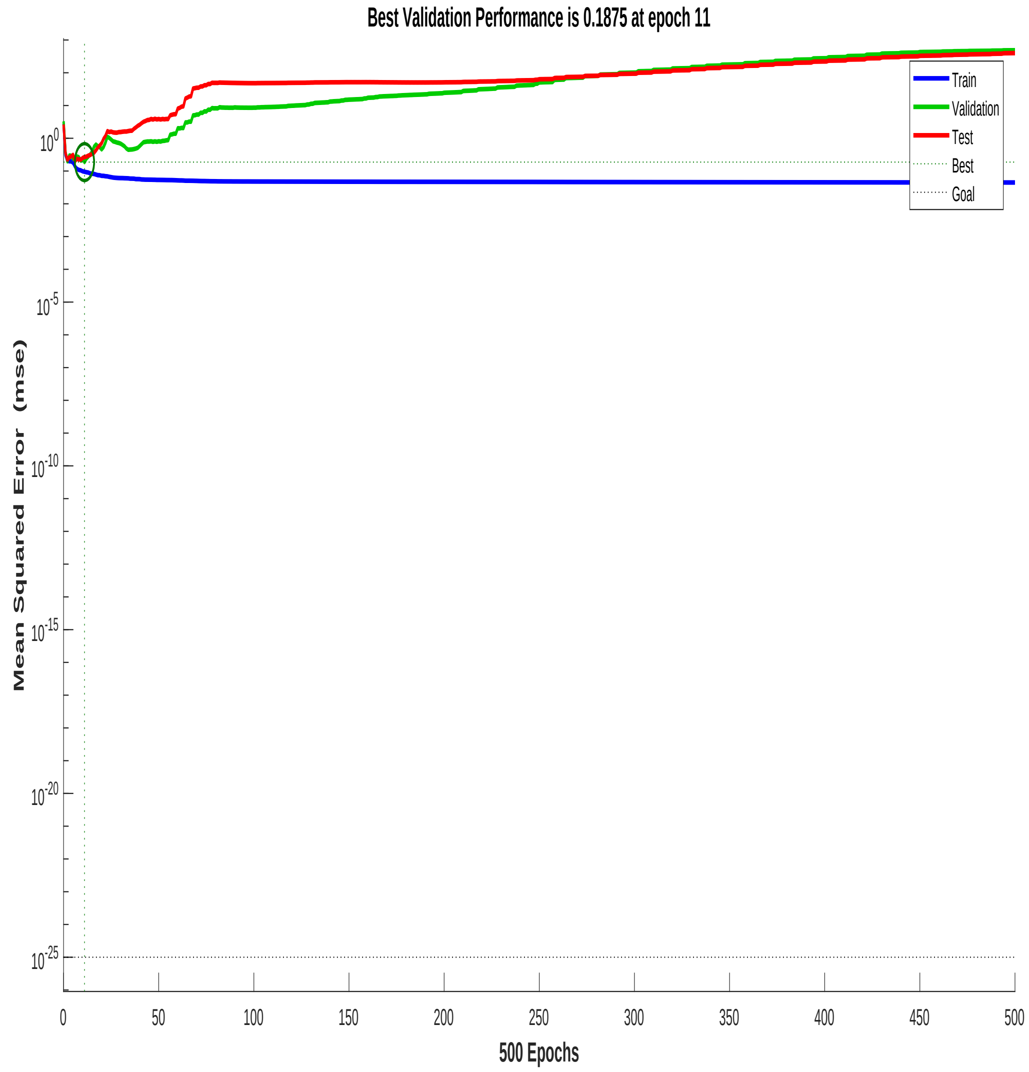


Figure 5.9: NN Model Performance

axis is the output of the neural network. The best line to create process is dotted lines but the line that created the neural network process in artificial neural network is respectively, blue lines with slope of 0.77 0.39 multiplied by the target plus 0.4, red with a slope of 0 multiplied by the target plus0.029, green with slope of 0.39 multiplied by the target plus 0.77, black with slope of 0.4 multiplied by the target plus 0.82. Figure 8 likewise shows the logistic regression of combining the artificial neural network with logistic regression and the line that this process could create is respectively, blue lines with slope of 1 multiplied by the target plus 0.00026, red with a slope of 1 multiplied by the target plus 0.0026, green with slope of 1 multiplied by the target plus 0.0037 and black with slope of 1 multiplied by the target plus 0.00088.

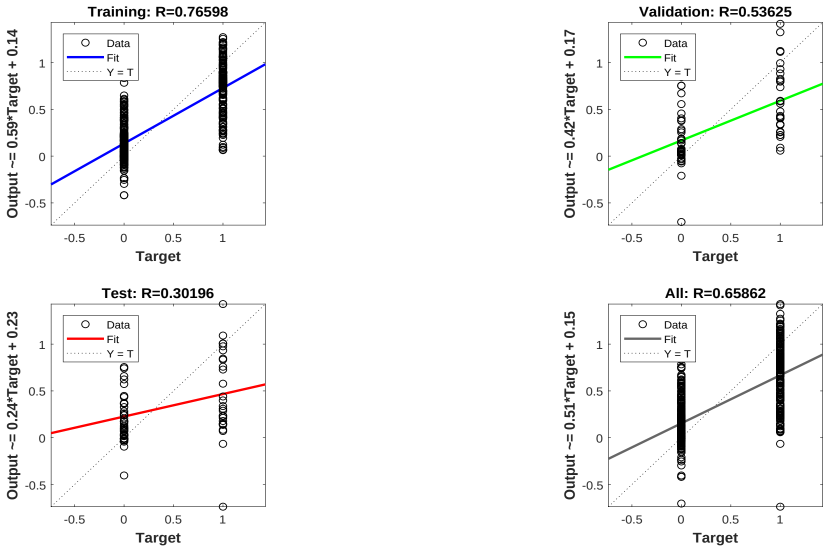


Figure 5.10: NN Model Training Regression

5.3.3 Fuzzy model

Fuzzy logic classifiers group data elements with a degree of membership in multiple classes by defining membership functions for each attribute. Various methods have been proposed to determine the partitioning of membership functions in a fuzzy logic inference system.

Experimental Setup Table 5.8 shows the parameters of the Fuzzy model. Figure 5.11 and figure 5.12 shows the training and testing of the fuzzy model respectively. The membership functions for every feature describe in figure 5.13.

Table 5.8: Fuzzy model Parameters

Number of cluster	8
Fuzzy Inference System	Sugeno-Takagi model
Training Set	500
Testing Set	268
Number of features	8

Table 5.9: Performance of the Fuzzy model with $x_1, \ldots, x_8$ as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.8927	0.8571
Specificity	0.6484	0.6512
Accuracy	0.8036	0.7910
Positive Predicted Value	0.8156	0.8387
Negative Predicted Value	0.7763	0.6829

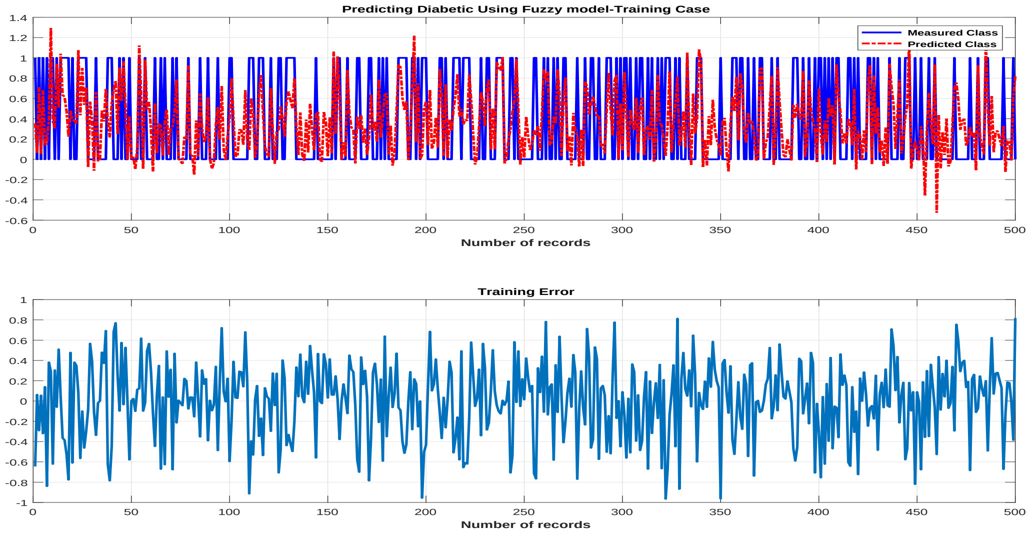


Figure 5.11: Fuzzy Model Training

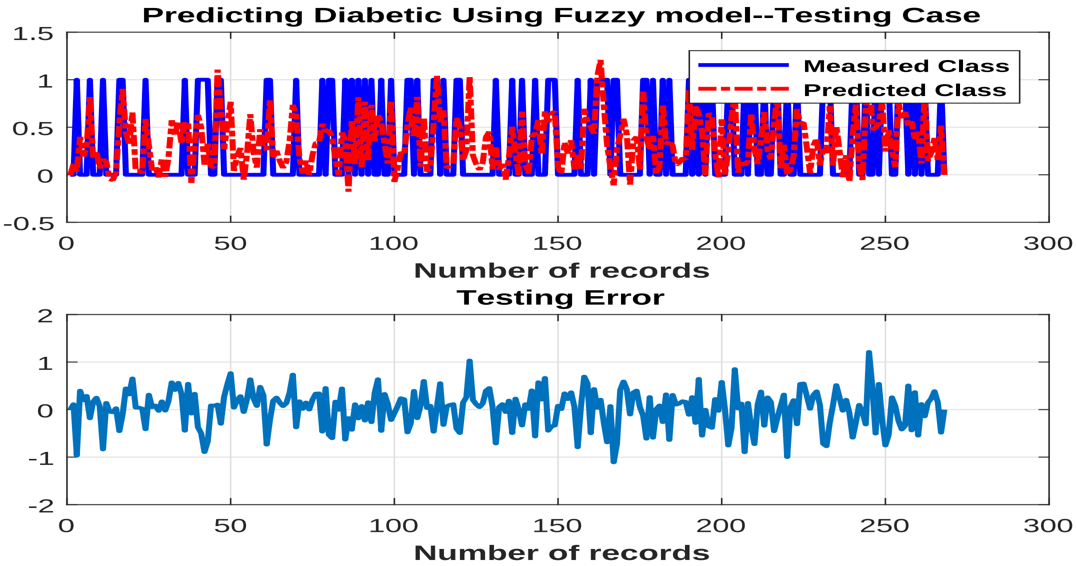


Figure 5.12: Fuzzy Model Testing

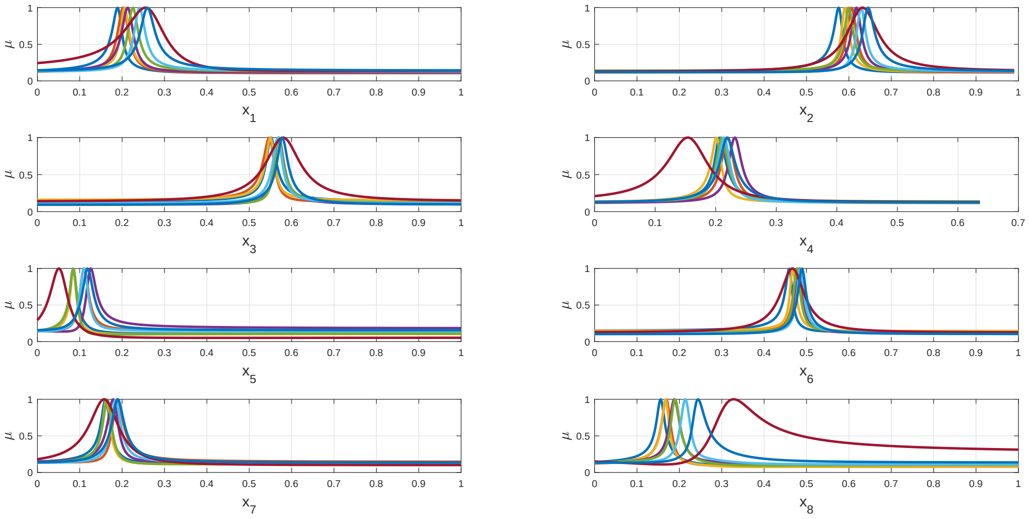


Figure 5.13: Membership Function of Fuzzy Model

5.4 Models of Classification Diabetes using Four Features

This section describe the result after select the four weighted features from the mathematical model that built with the genetic programming model. The selected features are shown in table 5.10.

Table 5.10: Inputs and Output for Diabetic Prediction Model

Inputs	The number of times pregnant	x_1
	The results of an oral glucose tolerance test	x_2
	Body mass index	x_6
	Age (year)	x_8
Output	Predicted class	y

5.4.1 Genetic Programming model

Figure 5.14, shows the convergence of GP over 500 generations. It can be clearly seen that the final model is a simple and compact mathematical model which is easy to evaluate. The performance measurements for the model was computed and summarized in Table 5.12. The best generated diabetic prediction Multigene GP model is given in Table 5.11.

Table 5.11: A GP model with Inputs: x_1, x_2, x_6, x_8

$$\begin{aligned}
y = \quad & 1.251\,{x_3}^2 - 2.502\,x_3 - 1.251\,x_2 + 1.251\,x_3\,(x_2 - x_3)^2\,(x_3 - x_4) - 1.251\,x_2\,x_3\,(x_1 - x_2) - 4.491 \\
+ \quad & 0.279\,{x_4}^2 - 1.279\,x_3 - 1.279\,x_4 - 1.535\,x_2 - 1.279\,x_4\,(x_1 - x_3)\left(-{x_3}^2 + x_3 + x_1\right) + 1.279\,x_1\,(x_1 + x_3)\,(x_1 - x_4) \\
+ \quad & 1.279\,{x_2}^2\,x_3\,(x_1 - x_3)\,(2\,x_3 - 2\,x_1 + x_2\,x_3) \\
+ \quad & 0.728\,x_2 + 2.728\,x_3 + 2.728\,x_4 - 2.728\,x_2\,x_4 + 2.728\,{x_2}^2\,(x_1 - x_3)\left(0.1005\,{x_2}^2 - 1.0\,x_3 + x_4\right) + 4.964\,x_2\,x_4\,(x_2 - x_4) \\
- \quad & 2.728\,{x_2}^3\,x_3\,(x_2 - 1)\,(x_4 + 1) + 4.536
\end{aligned}$$

Table 5.12: Performance of the GP model with x_1, x_2, x_6, x_8 as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.8868	0.8956
Specificity	0.5659	0.6512
Accuracy	0.7700	0.8172
Positive Predicted Value	0.7812	0.8446
Negative Predicted Value	0.7410	0.7467

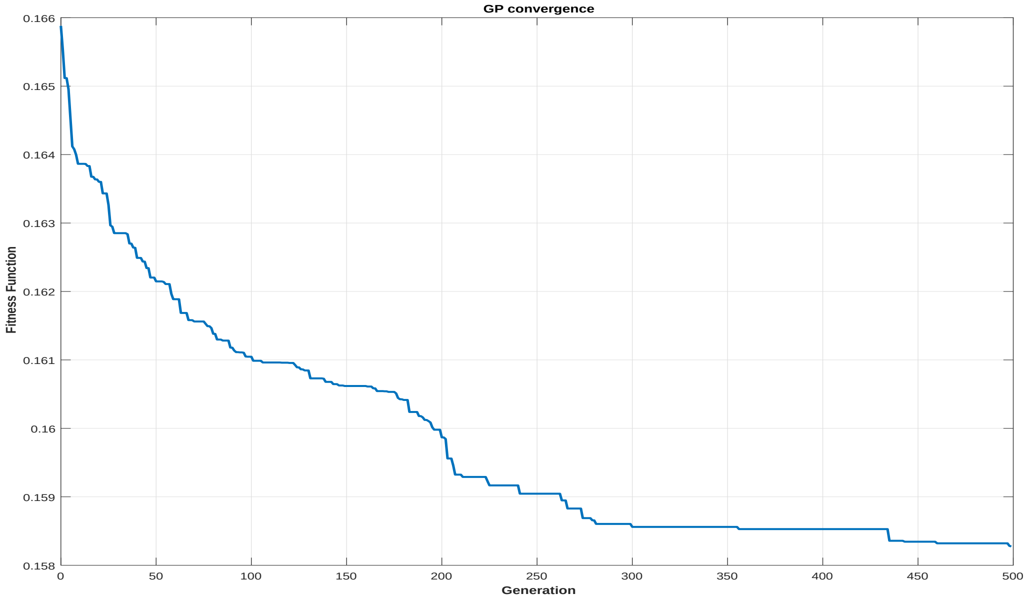
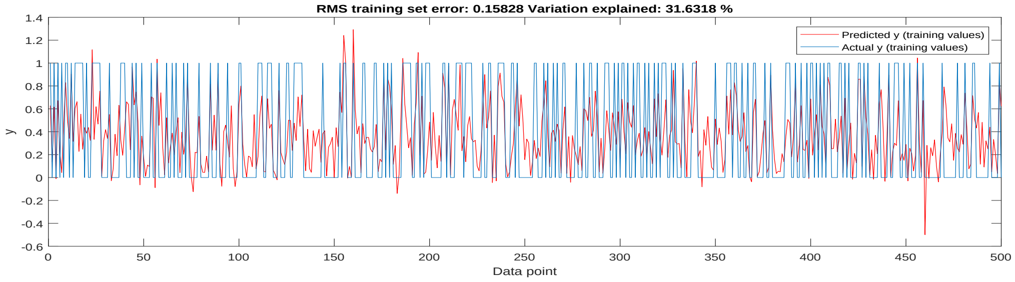


Figure 5.14: Convergence of the GP evolutionary process using Four Features



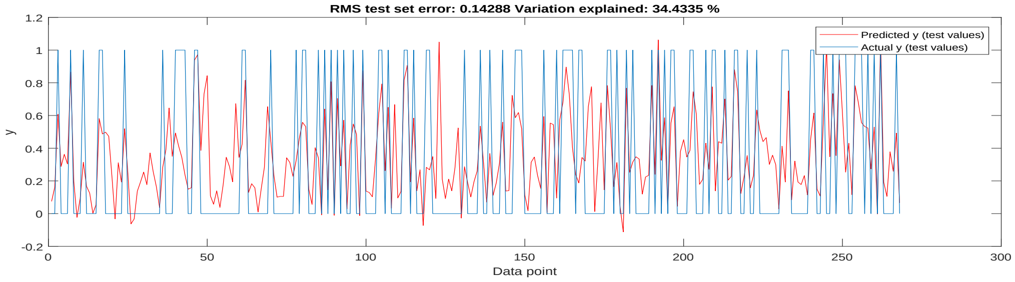
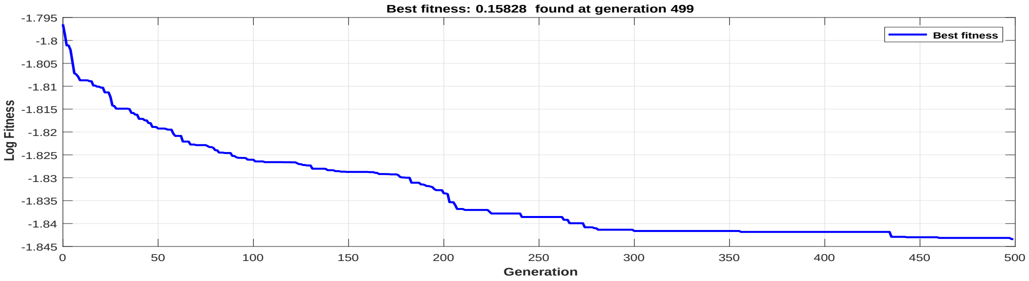


Figure 5.15: Measure class and Predicted Class in Training and Testing of GP Model using Four Features



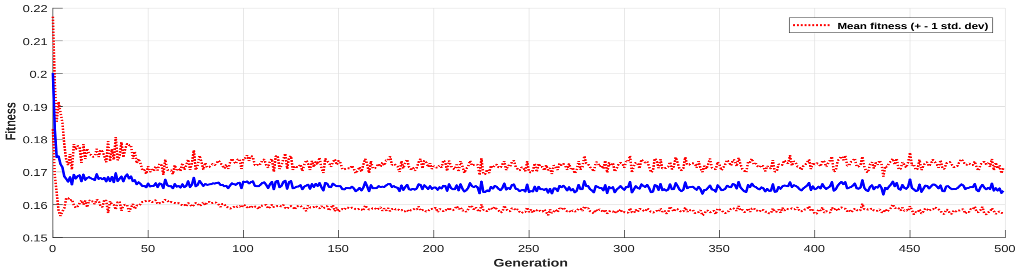


Figure 5.16: Best Fitness of the GP Model using Four Features

5.4.2 Neural Network model

Table 5.13 shows the performance measurement of neural network model. Figure 5.18, shows the convergence of Neural Network over 500 number of epochs. Figure 5.19 and figure 5.20 describe the measure class and predicted class in training and testing respectively. Figure 5.21 shows the training state of the model. Figure 5.22 describe the performance of the model.

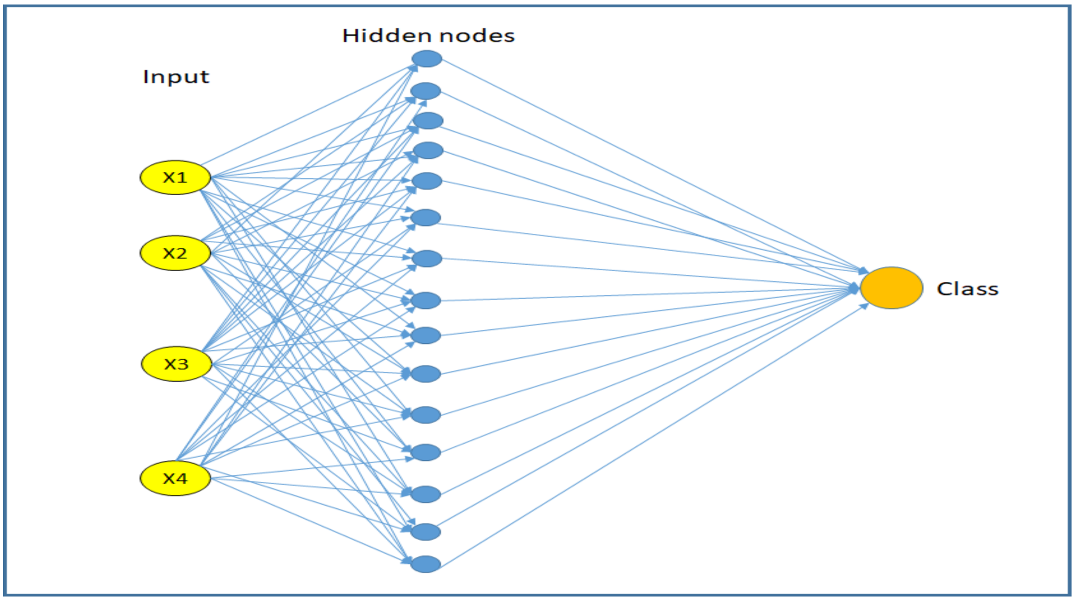


Figure 5.17: Artificial Neural Network Structure

Table 5.13: Performance of the NN model with x_1, x_2, x_6, x_8 as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.8797	0.8500
Specificity	0.6889	0.6706
Accuracy	0.8105	0.7925
Positive Predicted Value	0.8323	0.8453
Negative Predicted Value	0.7654	0.6786

5.4.3 Fuzzy model

Figure 5.24 and figure 5.25 shows the training and testing of the fuzzy model respectively. The membership functions for every feature describe in figure 5.26.

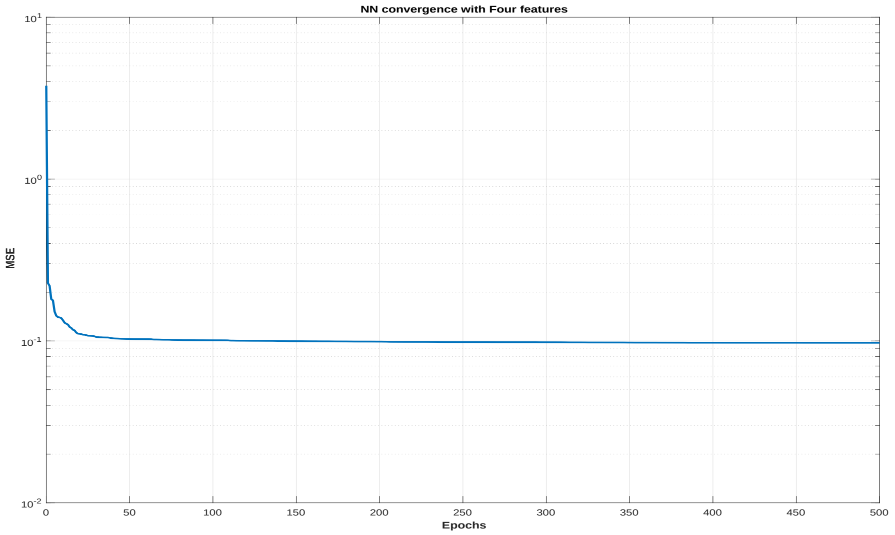


Figure 5.18: Convergence of the Artificial Neural Network Process

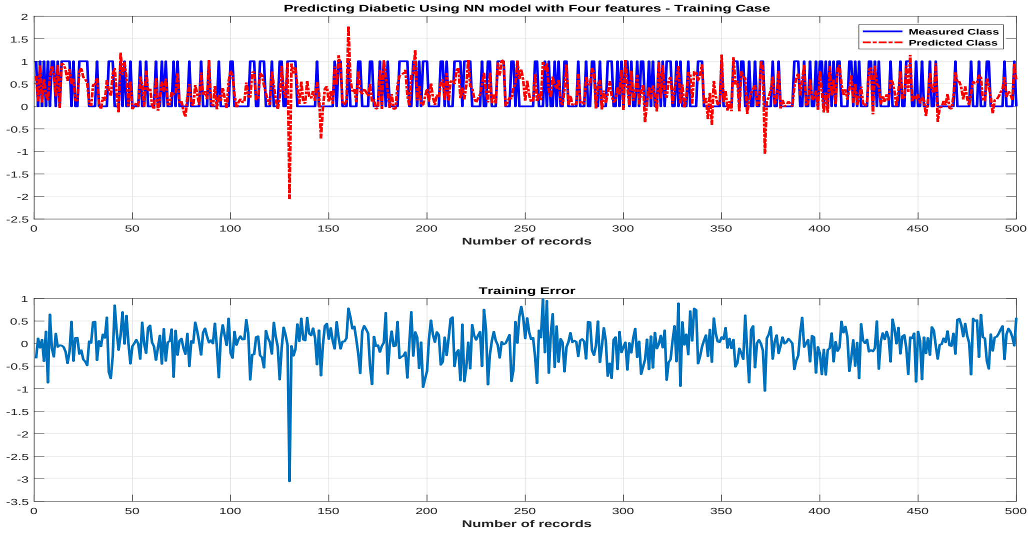


Figure 5.19: Measure class and Predicted Class in Training

Table 5.14: Performance of the Fuzzy model with x_1, x_2, x_6, x_8 as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.8836	0.8846
Specificity	0.6374	0.5698
Accuracy	0.7940	0.7836
Positive Predicted Value	0.8098	0.8131
Negative Predicted Value	0.7582	0.7000

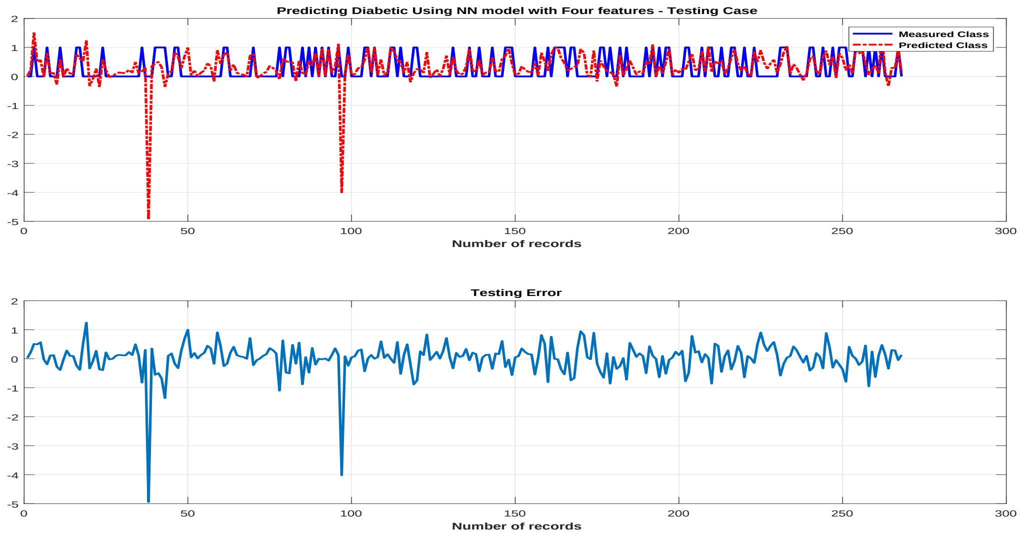


Figure 5.20: Measure class and Predicted Class in Testing

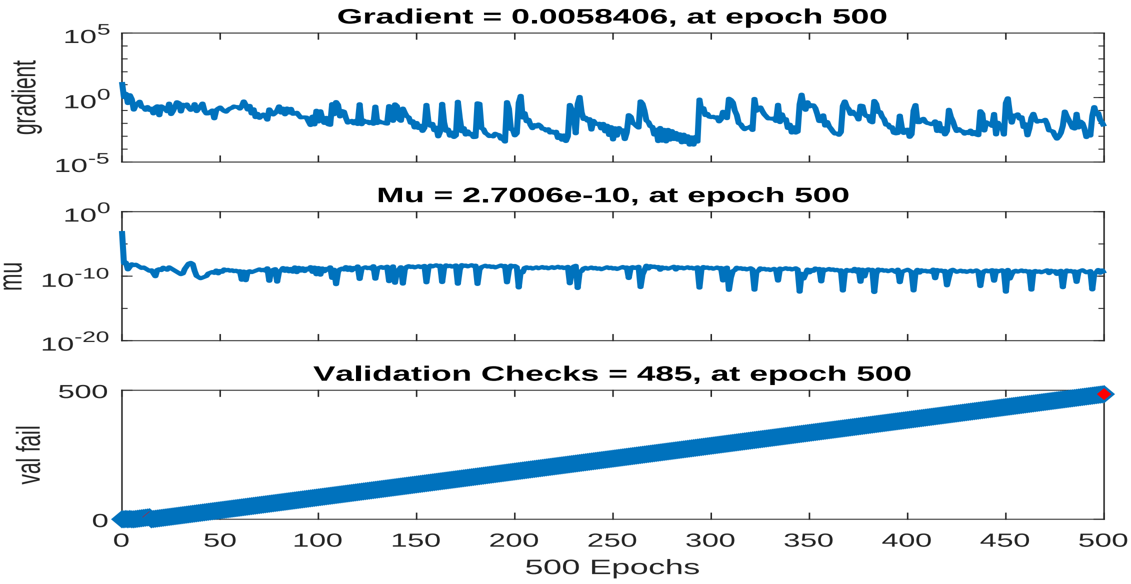


Figure 5.21: NN Model Training State

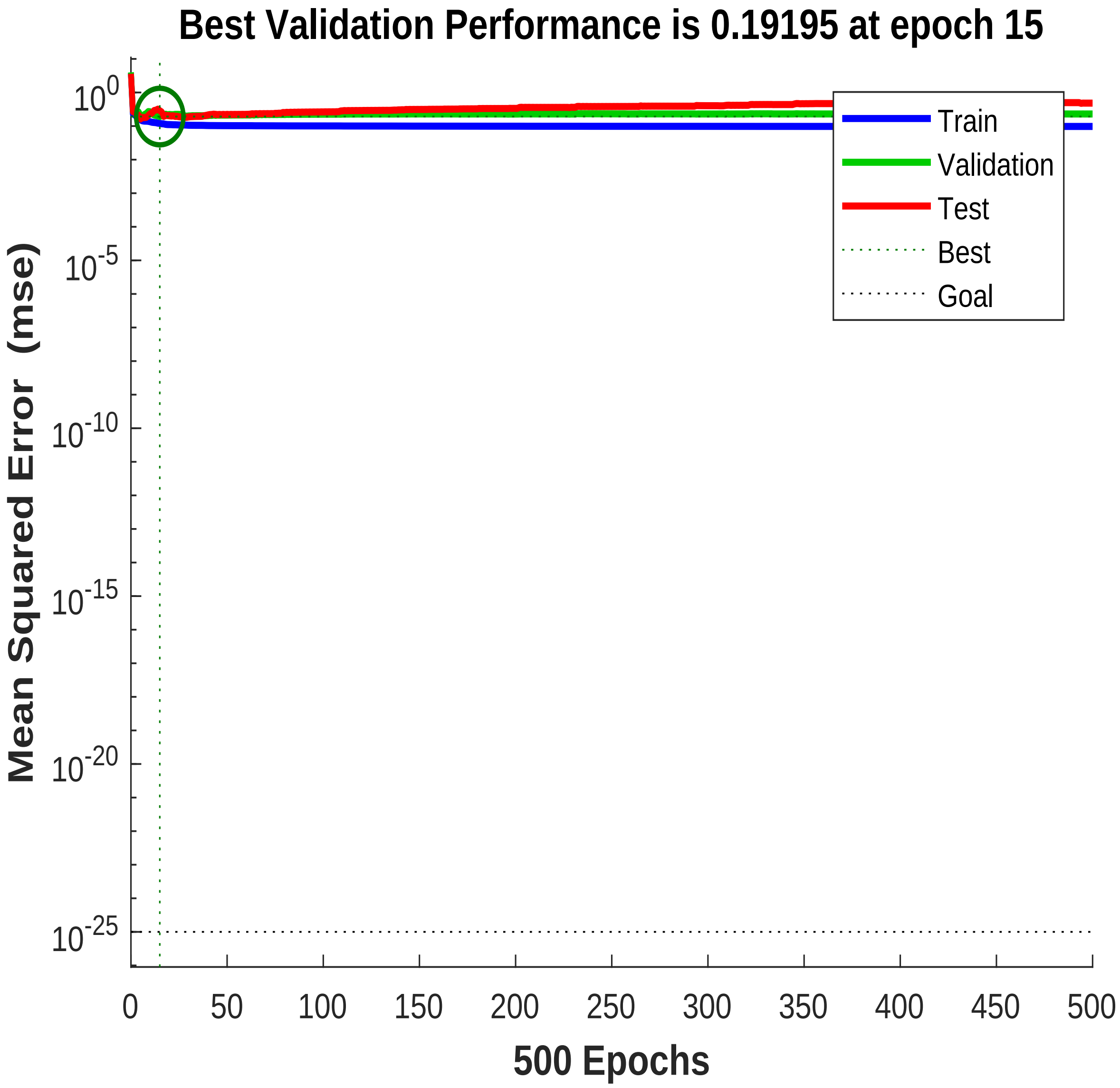


Figure 5.22: NN Model Performance

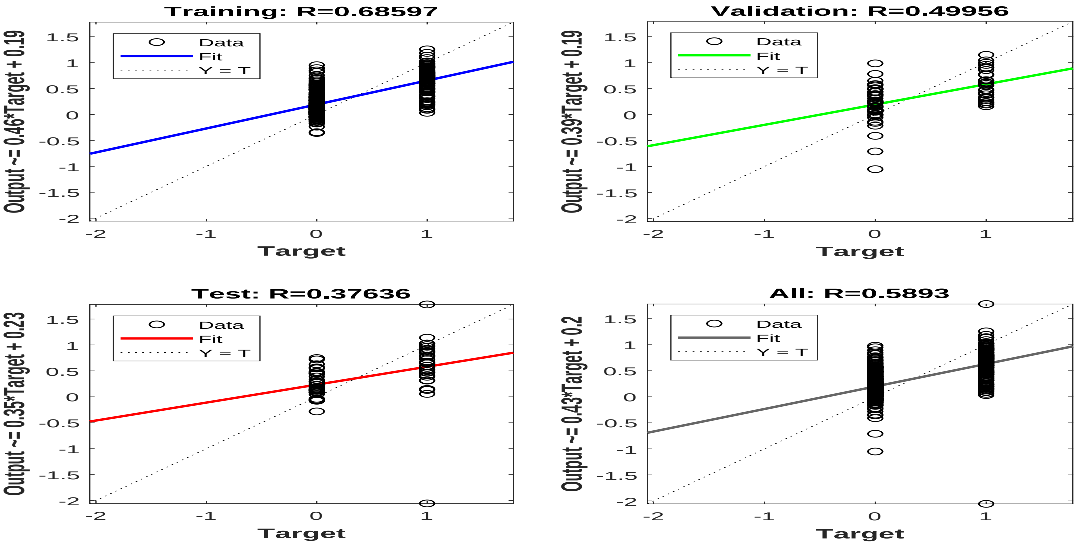


Figure 5.23: NN Model Training Regression

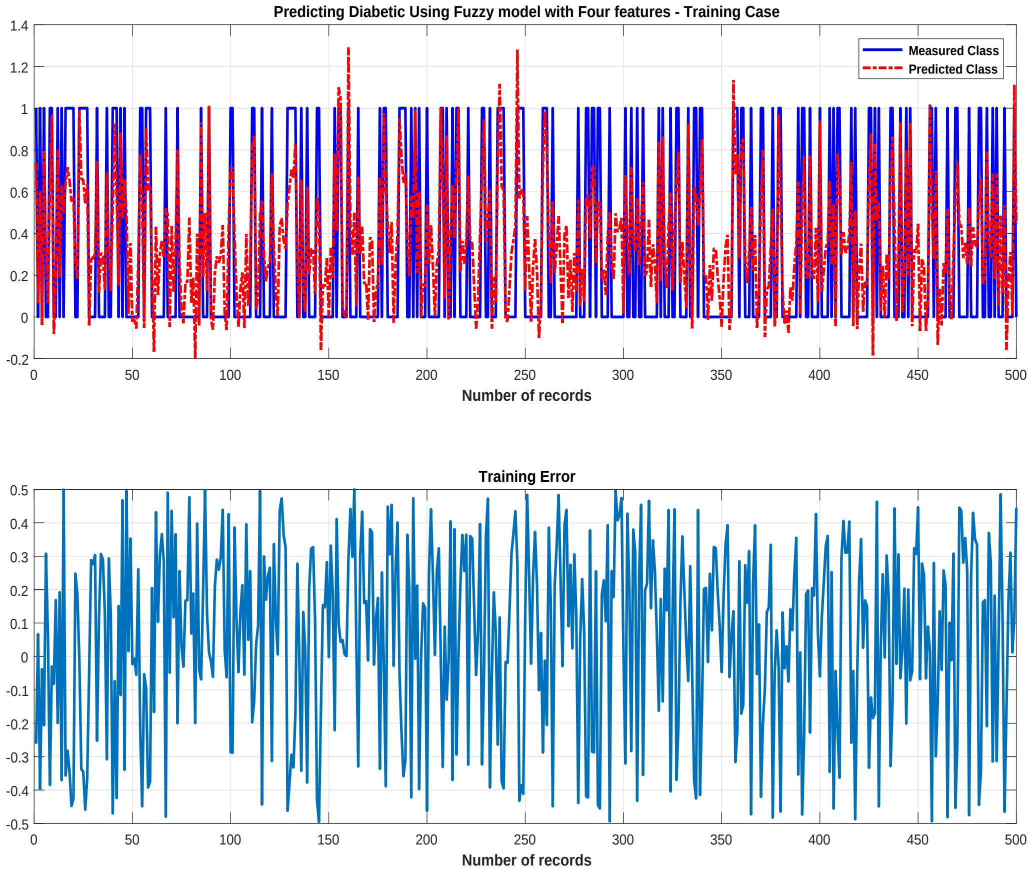


Figure 5.24: Fuzzy Model Training

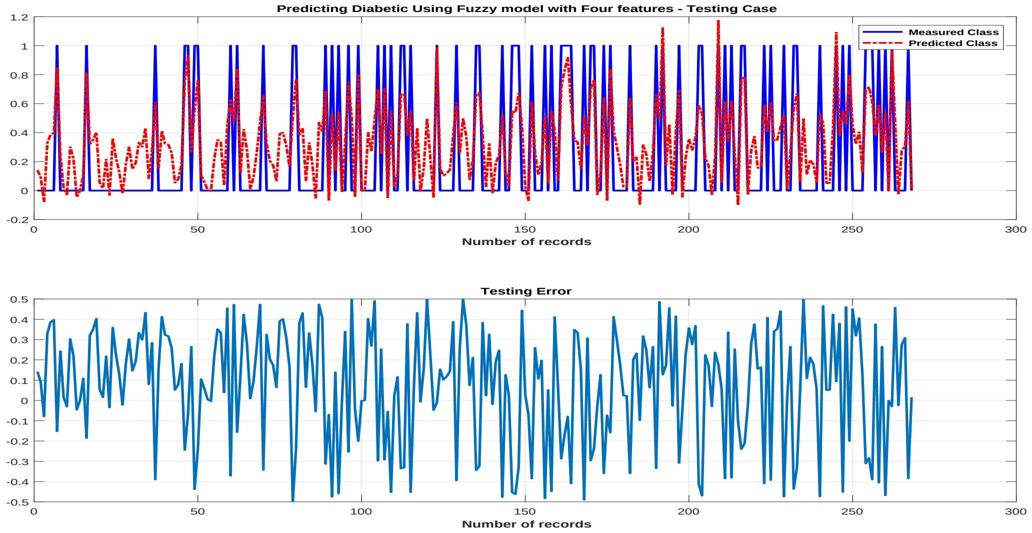


Figure 5.25: Fuzzy Model Testing

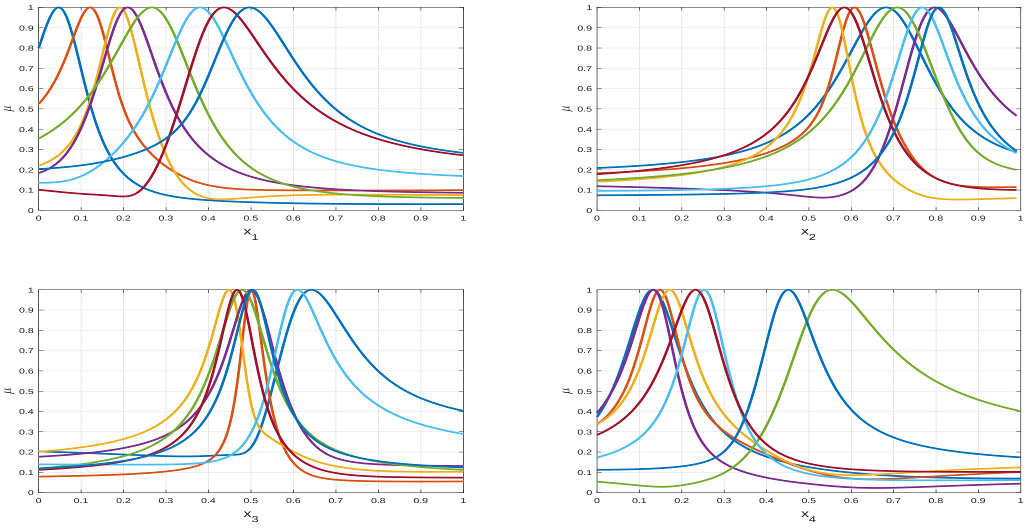


Figure 5.26: Membership Function of Fuzzy Model

5.5 Models of Classification Diabetes using Three Features

This section describe the result after select the three weighted features from the mathematical model that built with the genetic programming model. The selected features are shown in table 5.15.

Table 5.15: Inputs and Output for Diabetic Prediction Model

Inputs	The results of an oral glucose tolerance test	x_2
	Body mass index	x_6
	Age (year)	x_8
Output	Predicted class	y

5.5.1 Genetic Programming model

Figure 5.27, shows the convergence of GP over 500 generations. It can be clearly seen that the final model is a simple and compact mathematical model which is easy to evaluate. The performance measurements for the model was computed and summarized in Table 5.17. The best generated diabetic prediction Multigene GP model is given in Table 5.16.

Table 5.16: A GP model with Inputs: x_2, x_6, x_8

$$\begin{aligned}
y = \quad & 1.251\,{x_3}^2 - 2.502\,x_3 - 1.251\,x_2 + 1.251\,x_3\,(x_2 - x_3)^2\,(x_3 - x_4) - 1.251\,x_2\,x_3\,(x_1 - x_2) - 4.491 \\
+ \quad & 0.279\,{x_4}^2 - 1.279\,x_3 - 1.279\,x_4 - 1.535\,x_2 - 1.279\,x_4\,(x_1 - x_3)\left(-{x_3}^2 + x_3 + x_1\right) + 1.279\,x_1\,(x_1 + x_3)\,(x_1 - x_4) \\
+ \quad & 1.279\,{x_2}^2\,x_3\,(x_1 - x_3)\,(2\,x_3 - 2\,x_1 + x_2\,x_3) \\
+ \quad & 0.728\,x_2 + 2.728\,x_3 + 2.728\,x_4 - 2.728\,x_2\,x_4 + 2.728\,{x_2}^2\,(x_1 - x_3)\left(0.1005\,{x_2}^2 - 1.0\,x_3 + x_4\right) \\
+ \quad & 4.964\,x_2\,x_4\,(x_2 - x_4) - 2.728\,{x_2}^3\,x_3\,(x_2 - 1)\,(x_4 + 1) + 4.536
\end{aligned}$$

Table 5.17: Performance of the GP model with x_2, x_6, x_8 as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.8868	0.8956
Specificity	0.5659	0.6512
Accuracy	0.7700	0.8172
Positive Predicted Value	0.7812	0.8446
Negative Predicted Value	0.7410	0.7467

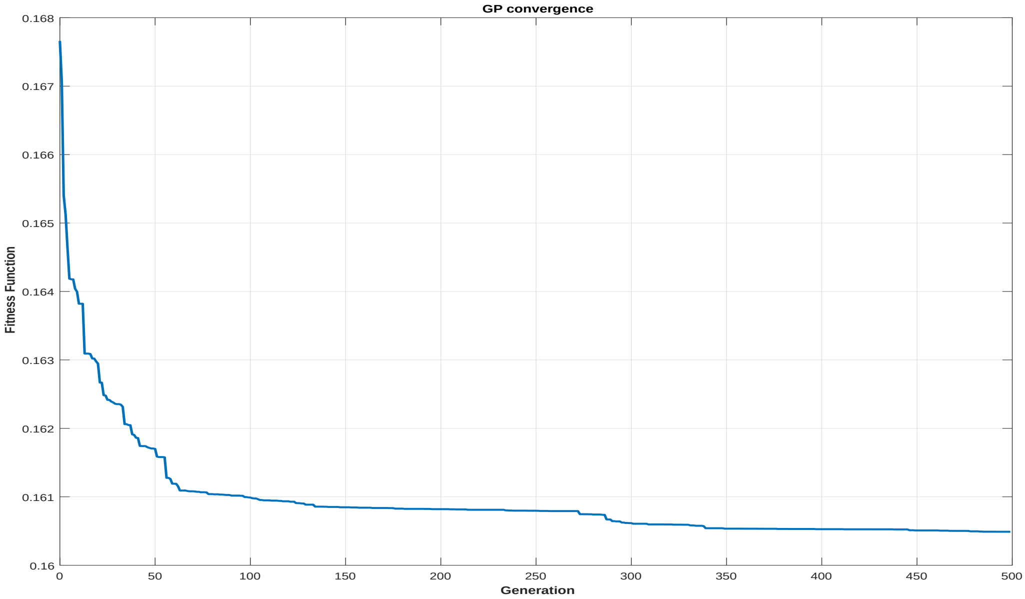


Figure 5.27: Convergence of the GP evolutionary process using Three Features

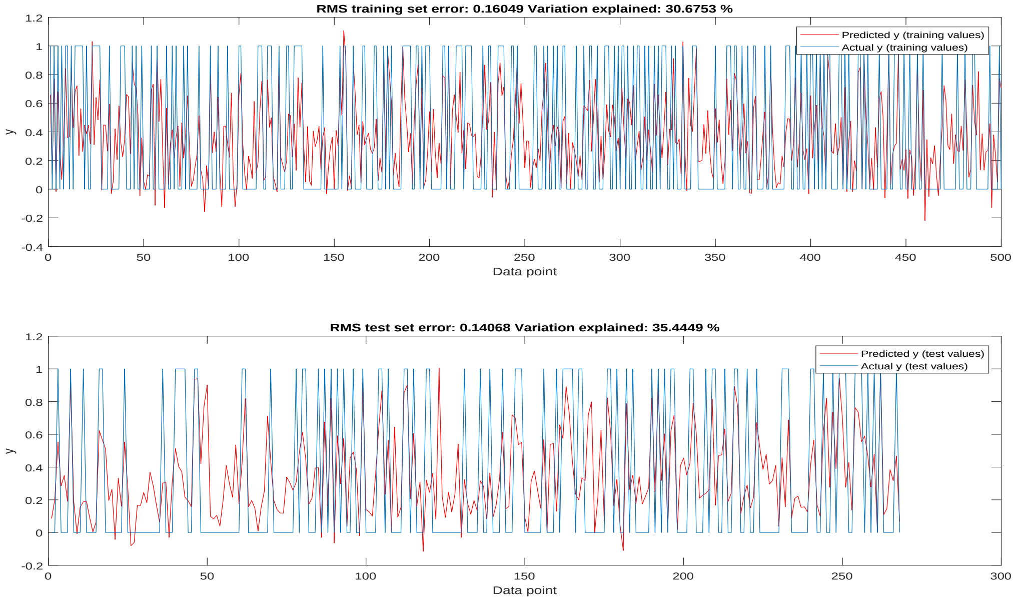


Figure 5.28: Measure class and Predicted Class in Training and Testing of GP Model using Three Features

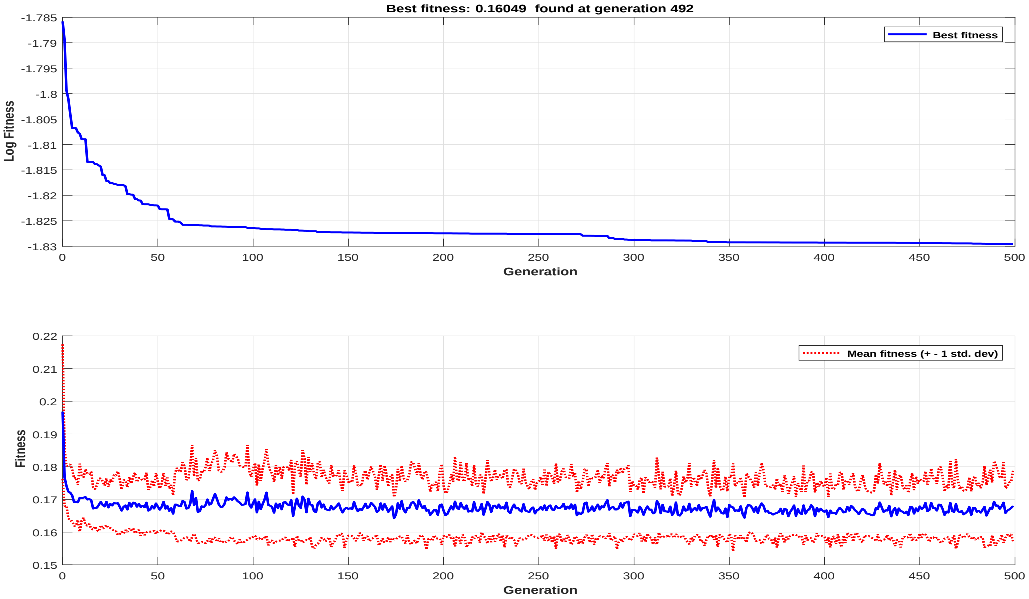


Figure 5.29: Best Fitness of the GP Model using Three Features

5.5.2 Neural Network model

Table 5.18 shows the performance measurement of neural network model. Figure 5.31, shows the convergence of Neural Network over 500 number of epochs. Figure 5.32 and figure 5.33 describe the measure class and predicted class in training and testing respectively. Figure 5.34 shows the training state of the model. Figure 5.35 describe the performance of the model.

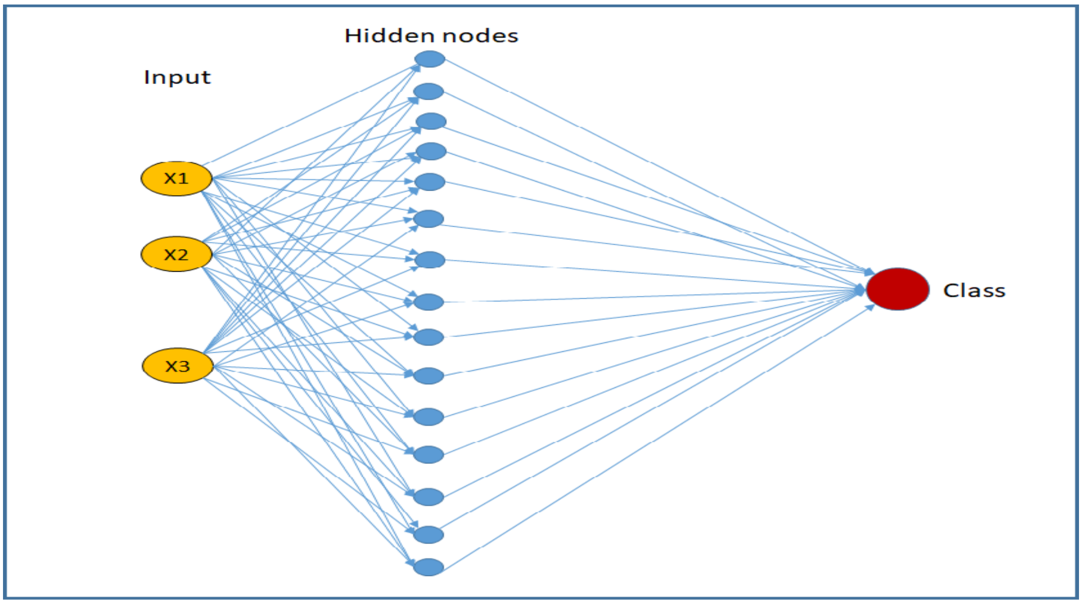


Figure 5.30: Artificial Neural Network Structure

Table 5.18: Performance of the NN model with x_2, x_6, x_8as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.8797	0.8500
Specificity	0.6889	0.6706
Accuracy	0.8105	0.7925
Positive Predicted Value	0.8323	0.8453
Negative Predicted Value	0.7654	0.6786

5.5.3 Fuzzy model

Figure 5.37 and figure 5.38 shows the training and testing of the fuzzy model respectively. The membership functions for every feature describe in figure 5.39.

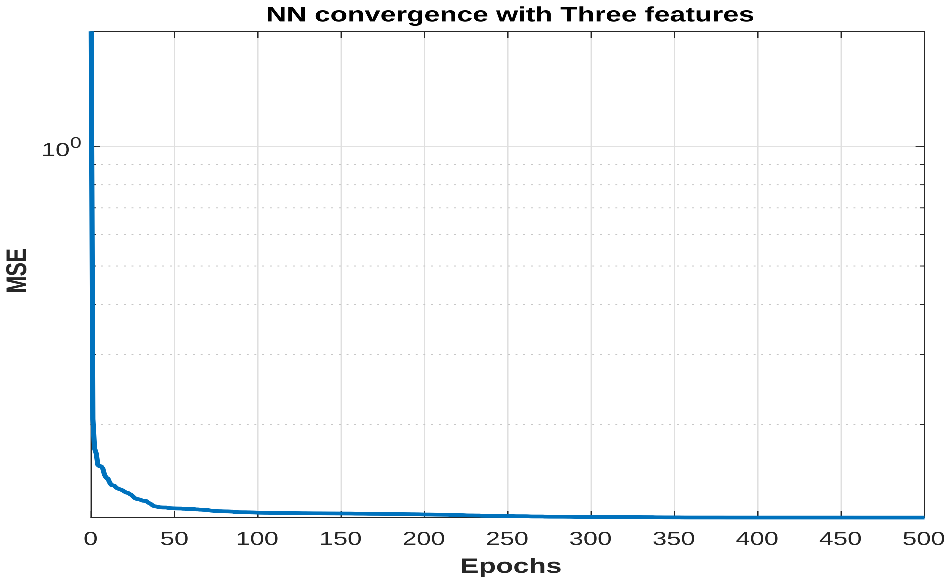


Figure 5.31: Convergence of the Artificial Neural Network Process

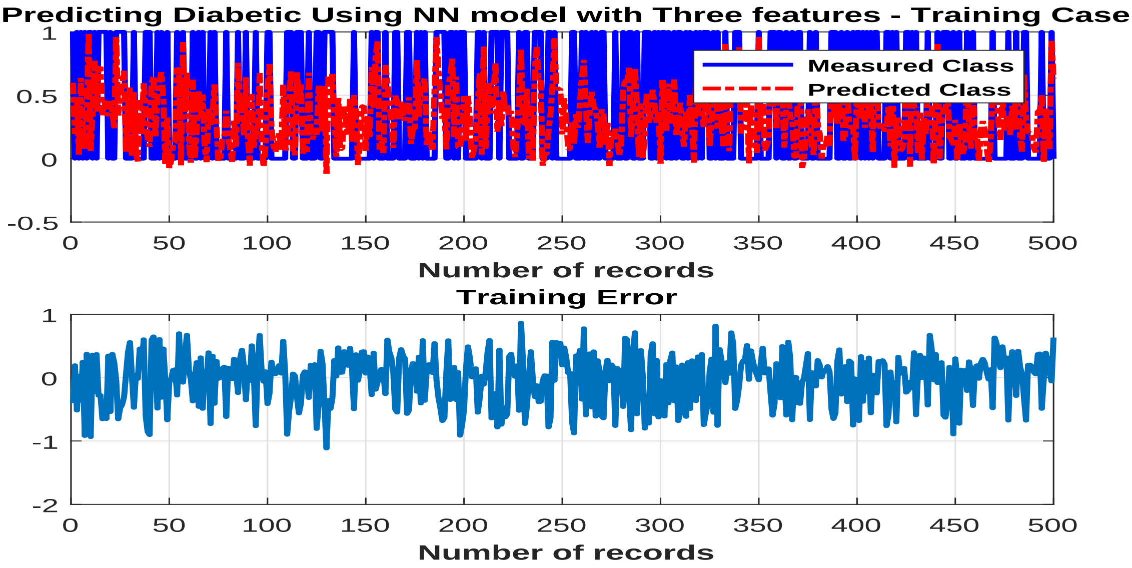


Figure 5.32: Measure class and Predicted Class in Training

Table 5.19: Performance of the Fuzzy model with x_2, x_6, x_8 as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.8962	0.8901
Specificity	0.5659	0.6395
Accuracy	0.7760	0.8097
Positive Predicted Value	0.7830	0.8394
Negative Predicted Value	0.7574	0.7333

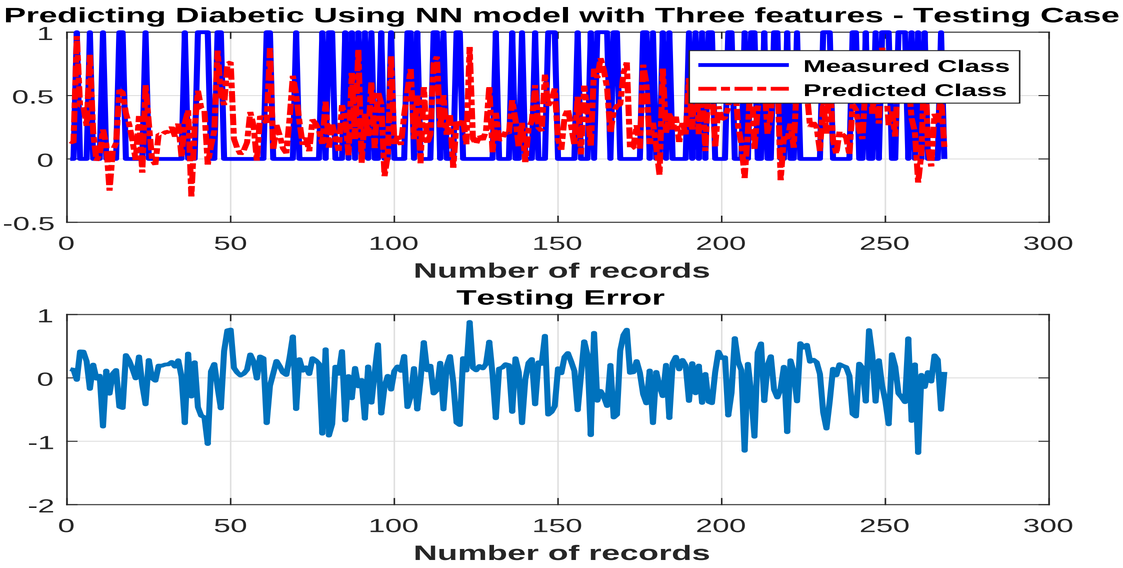


Figure 5.33: Measure class and Predicted Class in Testing

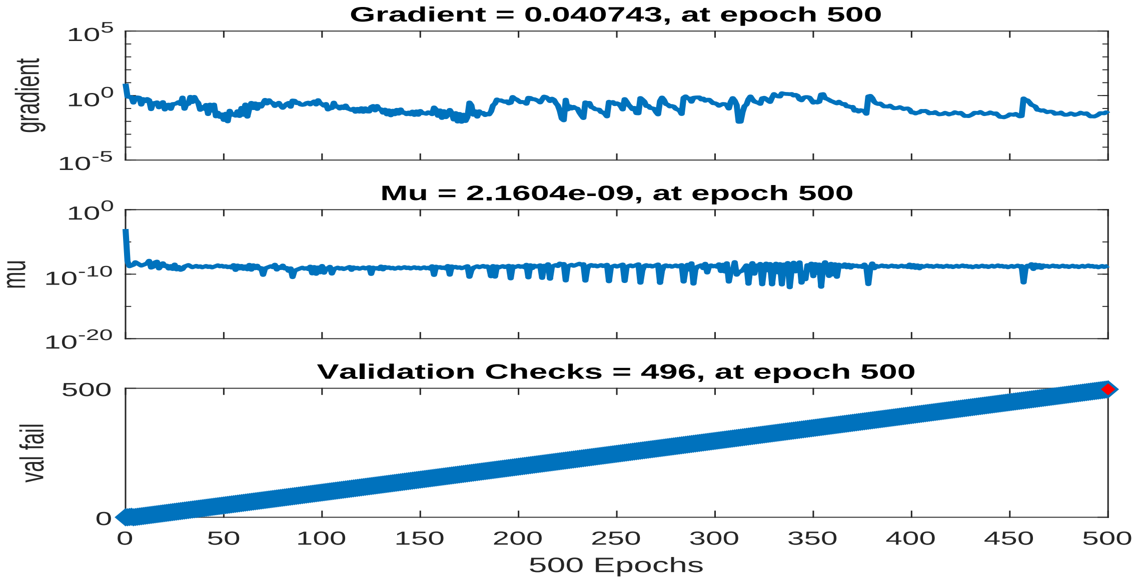


Figure 5.34: NN Model Training State

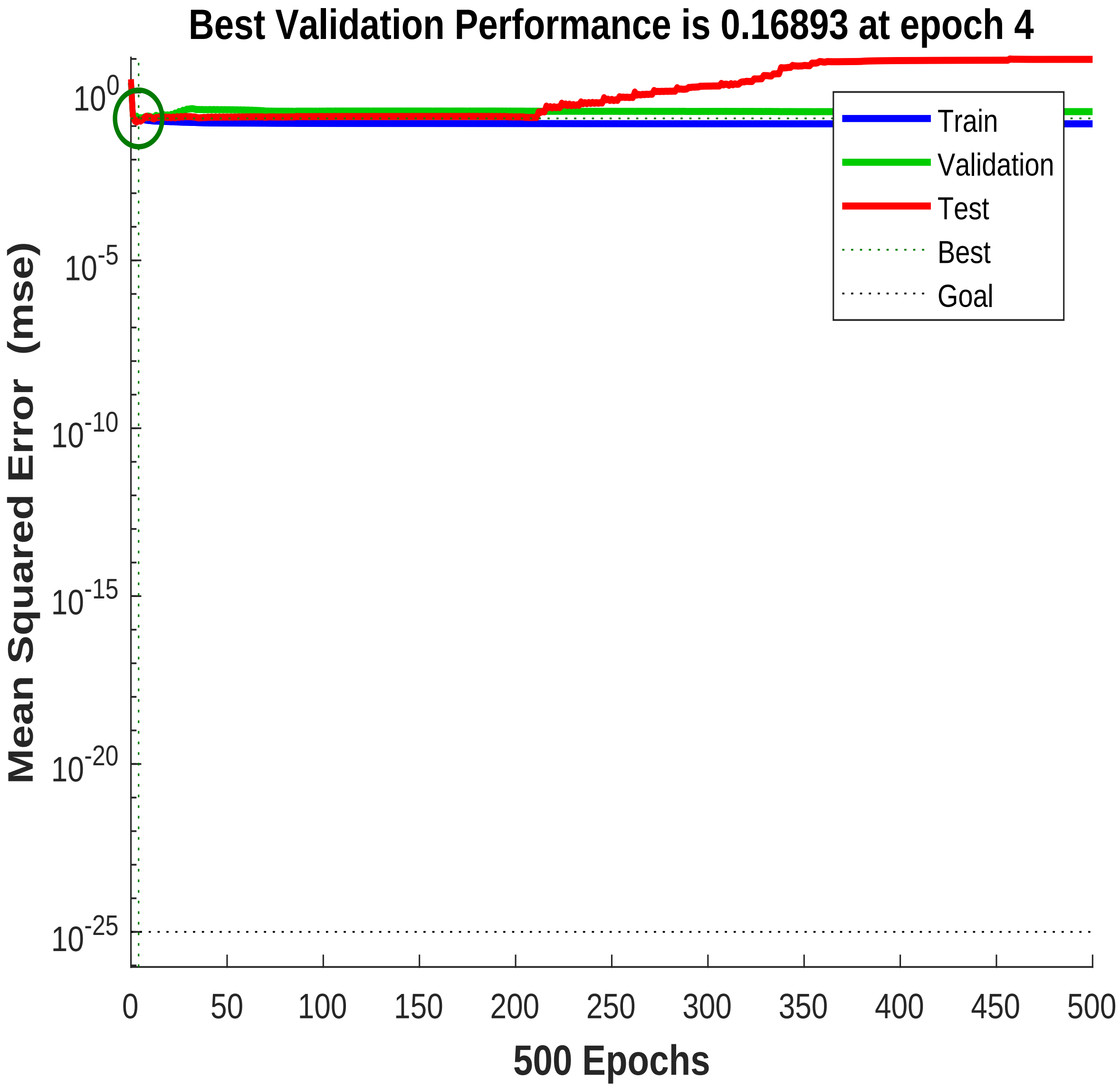


Figure 5.35: NN Model Performance

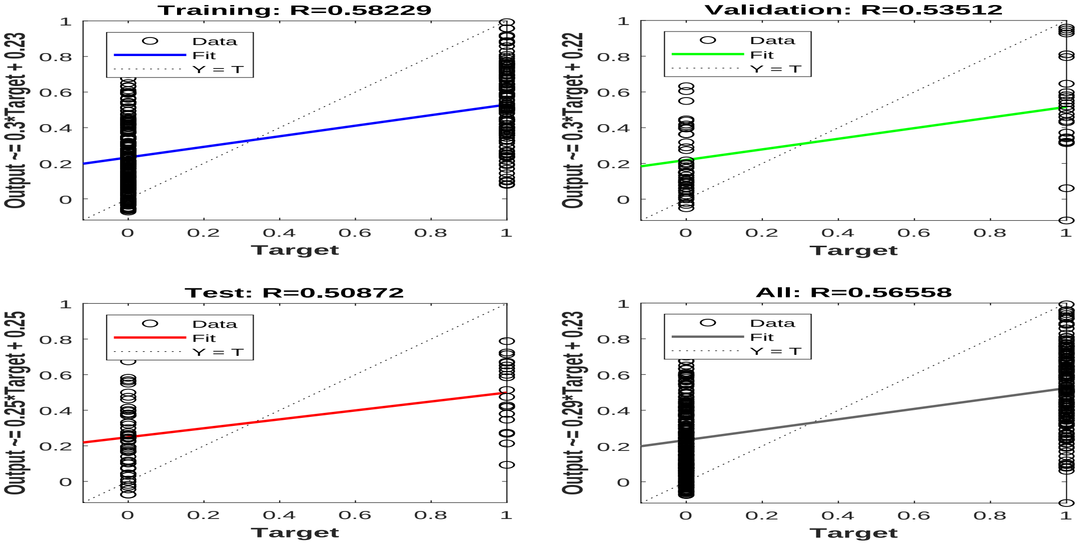


Figure 5.36: NN Model Training Regression

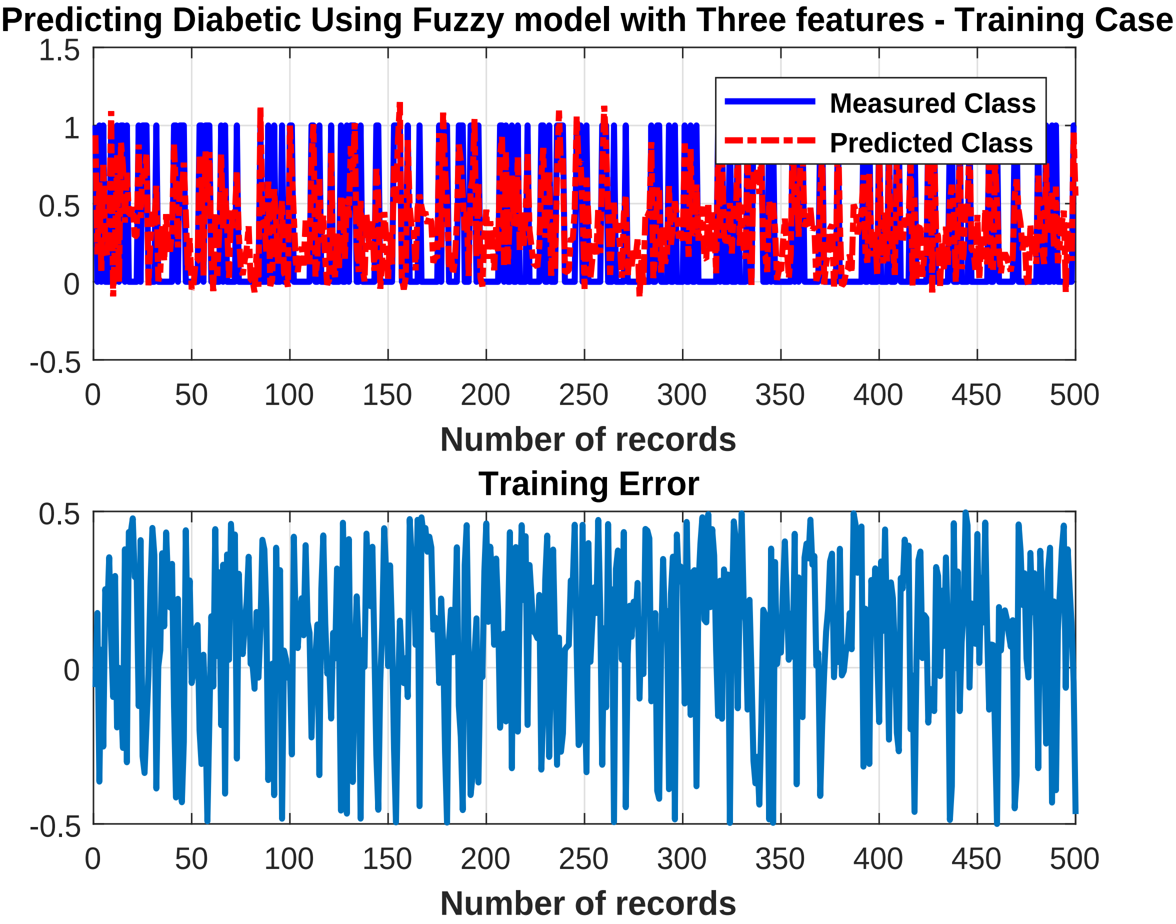


Figure 5.37: Fuzzy Model Training

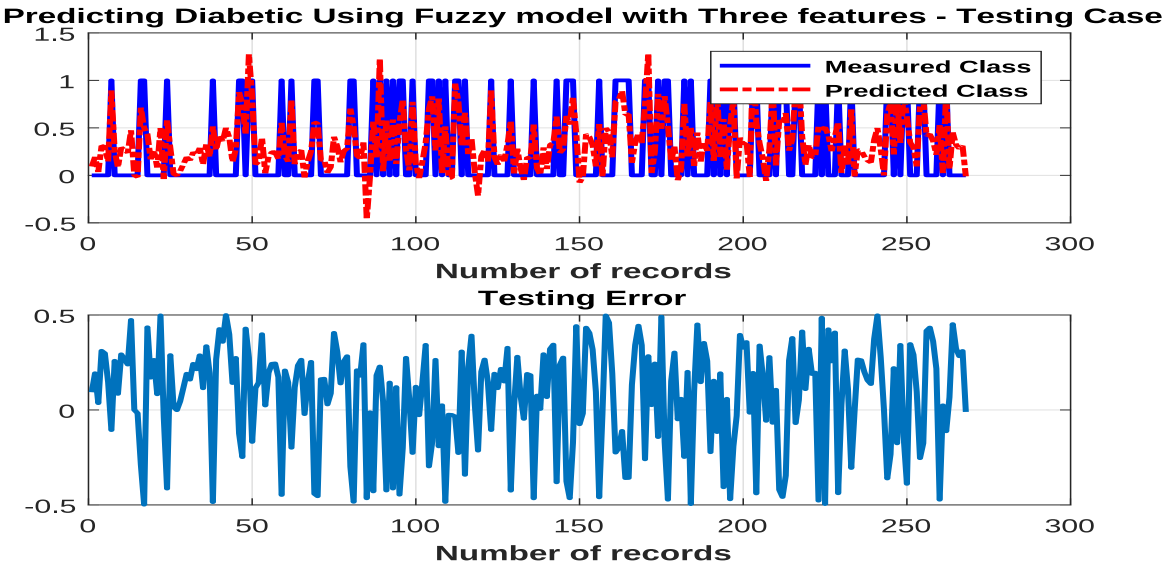


Figure 5.38: Fuzzy Model Testing

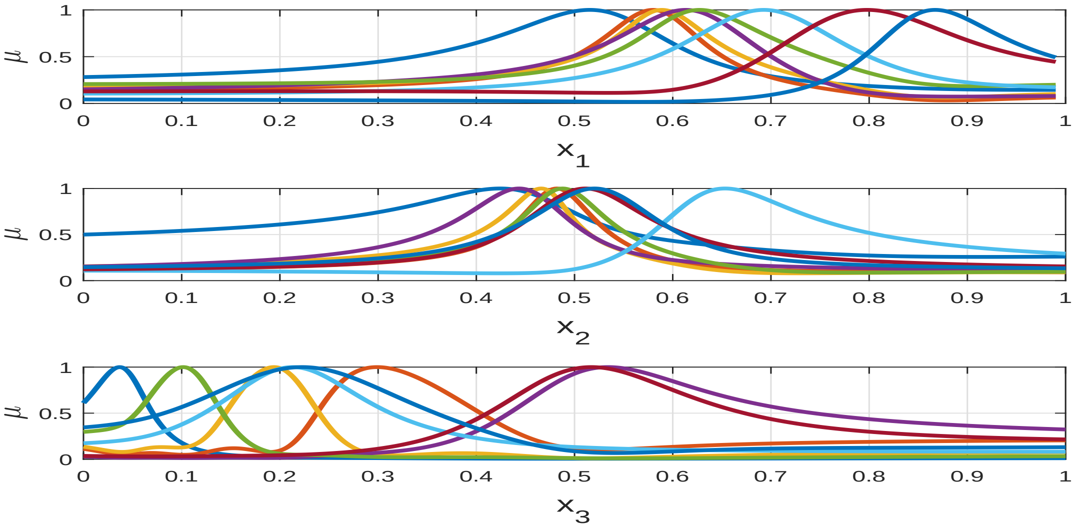


Figure 5.39: Membership Function of Fuzzy Model

5.6 Models of Classification Diabetes using Two Features

This section describe the result after select the two weighted features from the mathematical model that built with the genetic programming model. The selected features are shown in table 5.20.

Table 5.20: Inputs and Output for Diabetic Prediction Model

Inputs	The results of an oral glucose tolerance test	x_2
	Body mass index	x_6
Output	Predicted class	y

5.6.1 Genetic Programming model

Figure 5.40, shows the convergence of GP over 500 generations. It can be clearly seen that the final model is a simple and compact mathematical model which is easy to evaluate. The performance measurements for the model was computed and summarized in Table 5.22. The best generated diabetic prediction Multigene GP model is given in Table 5.21.

Table 5.21: A GP model with Inputs: x_2, x_6

$$\begin{aligned} y = \quad & 2.801\, {x_2}^6 - 11.82\, x_1 - 2.801\, x_1\, x_2\, \left(x_1 - x_2\right)\, \left(x_1 - 3\, x_2\right) + 28.18 \\ + \quad & .873\, x_1 - 3.291\, x_2 + 3.291\, x_1\, x_2 - 6.582\, x_1\, {x_2}^3 + 3.291\, {x_2}^2 - 27.59 \\ - \quad & 0.4318\, \left({x_1}^2 + {x_2}^2\right)\, \left(2\, {x_1}^3\, x_2 + x_1\, x_2\, \left(x_1 - x_2\right) - 4.639\right) \end{aligned}$$

5.6.2 Neural Network model

Table 5.23 shows the performance measurement of neural network model. Figure 5.44, shows the convergence of Neural Network over 500 number of

Table 5.22: Performance of the GP model with x_2, x_6 as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.8679	0.9176
Specificity	0.5714	0.5349
Accuracy	0.7600	0.7948
Positive Predicted Value	0.7797	0.8068
Negative Predicted Value	0.7123	0.7541

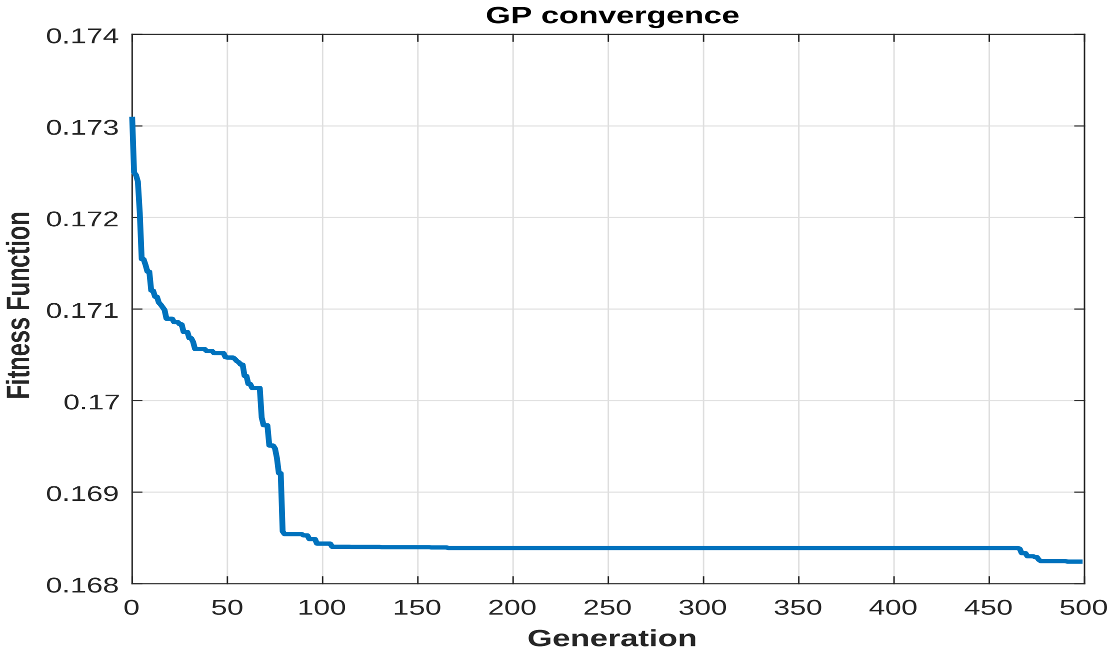


Figure 5.40: Convergence of the GP evolutionary process using Two Features

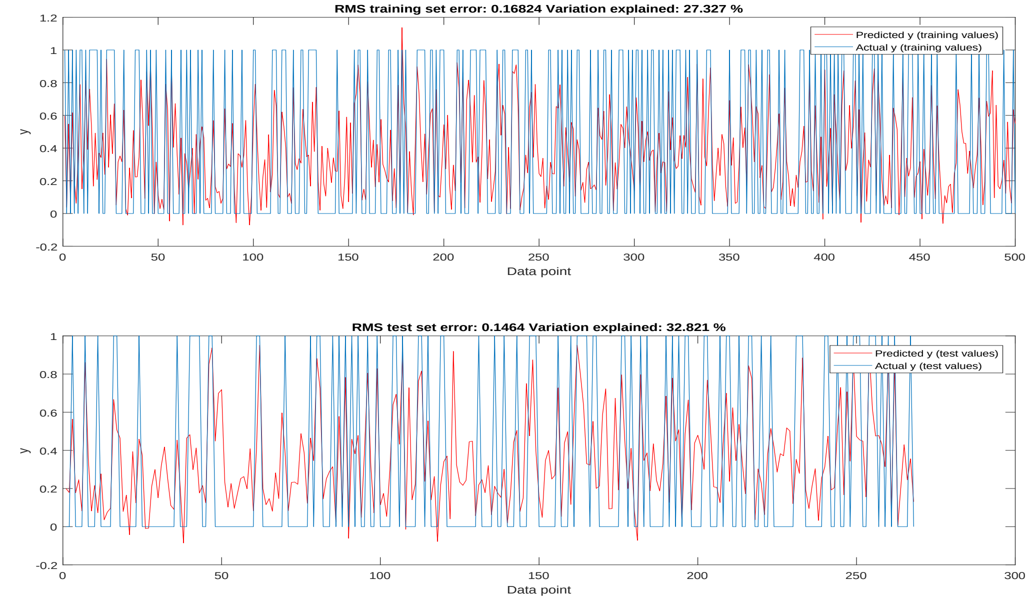


Figure 5.41: Measure class and Predicted Class in Training and Testing of GP Model using Two Features

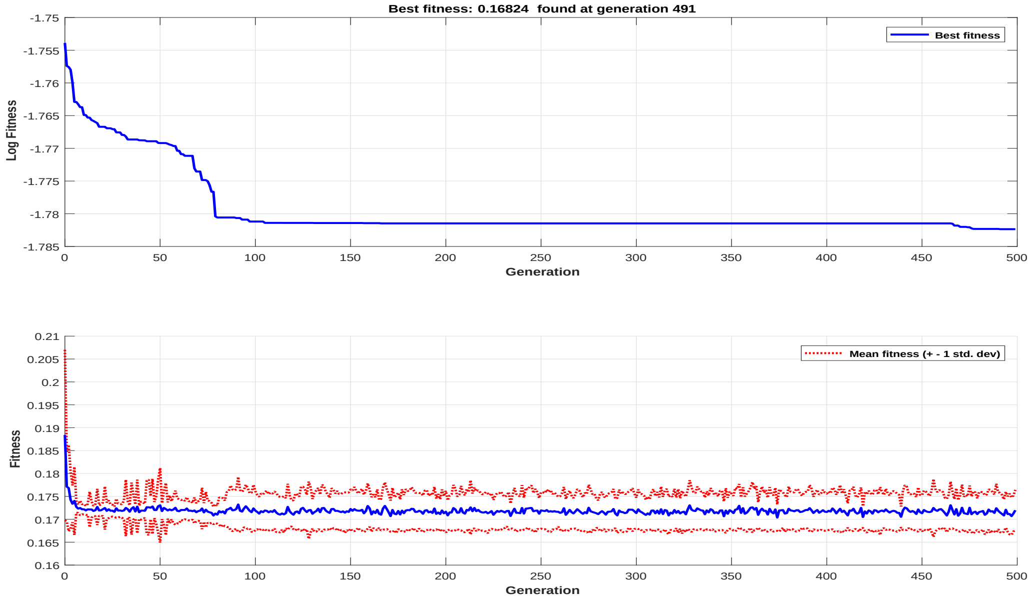


Figure 5.42: Best Fitness of the GP Model using Two Features

epochs. Figure 5.45 and figure 5.46 describe the measure class and predicted class in training and testing respectively. Figure 5.47 shows the training state of the model. Figure 5.48 describe the performance of the model.

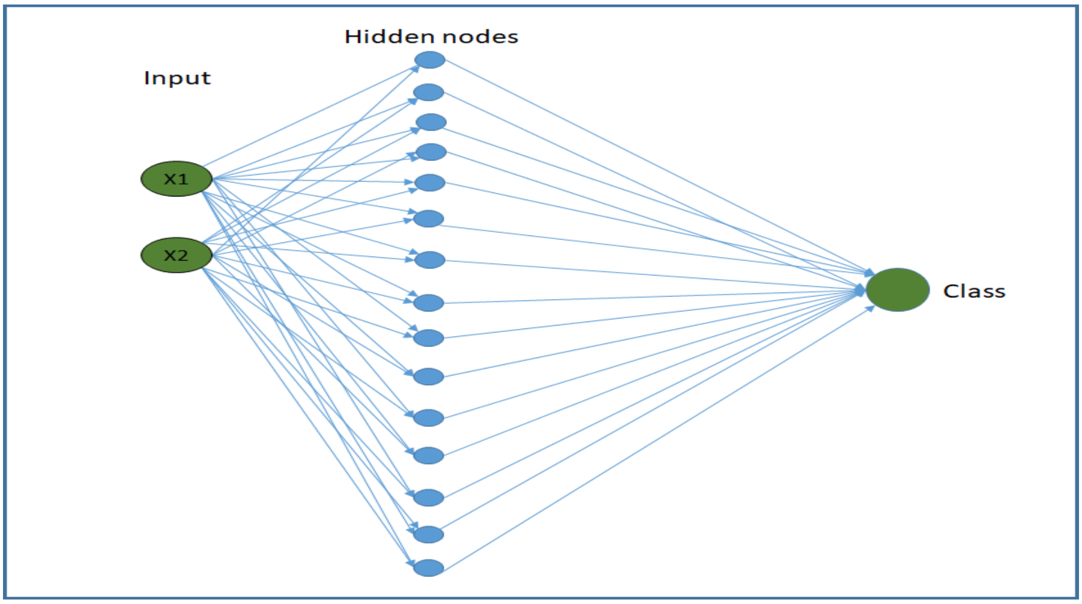


Figure 5.43: Artificial Neural Network Structure

Table 5.23: Performance of the NN model with x_2, x_6as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.9025	0.9231
Specificity	0.5330	0.5000
Accuracy	0.7680	0.7873
Positive Predicted Value	0.7715	0.7962
Negative Predicted Value	0.7578	0.7544

5.6.3 Fuzzy model

Figure 5.50 and figure 5.51 shows the training and testing of the fuzzy model respectively. The membership functions for every feature describe in figure 5.52.

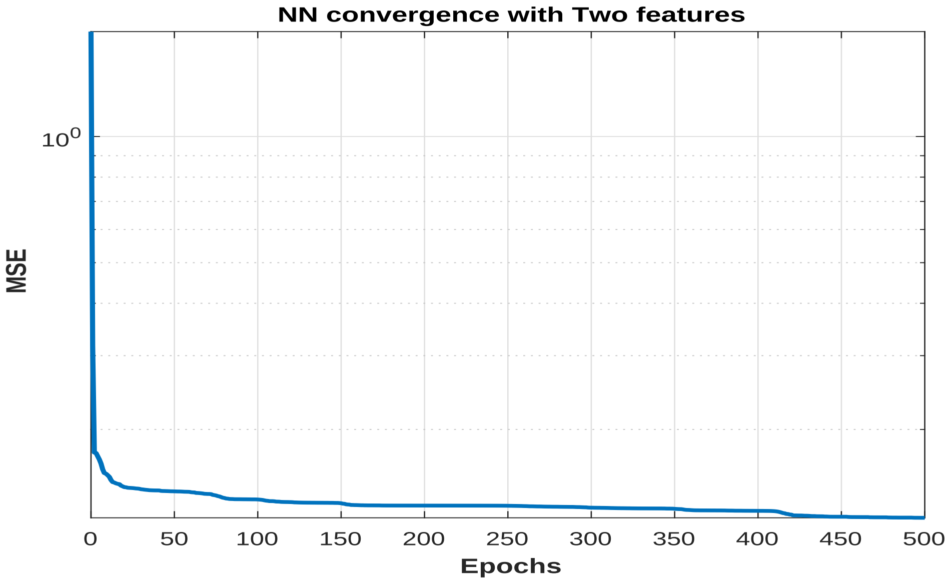


Figure 5.44: Convergence of the Artificial Neural Network Process

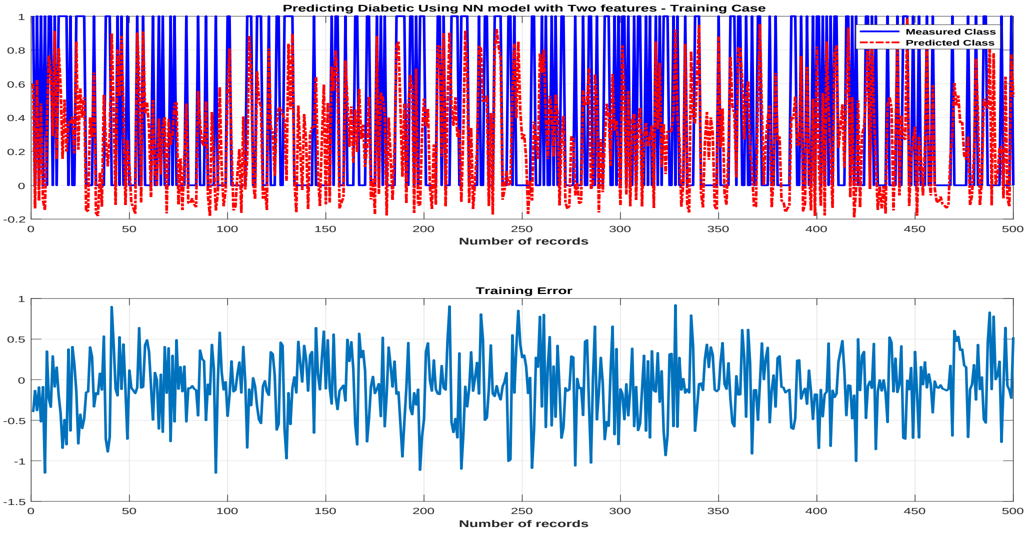


Figure 5.45: Measure class and Predicted Class in Training

Table 5.24: Performance of the Fuzzy model with x_2, x_6 as Inputs

Criteria	**Training**	**Testing**
Sensitivity	0.8239	0.8516
Specificity	0.6099	0.6163
Accuracy	0.7460	0.7761
Positive Predicted Value	0.7868	0.8245
Negative Predicted Value	0.6647	0.6625

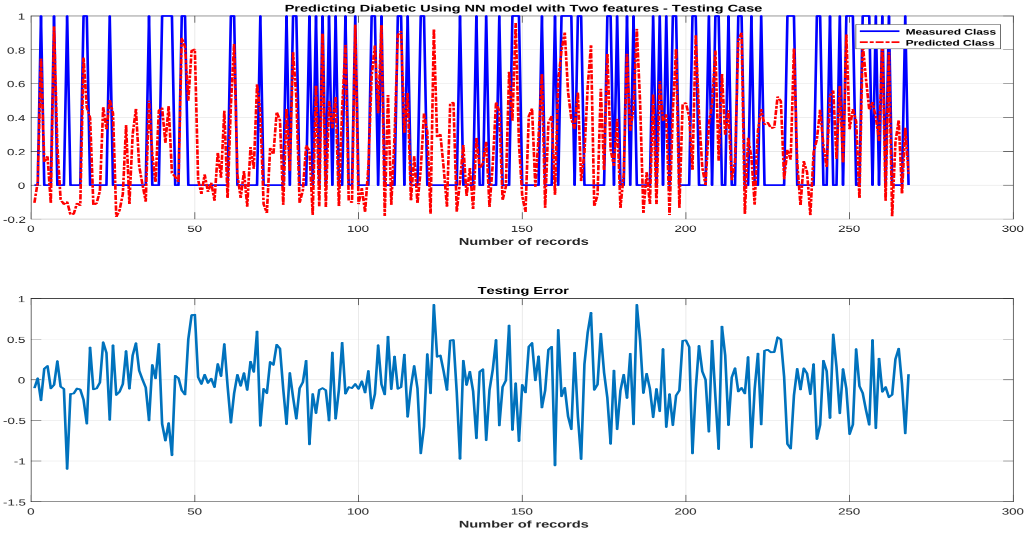


Figure 5.46: Measure class and Predicted Class in Testing

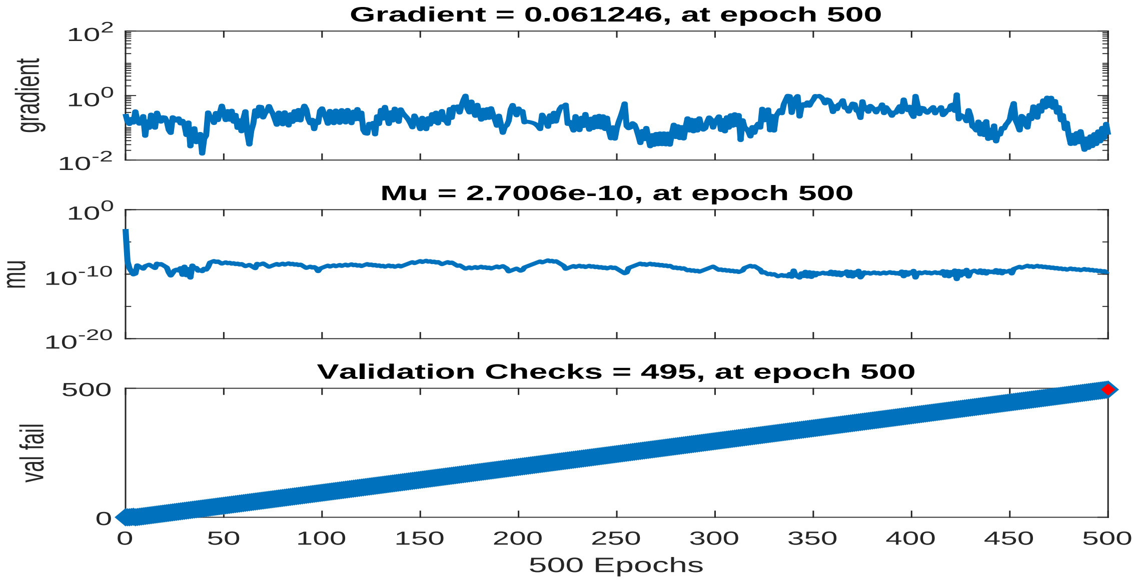


Figure 5.47: NN Model Training State

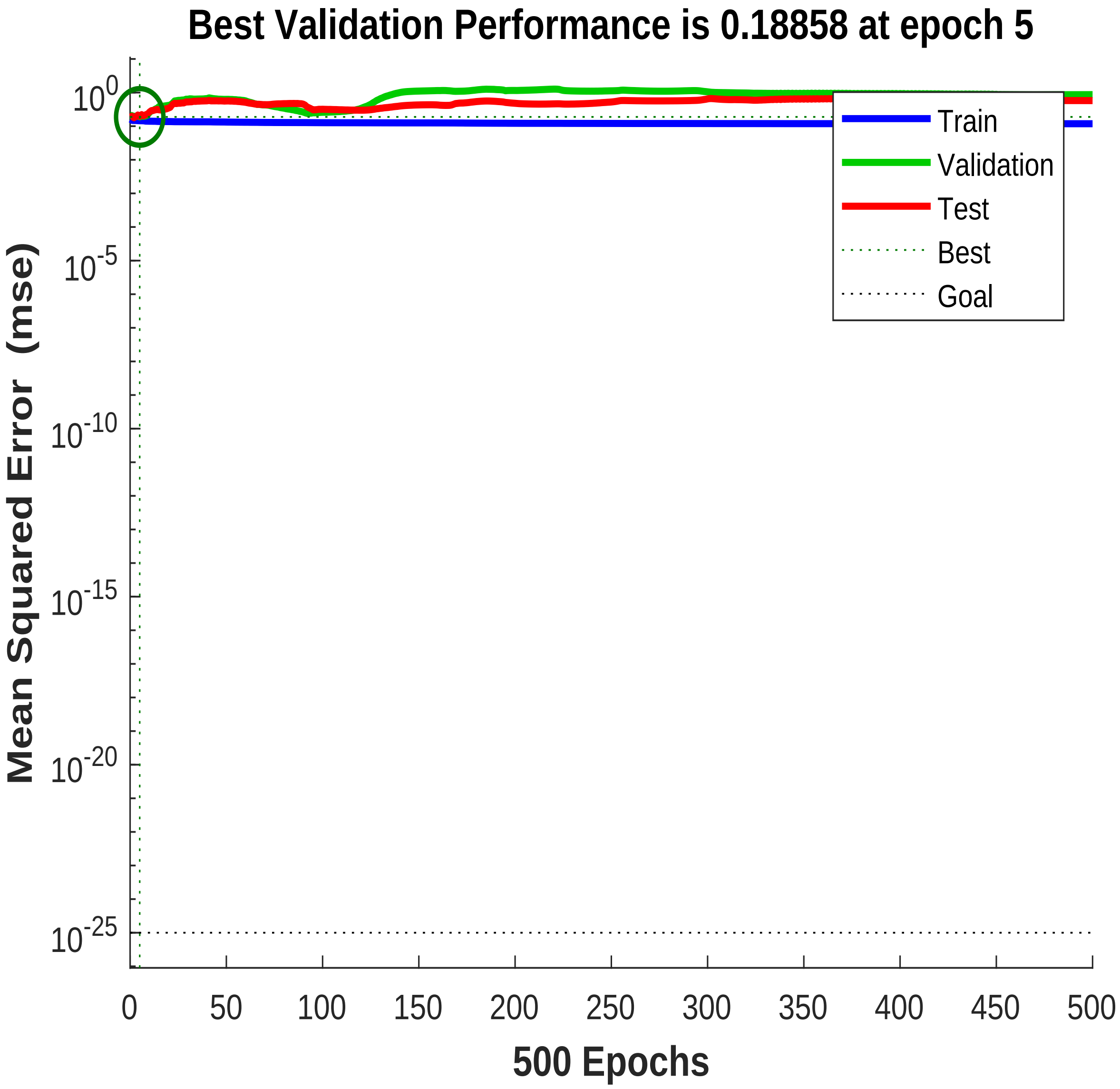


Figure 5.48: NN Model Performance

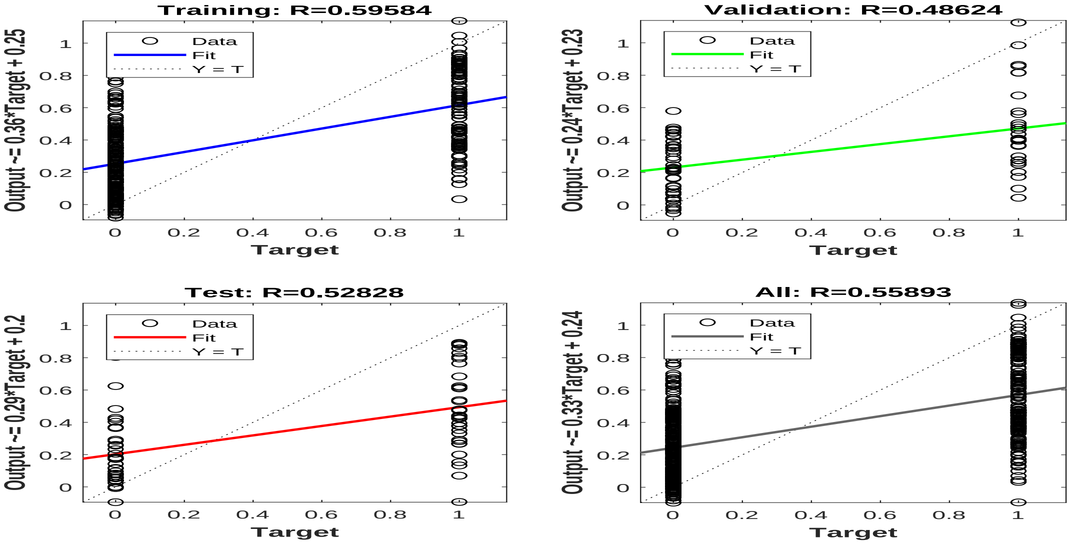


Figure 5.49: NN Model Training Regression

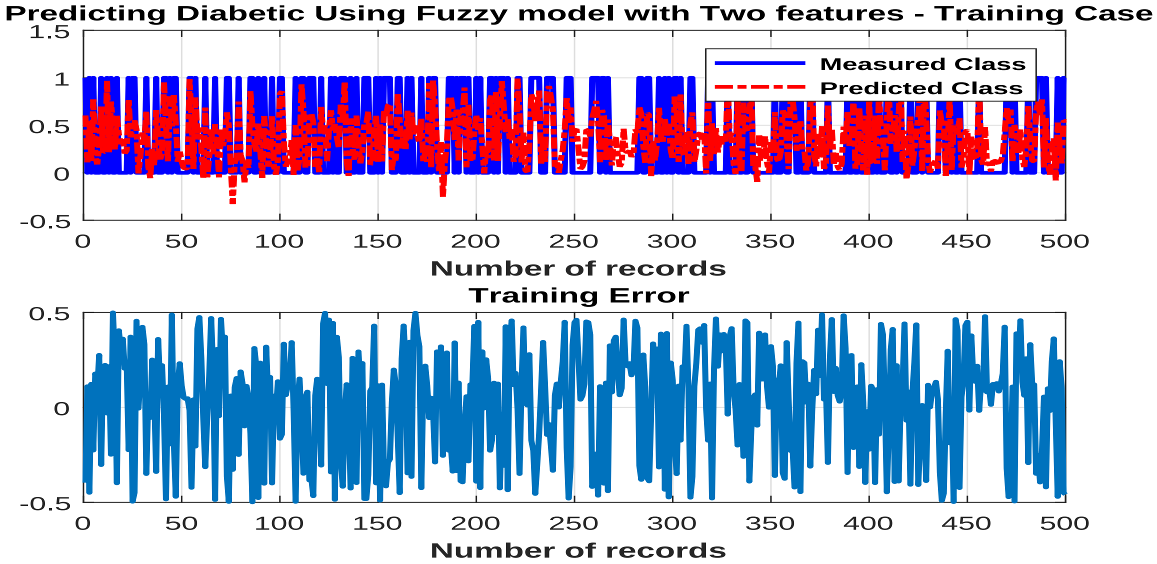


Figure 5.50: Fuzzy Model Training

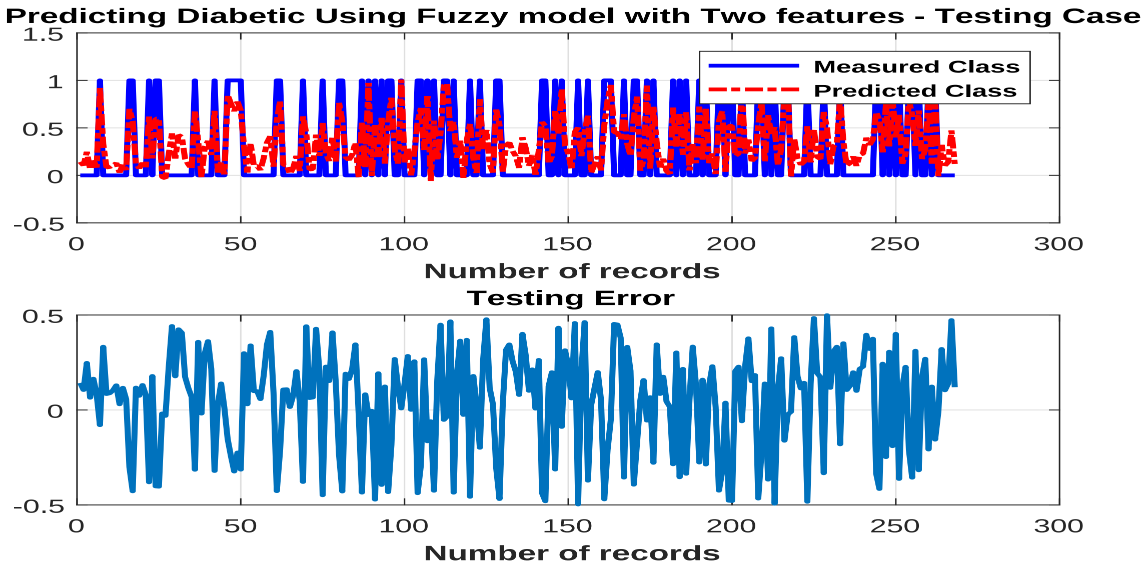


Figure 5.51: Fuzzy Model Testing

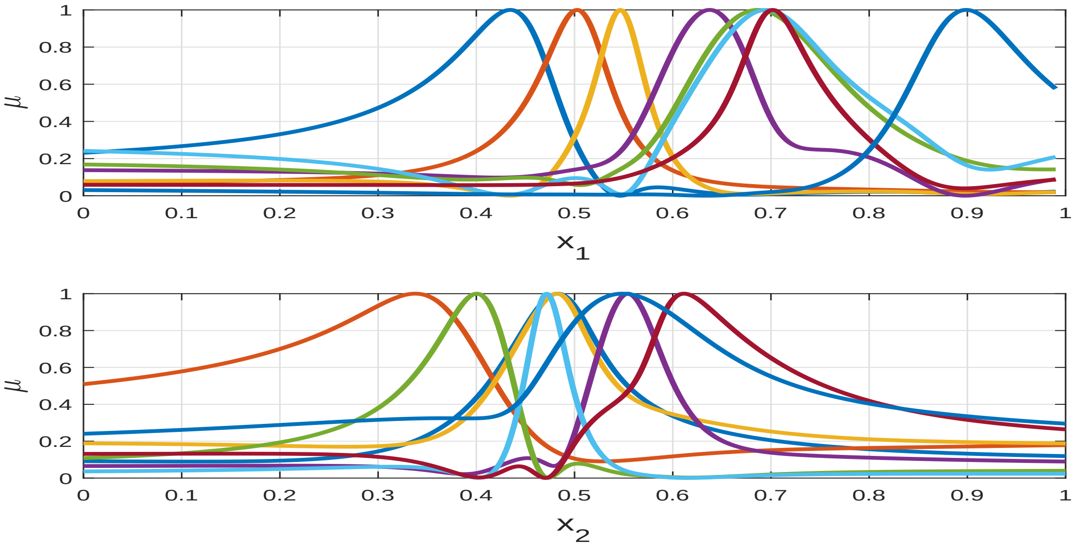


Figure 5.52: Membership Function of Fuzzy Model

5.7 Comparison of Different Classifiers

This section describe the comparison of different classifiers using different features. Table 5.25 shows the comparison.

Table 5.25: Performance comparison of different classifiers

No. of selected attribute	**NN**	**GP**	**Fuzzylogic**
Eight Features	0.8233	0.7740	0.8036
Four Features	0.8105	0.7700	0.7940
Three Features	0.8105	0.7700	0.7760
Two Features	0.7680	0.7600	0.7460

5.8 Features Analysis and Weighted

A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.

Several types of correlation coefficient exist, each with their own definition and own range of usability and characteristics. They all assume values in the range from 1 to +1, where +1 indicates the strongest possible agreement and 1 the strongest possible disagreement. As tools of analysis, correlation coefficients present certain problems, including the propensity of some types to be distorted by outliers and the possibility of incorrectly being used to infer a causal relationship between the variables. Table 5.26 shows the inputs and output for diabetic prediction model with their weights and 5.53 and 5.54 show the charts. The correlation coefficient of the features describe in figures 5.55,5.56,5.57 and 5.58.

Table 5.26: Inputs and Output for Diabetic Prediction Model with Their Weights

Feature	**Label**	**Weight**
The number of times pregnant	x_1	0.199
The results of an oral glucose tolerance test	x_2	0.476
Diastolic blood pressure (mm/Hg)	x_3	0.143
E-Triceps skin fold thickness (mm)	x_4	0.090
2-h serum insulin (micro U/ml)	x_5	0.066
Body mass index	x_6	0.310
Diabetes pedigree function	x_7	0.175
Age (year)	x_8	0.309

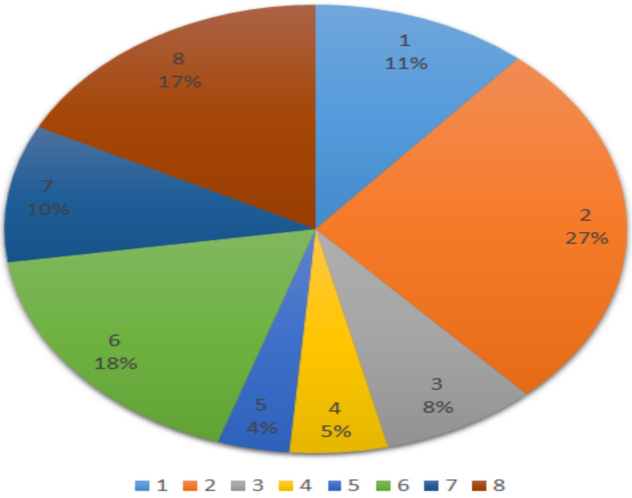


Figure 5.53: Features Chart depend on the Weight

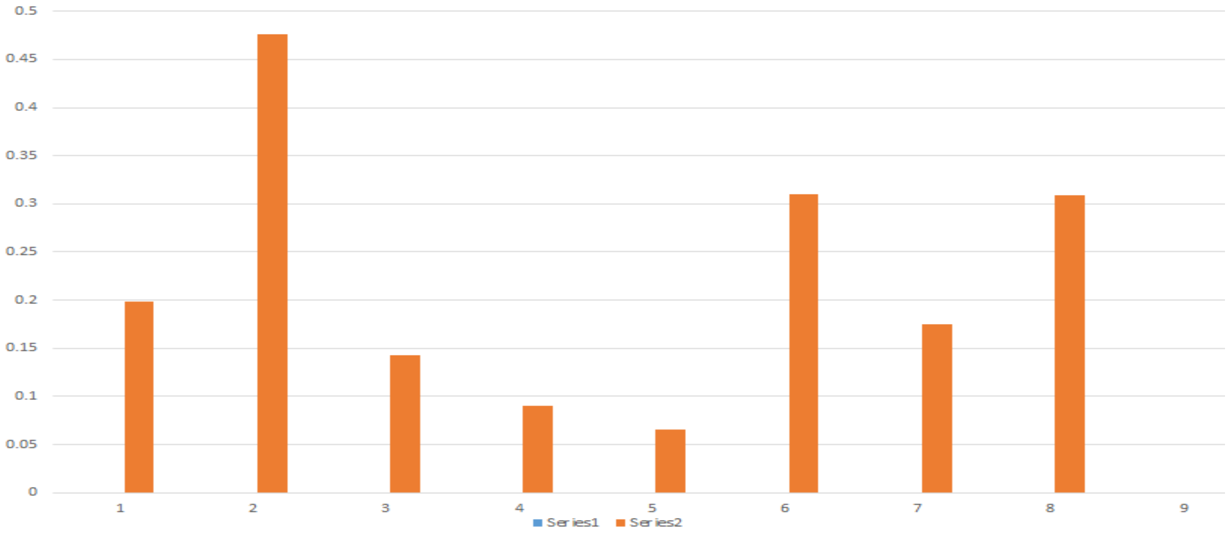
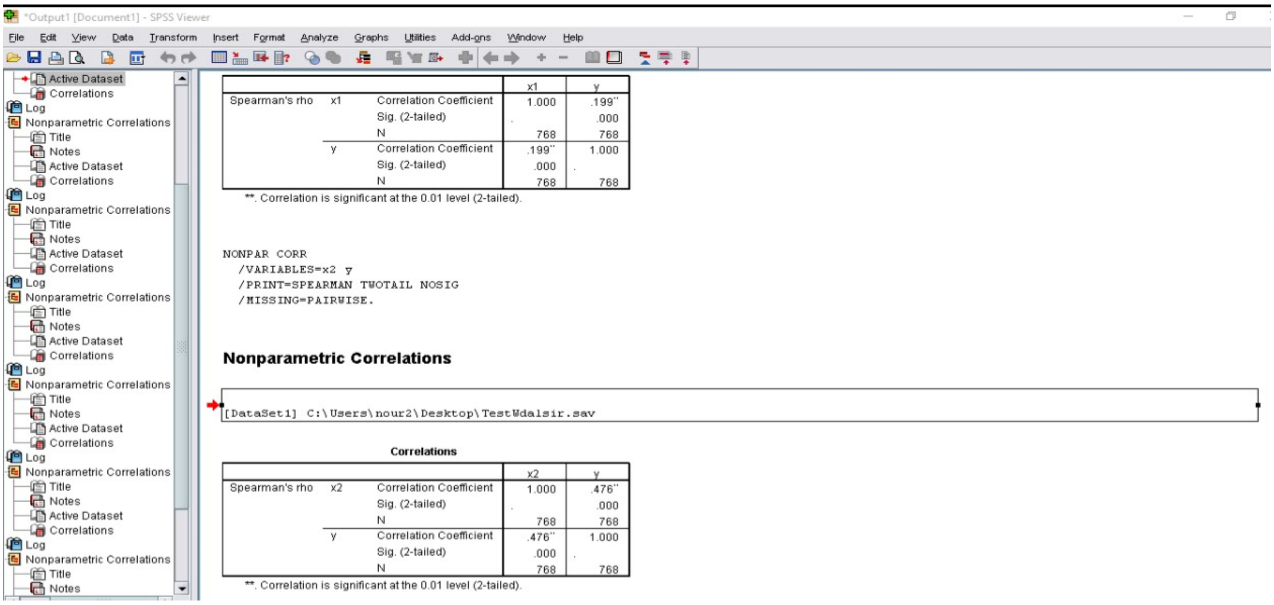


Figure 5.54: Features Chart depend on the Weight



			x1	y
Spearman's rho	x1	Correlation Coefficient	1.000	.199**
		Sig. (2-tailed)	.	.000
		N	768	768
	y	Correlation Coefficient	.199**	1.000
		Sig. (2-tailed)	.000	.
		N	768	768

**. Correlation is significant at the 0.01 level (2-tailed).

```
NONPAR CORR
  /VARIABLES=x2 y
  /PRINT=SPEARMAN TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Nonparametric Correlations

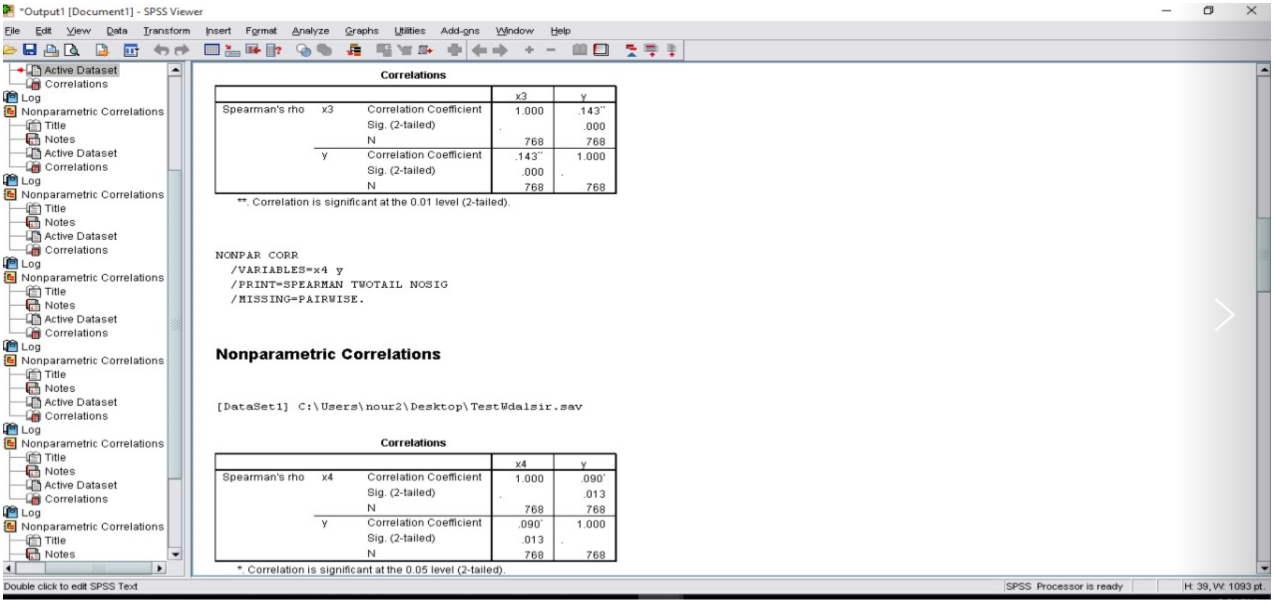
```
[DataSet1] C:\Users\nour2\Desktop\TestWdalsir.sav
```

Correlations

			x2	y
Spearman's rho	x2	Correlation Coefficient	1.000	.476**
		Sig. (2-tailed)	.	.000
		N	768	768
	y	Correlation Coefficient	.476**	1.000
		Sig. (2-tailed)	.000	.
		N	768	768

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 5.55: Correlation Coefficient of The Features x_1, x_2



Correlations

			x3	y
Spearman's rho	x3	Correlation Coefficient	1.000	.143**
		Sig. (2-tailed)	.	.000
		N	768	768
	y	Correlation Coefficient	.143**	1.000
		Sig. (2-tailed)	.000	.
		N	768	768

**. Correlation is significant at the 0.01 level (2-tailed).

```
NONPAR CORR
  /VARIABLES=x4 y
  /PRINT=SPEARMAN TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Nonparametric Correlations

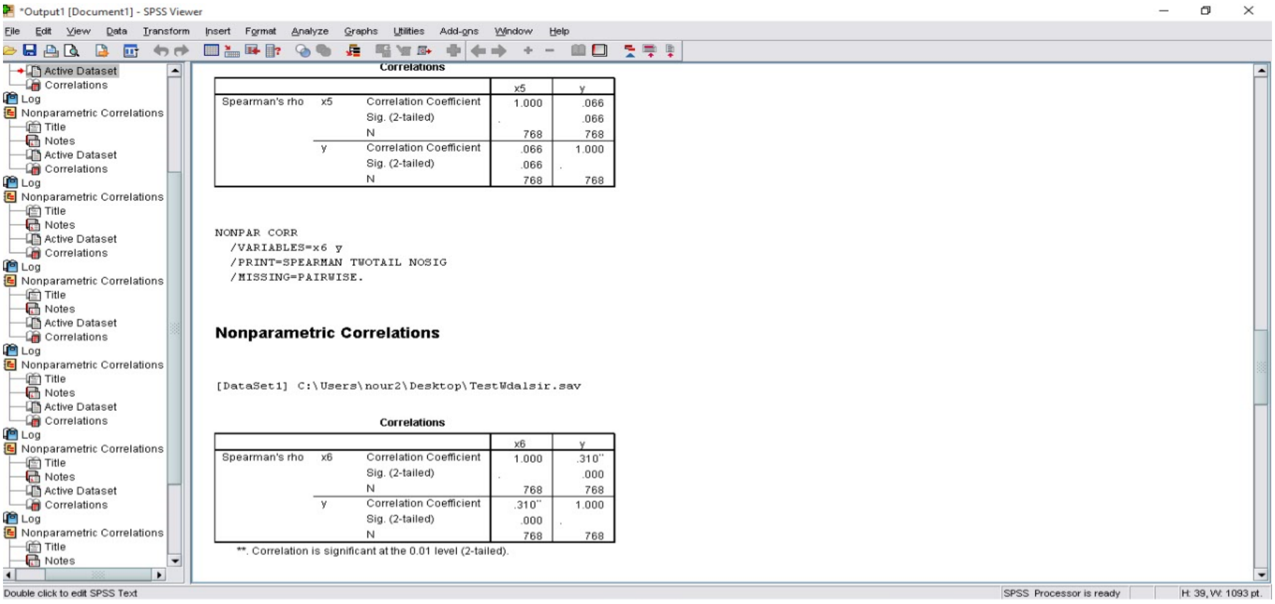
```
[DataSet1] C:\Users\nour2\Desktop\TestWdalsir.sav
```

Correlations

			x4	y
Spearman's rho	x4	Correlation Coefficient	1.000	.090*
		Sig. (2-tailed)	.	.013
		N	768	768
	y	Correlation Coefficient	.090*	1.000
		Sig. (2-tailed)	.013	.
		N	768	768

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 5.56: Correlation Coefficient of The Features x_3, x_4


*Output1 [Document1] - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Double click to edit SPSS Text
SPSS Processor is ready
H: 39, W: 1093 pt.

Correlations

			x5	y
Spearman's rho	x5	Correlation Coefficient	1.000	.066
		Sig. (2-tailed)	.	.066
		N	768	768
	y	Correlation Coefficient	.066	1.000
		Sig. (2-tailed)	.066	.
		N	768	768

```
NONPAR CORR
  /VARIABLES=x6 y
  /PRINT=SPEARMAN TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Nonparametric Correlations

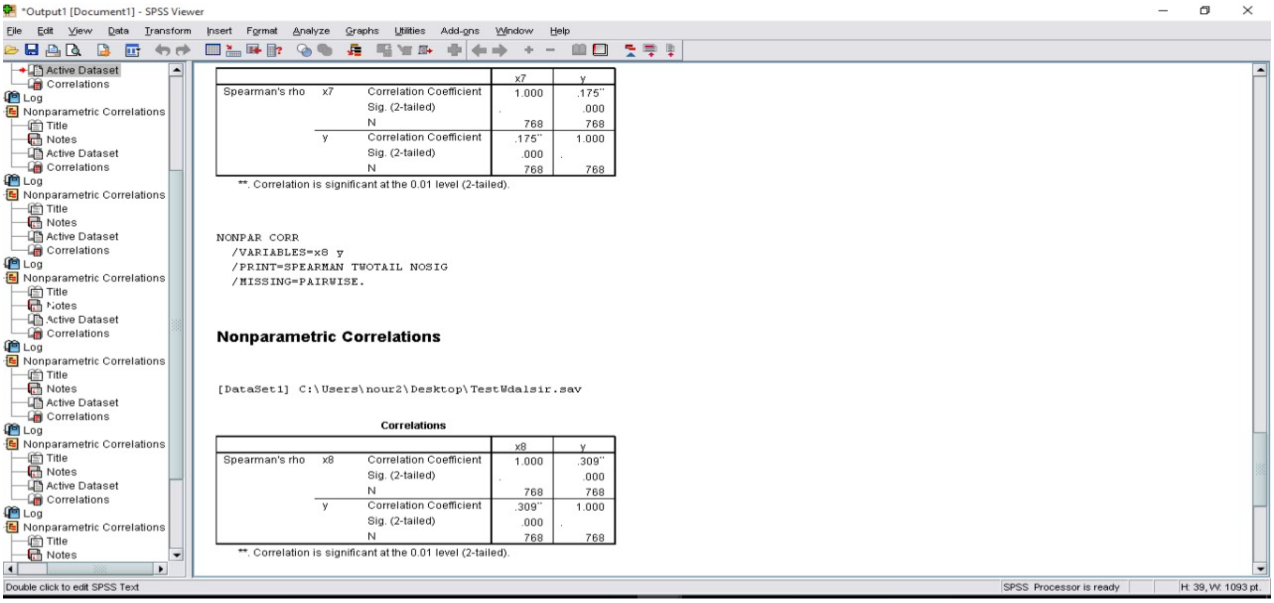
```
[DataSet1] C:\Users\nour2\Desktop\TestWdalsir.sav
```

Correlations

			x6	y
Spearman's rho	x6	Correlation Coefficient	1.000	.310**
		Sig. (2-tailed)	.	.000
		N	768	768
	y	Correlation Coefficient	.310**	1.000
		Sig. (2-tailed)	.000	.
		N	768	768

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 5.57: Correlation Coefficient of The Features x_5, x_6


*Output1 [Document1] - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Active Dataset
Correlations
Log
Nonparametric Correlations
Title
Notes
Double click to edit SPSS Text
SPSS Processor is ready
H: 39, W: 1093 pt.

			x7	y
Spearman's rho	x7	Correlation Coefficient	1.000	.175**
		Sig. (2-tailed)	.	.000
		N	768	768
	y	Correlation Coefficient	.175**	1.000
		Sig. (2-tailed)	.000	.
		N	768	768

**. Correlation is significant at the 0.01 level (2-tailed).

```
NONPAR CORR
  /VARIABLES=x8 y
  /PRINT=SPEARMAN TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Nonparametric Correlations

```
[DataSet1] C:\Users\nour2\Desktop\TestWdalsir.sav
```

Correlations

			x8	y
Spearman's rho	x8	Correlation Coefficient	1.000	.309**
		Sig. (2-tailed)	.	.000
		N	768	768
	y	Correlation Coefficient	.309**	1.000
		Sig. (2-tailed)	.000	.
		N	768	768

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 5.58: Correlation Coefficient of The Features x_7, x_8

5.9 Fuzzy Rules

The Fuzzy Rules using $x_1, \ldots, x_8$:

1. **If** x_1 is A_{11} **and** x_2 is A_{12} **and** x_3 is A_{13} **and** x_4 is A_{14}
   and x_5 is A_{15} **and** x_6 is A_{16} **and** x_7 is A_{17} **and** x_8 is A_{18} **then**
   $y(k) = 6.9 \cdot 10^{e+1} x_1 - 2.6 \cdot 10^{e+1} x_2 + 7.3 \cdot 10^{e+1} x_3 + 9.9 \cdot 10^{e+1} x_4$
   $+8.0 \cdot 10^{e+1} x_5 - 8.6 \cdot 10^{e+1} x_6 - 6.7 \cdot 10^{e+1} x_7 - 7.5 \cdot 10^{e+1} x_8 - 1.5 \cdot 10^{e+1}$
2. **If** x_1 is A_{21} **and** x_2 is A_{22} **and** x_3 is A_{23} **and** x_4 is A_{24} **and** x_5 is A_{25} **and** x_6 is A_{26} **and** x_7 is A_{27} **and** x_8 is A_{28} **then**
   $y(k) = -3.1 \cdot 10^{e+} x_1 - 1.5 \cdot 10^{e+1} x_2 - 1.8 \cdot 10^{e-1} x_3 - 9.8 \cdot 10^{e+1} x_4 + 2.9 \cdot 10^{e+} x_5 + 8.3 \cdot 10^{e+1} x_6 + 4.3 \cdot 10^{e+1} x_7 - 1.9 \cdot 10^{e+1} x_8 - 9.7 \cdot 10^{e-1}$

3. **If** x_1 is A_{31} **and** x_2 is A_{32} **and** x_3 is A_{33} **and** x_4 is A_{34} **and** x_5 is A_{35} **and** x_6 is A_{36} **and** x_7 is A_{37} **and** x_8 is A_{38} **then**
   $y(k) = -6.9 \cdot 10^{e+1} x_1 + 2.9 \cdot 10^{e+1} x_2 - 3.5 \cdot 10^{e+1} x_3 - 7.1 \cdot 10^{e+1} x_4 - 1.4 \cdot 10^{e+2} x_5 + 3.1 \cdot 10^{e+1} x_6 + 6.8 \cdot 10^{e+1} x_7 + 9.7 \cdot 10^{e+1} x_8 + 8.9 \cdot 10^{e+}$

4. **If** x_1 is A_{41} **and** x_2 is A_{42} **and** x_3 is A_{43} **and** x_4 is A_{44} **and** x_5 is A_{45} **and** x_6 is A_{46} **and** x_7 is A_{47} **and** x_8 is A_{48} **then**
   $y(k) = -5.4 \cdot 10^{e+1} x_1 + 1.7 \cdot 10^{e+1} x_2 - 5.5 \cdot 10^{e+1} x_3 + 7.0 \cdot 10^{e+} x_4 - 3.7 \cdot 10^{e+} x_5 - 1.8 \cdot 10^{e+1} x_6 - 1.3 \cdot 10^{e+} x_7 + 7.4 \cdot 10^{e+1} x_8 + 2.8 \cdot 10^{e+1}$

5. **If** x_1 is A_{51} **and** x_2 is A_{52} **and** x_3 is A_{53} **and** x_4 is A_{54} **and** x_5 is A_{55} **and** x_6 is A_{56} **and** x_7 is A_{57} **and** x_8 is A_{58} **then**
   $y(k) = 2.6 \cdot 10^{e-1} x_1 - 9.4 \cdot 10^{e+} x_2 - 3.6 \cdot 10^{e+1} x_3 - 8.8 \cdot 10^{e+1} x_4 + 1.9 \cdot 10^{e+1} x_5 + 8.9 \cdot 10^{e+1} x_6 + 3.5 \cdot 10^{e+1} x_7 - 4.1 \cdot 10^{e+1} x_8 + 6.8 \cdot 10^{e+}$

6. **If** x_1 is A_{61} **and** x_2 is A_{62} **and** x_3 is A_{63} **and** x_4 is A_{64} **and** x_5 is A_{65} **and** x_6 is A_{66} **and** x_7 is A_{67} **and** x_8 is A_{68} **then**
   $y(k) = 7.2 \cdot 10^{e+1} x_1 + 2.2 \cdot 10^{e+1} x_2 - 3.3 \cdot 10^{e+1} x_3 + 3.0 \cdot 10^{e+2} x_4 + 8.7 \cdot 10^{e+1} x_5 - 1.6 \cdot 10^{e+2} x_6 - 1.6 \cdot 10^{e+2} x_7 + 2.9 \cdot 10^{e+1} x_8 - 1.2 \cdot 10^{e+1}$

7. **If** x_1 is A_{71} **and** x_2 is A_{72} **and** x_3 is A_{73} **and** x_4 is A_{74} **and** x_5 is A_{75} **and** x_6 is A_{76} **and** x_7 is A_{77} **and** x_8 is A_{78} **then**
   $y(k) = -9.5 \cdot 10^{e+} x_1 - 3.4 \cdot 10^{e+} x_2 - 2.5 \cdot 10^{e+} x_3 - 1.1 \cdot 10^{e+1} x_4 + 2.9 \cdot 10^{e+1} x_5 + 2.9 \cdot 10^{e+} x_6 + 5.8 \cdot 10^{e+} x_7 + 2.5 \cdot 10^{e+} x_8 + 5.4 \cdot 10^{e+}$

8. **If** x_1 is A_{81} **and** x_2 is A_{82} **and** x_3 is A_{83} **and** x_4 is A_{84} **and** x_5 is A_{85} **and** x_6 is A_{86} **and** x_7 is A_{87} **and** x_8 is A_{88} **then**
   $y(k) = -5.6 \cdot 10^{e+} x_1 - 8.9 \cdot 10^{e+} x_2 + 9.3 \cdot 10^{e+1} x_3 - 1.4 \cdot 10^{e+2} x_4 - 6.8 \cdot 10^{e+1} x_5 + 6.2 \cdot 10^{e+1} x_6 + 7.8 \cdot 10^{e+1} x_7 - 6.3 \cdot 10^{e+1} x_8 - 2.4 \cdot 10^{e+1}$

The Fuzzy Rules using $x_1, x_2, x6, x_8$:

1. **If** x_1 is A_{11} **and** x_2 is A_{12} **and** x_3 is A_{13} **and** x_4 is A_{14} **then**
   $y(k) = -3.2 \cdot 10^{e+} x_1 - 4.7 \cdot 10^{e-1} x_2 + 8.3 \cdot 10^{e-2} x_3 + 5.0 \cdot 10^{e+} x_4 + 3.2 \cdot 10^{e-1}$

2. **If** x_1 is A_{21} **and** x_2 is A_{22} **and** x_3 is A_{23} **and** x_4 is A_{24} **then**
   $y(k) = -4.8 \cdot 10^{e+} x_1 + 1.0 \cdot 10^{e+} x_2 - 2.1 \cdot 10^{e+} x_3 + 2.2 \cdot 10^{e+} x_4 + 1.3 \cdot 10^{e+}$

3. **If** x_1 is A_{31} **and** x_2 is A_{32} **and** x_3 is A_{33} **and** x_4 is A_{34} **then**
   $y(k) = 4.1 \cdot 10^{e+} x_1 + 2.0 \cdot 10^{e+} x_2 - 3.3 \cdot 10^{e-1} x_3 - 2.0 \cdot 10^{e+} x_4 - 1.5 \cdot 10^{e+}$

4. **If** x_1 is A_{41} **and** x_2 is A_{42} **and** x_3 is A_{43} **and** x_4 is A_{44} **then**
   $y(k) = 6.6 \cdot 10^{e-1} x_1 + 3.5 \cdot 10^{e+} x_2 + 3.0 \cdot 10^{e-2} x_3 - 7.6 \cdot 10^{e-1} x_4 - 2.0 \cdot 10^{e+}$

5. **If** x_1 is A_{51} **and** x_2 is A_{52} **and** x_3 is A_{53} **and** x_4 is A_{54} **then**
   $y(k) = -1.7 \cdot 10^{e+} x_1 - 1.4 \cdot 10^{e+} x_2 + 1.9 \cdot 10^{e+} x_3 - 2.9 \cdot 10^{e+} x_4 + 3.2 \cdot 10^{e+}$

6. **If** x_1 is A_{61} **and** x_2 is A_{62} **and** x_3 is A_{63} **and** x_4 is A_{64} **then**
   $y(k) = -2.2 \cdot 10^{e+} x_1 - 1.9 \cdot 10^{e+} x_2 + 5.4 \cdot 10^{e+} x_3 + 2.0 \cdot 10^{e+} x_4 - 8.8 \cdot 10^{e-1}$

7. **If** x_1 is A_{71} **and** x_2 is A_{72} **and** x_3 is A_{73} **and** x_4 is A_{74} **then**
   $y(k) = 3.1 \cdot 10^{e+} x_1 - 2.4 \cdot 10^{e+} x_2 + 2.0 \cdot 10^{e-1} x_3 - 1.4 \cdot 10^{e+} x_4 + 9.5 \cdot 10^{e-1}$

8. **If** x_1 is A_{81} **and** x_2 is A_{82} **and** x_3 is A_{83} **and** x_4 is A_{84} **then**
   $y(k) = 4.2 \cdot 10^{e+} x_1 + 1.0 \cdot 10^{e+1} x_2 + 1.7 \cdot 10^{e+} x_3 - 3.0 \cdot 10^{e+} x_4 - 8.6 \cdot 10^{e+}$

The Fuzzy Rules using $x_2, x6, x_8$:

1. **If** x_1 is A_{11} **and** x_2 is A_{12} **and** x_3 is A_{13} **then**
   $y(k) = -5.6 \cdot 10^{e-1}x_1 - 1.8 \cdot 10^{e-1}x_2 + 1.9 \cdot 10^{e-1}x_3 + 2.5 \cdot 10^{e-1}$

2. **If** x_1 is A_{21} **and** x_2 is A_{22} **and** x_3 is A_{23} **then**
   $y(k) = -1.9 \cdot 10^{e+}x_1 - 6.1 \cdot 10^{e+}x_2 + 6.1 \cdot 10^{e+}x_3 + 2.9 \cdot 10^{e+}$

3. **If** x_1 is A_{31} **and** x_2 is A_{32} **and** x_3 is A_{33} **then**
   $y(k) = -5.2 \cdot 10^{e-2}x_1 + 6.5 \cdot 10^{e+}x_2 - 1.5 \cdot 10^{e+}x_3 - 2.0 \cdot 10^{e+}$

4. **If** x_1 is A_{41} **and** x_2 is A_{42} **and** x_3 is A_{43} **then**
   $y(k) = 1.2 \cdot 10^{e+}x_1 + 2.0 \cdot 10^{e+}x_2 + 1.7 \cdot 10^{e+}x_3 - 3.1 \cdot 10^{e+}$

5. **If** x_1 is A_{51} **and** x_2 is A_{52} **and** x_3 is A_{53} **then**
   $y(k) = 6.8 \cdot 10^{e-2}x_1 - 1.5 \cdot 10^{e+}x_2 - 3.0 \cdot 10^{e+}x_3 + 1.5 \cdot 10^{e+}$

6. **If** x_1 is A_{61} **and** x_2 is A_{62} **and** x_3 is A_{63} **then**
   $y(k) = 3.5 \cdot 10^{e-1}x_1 + 7.5 \cdot 10^{e+}x_2 + 2.2 \cdot 10^{e+}x_3 - 4.7 \cdot 10^{e+}$

7. **If** x_1 is A_{71} **and** x_2 is A_{72} **and** x_3 is A_{73} **then**
   $y(k) = 3.9 \cdot 10^{e+}x_1 - 2.5 \cdot 10^{e+}x_2 - 5.2 \cdot 10^{e+}x_3 + 2.0 \cdot 10^{e+}$

8. **If** x_1 is A_{81} **and** x_2 is A_{82} **and** x_3 is A_{83} **then**
   $y(k) = 5.6 \cdot 10^{e-1}x_1 - 2.5 \cdot 10^{e+}x_2 - 2.6 \cdot 10^{e+}x_3 + 2.4 \cdot 10^{e+}$

The Fuzzy Rules using $x_2, x6$:

1. **If** x_1 is A_{11} **and** x_2 is A_{12} **then**

   $y(k) = -8.3 \cdot 10^{e-1}x_1 - 2.1 \cdot 10^{e+}x_2 + 1.3 \cdot 10^{e+}$

2. **If** x_1 is A_{21} **and** x_2 is A_{22} **then**

   $y(k) = 2.8 \cdot 10^{e+}x_1 + 3.6 \cdot 10^{e-1}x_2 - 1.5 \cdot 10^{e+}$

3. **If** x_1 is A_{31} **and** x_2 is A_{32} **then**

   $y(k) = -1.4 \cdot 10^{e+}x_1 + 2.1 \cdot 10^{e-1}x_2 + 1.2 \cdot 10^{e+}$

4. **If** x_1 is A_{41} **and** x_2 is A_{42} **then**

   $y(k) = 7.0 \cdot 10^{e+}x_1 - 5.4 \cdot 10^{e+}x_2 - 1.2 \cdot 10^{e+}$

5. **If** x_1 is A_{51} **and** x_2 is A_{52} **then**

   $y(k) = 2.3 \cdot 10^{e+}x_1 + 5.3 \cdot 10^{e+}x_2 - 3.6 \cdot 10^{e+}$

6. **If** x_1 is A_{61} **and** x_2 is A_{62} **then**

   $y(k) = -3.0 \cdot 10^{e-1}x_1 + 3.1 \cdot 10^{e+}x_2 - 5.8 \cdot 10^{e-1}$

7. **If** x_1 is A_{71} **and** x_2 is A_{72} **then**

   $y(k) = -5.2 \cdot 10^{e+}x_1 + 4.6 \cdot 10^{e+}x_2 + 1.4 \cdot 10^{e+}$

8. **If** x_1 is A_{81} **and** x_2 is A_{82} **then**

   $y(k) = -3.7 \cdot 10^{e+}x_1 - 1.4 \cdot 10^{e+}x_2 + 5.0 \cdot 10^{e+}$

Chapter Six

Conclusion and Recommendation

6.1 Introduction

Diabetes is a disease that occurs when the insulin production in the body is inadequate or the body is unable to use the produced insulin in a proper manner, as a result, this leads to high blood glucose. The body cells break down the food into glucose and this glucose needs to be transported to all the cells of the body. The insulin is the hormone that directs the glucose that is produced by breaking down the food into the body cells. Any change in the production of insulin leads to an increase in the blood sugar levels and this can lead to damage to the tissues and failure of the organs. Generally a person is considered to be suffering from diabetes, when blood sugar levels are above normal (4.4 to 6.1 mmol/L). There are three main types of diabetes, Type I, Type II and Gestational. The automatic diagnosis of diabetes is an important real-world medical problem. Detection of diabetes in its early stages is the key for treatment. This desertion shows how machine learning are used to model actual diagnosis of diabetes for local and systematic treatment, along with presenting related work in the field. Experimental results show the effectiveness of the proposed model. The performance of the techniques was investigated for the diabetes diagnosis problem. Experimental results demonstrate the adequacy of the proposed model.

6.2 Thesis Contribution

The contribution of this thesis

- A model is constructed to solve the diabetics classification and diagnosis.
- Simulation for model that used.
- The best attributes that affect on the performance of prediction and classification of diabetic patients is selected.
- Compare results with the benchmark machine learning techniques.
- Publishing articles related to this field in academic journals.

6.3 Recommended Future Work

There are some limitations of this study.

Considering the diabetes dataset, there might be other risk factors that the data collections did not consider. According to , other important factors include gestational diabetes, family history, metabolic syndrome, smoking, inactive lifestyles, certain dietary patterns etc. The proper prediction model would need more data gathering to make it more accurate. This can be achieved by collecting diabetes datasets from multiple sources, generating a model from each dataset.

This study used Genetic Programming,Neural network,Fuzzy logic to predict diabetes class. In order to find a best prediction model, other machine learning optimization methods such as Grey Wolf Optimization, Particle Swarm Optimization will be used to improve the predicting accuracy.

Developing a series of rules and standards is a valid method to prevent people from developing DM. Based on that, a more effective model for predicting DM and grading potential patients is presented. This will help to lower the growth rate of diabetes and eventually decrease the risk of developing DM.

The work can be extended and improved for the automation of diabetes analysis.

Reference

References

Al-Afeef, A., A. F. Sheta, and A. Al-Rabea (2010). Image reconstruction of a metal fill industrial process using genetic programming. In *International Conference on Intelligent Systems Design and Applications, November 29 - December 1, 2010, Cairo, Egypt*, pp. 12–17. IEEE.

Alaydie, N., C. Reddy, and F. Fotouhi (2012). Exploiting label dependency for hierarchical multi-label classification. In P.-N. Tan, S. Chawla, C. Ho, and J. Bailey (Eds.), *Advances in Knowledge Discovery and Data Mining*, Volume 7301 of *Lecture Notes in Computer Science*, pp. 294–305. Springer Berlin Heidelberg.

Allam, F., Z. Nossai, H. Gomma, I. Ibrahim, and M. Abdelsalam (2011). A recurrent neural network approach for predicting glucose concentration in type-I diabetic patients. In L. S. Iliadis and C. Jayne (Eds.), *Engineering Applications of Neural Networks*, Volume 363 of *IFIP Advances in Information and Communication Technology*, pp. 254–259. Springer.

Andrew, A. M. (2000). An introduction to support vector machines and other kernel-based learning methods by nello christianini and john shawe-taylor, cambridge university press, cambridge, 2000, 189 pp., isbn 0-521-78019-5. *Robotica 18*(6), 687–689.

Aslam, M. W. and A. K. Nandi (2010, August). Detection of diabetics using genetic programming. In *European Signal Processing Conference*, Number 18, Aalborg, Denmark.

Aslam, M. W., Z. Zhu, and A. K. Nandi (2013). Feature generation using genetic programming with comparative partner selection for diabetes classification. *Expert Systems with Applications 40*(13), 5402 – 5412.

Benardos, P. G. and G.-C. Vosniakos (2007). Optimizing feedforward artificial neural network architecture. *Eng. Appl. of AI 20*(3), 365–382.

beyli, E. D. (2009). Combined neural networks for diagnosis of erythemato-squamous diseases. *Expert Syst. Appl. 36*(3), 5107–5112.

Çalisir, D. and E. Dogantekin (2011). An automatic diabetes diagnosis system based on lda-wavelet support vector machine classifier. *Expert Syst. Appl. 38*(7), 8311–8315.

Chaovalitwongse, W. A., Y.-J. Fan, and R. C. Sachdeo (2007). On the time series k-nearest neighbor classification of abnormal brain activity. *IEEE Transactions on Systems, Man, and Cybernetics, Part A 37*(6), 1005–1016.

Elkamel, A., S. Abdul-Wahab, W. Bouhamra, and E. Alper (2001). Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach. *Advances in Environmental Research 5*(1), 47 – 59.

Enas, G. G. and S. C. Choi (1986). Choice of the smoothing parameter and efficiency of k-nearest neighbor classification. *Computers Mathematics with Applications 12*(2), 235 – 244.

Estrada, G. E. C., L. del Re, and E. Renard (2010). Nonlinear gain in online prediction of blood glucose profile in type 1 diabetic patients. In *Proceedings of the 49th IEEE Conference on Decision and Control, CDC 2010, December 15-17, 2010, Atlanta, Georgia, USA*, pp. 1668–1673. IEEE.

Hinchliffe, M., H. Hiden, B. McKay, M. Willis, M. Tham, and G. Barton (1996a, 28–31 July). Modelling chemical process systems using a multigene genetic programming algorithm. In J. R. Koza (Ed.), *Late Breaking Papers at the Genetic Programming 1996 Conference Stanford Univer-*

sity July 28-31, 1996, Stanford University, CA, USA, pp. 56–65. Stanford Bookstore.

Hinchliffe, M., H. Hiden, B. McKay, M. Willis, M. Tham, and G. Barton (1996b, 28–31 July). Modelling chemical process systems using a multigene genetic programming algorithm. In J. R. Koza (Ed.), *Late Breaking Papers at the Genetic Programming 1996 Conference Stanford University July 28-31, 1996*, Stanford University, CA, USA, pp. 56–65. Stanford Bookstore.

Hinchliffe, M. P. and M. J. Willis (2003). Dynamic systems modelling using genetic programming. *Computers & Chemical Engineering 27*(12), 1841–1854.

Huang, Y., P. J. McCullagh, N. D. Black, and R. Harper (2007). Feature selection and classification model construction on type 2 diabetic patients' data. *Artificial Intelligence in Medicine 41*(3), 251–262.

Iyer, A., S. Jeyalatha, and R. Sumbaly (2015). Diagnosis of diabetes using classification mining techniques. *CoRR abs/1502.03774*.

Jain, A. K. and J. Mao (1997). Guest editorial special issue on artificial neural networks and statistical pattern recognition. *IEEE Trans. Neural Netw. Learning Syst. 8*(1), 1–4.

Kahramanli, H. and N. Allahverdi (2008). Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications 35*(12), 82 – 89.

Khashei, M. and M. Bijari (2010). An artificial neural network model for timeseries forecasting. *Expert Syst. Appl. 37*(1), 479–489.

Koza, J. (1991). Evolving a computer program to generate random numbers using the genetic programming paradigm. In *Proceedings of the Fourth International Conference on Genetic Algorithms*. Morgan Kaufmann, La Jolla,CA.

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. The MIT Press.

Kumar, D. K. R. and V. (2014, 06). Analysis of feature selection algorithms on classification: A survey. *96*, 28–35.

Kumari, V. A. and R. Chitra (2013). Classification of diabetes disease using support vector machine.

Lekkas, S. and L. Mikhailov (2010). Evolving fuzzy medical diagnosis of pima indians diabetes and of dermatological diseases. *Artificial Intelligence in Medicine 50*(2), 117–126.

Luke, S. and L. Panait (2002). Lexicographic parsimony pressure. In *GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference, New York, USA, 9-13 July 2002*, pp. 829–836.

Mugambi, E. and A. Hunter (2003). Multi-objective genetic programming optimization of decision trees for classifying medical data. In V. Palade, R. Howlett, and L. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems*, Volume 2773 of *Lecture Notes in Computer Science*, pp. 293–299. Springer Berlin Heidelberg.

Muhammad Waqar Aslam, A. K. N. (2010, August). Detection of diabetes using genetic programming. In *European Signal Processing Conference*, Number 18, Aalborg, Denmark.

Munandar, T. A. and E. Winarko (2015, March). Article: Regional development classification model using decision tree approach. *International Journal of Computer Applications 114*(8), 28–33. Full text available.

Murtagh, F. and M. M. Farid (2001). Pattern classification. *J. Classification 18*(2), 273–275.

Parashar, A., K. Burse, and K. Rawat (2014). A comparative approach for pima indians diabetes diagnosis using lda-support vector machine and feed forward neural network. *International Journal of Advanced Research in Computer Science and Software Engineering 4*, 378–383.

Patil, B. M., R. C. Joshi, and D. Toshniwal (2010). Hybrid prediction model for type-2 diabetic patients. *Expert Syst. Appl. 37*(12), 8102–8108.

Polat, K., S. Güneş, and A. Arslan (2008, January). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert Syst. Appl. 34*(1), 482–487.

Pradhan, P. M. A., V. Tribhuvan, K. B. Jadhav, V. Chabukswar, and V. Dhobale (2012). A genetic programming approach for detection of diabetes.

Rahamneh, Z. A., M. Reyalat, A. F. Sheta, and S. Aljahdali (2010). Forecasting stock exchange using soft computing techniques. In *International Conference on Computer Systems and Applications, Hammamet, Tunisia, May 16-19, 2010*, pp. 1–5. IEEE.

Sapna, S. and A. Tamilarasi (2009). Fuzzy relational equation in preventing diabetic heart attack. In *International Conference on Advances in Recent Technologies in Communication and Computing*, pp. 635–637. IEEE Computer Society.

Sc, M. S. S. B., M. Phil, and P. D. Data mining fuzzy neural genetic algorithm in predicting diabetes.

Searson, D. P. (2015). GPTIPS 2: An open-source software platform for symbolic data mining. In *Handbook of Genetic Programming Applications*, pp. 551–573.

Searson, D. P., D. E. Leahy, and M. J. Willis (2010, 17-19 March). GPTIPS : An open source genetic programming toolbox for multigene symbolic regression. In *Proceedings of the International Multi-conference of Engineers and Computer Scientists 2010 (IMECS 2010)*, Volume 1, Hong Kong, pp. 77–80.

Selvaraj, R. S., K. Elampari, R. GAYATHRI, and S. J. JEYAKUMAR (2010). A neural network model for short term prediction of surface ozone at tropical city. *International Journal of Engineering Science and Technology 2*(10), 5306–5312.

Sheta, A. F., M. Braik, and H. Al-Hiary (2009). Identification and model predictive controller design of the tennessee eastman chemical process using ann. In H. R. Arabnia, D. de la Fuente, and J. A. Olivas (Eds.), ””, pp. 25–31. CSREA Press.

Sheta, A. F., H. Faris, and E. Öznergiz (2014, June). Improving production quality of a hot-rolling industrial process via genetic programming model. *Int. J. Comput. Appl. Technol. 49*(3/4), 239–250.

Silva, L. M., J. M. de Sá, and L. A. Alexandre (2008a). Data classification with multilayer perceptrons using a generalized error function. *Neural Networks 21*(9), 1302–1310.

Silva, L. M., J. M. de Sá, and L. A. Alexandre (2008b). Data classification with multilayer perceptrons using a generalized error function. *Neural Networks 21*(9), 1302–1310.

Su, C.-T., C.-H. Yang, K.-H. Hsu, and W.-K. Chiou (2006). Data mining for the diagnosis of type ii diabetes from three-dimensional body surface anthropometrical scanning data. *Computers & Mathematics with Applications 51*(6-7), 1075–1092.

Tennis, J. T. (2002). Three spheres of classification research: Emergence, encyclopedism, and ecology. In J.-E. Mai, C. Beghtol, J. Furner, and B. H. Kwasnik (Eds.), *Classification Research Workshop*. Information Today.

Tresp, V., T. Briegel, and J. Moody (1999). Neural-network models for the blood glucose metabolism of a diabetic. *IEEE Transactions on Neural Networks 10*(5), 1204–1213.

Tsirogiannis, G. L., D. S. Frossyniotis, K. S. Nikita, and A. Stafylopatis (2004). A meta-classifier approach for medical diagnosis. In G. A. Vouros and T. Panayiotopoulos (Eds.), *SETN*, Volume 3025 of *Lecture Notes in Computer Science*, pp. 154–163. Springer.

Tsypin, M. and H. Röder (2011, September 20). Method for reliable classification of samples in clinical diagnostics using an improved method of classification. US Patent 8,024,282.

Vapnik, V. (1998). *Statistical learning theory*. Wiley.

WHO (2016). World health organization.

Zecchin, C., A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli (2011). A new neural network approach for short-term glucose prediction using continuous glucose monitoring time-series and meal information. *Conf Proc IEEE Eng Med Biol Soc 2011*, 5653–6.

Zhang, G. P. (2001). An investigation of neural networks for linear time-series forecasting. *Computers & OR 28*(12), 1183–1202.

Zitar, R. A. and A. Al-Jabali (2005). Towards neural network model for insulin/glucose in diabetics-ii. *Informatica (Slovenia) 29*(2), 227–232.