A Plasma glucose prediction tool based on dietary assessment: a case of type 2 diabetes patient

Alex W. Njihia
Faculty of Information Technology (FIT)
Strathmore University

A Plasma Glucose Prediction Tool Based on Dietary Assessment: A Case of Type 2 Diabetes Patient

By

Alex Njihia

A Thesis Submitted in Partial Fulfillment of the Requirements for the ward of the Degree of Master of Science in Information Technology at Strathmore University

Faculty of Information Technology
Strathmore University
Nairobi, Kenya

May, 2018

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. The thesis contains no materials previously published or written by another person except where due reference is made in the thesis itself.
No part of this may be produced without the permission of the Author and Strathmore University.

Alex Wainaina Njihia

Signature: _____

Date: _____

050712

The thesis of **ALEX WAINAINA NJIHIA** was received and approved by the following:

Senior Lecturer, Faculty of Information Technology

Approval

Dr, Vincent Omwenga (PHD)

Supervisor

Dr.Humphrey Njogu,

Strathmore University

Dr. Joseph Orero,

Dean, Faculty of Information Technology

Strathmore University

Professor, Ruth Kiraka

Dean, School of Graduate studies,

Strathmore University

Abstract

Management and control of blood sugar using dietary intervention has for a long time been considered to be important. The caregivers have always advised diabetic patients to moderate the amount of carbohydrates intake. The approach here has always been reduction in the amount of carbohydrates, unfortunately this does not translate to the reduction on the blood sugar in some case. This is explained by the fact that what determines the sugar levels in the blood has to do more with the glycemic load of the carbohydrates consumed which is dependent on the glycemic index of the food item consumed. Though the amount of carbohydrates taken by the patient has a role to play, it is rather indirect. The study, sought to develop a tool for the computation of the glycemic load of the food item consumed by an individual by aggregating the various meals parameters. The tool has been developed by analyzing the dietary factors that affect the glycemic load and using these factors has the regressing variables. The algorithms used in the development of the dietary assessment tool have been used to map and mine the standard glycemic index of

individual food item and to estimate individual patient meal item glycemic using regression analysis approach. Experimental data results indicate the tool can compute the glycemic load of the food item which is comparable to the standard glycemic load values and it also gives plasma glucose prediction trajectories which mirrors those obtained from existing clinical trial dataset.

Table of content


VT OMNES
VNVM SINT

List of Figures

List of Tables

Abbreviations/ Acronyms and Definition of Terms

ADA: American Diabetes Association

BMI: Body Mass Index

CVD: Cardiovascular Disease

Diabetes mellitus: This is a disease considered to be a metabolic disorder and is associated with chronic hyperglycaemia or hypoglycaemia disturbances **Euglycemia:** Normal concentration of glucose in the blood

Glycemia: refers to the concentration of sugar or glucose in the blood

GI: Glycaemic Index

Hyperglycaemia: is a condition in which an excessive amount of glucose circulates in

Isocaloric: having the same or a similar calorific value

KDD: Knowledge Discovery Process

HDL: low-density lipoprotein the

blood plasma

LDL: low-density lipoprotein

RCTs: Randomized Controlled Trials

SVM: Support Vector Machine

WHO: World Health Org~~aniza~~tion

Chapter 1

Introduction

1.1 Background to the study

In 2002, the World Health Organization (WHO) estimated the worldwide prevalence of diabetes in adults to be around 173 million and predicted that there will be at least 350 million people with Type 2 diabetes by 2030. Most of these cases are reported in developing countries and the majority of new cases will originate from these areas. The global increase in the incidence of diabetes is related to high levels of obesity associated with a change from traditional diets, diminishing levels of physical activity, population ageing and increasing urbanization.

The greater burdens of diabetes incidences have been bone by people in the lower socio-economic levels in the society. This could be explained by the fact that they have limited access to health care services and limited nutrition options. In developing countries where resources are limited, the situation is more confounded. The greater part of the costs are related directly to treatment of the disease and its complications, and indirectly to loss of earning power in those affected.

World Health Organization (WHO) has defined diabetes mellitus as a metabolic disorder often noticeable by hyperglycaemia due to higher plasma glucose levels. Some of the causal factors may include excessive consumption of carbohydrate, unhealthy fat and protein metabolism, which may cause difficulties in the secretion of insulin, insulin inaction, or both. Some of the symptoms of the disease include vision impairment, loss of weight, polyuria among others.

There are different categories of diabetes mellitus. They include:

1. Type 1 Diabetes Mellitus

This disease is found among the kids. It occurs abruptly and is characterized by ketoacidosis. Type 1 patients are in most case insulin dependent.

2. Type 2 Diabetes Mellitus

This is the most common disease and is associated with adults. Some of the predisposing factors include obesity, inactive life-style, poor eating habits and familiar history among others.

3. Gestational Diabetes

It is a kind of diabetic condition that develops when there is a higher level of glucose intolerance during pregnancy.

It is argued that a change on the lifestyle during the state of impaired glucose tolerance (commonly described as pre-diabetes) will significantly reduce the risk factor of developing diabetic disease by over 60%. Similarly, a change in the lifestyle is considered critical in the management and control of diabetes. A review of one's eating behavior and dietary intake is one of the lifestyle audits that may be considered. Therefore, dietary management is considered key in this by tracking the

nutritional needs of the affected persons over a period of time. Keeping profile of the diet of an individual over a long period of time is not easy. But using data mining tools that have been considered to hold great potential for the healthcare industry can do this.

Data mining is the computational methodology of finding examples in substantial information sets inclu~~ding~~ routines (Balaji 2013) and ~~(Pang et~~ al 2005). Beside the crude investigation step, it includes database and information administration perspectives, information pre-processing, model and induction contemplations, interestingness measurements, intricacy contemplations, post-preparing of found structures and visualization. Moreover, recent advances in technology have resulted in electronic data to be captured, processed, analyzed and stored rather inexpensively. The large supplies of powerful and affordable computers have increased in the past decades due to the amazing progress in computer hardware technology. This technology provides a great boost to the database and information industry, and makes a huge number of databases and information repositories available for transaction management, information retrieval, and data analysis. The explosion and growth of data has also led to the creation of the need for the development of techniques that extract knowledge and information intelligently. The systematic development of data mining tools to turn otherwise neglected data into golden nuggets of knowledge has become an absolute necessity.

Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, scientific and medical research. The main objective in data mining is the extraction of examples and learning from huge measure of information, not the extraction of information itself. It is likewise a popular expression and is regularly connected to any type of extensive scale information or data handling (gathering, extraction, warehousing, investigation, and insights) and additionally any use of PC choice emotionally supportive network, including computerized reasoning, machine learning, and business knowledge. The real information-mining errand is the programmed or selfloader examination of

substantial amounts of information to concentrate already obscure fascinating examples, for example, gatherings of information records, uncommon records (oddity recognition) and conditions.

1.2 Problem statement

In recent years, there has been an increase in reported incidences of type 2 diabetes especially in developing and newly industrialized countries. The reported cases are often attributed to issues around urbanization, changing working patterns, that is, transition from heavy labour to inactive life styles, change in eating habits, change in food regimes and food products. The enhancement in technology associated with development implies that more processed food products are consumed across the society and with improved distribution chain of the production stream, more processed food is now available to a larger populace of the society. These processed food are often associated with high calorific contents, highly refined carbohydrates, large amounts of added sugars and unhealthy fats. Thus, more and more unhealthy foodstuffs have displaced fresh food items and are generally available across the food stream. It is therefore considered that food ingested is one of the single most critical factors in the upsurge of the diabetes incidence.

Although much has been learnt about the role of various dietary factors in the management and control of diabetes, further studies are warranted to examine synergistic effects of individual components of various dietary patterns and to understand the biological mechanisms underlying the observed associations with respect to diabetes management.

Ezzati et al (2013) notes that a large percentage of the population in the globe have gained weight in the recent past due to changes in dietary patterns and decreased physical activities. These two are predisposition factors for one to develop diabetes. To overcome, this inevitable trend, Knowler et al (2002), argues that interventions around restricted calorific intake and involvement in physical exercise are key. They consider by doing this, the patient is likely to have tight control of the plasma glucose.

Unfortunately, in the developing countries, it is somewhat difficult to measure and even understand the correct calorific needs of an individual.

As a first step to solving this challenge, this study aims at developing a predictive tool that can be used to estimate likely ups of lows of the plasma glucose using the glycemic load of the ingested food. This study has utilized predictive algorithms in data mining to understand the underlying dietary patterns in order to predict the plasma glucose directions given the glycemic index of the food item consumed.

1.3Objectives General objective

The aim of the study was to develop a predictive tool for plasma glucose projections using dietary patterns for a type 2 diabetes patient.

Specific objectives

1. To analyze the dietary factors for management of type 2 diabetes
2. To analyze the various approaches used for data mining in health
3. To develop a predictive algorithm for associating meals with the glycogenic levels in diabetic patients
4. To evaluate predicting power of the proposed system in determining the dietary needs of a diabetic patient

1.4Research Questions

1 What are the dietary factors that influence management of type 2 diabetes?

2 What are the approaches used for data mining in health?

3 How would the dietary factors influencing management of type 2 diabetes do be used to predict the glycogenic levels in diabetic patients?

4 How does the predictive algorithm performance relate to dietary needs of a diabetic patient?

4.1Justification

By applying such a tailored algorithm to the data, the caregivers will be able to pinpoint which patients' dietary needs. Importantly, the caregivers will integrate this

insight into its workflow with a simple ranking of priority patients. A personalized nutritional intervention for the type 2 diabetes can be instituted.

Other benefits likely to come from this study include: improved early identification of and intervention for diabetes and pre-diabetes, improved patient engagement with self-management, education, and routine follow-up and improved use of appropriate medication therapy for patients and increased rates of compliance to therapy

4.1Scope

The study will focus on type 2 diabetes and more specifically on dietary intervention on diabetes management. It will also be limited to exogenous factors on the effects of dietary intervention.

Chapter 2 Literature Review

2.1 Introduction

Management and control of plasma glucose in a diabetes patient needs an allinclusive approach to avoid complications associated with the disease. The commonly used approach has been through nutritional management. Nutritional therapy has been in use for many years by both pre-diabetic and diabetic patients. The nutritional management has always been following a predetermined and prescriptive nutritional regime by a patient. This regime are developed by a dietician or nutritionist based on the patient dietary assessment. Stephenson et al. (2014), notes that a good balance between the nutritional management and other therapies like use of medication will give a good outcome for a type 1 diabetic patient.

2.2 Dietary interventions in diabetes management

Dietary intervention in the management of diabetes can be viewed from two perspectives; prevention and control. Prevention will focus more on the populace predisposed of developing diabetes due to the kind of dietary they are mostly exposed to. Control will focus more on the populace already diagnosed with diabetic condition. The aim at this stage is to control the sugar level.

There are three categories of prevention.

1. Basic/Primary prevention: The aim at this stage is to identify the group at the greatest risk of developing diabetes. This is a group characterized by BMI>25 or obese or pre-diabetic cases. The target is to institute complete eating/dietary lifestyle change.
2. Secondary prevention: The focus here is to use controlled dietary controls to manage the plasma glucose levels. The target at this category are those patients diagnosed diabetic condition.
3. Tertiary prevention: At this level, the focus is to use nutrition to manage diabetic complications.

To understand the nutritional needs of a patient; food is grouped according to macronutrients and micronutrients. A mix of the two is vital in providing the required nutritional values to the patients. It is critical to note that different populace will require different dietary mix and therefore dietary requirements for an individual needs to be determined. It generally accepted that reduction in fats more so saturated fats ingested by diabetes is critical in decreasing heart related complications (Fuller et al. 2015).

It is observed that dietary rich in low carbohydrates (20-120 g/d) and low fat have been used for weight loss. A close monitoring the lipid profile and renal function tests is needed. Dietary rich in low carbohydrates are good in reducing the values of fasting plasma glucose by about 21-28 mg/dL (Liu et al 2015), (Gannon and Nuttall 2004).

2.3Dietary factors for the management of diabetes a. Macronutrients

It is noted that a mix macronutrient will produce the desired results in the management of diabetes. It is however, difficult to obtain the optimal mixing ratio of the macronutrients. An appropriate mix may be obtained by having a variety of macronutrients having proteins, good fats and carbohydrates. This is depended on the individual patient (Institute of Medicine, 2002). A report by DRI recommends the following ratios for a healthy adults should consume 45–65% of total energy from carbohydrate, 20–35% from fat, and 10–35% from protein. The aim of any kix is to achieve a total caloric intake that is good for weight control. The range given by DRI

can be adjusted to fit individual groups. The American Diabetes Association (ADA) recommends tailoring macronutrient requirements considering the metabolic goals, preferences and eating patterns of an individual (Parker et al.
2002).

b. Quality and quantity of carbohydrates

A dietary pattern that includes carbohydrate from fruits, vegetables, whole grains, legumes, and low-fat milk is encouraged for good health. It is also important to note that low-carbohydrate diets, restricting total carbohydrate to <130 g/day, are not recommended in the management of diabetes. Monitoring carbohydrate, whether by carbohydrate counting, exchanges, or experienced-based estimation remains a key strategy in achieving glycemic control. The use of glycemic index (GI) and glycemic load (GL) may provide a modest additional benefit over that observed when total carbohydrate is considered alone. This study will focus on associating GI to dietary needs of diabetic patients.

The glycemic index (GI) of foods was developed to compare the postprandial responses to constant amounts of different carbohydrate-containing foods (Jenkins et al., 1981). Glycemic index for a food item is measured as the amount of an increase of the plasma glucose area above the fasting levels within 2 hours postprandial (that is, incremental area under the curve (iAUC)) (after ingesting food) of constant quantity of food (50g of carbohydrates) divided by a reference food (glucose or white bread). This is given by:

$$GI = ((iAUC\ test\ food // (iAUC\ reference\ food)) * 100))$$

The other measure on the quality of food is the glycemic load (GL) of the food item. It is computed as:

$$Gl = \left(GI * (meal\ item\ carbohydrates\ size\ in\ grams / 100)\right)$$

A higher GL depicts food that is likely to cause higher plasma glucose. And it is therefore important to select food with low GL.

Mayer et al. (2006) notes that food rich in fiber, fructose and good fats are dietary that components that may lower the glycemic levels. It is also good to select food ith lower glycemic indexes.

In general, the amount of carbohydrate ingested is usually the primary determinant of postprandial response, but the type of carbohydrate also affects this response. Intrinsic variables that influence the effect of carbohydrate-containing foods on blood glucose response include the specific type of food ingested, type of starch (amylose versus amylopectin), style of preparation (cooking method and time, amount of heat or moisture used), ripeness, and degree of processing. Extrinsic variables that may influence glucose response include fasting or preprandial blood glucose level, macronutrient distribution of the meal in which the food is consumed, available insulin, and degree of insulin resistance.

c. Protein

The dietary intake of protein for individuals with diabetes is similar to that of the general public and usually does not exceed 20% of energy intake. A number of studies in healthy individuals and in individuals with type 2 diabetes have demonstrated that glucose produced from ingested protein does not increase plasma glucose concentration but does produce increases in serum insulin responses (Franz et at. 2002, Gannon et al. 2001). Abnormalities in protein metabolism may be caused by insulin deficiency and insulin resistance; however, these are usually corrected with good blood glucose control (Gougeon et at. 2000).

Small, short-term studies in diabetes suggest that diets with protein content >20% of total energy reduce glucose and insulin concentrations, reduce appetite, and increase satiety (Gannon and Nuttall 2004, Gannon and Nuttall 2003). However, the effects of high-protein diets on long-term regulation of energy intake, satiety, weight, and the ability of individuals to follow such diets long term have not been adequately studied.

Parker et al (2002) and Hu et al (2001) have observed that most nutrition recommendations do not restrict protein intake for adults with type 2 diabetes. This is aimed at avoiding experiencing protein inadequacies by the patient who is on energy-reduced diets for weight loss (Hu et al. 2001).

d. Fat

The type of fat ingested by a person is important as compared to the amount of fat consumed. This is because some of the fats consumed are good in supporting metabolic functions (Parker et al. 2002, Kris-Etherton 2004). Therefore, the focus should be on limiting the amount of bad fats consumed. For instance, diabetic patient should avoid saturated fatty acids, trans-fatty acids and fats having cholesterol. (Summers et al. 2002, Parker et al. 2002).

In non-diabetic individuals, reducing saturated and Trans-fatty acids and cholesterol intakes decreases plasma total and LDL cholesterol. Reducing saturated fatty acids may also reduce HDL cholesterol. Importantly, the ratio of LDL cholesterol to HDL cholesterol is not adversely affected. Studies in individuals with diabetes demonstrating the effects of specific percentages of dietary saturated and *trans*-fatty acids and specific amounts of dietary cholesterol on plasma lipids are not available. Therefore, because of a lack of specific information, it is recommended that the dietary goals for individuals with diabetes be the same as for individuals with preexisting CVD, since the two groups appear to have equivalent cardiovascular risk. Thus, saturated fatty acids <7% of total energy, minimal intake of Trans- fatty acids, and cholesterol intake <200 mg daily are recommended.

e. Dietary patterns

Combinations of different foods and food groups have proven to be helpful in the management of diabetes. More so when the different food combinations are taken based on the dietician recommendations with consideration of the metabolic needs of an individual (Parker et al. 2002, Hu et al. 2001).

To achieve maximum effect from the food taken, it is recommended to have a pattern that incorporates a mix of food that has low glycemic index, low in carbohydrates and rich in proteins patients (Wasserman and Zinman 1994).

f. Vitamin and mineral supplementation

Deficiencies in the micronutrients have been associated to cases of uncontrolled diabetes (Mooradian, 1999). It is therefore important for diabetic patients to be aware of the daily micronutrients requirements. In developing countries the source of such micronutrients, that is, vitamins and mineral needs are food items consumed. Unfortunately, the populace do not know which food items will provide the micronutrients they need. The common mode of providing the populace with nutritional information is through the use of nutritional counseling clinics that are barely enough.

g. Fiber

People with diabetes are encouraged to choose a variety of fiber-containing foods such as legumes, fiber-rich cereals (≥5 g fiber/serving), fruits, vegetables, and whole grain products because they provide vitamins, minerals, and other substances important for good health. Moreover, there are data suggesting that consuming a high-fiber diet (~50 g fiber/day) reduces glycemia in subjects with type 1 diabetes and glycemia, hyperinsulinemia, and lipemia in subjects with type 2 diabetes (Franz et al. 2002). Palatability, limited food choices, and gastrointestinal side effects are potential barriers to achieving such high-fiber intakes. However, increased fiber intake appears to be desirable for people with diabetes, and a first priority might be to encourage them to achieve the fiber intake goals set for the general population of 14 g/1,000 kcal (Institute of Medicine).

h. Sweeteners

Substantial evidence from various clinical studies demonstrates that dietary sucrose does not increase glycemia more than isocaloric amounts of starch (Franz et al. 2002). Diabetes patients though need to consume food items rich in sucrose in moderation

without fear of hyperglycemia episodes. It is also important to take food which contain sucrose to avoid excessive energy intake.

2.4Approaches used for data mining in health

Data mining has been considered to be the process of the extraction of the hidden pattern from the data available. Data mining approaches are broadly categorized into descriptive and predictive models. Application of data mining techniques in health sub-sector focuses on summarizing, analyzing association, data classification, clustering data, trends generation, and generation of regressions analysis.

(i) Summarizing data: This involves the abstraction of the data results into smaller dataset. Through summarization different patterns from the data sets may be inferred (Gheware et al 2014).

(ii) Association analysis: This is aimed at establishing how key data objects are interconnected in a large database. This is done by defining the association rules for the database objects. The aim is to understand the correlations that may exist between the interacting data objects.

(iii) Data classification: It involves dividing data sets into defined classes using classification techniques. The defined classes are usually defined through the data instances. Classifiers are used to achieve the objectives of data classification. The classifier is trained using the training data set and then tested using the testing dataset. The concept of data classification has been applied in healthcare to predict cost of treatment of diseases and also in mining different healthcare issues (Divya and Sonali, 2013 Beller and Nucl, 2008).

Various classification approaches (Soni, Ansari and Sharma, 2011) have been implemented in data mining process. These approaches have been used to divide the data into different sets so that easily relation between different attributes can be identified. Different data mining techniques have been used to help health care professionals in the diagnosis of Diabetes disease (Mary and Mat 2004). Those most frequently used focus on classification for instance; naïve Bayes decision tree, and

neural network (Hian and Gerald, 2005). This study will focus on classification task. *(iv) Clustering of data*: Clustering of data involves the grouping of similar data elements into the same group called a cluster (Pang-Ning, Michael and Vipin 2002). It starts by first forming the clusters and then assigning instances to it (Gosian and Kumar 2009). It therefore does not require any data to partition the data.

(v) Trend generation: This involves the generation of graphs from time dependent data. Trends are used to show the patterns in a data set.

(vi) Regression analysis: It involves the formation of a regression function which is used to map data into prediction values (Dunham 2002). It is aso used to estbalish relationships between dependent and independent variables.

2.4.1 Data mining Techniques in Health

A number of data mining techniques have been used in health. Some of the techniques commonly used include:

a. Neural Networks

The neural networks are designed to mimic the human brain. In medicine, the neural networks have been used to develop decision support systems. The neural networks are constructed based on very many artificial neurons that are interconnected. These neurons are assigned specific weights or roles. The assigned weights are used to control errors by regularly adjusting them (Parvez, Saqib and Syed, 2015). Due to the interconnections among the neurons, they are expected to work together to produce the desired output through the output function. For a long time the neural networks have been the standard classification (Obenshian, 2004). Neural networks have also been used for pattern recognition and classification (Dunham, 2002). Due their non-linear attributes, they are considered adaptable unlike the ordinary modeling approaches (Shameem et al., 2011). Neural networks in medicine have been used in the classification of tissues, prediction of diseases and guiding in drugs production. Rani (2011) notes that neural networks have been used in the prediction of heart diseases.

Some architectures of neural networks include Multi-layered neural network (MLNN) and Polynomial neural network (PNN). The MLNN is used in solving classification challenges associated with non-linear data sets (Haykin 1999).

b. Decision Tree

Decision trees consist of non-terminal and terminal nodes. A node is made of a test condition applied on the data set. They are used to classify object instances by categorizing them into different nodes starting from the non-terminal nodes (Emina and Abdulhamit 2011). The terminal nodes in a tree are called the leaves.

What is assigned to a branch of a tree is depended on the test condition on the nonterminal node. For instance, readmission decision of a patient can be determined using the decision tree on evaluating a set of conditions which decides whether to readmit or not (Parvez, Sqib and Syed, 2015). A decision tree provides a visual diagrammatic representation of various options available depending of the test condition in a non-terminal node (Anto and Chandramathi 2011).

Other application of decision trees are in operation research analysis for computing conditional probabilities (Goharian and Grossman 2003). Apte and Weiss (1997) notes that selection of the best alternatives can be chosen with the help of decision trees and based on maximum information gained by traversing the tree from root to leaf node. Gayathri, Mona and Chitra (2014) notes other applications of decision trees with enormous potential to be in data mining for marketing.

c. Fuzzy Sets

The concept of Fuzzy sets theory was founded by Zadeh in 1998 to handled nonstructured or vague dataset. It is commonly used in mining data having a lot of uncertainties, imperfections or noisy data (Shameem et at. 2011).

Many expert systems have been build using the concepts of fuzzy sets and fuzzy logic. Fuzzy logic are used in the expert system to compute probabilities of a test case given conditions for a cluster where the test case falls (Kalyani and Samayita, 2013).

d. Support Vector Machine (SVM)

SVM was used for the first time by (Vapnik 1998a 1998b). It is one of the classification techniques that is based on the statistical learning theory (Vapnik 1998b, Cristianini and Shawe-Taylor, 1995). It has majorly been used by a number of kernels for universal approximation (Hammer and Gersmann, 2003).

In performing classification it uses as subset of the learning dataset called support vector to define a support vector machines. The training data is represented using the SVM model using the support vector. By using the training dataset a summarised dataset can be extracted (Vapnik, 2005). It has been extensively used in the classification of dataset which requires binary classification (Cristianini and ShaweTaylor, 2000). Some of the notable binary classification where SVM has been used in health is in the classification and predicting of the ICU mortality.

SVM can be implemented using two different approaches. One of the approach involves the mathematical programming and the other approach uses the kernel functions. The main focus in each of these methods is to apply the concept of creation of hyper-planes and then trying to minimize the separation of the data points (Parvez 2015). The use of the hyper-plane in SVM has been noted to help in reducing the effect of ~~noisy~~ data points in a dataset.

SVM are also categorized to be linear or non-linear. Linear SVM provides a separation of the data points using the linear decision boundary (Kumar et at 2014). Whereas the non-linear SVM uses the non-linear decision boundary to separate the data points (Kumar et at 2014). The non-Linear SVM is considered to be the most powerful algorithm that can be used to obtain maximal generalization and it gives accurate results compared to other algorithms. In health non-linear SVM has been used in the classification of the heart beats and heart valve classifications (Argyro et al., 1995 and Ilias et al., 2009).

e. Bayesian Networks

The Bayesian network uses the concept of conditional probability and is used to represent information/knowledge which contains uncertainties. It is a probabilistic model. An outcome in a Bayesian network is denoted by a node which represent an event variable. The edges of the network denotes the probability associated with an outcome of the event variable (Friedman and Goldszidt, 1997, Finn, 1996). The given probabilities on the Bayesian network will denote the realization or success and therefore a probability associated by a non-realization or failure can be computed. Therefore, the Bayesian network can provide two separate information about an event variable (Sebe et al. 2005).

2.4.2 Data mining and Knowledge Discovery Process (KDD)

In the medical field and in general the healthcare, the utilization of the concept of data mining has been slow. But in the recent past a number of application have been developed employing data mining techniques that are used by medical practitioners. For instance, data mining tools have been used to detect fraud associated with health insurance claims. Other areas where the power of data mining has been used in health include prediction of the correct treatment method for given medical condition and creating a profile of causes of a disease. Data mining has thus provided a broader base for knowledge discovery in medical field.

Knowledge Discovery in Databases (KDD) (Usama and Gregory, 1996) is an iterative process that transforms raw data into useful information. Different steps of KDD are:
Understanding: The first step is to understand requirements. This is necessary to have a clear understanding about the concept of application domain and objectives, whether it is to improve sales, predict stock market etc. Also know whether going to describe data or predict information.

Selection of data set: Data mining is done on current or past records. Select a data set or subset of data, in other words data samples, on which need to perform data analysis and get useful knowledge. Need large data set or we can say that ample amount of data to perform data mining.

Data Cleaning: Data cleaning is the step where noise and irrelevant data are removed from the large data set. This is a very compulsory step because output would be based on the quality of selected data. As part of data cleaning, might have to remove redundant records, enter logically correct values for empty or null records, remove extra data fields, standardize data format, and update data in a timely manner and so on.

Data Transformation: With the help of dimensionality reduction or transformation methods, data is transformed into appropriate form making it ready for data mining step.

Selection of data mining algorithm: Appropriate method is to be selected for looking for patterns from the data. Need to decide the model and parameters that might be appropriate for the method. Some popular data mining methods are decision trees and rules, relational learning models, example based methods etc.

Data mining: Data mining is the actual search for patterns from the data available using the selected data mining method.

Pattern evaluation: This is a post processing step in KDD which interprets mined patterns and relationships. If the pattern evaluated is ~~not use~~ful, then the process might again start from any of the previous steps, thus making KDD an iterative process.

Knowledge presentation: This is the final step in KDD. The knowledge discovered is consolidated and represented to the user in a simple and easy to understand format. Most of the times visualization techniques are being used to make information understandable by the users as well as by the interpreters.

Before information mining calculations can be utilized, target information set must be gathered. As information mining can just reveal examples really introduce in the information, the target information set must be sufficiently substantial to contain these examples while staying sufficiently succinct to be mined inside an adequate time limit. A typical hotspot for information is an information bazaar or information

distribution center. Pre-processing is crucial to investigate the multivariate information sets before information mining. The target set is then refined. Information cleaning uproots the perceptions containing clamor and those with missing information.

2.5A Predictive tool using classification approaches in Data Mining

Of the various data analysis technique, classification is a supervised machine learning technique which makes predictions about the future class instances by mapping instances of testing data to the predefined class labels which is learnt from the supplied instances of classes with class labels. There are several models in classifications such as probabilistic model, evolutionary algorithmic model etc.

Classification consists of predicting a certain outcome based on a given input (Vargas, Yu and Jiang, 2007). In order to predict the outcome, the algorithm process a training set, containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes, which would predict the outcome. In algorithm, a given data set which is not seen before is called prediction set, which contains the same set of attributes, except the prediction attribute, which are not yet known. Algorithm analysis the input, and produces a prediction. The accuracy of prediction is defined by the goodness of the algorithm used. Classification technique is capable of processing a wider variety of data, than regression and it is growing its popularity (Pooja, Jayanthy and Kliwad, 2005).

The objective of a classifier is not to explore the data to discover interesting segments, but rather to decide how new records should be classified (Namrata, 2012). Classification routines in data mining also use a variety of algorithms.

Steps in classification models development

1. The first step is identifying a set of subjects with a known behavior. All the inputs are known, as well as the target classes.

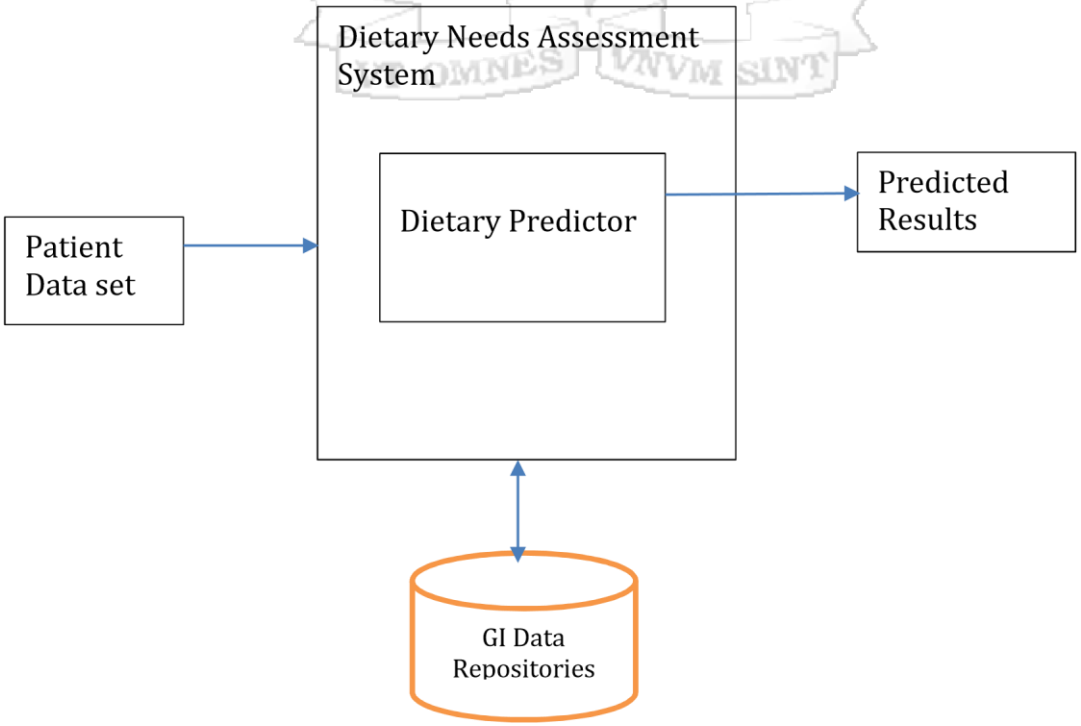
2. The second step is data preparation – it includes data cleaning, feature selections, data transformation.
3. The third step is training the model. This process uses about 60% of all the subjects identified in the first step to identify the relationships between the inputs and the target data. These relationships are dependent on the classification algorithm used. In this study hierarchical clustering algorithm was used.
4. The fourth step is testing the model. This step uses the remaining 40% of subjects to test the relationships identified in the previous step. This is where the accuracy of the model is checked.

5.

The refined model is then used to classify the data for storage in the database which is then used to classify the glycemic load levels.

2.6 Proposed System

System design is the process of defining system architecture, modules and interfaces for the proposed system to satisfy specified requirements. The System architecture is shown in the figure 2.1.



The fifth step is tweaking and refining the model – eliminate over fitting, regularization, validation, etc.

Figure 2. 1 A Framework for a Plasma Glucose Prediction Tool

The GI Data set is given as input to the Dietary Assessment and Awareness system. The dietary predictor helps in prediction of possible contribution of dietary to plasma glucose levels based on the GI provided.

Chapter 3
Methodology

3.1Introduction

Research Methodology is defined as the process of systematically solving problems. It can be considered as the science of doing research (Bhatnagar & Singh, 2013). The research was guided by the objectives the author proposed to meet at the end of the research. It was also guided by the nature of the problem being studied, research designs that have been used in the related work reviewed in chapter 2. Initial statistical analyses were done to determine how the data fits into the developed model to give a predicted outcome. A classifier for categorizing the data was identified and the validity of the classifier was determined. The plasma glucose prediction tool was then developed by employing Structured system analysis and Design together Object-oriented systems analysis and design approaches.

3.2 Research Design

The study is quasi-experimental in nature due to the processes involved that entails data/variable manipulation to give an outcome. The research employed an exploratory approach ~~with~~h the use of test-bed impl~~ementin~~g Machine Learning. Collected data was u~~sed to verify the existing theory, and the~~ research questions that were formulated at the beginning of the research.

In this study a purposed system data mining approach was done in different phases by dividing data set into different training and testing dataset. These different phases are explained as:

Pre-processing: In this phase dataset was pre-processed for the removal of empty sets and anomalies available in the dataset. Data cleaning was done to remove the observations containing noise and those with missing data.

Feature Extraction: Data mining approaches was implemented in the data set to develop the rule that could be implemented on the dataset to identify the hidden pattern from the dataset that was utilized on the time of classification. Data mining techniques are operated on large volumes of data to explore hidden patterns and relationships that are further helpful in decision making (Usama et al., 1996).

Parameter Evaluation: After extraction of the features from the dataset different results were evaluated by using training and testing samples. On the basis of these samples various parameters were evaluated for performance evaluation and validation of the proposed work.

3.3 Agile Development Systems Methodology

Agile development refers to the various iterative and incremental software development methodologies which incorporates methods like scrum, crystal, Lean development and Extreme Programming.

The adopted approach in this study incorporate iteration to enable continued feedback during development so as to successively refine and deliver on a software system. The iterations involved continuous planning, continuous testing, continuous integration and continuous evolution. The diagram below visualizes the processes to be taken to achieve the set objectives.

Agile methods appropriately was also suited in this study due to its ability to create and respond to change. Moreover, it encourages participation of users enabling creation of what users want leading to creation of better solutions. Figure 3.1 exhibits the stages of agile development that were followed to attain the set objectives for this research. These phases include planning, requirements analysis, design, building and testing.

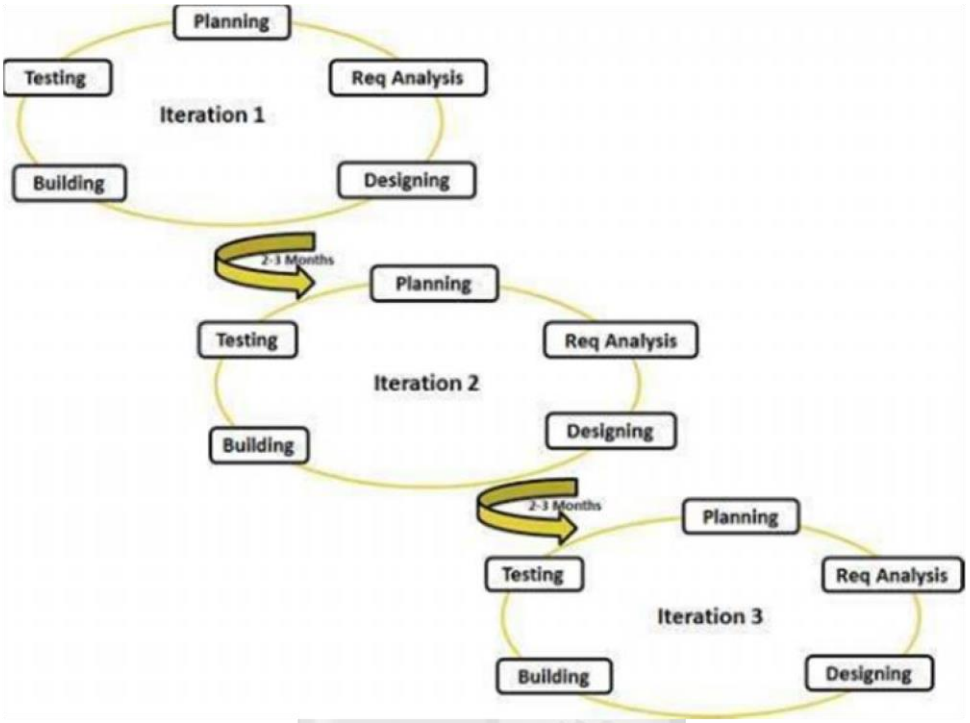


Figure 3. 1 Agile Software Development Methodology (Steljes, 2012)

3.4 Target population and sampling strategy

The target population in this study was the population that is at risk of creeping into type 2 diabetic state. This population is characterized by normal sugar levels to slightly elevated sugar levels (pre-diabetic cases). These is largely adults who have sedentary lifestyle that is considered to be the pre-causal for diabetes. Purposeful sampling was therefore key in this study as it helped the researcher identify the ideal group to participate in this study.

3.5 Data and Data Collection Methods

Largely data that was used in this study was secondary data and therefore secondary data collection instruments and methods were utilized. Data used in this study was sourced from existing databases and dataset. More specifically, data on GI of food was sourced from *International Table of Glycemic Index and Glycemic load value* (Foster-Powell et al. 2002) (see appendix A).

3.6 Data Analysis and model development

The main data analysis approach used in this study was classification techniques. Classification of the attributes and labels were done using a statistical analysis methods. The advantage of this method is that there is no bias in the data that is obtained. Classification is one kind of predictive modeling (Jian and Yongheng 2010, Pooja et al., 2011). More specifically, classification is the process of assigning new objects to predefined categories or classes: given a set of labeled records, build a model such as a decision tree (Yu et al., 2005), and predict labels for future unlabeled records. Model building in the classification process is a supervised learning problem. Training examples were described in terms of (1) attributes, which were categorical (i.e. unordered symbolic value or numeric) and (2) class label, which is also called the predicted or output attribute. The training examples were processed using some machine learning algorithm more specifically the decision tree algorithms to predict labels of new data.

Waikato Environment for Knowledge Analysis (WEKA) which is an open source tool for data mining was used. The researcher leveraged on WEKA's ability/tools for data pre-processing, classification, regression, clustering, association rules, and visualization. The application component of the model was developed using Oracle Application Express (~~APEX~~) which is an oracle applicat~~ion devel~~opment platform.

3.7 System analysis and Design

3.7.1 System Analysis

The system analysis involved the system requirements analysis and the representation of the identified system requirements using the various system

analysis tools. The method tool used in requirements gathering was document review approach. The system requirements were categorized into functional and non-functional requirements. The system analysis approaches used in this study adopted the UML- based specification. The UML syntax provides flexibility of the representing system components, and it applies various approaches to the synthesis of models (Giese, 2018).

3.7.2 Designing Phase

The approaches used in system design were largely object-oriented and dataoriented. Unlike data-oriented and process-oriented approaches which put weight on data and process respectively, object-oriented approach combines both processes and data into one entity called object (Joshi, 2013). Because of the above reason, object-oriented design (OOD) approaches was used to fine-tune the object requirements definition pinpointed in requirements analysis and to determine design specific objects.

System design aims at defining the architecture, components, modules, interfaces and data for a system to satisfy specified requirements (Waldo, 2006). The study made use of UML (unified modeling language), sequence diagrams, use-case and activity diagrams to design the system.

Use Case Diagram

Use Cases were used to pin point and divide the system actors and processes. Various manufacturers, distributors, retailers and customers interacting to make the LPG tracing and authentication solution a success were the actors of the proposed system. The Use Case was going to be depicted as text that outlines the effect the user was going to have on the system (Malan, 2001).

System Sequence Diagram

The System Sequence ~~Diagram was used to breakdown all the Use-~~Cases and summarize them. It illustrated the order of events that an external actor was going to generate, how those actions followed each other and any other interaction with external systems (Zhang, 2007).

Data flow Diagram

The data flow diagram was used to depict the movement of data from one system entity to the next and on how data was stored in the various data points.

Entity Relationship Diagram (ERD)

An Entity Relationship Diagram was used during database design. Entities, attributes and relationships between entities were identified and represented in the (E-R) Diagram.

3.7.3 Prototype, Testing and Validation

This phase focused on the system implementation. At this stage the aim is ensure the needed functionalities are working as required with regards to the set objectives. Various testing approaches were used. Unit testing was used to test each module, Integration testing for the integrated modules, System testing involved the entire system, Audit testing ensured there were no errors and finally Load testing was done to measure the amount of time the application took to process a request. Compatibility test was carried out to make sure that the web and mobile applications are compatible with existing platforms. Validation testing was done to check on if the developed application responded to the set targets and expectations.

3.8 Research Quality

The research quality was founded on the following aspects; accuracy, precision, recall ratio and F-measures.

Accuracy

The accuracy (AC) is defined as the proportion of the total number of predictions that were correct. tp represents the true positive, tn represents the true negative, fp represents the false positive and fn represents the false negative in the equations that were used to measure performance. Equation 3.1 will be used to determine accuracy

Accuracy (AC) = $\frac{tp+tn}{tp+tn+fp+fn}$ *(3.1)*

Recall ratio

The recall or true positive rate (TP) is the proportion of positive cases that were correctly identified as shown in the equation 3.2

Recall ratio = $\frac{tn}{tn+fp}$ (3.2)

Precision

Precision (P) is the proportion of the predicted positive cases that were correct, as computed using the equation, 3.3

Precision (P) = $\frac{fp}{fn+fp}$ (3.3)

F_Measure

The F-measure computes some average of the information retrieval precision and recall metrics.

3.9Ethical Considerations

The researcher ensured use anonymous data as obtained from reliable databases and datasets to ensure confidentiality. Since, the outcome of the study touches on human life, the accuracy of the outcome from the study were tested for plausibility through existing theories.

Chapter 4
System Design and Architecture

4.1Introduction

The aim of the study was to develop an application for the prediction of possible glucose levels categorization by analyzing the glycemic load levels of the food item

consumed by an individual. This chapter presents the system analysis and system design models that were adopted in the development of the application.

4.2System Requirements analysis

The system requirements analysis provides the basic building blocks for a system. They are categorized into two; functional and non-functional requirements.

Secondary data was used in this study and was specifically obtained from the *International Table of Glycemic Index and Glycemic load value* (Foster-Powell et al. 2002). The data obtained was grouped into 6 food categories. Figure 4.1 gives the categorization of the food items based the EM clustering algorithm in WEKA.

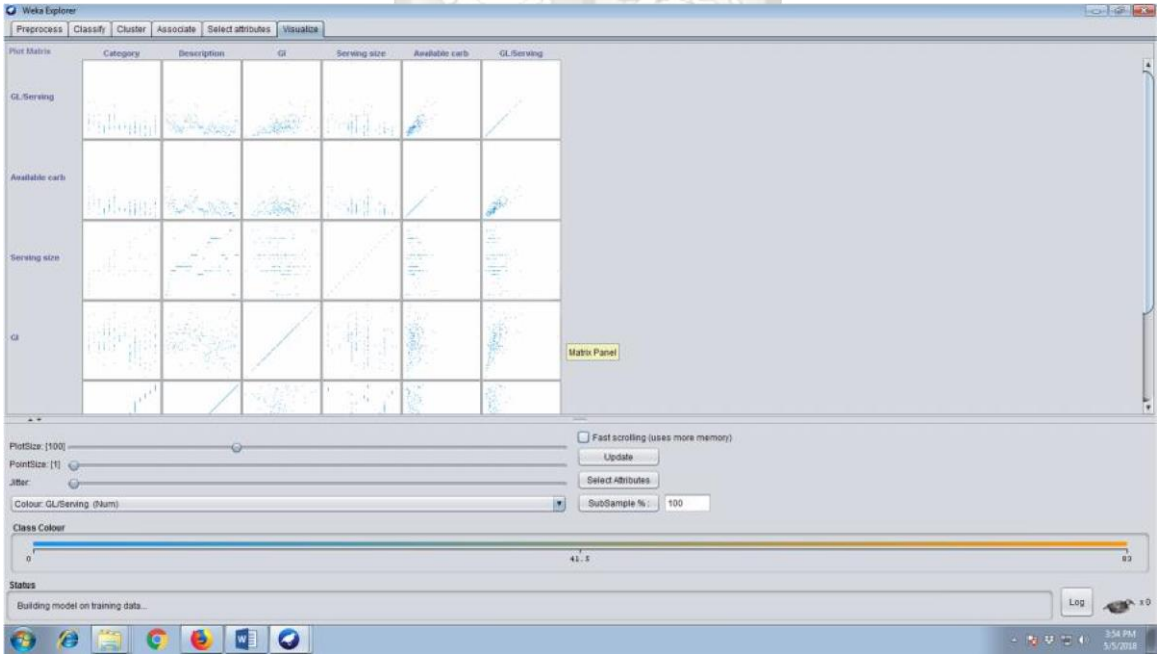


Figure 4. 1: Run information

The clustering results

Instances used in the clustering: 158
Attributes used in the clustering: 6

Category
Description
GI
Serving size
Available carb
GL/Serving

Test mode: split 66.0% train, remainder test

Classifier model (full training set)

Dictionary size: 12
The independent frequency of a class

Cakes	11.0
Beverages	5.0
Juices	8.0
Breads	18.0
Breakfast cereals	17.0
Cereal Grains	16.0
Cookies	8.0
Crackers	6.0
Dairy Products	17.0
Fruit and Fruit products	25.0
Legumes and Nuts	9.0
Snack foods and confectionary	10.0
Sugar and Sugar Alcohols	8.0
Vegetables	14.0

The frequency of a word given the class is given in appendix C.

Evaluation on test split Summary

Correctly Classified Instances	21	38.8889 %
Incorrectly Classified Instances	33	61.1111 %
Kappa statistic	0.3122	
Mean absolute error	0.115	
Root mean squared error	0.2322	
Relative absolute error	87.8569 %	
Root relative squared error	90.712 %	
Total Number of Instances	54	

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.400	0.327	0.111	0.400	0.174	0.045	0.527	0.200	Cakes
0.000	0.000	?	0.000	?	?	0.923	0.200	Beverages
0.000	0.000	?	0.000	?	?	0.931	0.300	Juices
1.000	0.104	0.545	1.00	0.706	0.699	0.948	0.545	Breads
0.000	0.000	?	0.000	?	?	0.753	0.226	Breakfast cereals
0.714	0.064	0.625	0.714	0.667	0.615	0.834	0.532	Cereal Grains
1.00	0.075	0.200	1.00	0.333	0.430	0.962	0.200	Cookies
0.000	0.000	?	0.000	?	?	0.971	0.400	Crackers
0.000	0.000	?	0.000	?	?	0.815	0.578	Dairy Products
0.778	0.11	0.583	0.778	0.667	0.598	0.930	0.848	Fruit and fruit products
0.000	0.000	?	0.000	?	?	0.951	0.375	Legumes and Nuts
0.000	0.000	?	0.000	?	?	0.870	0.309	Snack foods and confectionary
0.000	0.000	?	0.000	?	?	0.585	0.040	Sugar and Sugar Alcohols
0.000	0.000	?	0.000	?	?	0.827	0.167	vegetables
Weighted Averages								
0.389	0.070	?	0.389	?	?	0.858	0.463	

Confusion Matrix

a b c d e f g h i j k l m ~~n <--~~
classified as 2 0 0 0 0 0 1 0
0 2 0 0 0 0 | a = Cakes
2 0 0 0 0 0 0 0 0 0 0 0 0 | b = Beverages
3 0 0 0 0 0 0 0 0 0 0 0 0 | c = Juices
0 0 0 6 0 0 0 0 0 0 0 0 0 | d = Breads
1 0 0 3 0 0 0 0 0 0 0 0 0 | e = Breakfast cereals
0 0 0 1 0 5 0 0 0 1 0 0 0 0 | f = Cereal Grains
0 0 0 0 0 0 1 0 0 0 0 0 0 0 | g = Cookies
0 0 0 0 0 0 2 0 0 0 0 0 0 0 | h = Crackers
6 0 0 0 0 0 0 0 0 1 0 0 0 0 | i = Dairy Products
1 0 0 1 0 0 0 0 0 7 0 0 0 0 | j = Fruit and Fruit products
0 0 0 0 0 3 0 0 0 0 0 0 0 0 | k = Legumes and Nuts
1 0 0 0 0 0 0 0 0 1 0 0 0 0 | l = Snack foods and confectionary
0 0 0 0 0 0 1 0 0 0 0 0 0 0 | m = Sugar and Sugar Alcohols
2 0 0 0 0 0 0 0 0 0 0 0 0 0 | n = Vegetables

4.2.1 Functional requirements

The functional requirements are derived from the basic processes and functions that will be performed by the systems once implemented. The application functional requirements include inputting food list details (the food item name and food group), capturing the meal details (meal-type, preparation method, duration of preparation, time when meal is taken, nature of the meal taken and the size of the meal item), recording of patient biomarkers (age, weight and height) as independent variables to be used in the prediction of the glucose level, viewing the predicted output from the system after running the predictor module of the application and managing of the food item details by the administrator among other basic functionalities.

4.2.2 Non-functional requirements

These are the characteristic features of the system that will enable the developed application function optimally. These characteristic features include: data security, integrity and accuracy (this is enforced by using system authentication mechanisms through authorized logins), error handling mechanism through the use of exception handling approaches, system availability which is handled through appropriate sessions, system performance, and the usability of the system that has been tackled through the principles of usability which includes easy navigability, user friendly interface designs.

4.3Use case Diagram

Figure 4.1 illustrates the major interactions between the external actors and the system processes.

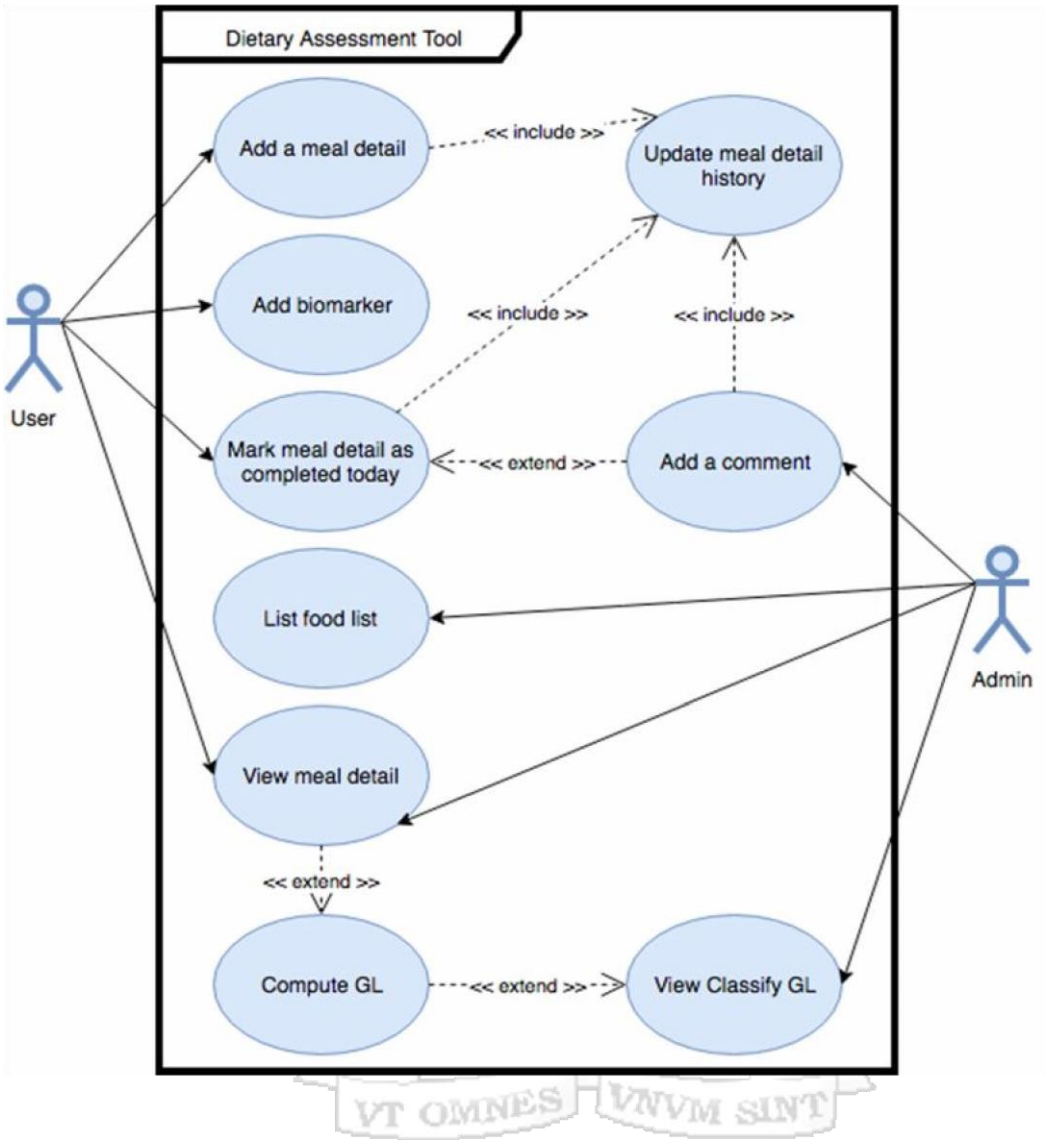


Figure 4. 2: Use case diagram for the dietary assessment tool

The following is the use case description of the application

Table 4. 1: Verification of food item details

Use Case: Verification of food item details	
Primary actor:	User
Pre-condition	Verify the food item details
Post-condition	Correct matching of food item details with existing international standards
Main success scenario	

User	System
1. User adds meal details 2. User adds biomarker information	1. System compares food item details to existing food list

Use Case: Compute GL	
Primary actor:	User, System
Pre-condition	Confirm the completeness of the Verify the food item details
Post-condition	Correct computation of the GL
Main success scenario	
User	**System**
User checks for completeness and correctness of the food item details	System computes meal item GL

Table 4. 2: Computing GL

Table 4. 3: User details and biomarker confirmation

Use Case: User details confirmation	
Primary actor:	User, System administrator
Pre-condition	Administrator logged in
Post-condition	User registered in the system
Main success scenario	
User	**System**
1. Administrator activates the registration 2. User adds biomarker information	1. System accepts the registration details of users

Table 4. 4: Analysis of the GL

Use Case: Analysis of GL	
Primary actor:	User, System
Pre-condition	Computed GL
Post-condition	Correct classification of GL levels
Main success scenario	
User	**System**
1. User checks for correctness of GL values	1. System generates the classification of GL levels

A sequence diagram is used to show the flow of information between the various interacting entities of the system. Figure 4. 2 shows the interactions of the various entities in the dietary assessment tool. It depicts the follow of requests and feedbacks to and from the application.

4.4Sequence Diagram

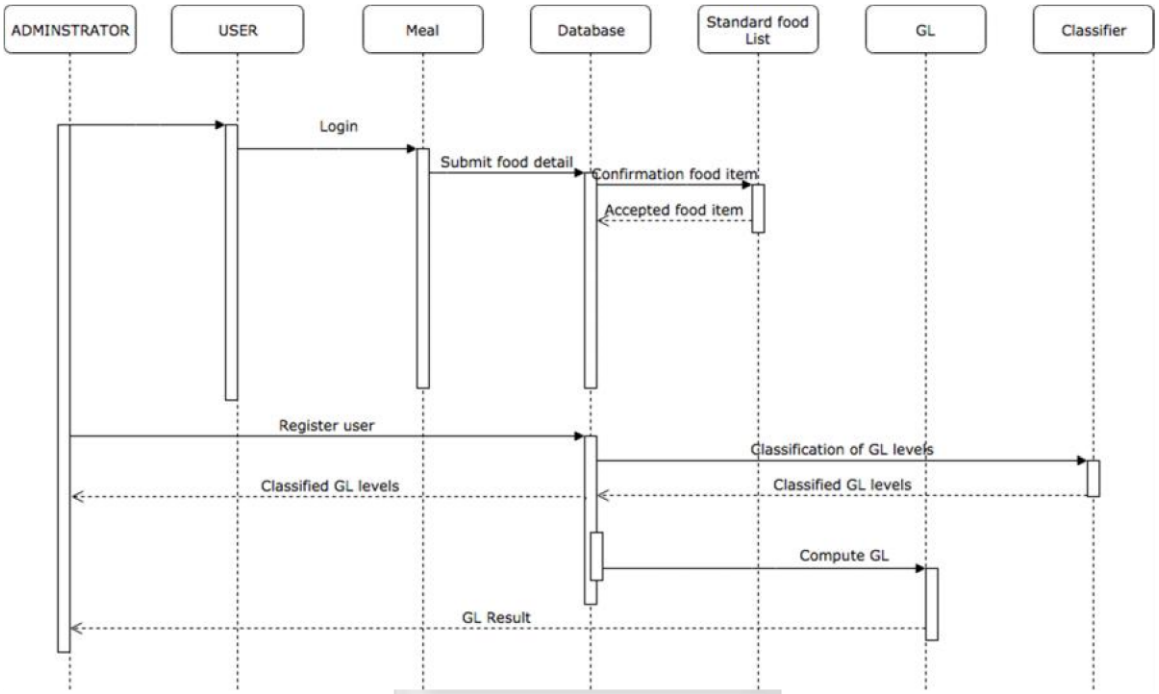


Figure 4. 3: Sequence Diagram for the predictive tool

4.5 Description of the System development interactions

The application developed has the following key actors: the system user (this is anybody who signs up for the service), the application administrator (this is the person who handles all the system management issues). The application is built on an online database application development environment and therefore can be accessed via any web-browser and is adaptable.

The procedure followed in designing the glucose predictive application consisted of six stages, which included:

1. Data identification: This was the first stage in the data mining process for the purposes of predicting the possibility of a person having high or low sugar levels. The data used in the study was obtained from *International Table of Glycemic Index and Glycemic load value.* It was selected for this study due to its consistence and completeness.
2. Data preparation: This involved the organization and conversion of the data into the correct formats for entry into the oracle database. It also involved the classification of the data using WEKA functionalities.

3. Data analysis: At this stage the data was analyzed to obtain the desired research outcome. At this stage various data analysis and mining techniques were used. For example, clustering technique was used to cluster the dataset into two broad clusters of low and high glycemic index. A regression analysis technique was also used to give predicted values.
4. Database results: The Apex Oracle database was used to publish the results from the database.
5. Prediction and Knowledge Evaluation: The predicted values obtained from the data analysis were evaluated to extract new knowledge from the results in the database. This stage offered a chance to conduct forecasting using the predicted values.
6. Implementation and deployment: This involved the test running of the developed algorithms and comparing it with other existing literature data on the glucose level.

4.6Process control

The mapping of the flow of data from one entity to another and across the system processes, the data flow diagram (DFD) was used. A DFD was used in this study to specially show the movement of information within the glucose control system using dietary management approach. It provided a means of visualizing the transfer of data between the interacting system processes, data stores and entities external to the system. Figure 4.2 shows the main processes involved in the application.

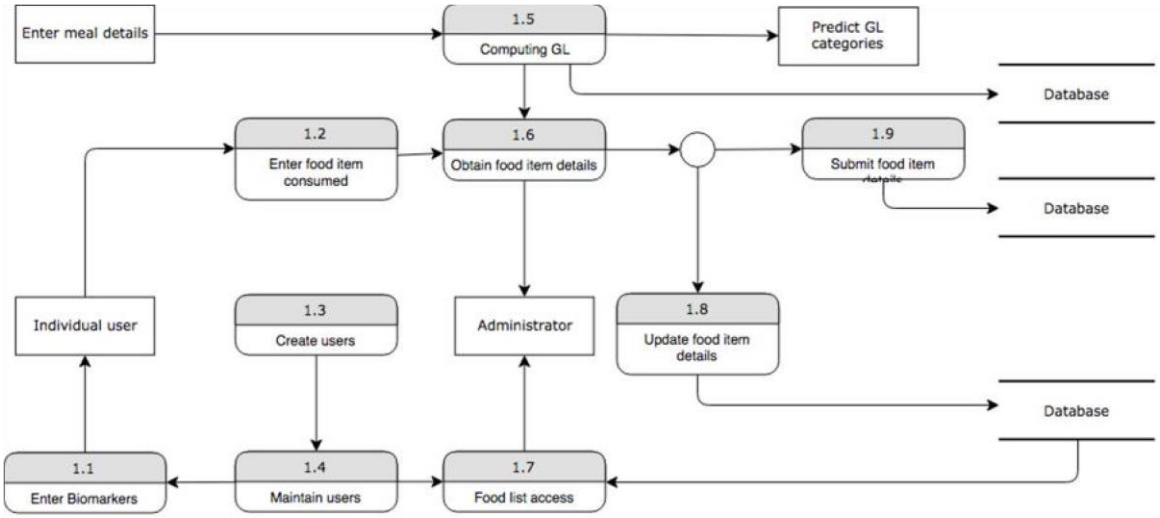


Figure 4. 4: Data Flow Diagram for the application

4.7Entity Relationship Diagram

Entity relationship diagram (ERD) is used to represent the conceptual data model of an information system. It provides a basic building block on how the database structure will be build. It also shows the interaction between the various entities of the application. Figure 4.3 shows the ERD for the proposed system after normalization.

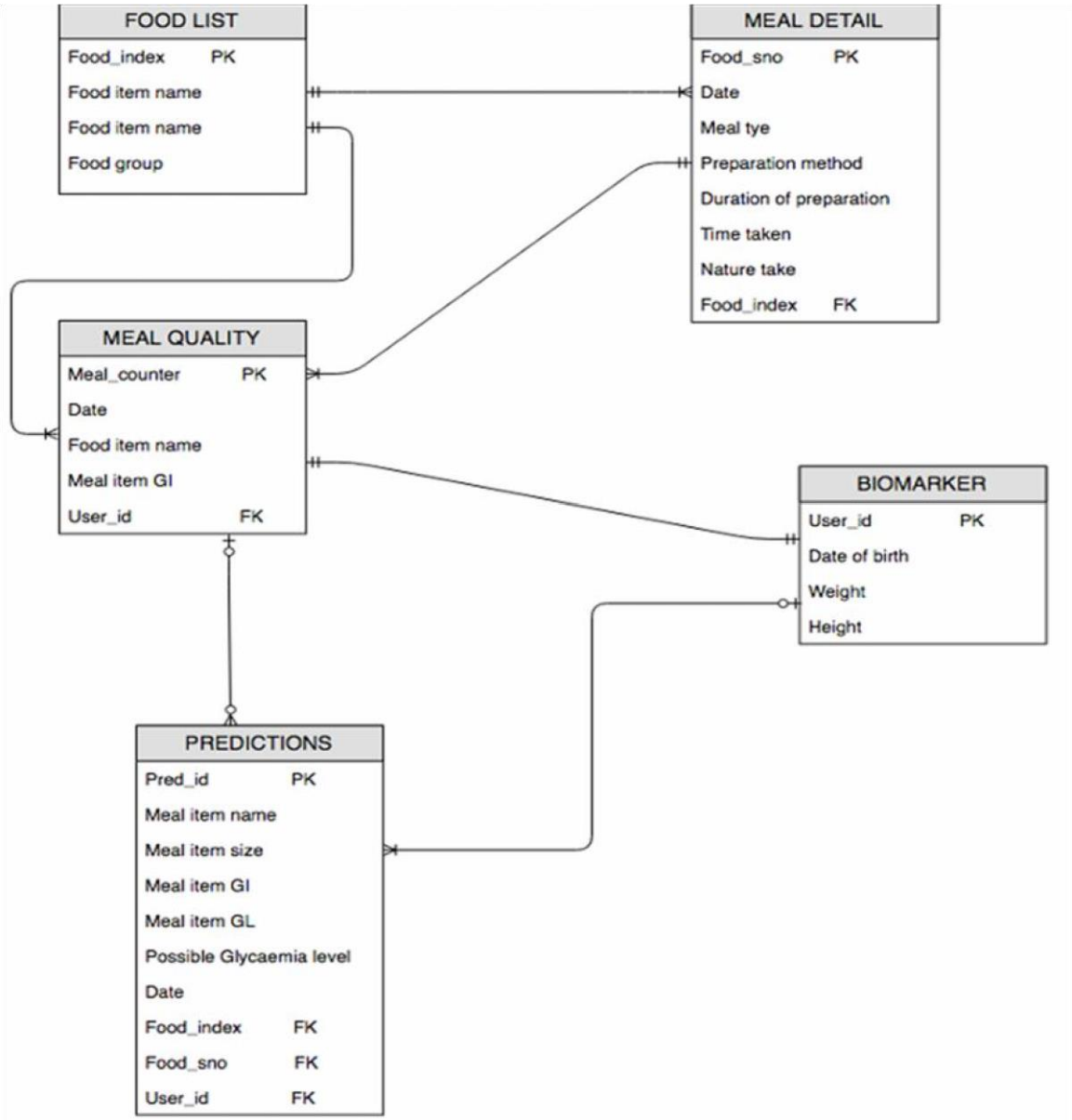


Figure 4. 5 Entity Relationship Diagram for the application **Chapter 5**

System Implementation and Testing

5.1 Introduction

The chapter provides detailed explanation on how the application was developed and tested. It offers diagrammatic representation of the various system modules and screenshots to complement on the explanations.

5.2 System development

The system development started with the categorization or clustering of the food items using Hierarchical Clusterer -N 2 -L SINGLE -P -A "weka.core.Euclidean Distance -R first-last" algorithm from the WEKA data mining tool. The dataset items used in the study were assigned the formed clusters according to the food groups.

The application consists of a front-end and back-end sub-systems both available as web applications. The application was built on the Oracle database using Oracle application Express platform. The predictive algorithms used in the categorization of the predicated plasma glucose used the evaluative criteria given by:

$$\begin{aligned} &\text{if computed } GL \geq 20 \text{ return high plasma glucose levels expected} \\ &\text{else if computed } GL > 10 \text{ and computed } GL < 20 \text{ return medium plasma glucose level} \\ &\text{else return low plasma glucose level} \end{aligned}$$

These algorithms were coded into the application functionalities as SQL statements.

5.3Application Front-end

The front-end of the application was designed to be used by the any registered users. The web interfaces of the application are arranged into modules and to enhance navigability a menu layout structure was adopted as shown in figure 5.1 once a user is logged in.

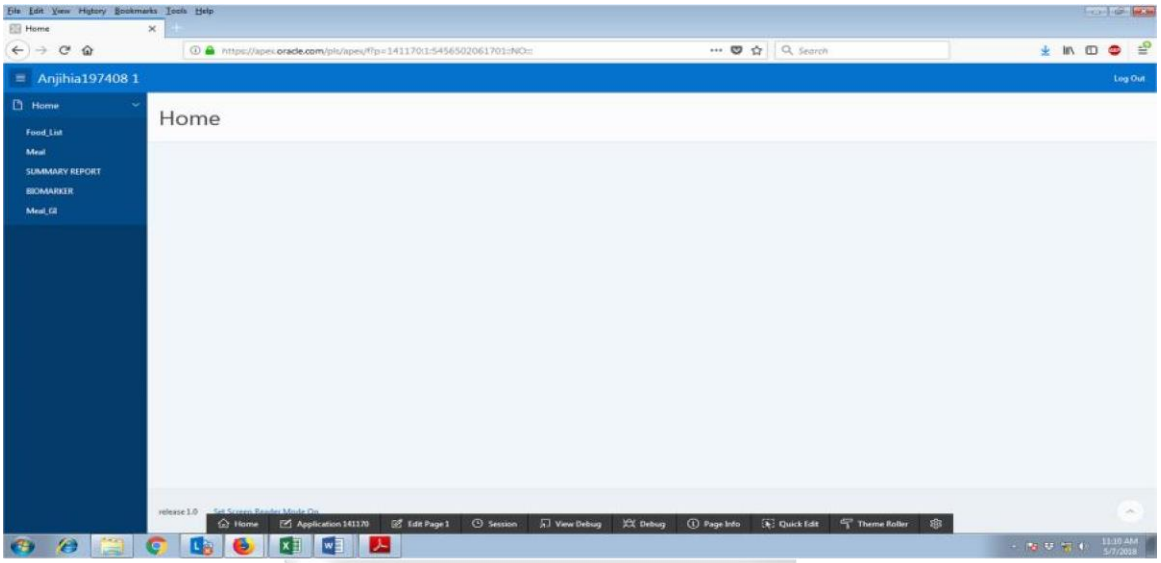


Figure 5. 1 Home page for the dietary Assessment tool

5.4 Data capture modules

Figures 5.2-5.5 show the data capture interfaces the users are prompted to enter the required data details.

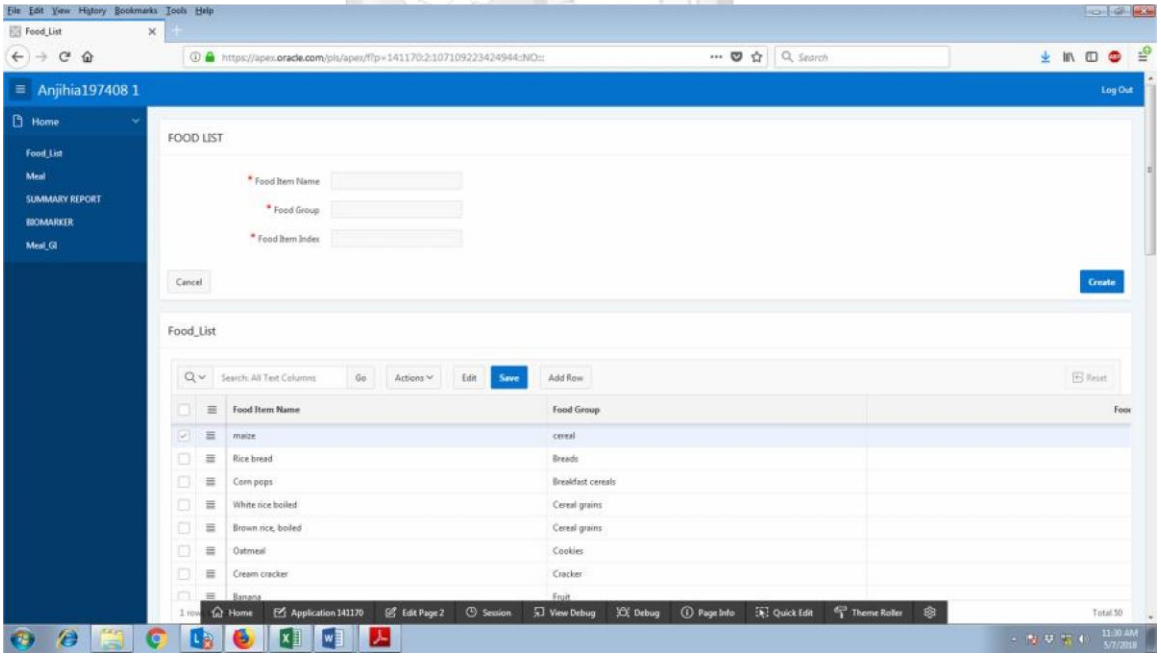


Figure 5. 2: ***Food list capture module***

This page provides an interface for the system administrator to enter the food item basic details. The data entered in through this page should be in conformity with the

FAO standards as documented in the *International Table of Glycemic Index and Glycemic load value* (Foster-Powell et al. 2002).

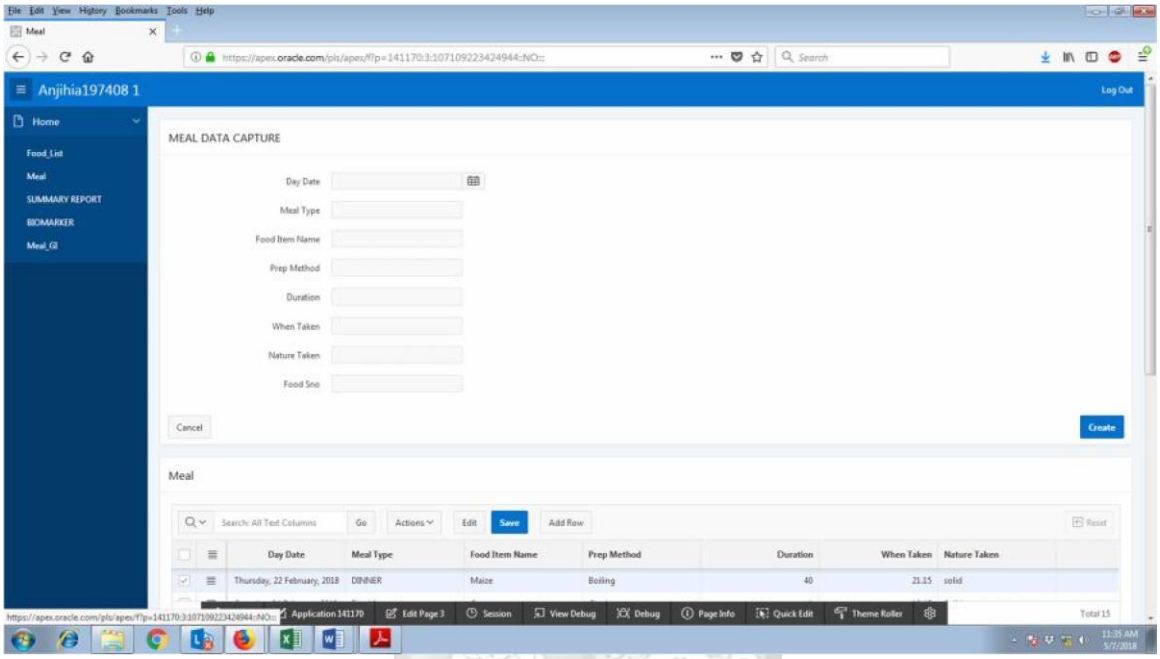


Figure 5. 3: ***Meal Data recording module***

This module provides an interface for the system user (patient) to enter the details of the food consumed. The module is linked with the food list module so as to import some important characteristics that will be relevant in determining the type of food item the patient may have consumed.

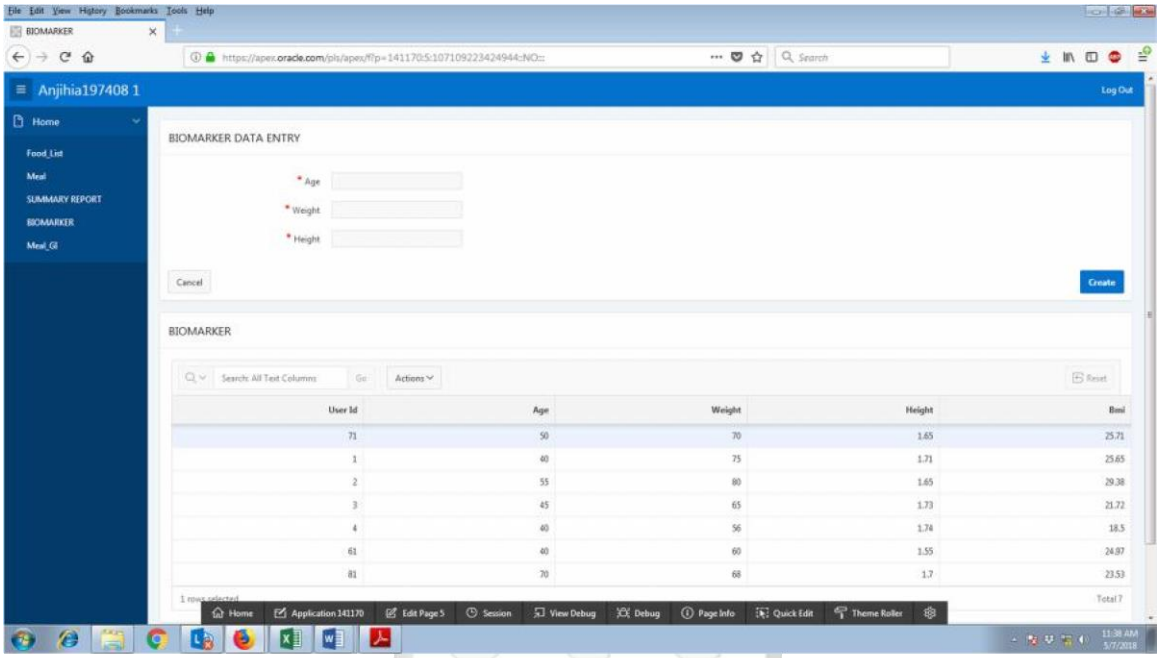


Figure 5. 4: ***Biomarker capture***

Through this module basic information relating to the patient biomarkers are recorded. This module is critical in determining the basic biological state of the patient more specifically the BMI and age of the patient which are critical estimation of the dietary needs of an individual.

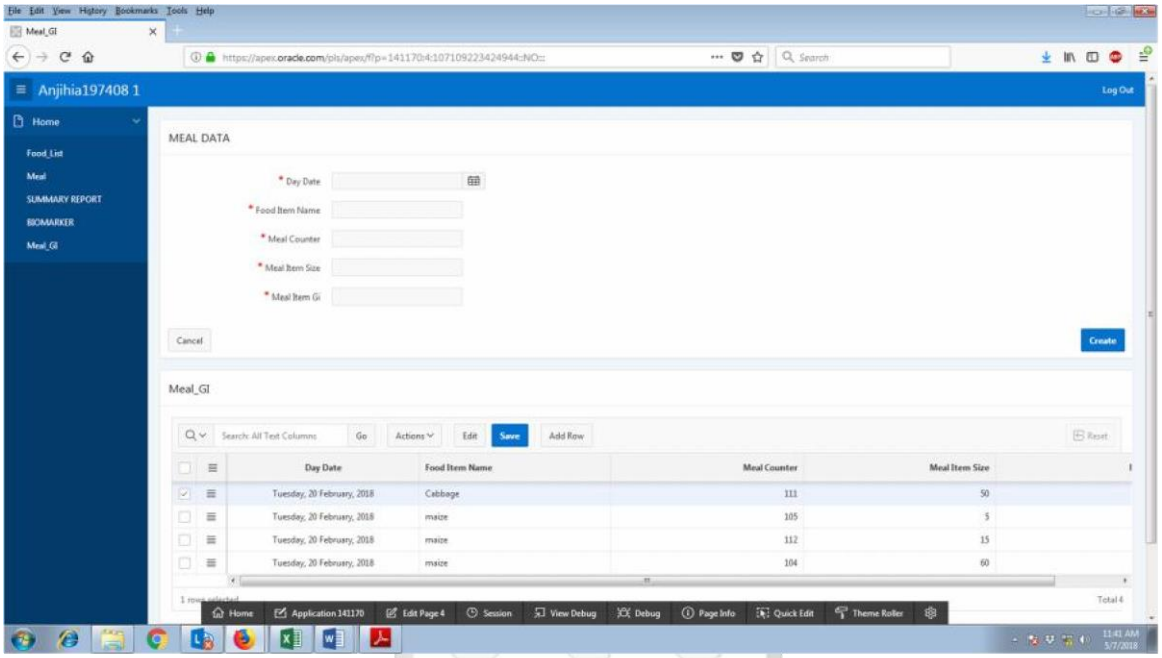


Figure 5. 5 ***Meal analysis module***

This interface page provides both a data capture and analysis functionality. It is through this module the patient is able to estimate the Glycemic Load (GL) of the food item they may have consumed given the meal item size and food item name. The main algorithmic function used in the estimation of the GL is based on the mathematical function given by:

$$Computed\ Gl = \left(GI * (meal\ item\ carbohydrates\ size\ in\ grams/100)\right)$$

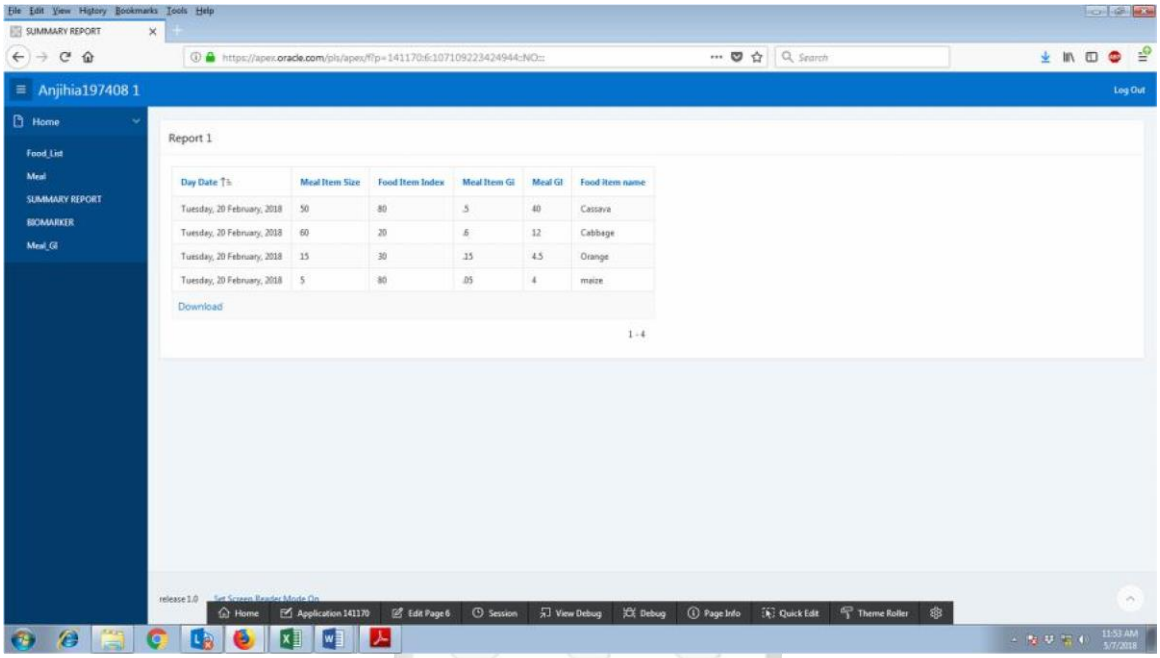


Figure 5. 6: ***Summary report and Analytics***

Figure 5.6 shows the analytics of the data provided by the patient as compared with the *International Table of Glycemic Index and Glycemic load value* for the estimation of the plasma glucose trajectories for the patient.

5.5 System testing

Agile system testing approach was used in testing on the performance of the developed application. It was the ideal approach because it focuses on analyzing the performance issues or/and bugs in the context of the system operational work flow. It was applied throughout the system development period.

Some of the performance aspects analyzed include

a) Compatibility testing

Compatibility test was conducted to ensure the developed application was compatible with various platforms where it could be used. The tests results are captured in table 5.1.

Table 5.1: Web platform compatibility test

Type of web browser	Compatibility acceptance
Mozilla Firefox (version 8.0 or higher)	Yes
Google Chrome	Yes

b) Usability testing

The usability testing was conducted to check on the predictive power of the application by comparing the predicted results from GL with the known standard data results as mined and analyzed by WEKA. The outcome of the comparison is given in table 5.2

Accuracy

Table 5.2: Predicted results vs mined data results

Food	**Food size (Estimated)**	**Computed GL**	**GL**	**Standard predicted**	**Application results name**	**Mined item**
Banana	100	17	16	Medium	medium cake	
White 150	40	36	High	High rice		
Wheat 30	7.5	6	Low	Low bread		
Milk	250ml	3.5		3	Low	Low
Orange 120 6 5 Low Low Maize 150 12.5 9 Medium Low						

The predicted results compares closely with the mined data results and therefore the application gives ~~fairly~~ plausible results. The categorization of the levels of GL is based on a scale ado~~pted from the work of Kouame et al 201~~5, that is, high>= 20, medium> 10 -<20 and low =<10.

Chapter 6
Discussion

6.1Introduction

This chapter presents an overview on the study objectives from the data analysis and the reviewed literature. It provides an analysis on the outcome of the study as compared to the reviewed methodology from the reviewed literature

6.2A Review of the Research objectives

From section 1.3, the research focused on analyzing the dietary factors for management of type 2 diabetes, analyzing the various approaches used for data

mining in health, developing a predictive algorithm for associating meals with the glycogenic levels in diabetic patients and evaluating the performance of the developed application in predicting the glycemic levels.

From the reviewed literature, it was found that dietary plays a very critical role in the management and control of plasma glucose. It was also found that the type of food consumed, quantity consumed, way of preparation and its glycemic index play a significant role in determining the plasma sugar levels.

The study established that data mining techniques have been substantially used in health. Some of the techniques identified include genetic algorithms, rough set techniques, Bayesian networks, Support Vector machine, Fuzzy sets, decision trees and neural networks. The main focus in all these techniques is to aid in the knowledge discovery process (Usama and Gregory, 1996) and to support in management. Almost all of the techniques have also been used in one or all of the following tasks: summarization, association, classification, clustering, trend analysis and regression.

The development of the predictive algorithm required the determination of the appropriate data mining algorithms. A developed algorithm leveraged on the regression attributes to make predictions based on the available and apriori dataset.

6.3Advantages of the developed application

The developed application offers the patient's ability to predict the possible future plasma glucose levels and therefore institute corrective measures. It will also help the patient to track the food item consumed and possibly isolate the food regime that is likely to cause problems way in advance.

6.4Application limitation

Since the application does not make use other physiological characteristics, it cannot measure the exact plasma glucose values. It therefore gives a pointer to

the trajectories the plasma glucose is likely to take in descriptive terms without exact quantitative values.

Chapter 7
Conclusion, Recommendation and Future work

7.1Introduction

The chapter offers an overview of the study objectives by reporting on the key outcomes of the study objectives. It also provides recommendation on the usage of the developed application and finally it provides possible extensions that can be made on the applications as future work.

7.2Conclusions

From the study, it was established that management and control of blood sugar is dependent on many factors chief of which is the amount of carbohydrates consumed per meal by an individual. This is also dependent on the nature of the carbohydrate ingested which is influenced by glycemic index of the food item. It is therefore important for the diabetic patient to understand the glycemic index of the food item they are consuming.

The concept of data mining was found to be in use in health sector. Specifically, it has been used in the discovery of the disease patterns and also in the development of knowledge based health system.

A predictive tool f~~or the determination of the glycemic levels~~ has been developed to use data inputs from the patient instead of relying on the historical data to determine the future blood sugar trajectories for the patient. Through the prediction of the possible ups or lows of the sugar, the patient is better prepared to make informed choices of the food item that is available in their environment.

7.3Recommendations

From the study outcome, it is observed that for proper functioning of the application, the food item to be analyzed must be that which has been tested by FAO and whose GI has been scientifically determined. It is therefore recommended that the users of the application confirms the correct food item

nomenclature before entering the food item to avoid misclassification of the food item.

7.4Future work

Future work on this application could focus on analyzing the local food items and developing a local food corpus using the local dialects. This will eliminate the bottleneck associated with scientific naming of the food items which may often lead to wrong classifications. Also more parameters could be considered to make the prediction tool more robust.

REFERENCES

Ankita A. (2012). "Secret Key Encryption algorithm using genetic algorithm", vol.-2, no.-4, ISSN: 2277 128X, IJARCSSE, pp. 57-61.

Anto S. & Chandramathi S. (2011). Supervised Machine Learning Approaches for Medical Data Set Classification - A Review, IJCST Vol. 2, Issue 4.

Apte & Weiss S.M. (1997). Data Mining with Decision Trees and Decision Rules, T.J. Watson Research Center. http://www.research.ibm.com/dar/papers/pdf/fgcsapteweissue_with_cover.pdf.

Argyro K., Christophoros N. & George M. (1995). "Robustness of Support Vector Machine-based Classification of Heart Rate Signals", Proceedings of the 28th IEEE, EMBS Annual International Conference, New York, USA, Aug30-sep 3, 2006, 1995.

Balaji P. (2013). "Data Mining Overview and Optimization Opportunities". Microsoft Research Report MSR-TR-98-04.

Beller G. & Nucl V. C~~ardio~~l. (2008). "The rising cost ~~of healt~~h care in the United States: is ~~it making the United States globally n~~oncompetitive?" vol. 15, no. 4, pp. 481-482.

Bhatnagar M. & Singh K. P. (2013). Research Methodology as SDLC Process in Image Processing. *International Journal of Computer Applications*. Vol. 77 No 2

Cristianini N. & Shawe-taylor J. (1995). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

Divya T. & Sonali A. (2013). A survey on data mining approaches for healthcare, International Journal of Bio-Science and Bio-Technology Vol.No.5, pp. 241-266.

Emina A. & Abdulhamit S. (2011). Data Mining Techniques for Medical Data Classification. The International Arab Conference on Information Technology (ACIT).

Ezzati M. & Riboli E. (2013). Behavioral and dietary risk factors for noncommunicable diseases. N Engl J Med. 2013; 369(10):954–64.

Finn V. J. (1996). An Introduction to Bayesian Networks, Springer, New York, 1996.

Franz M. J., Bantle J.P., Beebe C.A., Brunzell J.D., Chiasson J.L., Garg A., Holzmeister

L.A., Hoogwerf B., Mayer-Davis E., Mooradian A.D., Purnell J.Q. & Wheeler M. (2002). Evidence-based nutrition principles and recommendations for the treatment and prevention of diabetes and related complications. Diabetes Care 25*:* 148–*198.*

Friedman N., Geiger, D.Goldszmidt M. (1997). "Bayesian network classifiers. Machine Learning 29: pp. 131-163.

Fuller N.R., Sainsbury A, Caterson I.D. & Markovic T.P. (2015). Egg consumption and human cardio-metabolic health in people with and without diabetes. *Nutrients*. 2015 Sep 3. 7(9):7399-420.

Gannon M. C. Nuttall F.Q., Saeed A., Jordan K. & Hoover H. (2003). An increase in dietary protein improves the blood glucose response in persons with type 2 diabetes. Am J Clin Nutr 78:734–741.

Gannon M.C. & Nuttall F.Q. (2004). Effect of a high-protein, low-carbohydrate diet on blood glucose control in people with type 2 diabetes. *Diabetes*. 2004 Sep. 53(9):2375-82.

Gannon M.C., Nuttall J.A., Damberg G., Gupta V. & Nuttall F.Q. (2001). Effect of protein ingestion on the glucose appearance rate in people with type 2 diabetes. J Clin Endocrinol Metab 86:1040–1047.

Gayathri V., Chanda M.M & Banu S. C. (2014). A survey of data mining techniques on medical diagnosis and research. International Journal of Data Engineering (IJDE) Singaporean Journal of Scientific Research (SJSR) Vol.6.No.6 2014.

Giese, P. D. (2018). Systems Analysis and modelling.

Gheware S.D., Kejkar A.S. & Tondare S.M. (2014). Data Mining: Task, Tools, Techniques and Applications, International Journal of Advanced

Research in Computer and Communication Engineering Vol. 3, Issue 10, October 2014.

Goharian & Grossman (2003). Data Mining Classification, Illinois Institute of Technology. http://ir.iit.edu/~nazli/cs422/CS422-Slides/DMClassification

Gosain, A. & Kumar, ~~A. (2~~009). "Analysis of health care ~~data~~ using different data mining ~~techniques," Intelligent Agent & Multi-A~~gent Systems, 2009. IAMA 2009, International Conference on, vol. no., pp.1,6, 22-24 July 2009.

Gougeon R., Styhler K., Morais J.A., Jones P.J. & Marliss E.B. (2000). Effects of oral hypoglycemic agents and diet on protein metabolism in type 2 diabetes. Diabetes Care 23:1–8.

Hammer, B. & Gersmann, K. (2003). "A Note on the Universal Approximation Capability of SupportVector Machines", Neural Process Lett 17, pp. 1061 - 1085.

Haykin. S. (1999). Neural Networks: A Comprehensive Foundation. Prentice Hall.

Hian Chye K. & Gerald T. (2005). Data mining applications in healthcare, Journal of healthcare information management: JHIM.19 (2): 64-72.

Howard A.A., Arnsten J.H. & Gourevitch M.N. (2004): Effect of alcohol consumption on diabetes mellitus: a systematic review. Ann Intern Med 140*:*211– *219.*

Hu F.B., van Dam R.M. & Liu S. (2001) Diet and risk of type II diabetes: the role of types of fat and carbohydrate. Diabetologia 44:805–817.

Ilias M., Euripidis L., Elias Z. & Antonis S. (2009). "Support vector machine based identification of heart valve diseases using heart sounds", Elsevier, Computer Methods and Programs in Biomedicine, 95, pp. 47-61.

Institute of Medicine (2002): Dietary Reference Intakes: Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids. *Washington, DC, National Academies Press.*

Jenkins D.J., Wolever T.M., Taylor R.H., Barker H., Fielden H., Baldwin J.M., Bowling A.C., Newman H.C., Jenkins A.L. & Goff D.V. (1981). Glycemic index of foods: a ~~phys~~iological basis for carbohydrate ~~exch~~ange. Am J Clin Nutr 34:362–366

Jian C. & Yongheng Z. (2010). "Object-oriented Classification of High-resolution Remotely Sensed Imagery". IPCSIT Vol. 47, pp. no.123.

Jihoon Y. & Vasant H. (1998). Feature subset selection using Genetic Algorithm. IEEE Intelligent Systems.

Joshi, P. (2013). Programming Paradigms: Object-Oriented vs Data Oriented. OOP. Retrieved January 12, 2017, from https://prateekvjoshi.com/2013/11/30/programming-paradigmsobject-oriented-vs-data-oriented/

Kalyani M. & Samayita B. (2013). Soft computing on Medical -Data (SCOM) for a Countrywide Medical System using Data Mining and Cloud Computing Features, Global Journal of Computer Science and Technology Cloud and Distributed, Volume 13 Issue 3 Version 1.0.

Knowler W.C., Barrett-Connor E., Fowler S.E., et al. (2002). Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. N Engl J Med.; 346(6):393–403.

Koume A., Kouassi K., N'dri Y. & Amani N. (2015). Glycaemic Index and Load Values Tested in Normoglycemic Adults for Five Stable Foodstuffs: Pounded Yam, Pounded Cassava-Plantain, Placali, Attieke and Maize Meal Stiff Porridge. Nutrients; 1267-1281. Doi: 10.3390/nu7021267

Kris-Etherton P.M., Lichtenstein A.H., Howard B.V., Steinberg D. & Witztum J.L. (2004). Antioxidant vitamin supplements and cardiovascular disease. Circulation 110:637–641.

Liu Y., Cotillard A., Vatier C., et al. (2015) A dietary supplement containing cinnamon, chromium and carnosine decreases fasting plasma glucose and increases lean mass in overweight or obese pre-diabetic subjects: a randomized, placebo-controlled trial. *PLoS One*. 2015. 10(9): e0138646.

Dunham M.H. (2002). Data Mining, Introductory and Advanced Topics. Prentice Hall.

Malan, R. (2001). Functional Requirements and Use Cases. Architecture Resources. Retrieved January 16, 2017, from http://www.bredemeyer.com

Mary O. & Mat K. (2004). "Applications of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, August 2004.

Mayer-Davis E.J., Dhawan A., Liese A.D., Teff K & Schulz M. (2006). Towards understanding of glycaemic index and glycaemic load in habitual diet: associations with measures of glycaemia in the Insulin Resistance Atherosclerosis Study. Br J Nutr 95:397–405.

Mooradian A.D. (1999). Micronutrients in diabetes mellitus. Drugs, Diet and Disease 2:183–200.

Namrata C. M. (2012). "Text Classification and Classifiers a Survey". International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2.

Obenshain M. K. (2004). Application of data mining techniques to healthcare data Infect. Control Hosp. Epidemiol, 25(8):690–695.

Pang. T, Steinbach M. & Kumar V. (2005). “Introduction to Data Mining”. Addison Wesley, pp. no.769.

Pang-Ning T., Michael S. & Vipin K. (2002) "Introduction to Data Mining", Addison Wesley.

Parker B., Noakes M., Luscombe N., Clifton P. (2002). Effect of a high-protein, high–monounsaturated fat weight loss diet on glycemic control and lipid levels in type 2 diabetes. Diabetes Care 25:425–430.

Parvez A., Saqib Q. & Syed Q. R. (2015). Techniques of Data Mining in Healthcare: A Review, International Journal of Computer Applications (0975 – 8887) Volume 120 – No.15.

Pooja P., Jayanthv J. & Koliwad S. (2005). “Classification of RS data using Decision Tree Approach”. International Journal of Computer Applications,Vol. 23(3), pp. no.7-11

Ravi G. K., Ramachandra G.A. & Nagamani K. (2014). An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2.

Sang J. L & Keng S. (2001). A review of data mining techniques, Industrial Management & Data Systems, 101/1, MCB University Press, ISSN 0263-5577.

Sebe N., Ira C., Ashutosh G. & Thomas H. S. (2005). "Machine Learning in Computer Vision", Springer, Netherlands, pp. 130-133.

Shameem A., Manimegalai D. & Hundewale N. (2011). A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, ISSN (Online): 1694-0814.

Soni J., Ansari U. & Sharma D. (2011). "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, March 2011.

Steljes. (2012). Agile Methodologies. Retrieved January 13, 2017, from http://vi.sualize.us/1_minus_1_agile_methodologies_process_ui_diagrams_picture_pfgo.html

Stephenson E.J, Smiles W. & Hawley J.A. (2014). The relationship between exercise, nutrition and type 2 diabetes. *Med Sport Sci*. 2014. 60:1-10.

Summers L.K., Fielding B.A., Bradshaw H.A., Ilic V., Beysen C., Clark M.L., Moore N.R., & Frayn K.N. (2002): Substituting dietary saturatedfat with polyunsaturated fat changes abdominal fat distribution and improves insulin sensitivity. Diabetologia 45:369–377.

Turner B.C., Jenkins E., Kerr D., Sherwin R.S. & Cavan D.A. (2001). The effect of evening alcohol consumption on next-morning glucose control in type 1 diabetes. Diabetes Care 24:1888–1893.

Usama F. & Gregory P. S. (1996). "Knowledge Discovering and Data Mining: towards a Unifying Framework". In Proc. Of KDD-996, Vol. 2, pp. no.20 30.

Usha K. R. (2011). Analysis of Heart Diseases Dataset Using Neural Network Approach, International Journal of Data Mining & Knowledge Management Process (Ijdkp) Vol.1, No.5.

Vapnik V. (1998). "Statistical Learning Theory", Wiley, ISBN: 978-0-471-03003-4.

Vapnik V. (1998). "The support vector method of function estimation", AT & T Labs – Research, John Wiley and Sons, New York, USA.

Vapnik, V.N. (2005). "The Nature of Statistical Learning Theory", Springer, New York.

Vargas, J., Yu, H. & Jiang, X. (2007). "Privacy preserving SVM classification, Knowledge and Information Systems". Vol.14, pp. no.161-178.

Waldo, J. (2006). On Systems Design. Burlington: Sun Labs. Retrieved January 9, 2017, from https://scholar.harvard.edu/files/waldo/files/ps-20066.pdf

Wasserman D.H. & Zinman B. (1994). Exercise in individuals with IDDM. Diabetes Care 17:924–937.

Yongjian F. (ND). Data Mining: Tasks, Techniques and Applications. http://academic.csuohio.edu/fuy/Pub/pot97.pdf

Yu Q., HuiPingLiua, M., XiaoDong W. & XiaoLuo Z. (2005). "The Decision Tree Algorithm of urban extraction from Multisource Image Data". Vol. 3, pp. 301 308.

Zadeh L.A. (1998). "Some reflection on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent system", Soft computing, vol. 2.

Appendix A: Food List

Sno.	Category	Description	GI	Serving size	Available carb	GL per Serving
1	Cakes	Banana cake, with added sugar	47	80	38	18
2	Cakes	Banana cake, no added sugar	55	80	29	16
3	Cakes	Chocolate cake	38	111	52	20
4	Cakes	Vanilla cake	42	111	58	24
5	Cakes	Doughnut	76	47	23	17
6	Cakes	Muffins apple with sugar	44	60	29	13
7	Cakes	muffins apple without sugar	48	60	19	9
8	Cakes	Pancakes	67	80	58	39
9	Cakes	Pastry	59	57	26	15
10	Cakes	Scones plain	92	25	9	7
11	Beverages	Coca cola soft drink	53	250ml	26	14
12	Beverages	Fanta Orange	66	250ml	34	23
13	Beverages	Lucozade	95	250ml	42	40
14	Beverages	Lemon Squash	58	250ml	29	17
15	Juices	Apple unsweetened	44	250ml	30	13
16	Juices	Carrot freshly made	43	250ml	23	10
17	Juices	Cranberry juice cocktail	68	250ml	36	24
18	Juices	Grapefruit	48	250ml	22	11
19	Juices	Orange unsweetened	53	250ml	26	13
20	Juices	Pineapple	46	250ml	34	16
21	Juices	Tomato	38	250ml	9	4
22	Breads	Barley flour breads	67	30	13	9
23	Breads	Barley flour breads whole-meal	50	30	15	7
24	Breads	Barley flour breads whole-meal with lactic acid	66	30	19	12
25	Breads	Fruit bread	44	30	13	6
26	Breads	Oat bread	65	30	19	12
27	Breads	Gluten-free bread	79	30	13	10
28	Breads	Rice bread	72	30	12	8
29	Breads	Rye bread, whole-grain pumpernickel	46	30	11	5
30	Breads	Rye bread whole-meal	41	30	14	8
31	Breads	Wheat bread	52	30	12	6
32	Breads	Spelt wheat bread, white flour	71	30	13	9
33	Breads	white bread with enzyme inhibitors, acarbose	18	30	17	3
34	Breads	white bread with soluble fiber	41	30	17	7
35	Breads	white resistant starch-enriched bread, whole meal	52	30	12	6

36	Breads	Specialty wheat bread, 9-grainmult-grain	43	30	14	6
37	Breads	Unleavened bread, Amaranth: wheat	66	30	15	10
38	Breads	unleavened bread, wheat-flour	66	30	16	10
39	Breakfast cereals	Whole-meal barley flour porridge	68	14	34	23
40	Breakfast cereals	whole-meal high fiber barley flour porridge	55	14	15	8
41	Breakfast cereals	Cornflakes	72	30	25	18
42	Breakfast cereals	cornflakes, high-fiber	74	30	23	17
43	Breakfast cereals	Cornflakes, crunchy nut	72	30	24	17
44	Breakfast cereals	Corn pops	80	30	26	21
45	Breakfast cereals	Cream of wheat	66	250ml	26	17
46	Breakfast cereals	Energy mix	80	30	24	19
47	Breakfast cereals	Grapenuts flakes	80	30	22	17
48	Breakfast cereals	Porridge from rolled Oats	42	30	21	9
49	Breakfast cereals	Whole-meal oat-flour porridge	74	50	32	24
50	Breakfast cereals	Instant porridge	65	250ml	26	17
51	Breakfast cereals	Rice Bubbles	87	30	26	22
52	Breakfast cereals	Shredded wheat	67	30	20	13
53	Breakfast cereals	Hi-Bran Wheat-bix	61	30	17	10
54	Breakfast cereals	Crunchy nut cornflakes bar	72	30	26	19
55	Cereal Grains	Amaranth	97	30	22	21
56	Cereal Grains	Pearl Barley	22	150	42	11
57	Cereal Grains	Buckwheat	49	150	30	16
58	Cereal Grains	Maize	59	150	13	9
59	Cereal Grains	Cornmeal, boiled in salted water	109	150	13	9
60	Cereal Grains	cornmeal +margarine	69	150	12	9

61	Cereal Grains	Sweat corn on the cob	48	150	30	11
62	Cereal Grains	Sweat corn, frozen, reheated in microwave	47	150	33	16
63	Cereal Grains	Millet	71	150	36	25
64	Cereal Grains	White rice, boiled	69	150	53	36
65	Cereal Grains	White rice, long grain, boiled	41	150	40	16
66	Cereal Grains	white rice, basmati	58	150	38	22
67	Cereal Grains	Brown rice, boiled	72	150	33	16
67	Cereal Grains	Instant or puffed white rice	46	150	42	19
68	Cereal Grains	Wheat	90	68	38	34
69	Cookies	Arrowroot	63	25	20	13
70	Cookies	Milk arrowroot	62	25	18	11
71	Cookies	Digestives, gluten-free	58	25	17	10
72	Cookies	Golden Fruit	77	25	17	13
73	Cookies	Highland Oatmeal	55	25	18	10
74	Cookies	Highland Oatcakes	57	25	15	8
75	Cookies	Oatmeal	54	25	17	9
76	Crackers	Breton Wheat Cracker	67	25	14	10
77	Crackers	Cream Cracker	65	25	17	11
78	Crackers	Puffed Crispbread	81	25	19	15
79	Crackers	Rye crispbread	63	25	16	10
80	Crackers	Premium soda cracker	74	25	17	12
81	Dairy Products	Ice cream	61	50	13	8
82	Dairy Products	Ice cream, reduced or low-fat, vanilla	50	50	6	2
83	Dairy Products	Ice cream high-fat, chocolate	37	50	9	4
84	Dairy Products	Milk, full-fat	27	250ml	12	3
85	Dairy Products	Milk, skim	32	250ml	13	4
86	Dairy Products	Milk, condensed, sweetened	61	250ml	136	83
87	Dairy Products	Milk, low-fat chocolate with aspartame	24	250ml	15	3
88	Dairy Products	Milk, low-fat with sugar	34	250ml	26	9
89	Dairy Products	Pudding, instant, chocolate from powder and whole milk	47	100ml	16	7
90	Dairy Products	Yogurt	36	200ml	9	3
91	Dairy Products	Low-fat yogurt	14	200ml	13	2
92	Dairy Products	Reduced-fat yogurt	26	200ml	30	8

93	Dairy Products	Yogurt drink, reduced-fat, vaalia, tropical passion	38	200ml	29	11
94	Dairy Products	Soy smoothie drink, banana 1% fat	30	250ml	22	7
95	Dairy Products	Soy milk, full-fat	44	250ml	17	8
96	Dairy Products	Soy yogurt	50	200ml	26	13
97	Fruit and Fruit products	Apple, raw	38	120	15	6
98	Fruit and Fruit products	Apple juice	40	250ml	28	11
99	Fruit and Fruit products	Apricots, raw	57	120	9	5
100	Fruit and Fruit products	Banana, raw	70	120	23	16
101	Fruit and Fruit products	Banana ripe	51	120	25	13
102	Fruit and Fruit products	Banana, processed fruit finger	61	30	20	12
103	Fruit and Fruit products	Cherries	22	120	12	3
104	Fruit and Fruit products	Fruit cocktail, canned	55	120	16	9
105	Fruit and Fruit products	Grapefruit	46	120	18	8
106	Fruit and Fruit products	Mango	51	120	17	8
107	Fruit and Fruit products	Oranges	42	120	11	5
108	Fruit and Fruit products	Orange juice	52	250ml	23	12
109	Fruit and Fruit products	Paw Paw	59	120	17	10
110	Fruit and Fruit products	Peaches	42	120	11	5
111	Fruit and Fruit products	Peaches, canned in natural juice	38	120	11	4
112	Fruit and Fruit products	Pears	33	120	13	4
113	Fruit and Fruit products	Pears, canned in pear juice	44	120	11	5
114	Fruit and Fruit products	Pineapple raw	59	120	10	6
115	Fruit and Fruit products	Pineapple juice	46	250ml	34	15
116	Fruit and Fruit products	Plums raw	53	120	11	5
117	Fruit and Fruit products	Strawberry jam	51	30	20	10
118	Fruit and Fruit products	Strawberry processed fruit bar	90	30	26	23

119	Fruit and Fruit products	Tomato juice	30	250ml	9	4
120	Fruit and Fruit products	Watermelon raw	72	120	6	4
121	Legumes and Nuts	Baked beans	48	150	15	7
122	Legumes and Nuts	Beans, dried boiled	29	150	30	9
123	Legumes and Nuts	Black-eyed beans and peas	42	150	30	13
124	Legumes and Nuts	Butter beans	54	150	20	11
125	Legumes and Nuts	Chickpeas	28	150	30	8
126	Legumes and Nuts	Black beans	20	150	25	5
127	Legumes and Nuts	Lentils	29	150	18	5
128	Legumes and Nuts	Soya beans	18	150	6	1
129	Snack foods and confectionary	Burger Rings, barbeque-flavored	90	50	31	28
130	Snack foods and confectionary	Chocolate, milk, plain	43	50	28	12
131	Snack foods and confectionary	Cashew nuts	22	50	13	3
132	Snack foods and confectionary	Peanuts	14	50	6	1
133	Snack foods and confectionary	Popcorn	72	20	11	8
134	Snack foods and confectionary	Potato Crips	54	50	21	11
135	Snack foods and confectionary	Black beans soup	64	250ml	27	17
136	Snack foods and confectionary	Lentil, canned soup	44	250ml	21	9

137	Snack foods and confectionary	Noodle soup	1	250ml	9	0
138	Sugar and Sugar Alcohols	Blue agave cactus nectar, high-fructose	11	10	8	1
139	Sugar and Sugar Alcohols	Fructose 50g-portion	19	10	10	2
140	Sugar and Sugar Alcohols	Glucose 50-g portion	99	10	10	10
141	Sugar and Sugar Alcohols	Honey	55	25	18	10
142	Sugar and Sugar Alcohols	Lactose 50-g portion	46	10	10	5
143	Sugar and Sugar Alcohols	Maltose 50-g portion	105	10	10	5
144	Sugar and Sugar Alcohols	Sucrose 50-g portion	68	10	10	7
145	Vegetables	Broad beans	79	80	11	9
146	Vegetables	Green peas	48	80	7	3
147	Vegetables	Pumpkin	75	80	4	3
148	Vegetables	Beetroot	64	80	7	5
149	Vegetables	Carrots	47	80	6	3
150	Vegetables	Cassava, boiled with salt	46	100	27	12
151	Vegetables	Baked potatoes	60	150	30	18
152	Vegetables	Boiled potatoes	24	150	28	7
153	Vegetables	French fries	75	150	29	22
154	Vegetables	Instant mashed potato	85	150	20	17
155	Vegetables	Steamed potato	65	150	27	18
156	Vegetables	sweet potato	61	150	28	17
157	Vegetables	Yam	37	150	36	13

(Foster-Powell et al. 2002).

Source: Extracted from *International Table of Glycemic Index and Glycemic load value*

Appendix B: Frequency of a word given the class

Beve Vege s food	Juice rage and	Brea s table s	kfast ds u	Brea	Cere al c s	Dair Coo Grai S	Fruit Crac kies s	Legu y kers	Snac and Prod	Suga mes F	Cake k and	r	
3.0	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	111
<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	14 ace= 1>	ace= 1>	3.0 ace= 1>	ace=
<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	25 ace= 1>	ace= 1>	ace= 1>	8.0 6.0 ace=
<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	25 ace= 1>	5.0 8.0 ace= 1>	3.0 ace= 1>	ace= 1>	ace= 1>	8.0 5.0 ace= 1>	4.0 ace= 0 ml
<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	20 ace= 1>	ace= 6.0 ace= 0 ml
4.0	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	6.0	80
3.0	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	60
<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	50 ace= 1>	ace= 1>	ace= 1>	4.0 ace= 6.0 ace=
<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	14.0	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	9.0	<lapl ace= 1>	<lapl ace= 1>	8.0	150
<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	30 ace= 1>	ace= 1>	18.0 12.0 ace= 1>	ace= 4.0 ace=
<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	12 ace= 1>	ace= 1>	ace= 1>	ace= 18.0 ace= 0
<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	<lapl ace= 1>	10 ace= 1>	ace= 1>	ace= 1>	ace= 7.0 ace=