

DIET ANALYSIS FOR DIABETES USING MACHINE LEARNING

**Dissertation submitted in part fulfilment of the requirements for the degree of
Masters in Data Analytics**

At



ASHOK KUMAR MOVVA

10543339

MSc. Data Analytics

January 2021

Declaration

I, Ashok Kumar Movva (10543339), a student of Dublin Business School pursuing Masters of Science in Data Analytics declare that the work in this thesis titled “Diet Analysis for Diabetes Using Machine Learning” has been carried out by me in the Department Of Data Analytics. The information derived from the literature has been duly acknowledged in the text and a list of references provided. No part of this thesis was previously presented for another degree at this or any other institution.

Signed: Ashok Kumar Movva

Date: 11th January 2021

Acknowledgement

I would like to express sincere gratitude to my project guide Mrs. Shubham Sharma of Dublin Business School, for giving me invaluable support and guidance throughout this research. Her vision, and motivation have deeply inspired me, she has taught me methodology to carry out this research and to present it as clearly as possible. It was a great privilege and honor to work under her guidance. I would also like to thank her friendship, empathy and great sense of humor.

I am extending my heartfelt thanks to Mr. Abhishek Kaushik for his thoughtful insights during the discussion of the research I had with him on this work and thesis preparation.

I am extremely grateful to my mom for her love, caring during tough times, and special thanks to my dad for his interest shown to support me to do this research and to complete it successfully.

Contents

Introduction	2
1.1 Research Question	3
1.2 Aim and Objective	3
2. Literature Review	4
2.1 Blood Sugar Level Forecast	4
2.1.1 CGMS.....	4
2.2. Analysis on diabetes predication on different datasets	5
2.2.1 Pima Indian Diabetes	6
2.2.3 BGL diagnosis	6
2.2.4 Neuro-fuzzy Framework.....	7
2.2.5 Data Mining Strategies.....	7
2.2.6 Diabetes Prediction Using Health Risk Assessment (HRA) Questionnaires	8
2.3 Diabetes	9
2.3.1 Overview of Diabetes.....	10
2.3.2 Diabetes Support System Project	11
2.4 The Preliminary Study	12
2.4.1 The Second Study.....	12
2.4.2 The Third Study	13
2.5 Machine Learning.....	14
2.5.1	14
Support Vector Machine	14
3. Methodology.....	15
3.1 Dataset Selection	15
3.2 Data preprocessing	16
3.2.1 Missing Values Removal.....	16
3.3 Leveraging Machine Learning	17
3.4 Support Vector Machine	17
3.5 K-Nearest Neighbor.....	18
3.6 Logistic Regression	18
3.7 Random Forest.....	18
3.8 Model building	19

4.	Result Discussion.....	20
4.1	Sugar level and level of risk.....	20
4.2	Statistical Analysis.....	21
4.2.1	Analyzing how low spiciness influences blood sugar level	21
4.2.2	Analyzing how medium spiciness influences blood sugar level.....	23
4.2.3	Analyzing how high spiciness influences blood sugar level	24
4.3	Food distribution based on spiciness.....	25
4.4	Food consumed during morning and evening	26
4.5	Modelling	29
4.6	Finding outliers in the dataset	29
4.6	Machine learning algorithms and accuracies	30
4.6.1	KNN	30
4.6.2	Logistic Regression.....	31
4.6.3	Random Forest.....	31
4.6.4	SVM	32
	Output.....	32
5.	Statistical Testing.	33
6.	Discussion.....	34
7.	Limitations of work	35
8.	Conclusion and Future work	36
9.	Reference	37
	Appendix	40
	Appendix A.....	40
	Appendix B	41

List of figures

Figure 1: The process of metabolism (Péter Gyuk, 2019)	2
Figure 2: Sugar and spice level in the morning food	16
Figure 3: Blood sugar level and risk level	20
Figure 4: Blood sugar chart	21
Figure 5: Low spiciness vs high sugar level	22
Figure 6: Low spiciness food vs Normal/medium sugar level	22
Figure 7: Medium spiciness food vs high sugar level	23
Figure 8: Medium spiciness food vs normal/medium sugar level	24
Figure 9: High spiciness vs sugar level	25
Figure 10: Spice consumption based on the levels	25
Figure 11: Food consumed during morning	26
Figure 12: Sugar level in the morning food	26
Figure 13: Morning spiciness vs natural sugar level	27
Figure 14: Food consumed during evening	27
Figure 15: Natural sugar level in the evening	28
Figure 16: Food spiciness vs natural sugar in the evening food	28
Figure 17: Data outliers of the dataset	30

Abstract

The latest advancement in health sciences have prompted a need for creation of data, for example, health treatment information, produced in large volumes of health records. Machine learning techniques seems to be increasing every day, like never before, the motivation behind this work is to change all accessible data into significant data. Diabetes mellitus is a type of metabolic problem, creating an impact on human health around the world and the main cause for this is hereditary. Patients should know how much sugar content present in their meal and what provokes the sugar level. The motive of this thesis is to analyze the data and use machine learning to understand regarding 1) how spicy levels and natural sugars impacted sugar levels 2) how can a food impact sugar levels 3) Comparison of results with different classification algorithms. The best accuracy was obtained from SVM to classify the sugar level present in food.

Introduction

Diabetes mellitus, a metabolic illness, is a serious issue in present day medical care, as of now it hits over 8% of population in the age group of 20-79 years (Guariguata, L, 2014). Recent studies predict that the numbers may increase by 55 percent in next twenty years (Guariguata, L, 2014), this could eventually increase the rate of death and could add up difficulties by this illness. This proves the significance of finding a better approach to lead a diabetic life and improve treatment strategies. Current status of facilities and treatment in health sectors doesn't give a complete solution or a possible treatment for curing the disease, the patients need to live an uncommon way of life with an effective treatment for different sort of diabetes; There are three types of diabetes Type1 (T1D), Type2 (T2D) and Gestational diabetes, and these metabolic disorders can be described as, T1D is total lack of insulin production by pancreatic glands, and T2D is inability to produce sufficient amount of required insulin (Guariguata, L, 2014). This work focuses on identifying the sugar levels in food intake and also to instruct the patient about sugar levels that food contains. To perform this experiment, the data is collected from diabetes patient and also took survey on food intake, this helped to identify what kind of food causes the increase in the blood sugar level. Leveraging machine learning algorithms, the research is performed to find the best model that is suitable for this process by comparing the accuracies of different algorithms.

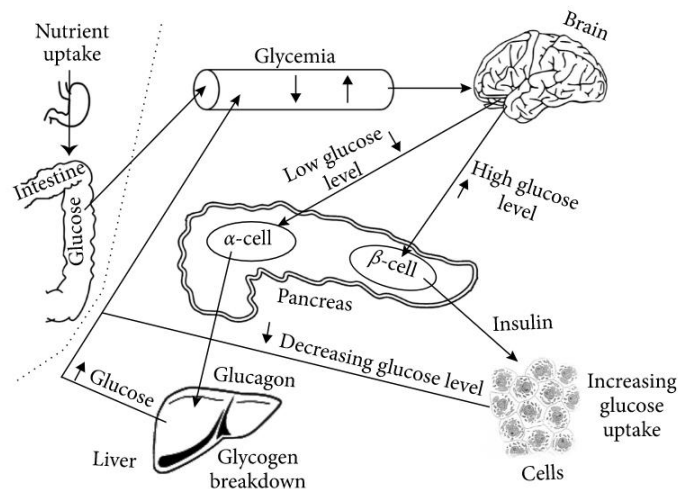


Figure 1: The process of metabolism (Péter Gyuk, 2019)

Checking blood sugar levels before each meal and different exercises is a day by day task for every diabetic patient (Florkowski, C., 2013). These methods are normally found on experiments and

predictions, which is now not so useful, bringing about high glycated hemoglobin (HbA1c) values. These perceptions helps in creating blood glucose forecast algorithms (Nicholas, J., 2013) (American Diabetes Association, 2015).

1.1 Research Question

- i. How spice levels can impact sugar levels in the food and what are the suitable method to identify it?
- ii. How natural sugars in the meal impact the blood sugar levels?
- iii. How to choose a suitable algorithm from the comparison of results with different classification algorithms?

1.2 Aim and Objective

The main aim of this research was to develop a model that can predict blood glucose development dependent on the consumption of the food in everyday life. Hence this research has been done based on the below aim and objectives.

- i. To identify the blood sugar level in the everyday food consumption of the diabetes patient
- ii. To detect the impact of spiciness in high and low blood sugar level
- iii. To effectively utilize machine learning algorithms for this process and identify the suitable algorithm
- iv. To develop a machine learning model for this process that can be effectively used in future applications

2. Literature Review

2.1 Blood Sugar Level Forecast

Eskaf, K., explained few strategies for BGL forecast found in the literature utilizing various kinds of numerical models and parameter ID techniques (Eskaf, K, 2014), yet many of these classification has impacted some significant factors, an example, food intake or activity., the result and findings of their work can be contrasted with outcomes of current work, the framework utilized by Stahl and Johansson (Ståhl, F, 2009) comprises of three sections that are independently displayed with linear models ,the framework proposed by the authors doesn't display the digestion, however the sugar intake was assessed for every meal.

2.1.1 CGMS

The BGL input data came from a sort 1 diabetic patient utilizing MiniMed Continuous Glucose Monitoring System (CGMS) during a 6-month time frame. In one of their different works (Ståhl, F., 2010), they utilized limited drive reaction models on 18 patients to appraise postprandial plasma glucose level. For the assessment, they utilized Clarke's Error Grid Analysis (EGA) (Ståhl, F., 2010). Robertson et al. (Robertson, 2011) exhibited Elman's repetitive artificial neural network (ANN) for meal and insulin intake. The data index began from a free, counterfeit numerical diabetes test system considered AIDA that demonstrated 28 days of estimations of a T1D quiet. With respect to meal intake, just starch amounts were considered, and the outcomes depend on the very restricted food retention displaying abilities of AIDA (Robertson, 2011). Another, neural organization-based arrangement is introduced by Shanthi and Kumar (Shanthi, S, 2012). The contrast between their work and the recently referenced ANN-based tests for this situation, the approval data history remembered patients for a medical clinic setting with various insulin treatments utilizing Medtronic's CGMS.

2.1.2 Different models proposed for BGL The point of the analysis detailed by Plis et al. (Plis, 2014) is to stay away from hypoglycemia during 30 minutes with BGL forecast. They utilized the support vector regression (SVR) and AutoRegressive Integrated Moving Average (ARIMA) models. The boundary ID was performed with an all-encompassing Kalman channel (Simon, D., 2006.). The strategy and the approval by Khaled et al. (Eskaf, K., 2014) completed a 30-minute

BGL forecast utilizing hereditary algorithms. The data collection was 1-hour long, and the approval proportion of the data is 2:1. The contribution of the approval came part from the AIDA test system and part from volunteers. The meal consumption for outpatients was demonstrated as a bolus infusion of glucose. There are different techniques detailed in the past research also (Chuah, Z.M., 2010), however they normally have more weaknesses as they utilize replicated input data or basic models that are less usable in outpatient care. The greatest deficiency of the referenced models is the absence of taking care of complex supplement intake and glucose digestion. Thus, despite the fact that there is promising classification, the ideal setting isn't yet found, and subsequently our vision is to discover or draw nearer to a stunningly better model.

2.2. Analysis on diabetes predication on different datasets

The various analysis has been conducted to obtain positive results on various tending datasets, for diabetics' predication and the diet plan associated to it. Various prediction models are created and upheld by various analyst mistreatment variations of data mining procedures, machine learning algorithm also have been blended for those procedures.

Dr Saravana Kumar N M, Eswari, Sam path P and Lavanya S (2015) leveraged framework Hadoop and scale back procedure for prediction of Diabetic data. This technique predicts type of diabetes and furthermore issues identified with it. The framework is Hadoop essentially based and is prudent for any organization.

Aiswarya Iyer (2015) used classification technique to check hidden patterns in diabetes dataset. Naïve Thomas Bayes and Decision Trees were utilized in this model (Iyer, A. 2015). Correlation was made for output of every algorithms and adequacy of every algorithms was appeared as a result. (Rajesh, K, 2019) K. Rajesh and V. Sangeetha (2012) utilized order strategy. They utilized C4.5 Decision Tree algorithm to extract the hidden features from dataset for grouping with efficiency. Humar Kahramanli and NovruzAllahverdi (2008) utilized artificial neural networks along with formal rationale to foresee diabetes, (Kahramanli, H, 2008). B.M. Patil, R.C. Joshi and Hindu divinity Toshniwal (2010) projected Hybrid Prediction Model which consolidates clear K-implies bunch algorithmic program, trailed by use of grouping algorithmic program to the outcome acquired from bunch algorithmic program. In order to make classifiers C4.5 Decision Tree algorithmic program is employed.

2.2.1 Pima Indian Diabetes

The majority of the business related to machine learning in the area of diabetes finding has focused on the analysis of the Pima Indian Diabetes dataset in the UCI archive. In this specific situation, Shanker (Shanker, M.S., 1996) utilized neural organizations to foresee the beginning of diabetes mellitus among the Pima Indian female population close to Phoenix, Arizona. This specific dataset has been generally utilized in machine learning analyzes and is as of now accessible by UCI dataset repositories. This analysis has been conducted and monitored ceaselessly by the National Institute of Diabetes, Digestive and Kidney Diseases inferable from the high occurrence of diabetes. The analysis picked 8 specific factors which were considered as risky factors in the event of diabetes (Type1 or Type2), Gravidity(number of pregnancies), plasma glucose levels in two hour interval in an oral glucose test(OGTT), diastolic circulatory strain, two hour insulin, loss of weight, diabetes family, and so on All the 768 models were haphazardly isolated into a preparation set of 576 cases (378 subjects without diabetes and 198 subjects with diabetes) and a test set of 192 cases (122 non diabetic subjects and 70 diabetic cases). Utilizing neural organizations with one concealed layer, Shanker (Shanker, M.S., 1996) acquired a general exactness of 81.25% which was higher than the forecast precision got utilizing a calculated regression technique (79.17%) and the ADAP model (76%). Numerous different papers have revealed results on this dataset. Research on diabetes data, identified with the utilization of AI strategies, has mostly centered on attempting to foresee and screen the Blood Glucose Levels (BGL) of diabetic patients (Sandham, 2018) or conceivable health risks of such patients. In (Sandham, 2018), a mix of Artificial Neural Networks (ANN) and a Neuro-Fuzzy Optimizer was utilized to predict the BGL of a diabetic patient in the new future and afterward a potential timetable of diet and exercise just as the dose of insulin for the patient was recommended. Despite the fact that the BGL forecasts were near the actual readings, the dataset was confined to just two Type 1 diabetic patients, which raises questions about its ease of use for huge gatherings

2.2.3 BGL diagnosis

In another analysis, by Karim Al Jabali (El-Jabali, A.K., 2005), artificial neural organizations were utilized to show and reenact the movement of Type 1 diabetes in patients just as to foresee the ideal (or sufficient) measurement of insulin that should be conveyed to keep up

the blood glucose level (BGL). The dataset was included 70 patients with 30,000 preparing cases and the properties considered were Previous Glucose Level, Short Term, Mid Term and Long-Term Insulin discharge just as some different highlights like exercise, meal, and so forth. A back engineering neural organization with four layers was utilized to reproduce the diabetic patient's digestion and furthermore mimic the regulators conveying insulin. The outcomes demonstrated that the utilization of complex neural organization designs could adequately copy the working of regulators that convey insulin to Type 1 diabetic patients.

2.2.4 Neuro-fuzzy Framework

Neuro-Fuzzy frameworks have likewise been utilized by Dazzi et al. (Dazzi, D., 2001) for the control of BGL in basic diabetic patients, with the fundamental goal of having the option to predict the specific dose of insulin with the most un-number of obtrusive blood tests. A mix of back spread (BEP) neural organizations and fluffy rationale were utilized to foresee the variety in insulin measurement. The neural organizations were utilized to find the connections among factors and locate the correct standards and change enrollment capacities. For preparing the neural organizations, a bunch of 1000 arbitrarily reproduced BG values were utilized, and the comparing insulin mixture rates noted. The prepared neural nets were then tried with a bunch of 400 concealed BG values and the predicted insulin mixture rates were checked and used to assemble a nomogram. The Neuro-fuzzy framework had the option to give tweaked varieties in insulin imbuelement because of little glyceimic varieties and keep up BGL better than traditional control frameworks. Another territory of analysis in Type 1 diabetes, utilizing AI strategies has been in the analysis of the hereditary data related with the event of Type-1 diabetes (T1DM). Various late analysis has pointed toward unwinding the hereditary premise of T1DM with an attention on entire genome screenings of families with influenced kin sets (ASPs).

2.2.5 Data Mining Strategies.

Pociot et al., (Pociot, F., 2004), considered the utilization of data mining strategies to recognize complex connections of qualities fundamental the beginning of Type 1 diabetes (for example non-straight communications between various characteristic loci). The dataset they examined, had the hereditary data from the analysis of 318 miniature satellite markers in 331 multiplex families. The subjects included 375 ASPs, 188 unaffected sib sets, 564 grating sib sets making up an aggregate of 1586 people. Decision trees

and neural organization approaches were utilized to dissect the data. Both these procedures were not just ready to distinguish all the significant linkage tops that were recognized by other non-parametric linkage (NPL) analysis, yet in addition discovered proof of some new districts of interest that influence the beginning of diabetes on certain chromosomes. The data mining strategies demonstrated vigorous to absent and incorrect data. Besides, these methodologies could foresee the Type 1 diabetic patients from the non-diabetics, with preparing utilizing sets of mixes of less markers. This analysis additionally stressed that acquired components impact both vulnerability and protection from the illness. Linkage analysis of ASPs couldn't recognize defensive quality variations, while data mining analysis with unaffected subjects had the option to distinguish certain mix decides that happened uniquely in non-diabetics. The standards on marker cooperation were created by decision trees which were approved utilizing neural organization analysis. For tests focusing on foreseeing potential health risks of diabetic patients, the AI algorithm of decision for most analysts is affiliation rule mining.

In (Zorman, M., 2004), the authors make a similar analysis of affiliation rules and decision trees to foresee the events of specific infections common in diabetic patients. In (Hsu, W., 2000), they manage affiliation rule mining on diabetes understanding data, to think of new principles for forecast of explicit infections in such patients. A Local Causal Discovery (LCD) algorithm (Silverstein, C., 2000) is utilized to concentrate how causal structures can be resolved from association rule and general rules to plan side effects to sicknesses. Additionally, exemption rule mining prompts more valuable principles from a clinical perspective.

2.2.6 Diabetes Prediction Using Health Risk Assessment (HRA) Questionnaires

In (Park, J., 2001) the forecast of diabetes from rehashed Health Risk Appraisal (HRA) polls of the members utilizing a neural organization model was contemplated. It utilized successive 6 multilayered perceptron (SMLP) with back spread and caught the time-affectability of the risk factors as a device for expectation of diabetes among the members. A chain of command of neural organizations was utilized, where each organization yields the likelihood of a subject getting diabetes in the next year. This likelihood esteem is then taken care of forward to the following neural organization alongside the HRA records for the following year. Results show improvement in precision after some time, for example the analysis of the risk factors after some time instead of at a specific moment, yields better outcomes. With the SMLP approach, the most extreme exactness of expectation acquired was 99.3% for non-diabetics and 83.6% for diabetics at a limit

(of yield likelihood from each neural organization in the progressive system) of 20%. While (Park, J., 2001) centers around the significance of time-affectability of the risk factors in diabetes forecasts utilizing just neural organizations, our analysis analyzes decision tree learning strategies and an outfit of neural organizations applied to a particular adolescent diabetes dataset. My analysis additionally contrasts from (Park, J., 2001) in the following point:

- (Park, J., 2001) utilized HRA records of representatives from an assembling firm with times of the subjects going from 45 to 64, while our subjects are altogether adolescents.
- The properties in the dataset in (Park, J., 2001) are general health boundaries like Body Mass Index (BMI), Alcohol Consumption, Back agony, Cholesterol, and so on which are totally not the same as the qualities that we manage, as Intravenous Glucose Tolerance, C-Peptides and other clinical tests that are explicit to Type 1 diabetes.

In the current analysis I have utilized data from the Diabetes Prevention Trial - Type 1 (Haller, M.J., 2015), which was the primary enormous scope preliminary in North America intended to test whether mediation with antigen-based treatments, parenteral insulin and oral insulin would forestall or postpone the beginning of diabetes. In (Chase,2001) it was indicated that a solid relationship between first-stage (1 moment + 3 moment) insulin (FPIR) creation during intravenous glucose resilience tests (IV-GTT) and risk factors for creating type 1 diabetes existed utilizing the DPT-1 data. In (Greenbaum, 2001) the asymptotic gathering of cases in the DPT-1 preliminary whose diabetes could be straightforwardly analyzed by the 2-h models on Oral Glucose Tolerance Test (OGTT) was considered. Both these analysis (Haller, M.J., 2015), distinguished the tests I utilized for our preparation data.

2.3 Diabetes

As indicated by the American Diabetes Association (ADA), "Diabetes mellitus (MEL-ih-tus), or essentially, diabetes, is a gathering of infections described by high blood glucose levels that outcome from deserts in the body's capacity to create or potentially use insulin" (American Diabetes Association, 2010a). Starting at 2011, 10.9 million individuals age 65 years or more seasoned (26.9%), and 25.6 million individuals age 20 or more established (11.3%) in the United States have diabetes mellitus (American Diabetes Association, 2011). The absolute expense

of diabetes in the United States for 2007 was \$174 billion dollars (American Diabetes Association, 2011). There are three types of diabetes out of which two types of diabetes are predominant, Type 1 (T1DM) and Type 2 (T2DM) and third type of diabetes is gestational diabetes. Type1 diabetes happens when the body can't or at a point unable to deliver insulin. Beginning of Type 1 diabetes is normal in adolescence; this illness used to be known as adolescent diabetes. This type of diabetes is more uncommon; just around 5-10% of individuals with diabetes have T1DM (American Diabetes Association, 2010b). T2DM happens when the body can't use the insulin delivered or insufficient insulin is created. T2DM is usually connected with heftiness; notwithstanding, stoutness isn't the solitary high-hazard factor. Certain nationalities are viewed as high risk gatherings, as enormous rates of those identities have diabetes (American Diabetes Association, 2010c).

2.3.1 Overview of Diabetes

Patients with diabetes need to control their blood glucose levels. Insulin or other medicine might be utilized to control blood glucose levels. In the event that blood glucose levels are not sufficiently controlled, the drawn-out difficulties can be very expensive as far as both health and money. Such intricacies incorporate expanded risk for coronary illness and stroke, visual impairment, kidney disappointment, and even demise (American Diabetes Association, 2011). Regularly, patients with T2DM can control their blood glucose levels through medicine, work out, and appropriate eating routine. Patients with T1DM expect insulin to endure, either from injection or from a siphon (Centers for Disease Control and Prevention, 2007). Numerous patients with T1DM utilize an insulin siphon related to Continuous Glucose Monitoring (CGM). The insulin siphon permits the patient to control any measure of insulin. For patients utilizing Medtronic siphons, this sum is picked with the assistance of the Bolus Wizard. There are numerous variables which impact the viability of insulin for the patient that the Bolus Wizard considers. Insulin affectability, which shifts from patient to understanding, is a proportion of the patient's responsiveness to insulin. The carb proportion, which is additionally persistent explicit, depicts the measure of insulin needed to cover sugars for a meal. While computing a suggested bolus sum, the Bolus Wizard utilizes the insulin affectability and carb proportion boundaries, alongside a new bolus history and the current blood glucose perusing. There are two significant issues in blood glucose control: hyperglycemia and hypoglycemia. Hyperglycemia, or high blood glucose levels, happens during diabetes without treatment.

In T1DM, it is articulated when the insulin siphon comes up short or when the patient doesn't regulate enough insulin. Hypoglycemia, or low blood glucose levels, happens when the patient overuses a lot of insulin. Late research shows that glycemic inconstancy, or variance among highs and lows, is a third issue adding to expanded risk of confusions (Ceriello and Ihnat, 2010; Hirsch and Brownlee, 2005; Kilpatrick et al., 2010; Kilpatrick et al., 2006; Monnier and Colette, 2008; Monnier et al., 2006). Patients with diabetes should consistently screen their blood glucose levels utilizing fingerstick. A fingerstick is acquired by drawing a limited quantity of blood for analysis by an individual glucose meter. Fingerstick are utilized by the Bolus Wizard® for suggesting insulin doses. For patients utilizing CGM, fingerstick are utilized to adjust the sensor. Patients are encouraged to adjust their sensor three times each day. This alignment may cause discontinuities in the CGM values. Nonetheless, fingerstick data is more exact than CGM data and is depended upon when readings oppose this idea. The CGM sensor records test blood glucose esteems at regular intervals, which permits the patient to intently screen their blood glucose levels. The CGM sensor slacks the actual blood glucose esteems by 10 to 15 minutes, giving qualities inside $\pm 20\%$ of the real qualities. (Mastrototaro et al., 2008). The arrangement of utilizing a CGM sensor and an insulin siphon to control blood glucose esteems is open circle; the patient should mediate with the framework for everything to be in motion. Shutting the circle with a counterfeit pancreas is a thought proposed by Dr. Arnold Kadish that goes back to 1964 (Juvenile Diabetes Research Foundation, 2010). On the off chance that a counterfeit pancreas could supply the patient with insulin with the end goal that the framework would not reason hypoglycemia or hyperglycemia, at that point it is conceivable to construct a shut circle framework. In any case, the significant test to building a shut circle framework is the elements of the viability of insulin. Each patient responds contrastingly to insulin. Indeed, even a similar patient may respond distinctively to insulin at various occasions. Elements known to impact the viability of insulin incorporate exercise, diet, stress and other life occasions. These components present numerous difficulties to open circle, just as shut circle, control.

2.3.2 Diabetes Support System Project

The work portrayed in this postulation was led inside the setting of the 4 Diabetes Support System (4DSS) venture. The 4DSS is a case-based decision emotionally supportive network intended to encourage both doctor and patient administration of T1DM. The 4DSS finishes this undertaking in 3 stages: recognizing issues in blood glucose control, producing answers for these

identified issues, and recalling which classification worked for future reference. 4DSS innovative work has been directed throughout the span of three clinical analysis contemplates, the third is as yet progressing. These analysis are portrayed straightaway.

2.4 The Preliminary Study

The reason for the main 4DSS analysis was to decide whether a decision emotionally supportive network could be created to help oversee patients with T1DM. Before the analysis was directed, it was endorsed by the Institutional Review Board (IRB) at Ohio University. Twenty human subjects with T1DM enlisted for a time of about a month and a half for every subject, and 12 subjects finished the whole convention. A collection of patient data was gathered, including: foundation data, insulin siphon data, CGM data, and everyday life occasion data. Day by day life occasion data included supper data, rest data, work data, stress, sickness, and other various data. Also, every patient rounded out a leave overview toward the finish of their investment in the analysis. Utilizing the gathered data, a 4DSS model was worked by data designs and assessed by both diabetes specialists and data engineers. This analysis demonstrated that a decision emotionally supportive network for T1DM would be practical. It recognized the necessities to address extra issues in blood glucose control and to diminish data passage time for patients (Marling et al., 2008;). Toward the finish of the starter study, the 4DSS model comprised of four distinct modules and a case base with 49 cases. These modules incorporated a site for data section, a 20 data set for recording tolerant data, circumstance appraisal for distinguishing issues, and a case recovery module for recognizing cases with comparative issues to those identified by circumstance evaluation. The site for data passage was created by Anthony (Maimone, 2006). The data base was created by Anthony Maimone with the assistance of Kathleen Evans-Romaine and Wesley Miller (Maimone, 2006). The circumstance appraisal module was created by Wesley (Miller, 2009). The case recovery module was created by Donald (Walker, 2007). The case-base was made by the diabetes specialists and data engineers, utilizing the gathered data.

2.4.1 The Second Study

The reason for the second 4DSS analysis was to assess the capacities of the circumstance evaluation and case recovery modules created during the starter study. The subsequent analysis got endorsement from the IRB at Ohio University before its start. 26 grown-up human subjects

with T1DM enlisted for a time of 5 weeks for every subject. 23 subjects finished the whole convention. Since the case base was assembled utilizing data from the primary analysis, patients who partook in the principal study didn't likewise partake in the subsequent analysis. This was done to forestall any inclination in the assessment. As in the primary analysis, understanding data was recorded in the data base (Schwartz et al., 2010). Assessment of the difficult recognition and the case recovery module were introduced in (Schwartz et al., 2010). For recognizing issues, this assessment indicated that the issues identified were right and valuable a dominant part of the time. The patients' own doctors assessed these recognitions and found that 97.9% of the identified issues were right, and 96.1% were valuable. Four diabetes specialists assessed the case recovery module. The specialists found that 79% of the cases recovered had issues that were like the difficult that was distinguished, and 82% 21 of the related classification were useful to the patients encountering the issues (Schwartz et al., 2010).

2.4.2 The Third Study

The assessment from the subsequent analysis demonstrated that there was opportunity to get better for the situation recovery module. Gathering more data and adding cases to the case base is one approach to improve this exhibition. This inspired the requirement for a third report. Likewise, with the initial two analysis, the third analysis was endorsed by the IRB at Ohio University. Up until this point, seventeen human subjects with T1DM enlisted for a time of 3 months for each subject, and twelve have finished the whole convention. This analysis prompted a few upgrades and expansions for the 4DSS venture. New cases were made from the gathered data and added to the case-base. This gave the case recovery module more cases to choose from. Be that as it may, to make recovered classification explicit to singular patients, classification should have been adjusted. This brought about the fifth module of the 4DSS task, which is transformation. This module was created by Tessa (Cooper, 2010). The motivation behind this module is to tailor the arrangement found by the case recovery module to the particular requirements of the patient. An illustration of an answer requiring transformation is one that proposes the patient should expand their basal rate before sleep time from 0.9 to 1.0 units. Notwithstanding, if the patient's present basal rate before sleep time is 0.6, changing it to 1.0 would not be ideal. The transformation module can tailor the exhortation with the end goal that a fitting basal rate is proposed.

2.5 Machine Learning

This part depicts machine learning strategies and details that were utilized for this work. These methods incorporate the machine learning algorithms Multilayer Perceptron (MP), Support Vector Machines (SVM) for order, and Support Vector Regression (SVR). The detailing of a period arrangement expectation issue is significant for predicting blood glucose esteems.

2.5.1 Support Vector Machine

SVM was first portrayed in 1979 in (Vapnik, 1979). The books (Vapnik, 1998) present a presentation and outline of SVMs. An extra ordinary instructional method on SVMs for design acknowledgment is yielded (Burgess, 1998) and the books (Theodoridis 2009) have chronicled huge numbers of the new improvements with 25 SVMs. In this work, SVMs are utilized for arranging glycemic changeability and predicting blood glucose esteems. A concise prologue to SVMs for grouping is given in the remainder of this part. Like perceptron, SVMs endeavor to discover a hyperplane that isolates the data; It is indicated that the distance between a given point x_n and the decision limit is given as: $t_n(x_n) = w \cdot \phi(x_n) + b$ where $t_n \in \{-1, 1\}$ and relates to the name for the n th model, x_n is the n th element vector,

$$\|w\| \equiv \sqrt{w_1^2 + \dots + w_n^2} \text{ and } y(x_n) = w \cdot \phi(x_n) + b$$

where $\phi(x_n)$ is a component space change and b is a balanced (Bishop, 2006). We need to discover the point x_n with the nearest opposite distance to the decision limit while upgrading the boundaries w and b to boost the distance of the edge.

SVMs were initially utilized for tackling order issues. They have since been stretched out to take care of regression and positioning issues. Regression analysis with SVMs is known as Support Vector Regression (SVR). Smola and Scholkopf have distributed a thorough instructional exercise clarifying SVR (Smola, A.J. also, Scholkopf, B., 2004). The utilization of SVR to tackle a period arrangement forecast issue has turned into a subject of interest over the previous decade (Sapankevych and Sankar, 2009). The forecast of future qualities is driven by the preparation data. In this work, SVR is utilized to predict future blood glucose esteems.

3. Methodology

The variation in glucose levels is reason for diabetes. Insulin adjusts the blood glucose level in the body, inadequacy of which cause diabetes. For the expectation of blood glucose levels AI is utilized, these have numerous means like picture pre-preparing/data preprocessing followed by an element extraction and afterward grouping. We can utilize any of the referenced AI classifiers to predict this illness. In the above area we have finding out about numerous arrangement algorithms, we can either utilize any of these to predict the disease or we can investigate the procedures to utilize the half breed philosophy to improve the precision over utilizing a solitary one.

Depending upon the application and nature of the dataset utilized we can utilize any of the algorithms referenced below. As there are various applications, we cannot decide which of the algorithms are prevalent or not. Every classifier has its own specific manner of working and classification. Each of these algorithms has been discussed in the later sections.

3.1 Dataset Selection

The Dataset has been collected from a diabetic patient (Type 2 diabetic) by recording day to day meal intake and blood sugar levels. The data has been collected over a period of 6 months. The motive behind this research is to identify the blood sugar level and the sugar level in the food consumed. Since data collected from an Asian origin individual, the dietary contains more spicy foods compared to other cuisines of the world. The intention of this work is to find the correlation between spiciness and sugar level in the food. Blood sugar levels are recorded in the morning and evening, and the meal intake is recorded in morning and evening as well, to predict at which time the sugar intake in more. Also, the amount of spiciness of consume in every meal. Below are the attributes present in the dataset.

- Date: Date in which sugar level is tested
- Morning Reading: Level of sugar tested during morning time
- Food: Food consumed for breakfast
- Morning Spicy level: Level of spiciness in breakfast
- Morning Sugar Level: Level of sugar in morning
- Morning Natural sugar: Natural sugar levels in the food

- Morning External sugar: Sugar due to external factors like artificial sweeteners.
- Evening Reading: Sugar level tested during evening time
- Evening food: Food consumed in the evening
- Evening Spicy level: Level of spiciness in evening food
- Evening Sugar Level: Level of sugar present in food consumed for dinner.
- Evening Natural sugar: Natural sugars present in food consumed in snack item or dinner
- Evening External Sugars: Sugar due to external factors like artificial sweeteners

	Date	Sugar_Reading	Food	SpicyLevel	SugarLevel	NaturalSugars	ExternalSugars	measured_time
0	2020-02-01	222.0	idly, yellow peas chutney	medium	low	no	no	morning
1	2020-02-02	237.0	chapati, potato curry	high	high	yes	no	morning
2	2020-02-03	252.0	dosa peanut chutney	medium	low	no	no	morning
3	2020-02-04	238.0	ragi dosa, chutney	high	low	no	no	morning
4	2020-02-05	222.0	pongal, sambar	medium	high	yes	no	morning

Figure 2: Sugar and spice level in the morning food

3.2 Data preprocessing

Data preprocessing is most important part of this analysis. Generally data from health care services contain missing values and different conventional values that may cause ambiguity in analysis of data. In order to improve quality and viability acquired subsequent to mining measure, data preprocessing is done on the dataset. To utilize Machine Learning techniques on the dataset feasibly a value is basic unit for precise outcome and effective forecast. For the dataset used in this research, data preprocessing is completed as two stage process.

3.2.1 Missing Values Removal

The dataset has rows that have missing values for food intake in morning or evening, and data with 0 as value in sugar level column, because having zero in the data is not useful in training the model and all these removed. Consequently, these occurrences are deleted from the dataset. By eliminating these kind of rows from the dataset is termed as highlights subset determination, this works helps in reducing the dimensionality of the data and help to work easily and quickly.

The part of data after removing all missing values (data cleansing), data is normalized for preparing and testing the machine learning model. At the point when data is divided into training and testing data, then the algorithm is trained on the training dataset and test dataset is kept aside. This preparation cycle will deliver the preparation model dependent on rationale, algorithms and values of the component in preparing data. The basic aim in normalizing the data is to eliminate the anomalies and bring all the data attributes in to one scale.

3.3 Leveraging Machine Learning

After removing missing values and normalizing the data, the data is ready to train the machine learning model, in this work diverse collection and methods are applied on the dataset to understand the impact of food on blood sugar levels. Aim of using machine learning algorithms is to analyze the data and use of strategies to find accuracy and improvement in the accuracy of the model, and furthermore to identify which foods are causing or impacting BGL. Following machine learning models are trained and tested.

3.4 Support Vector Machine

Support Vector Machine is a supervised machine learning algorithm that can be used to solve both regression and classification problems but mainly utilized for classification objectives. SVM is a machine learning method that looks at data and sorts it into one of two categories. This is one of the most effective classifier among those, which has a sort of linear. The mathematical intuition behind this support vector machine model is it has kernel functions. These functions are mainly able to handle certain cases where there is non-linearity by using this non-linear basis functions. This support vector machines have a clever way to prevent over fitting and this can work with relatively large number of features without requiring too much computation. The main objective of this support vector machine we can easily separate two classes using a hyperplane. This support vector machine makes sure that when there is a creation of hyperplane it tries to create two margin lines and these two margin lines will have some distance by which it is easy to classify the two classes. These two margin lines are parallel to the main hyper plane and this model makes sure that these lines will be passing through one of the nearest points. The distance between these two marginal lines is called as marginal distance.

3.5 K-Nearest Neighbor

K Nearest Neighbor Algorithm is a simple supervised machine learning model that utilizes entire dataset in its training place. This is mostly used for Classification models. Whenever prediction is required for unseen data, what it does is it searches through the entire training dataset for K-most similar instances and the data with the most similar instances is finally returned as prediction (that is it classifies the data points based on how its neighbors are classified). This is used generally in search applications when looking for similar items. The letter K in KNN denotes the number of nearest neighbors which are voting class of the new data or the testing data. This algorithm is based on feature similarity. Choosing the right value for K is called parameter tuning that can lead to better accuracy. Choosing the correct value for K is important because if the value is too low it leads to noise and if the value is too big then it leads to resource issue or processing issues. The common use of choosing the appropriate value for K is to the square root of n (where n is the total number of data points)

3.6 Logistic Regression

Regression analysis is a predictive modelling technique that tries to estimate the relationship between a dependent and an independent variable. This is one of the most popularly used machine learning model for binary classification. But this can be used for categorical variable (that is having more than 2 classes) which is known as multinomial logistic regression. This one is mainly used for predicting the discrete variables. Logistic Regression produces the results in a binary format which is used to predict the outcome of a categorical dependent variable. The outcome should be either in discrete or categorical form such as yes or no/ true or false etc., the logistic regression equation is derived from the straight-line equation. This logistic regression is mainly used for solving classification problems. This regression model is mainly used for predictive analysis. The main advantage of this model is it faster when compared to other classification models like kernel support vector machine etc.

3.7 Random Forest

It is sort of outfit learning strategy and furthermore utilized for arrangement and regression assignments. The precision it gives is grater at that point contrasted with different models. This strategy can undoubtedly deal with huge datasets. Random Forest is created by Leo Bremen. Random Forest Improve Performance of Decision Tree by decreasing change. It works by

developing a huge number of decision trees at preparing time and yields the class that is the method of the classes or classification or mean forecast (regression) of the individual trees

3.8 Model building

In this stage of development, the algorithms that were discussed in earlier chapters have been executed and sugar level prediction is found.

Technique of Proposed Methodology-

- Import required libraries, Import diabetes dataset.
- Pre-measure data to eliminate missing data.
- Divided the dataset into Training and Test set in the ratio of 8:2
- Algorithms like K-Nearest Neighbor, Support Vector Machine, Logistic regression, and Random Forest are used.
- Build the classifier model for the referenced machine learning algorithm dependent on training set.
- Test the Classifier model for the referenced machine learning algorithm dependent on test set.
- Perform Comparison Evaluation of the research analysis results acquired for every classifier.
- After breaking down dependent on different parameters and chose the best performing algorithm.

4. Result Discussion

Using Machine Learning Algorithms, the desired output was obtained for this dataset. The data is segregated based on the morning and evening food intake, the model identifies different parameters in the food, like, the spiciness, sugar level, natural sugar and the external sugar. In our regular dietary there are foods that contain natural sugar, which is highly dangerous for the patients to consume. Based on the food consumption the blood sugar rate changes which can be identified by the running different test. In order to prevent the patient from dangerous foods that might risk their health, it is prominent to identify the food that has high sugar level. Also, in the low sugar patients, doctor advised that they should consume more amount of sugar than the normal consumption.

4.1 Sugar level and level of risk

Diabetes comes with some extreme notable symptoms in the body, the patients must be aware what will happen if the blood sugar level increases in their body. For every level sugar the symptoms and the level risk vary. For the sugar level of 50 mg, it is considered as the low sugar level which is extremely dangerous, the patient is advised to seek medical help in this case, 90-120 is the normal blood sugar that everyone should maintain in their body. On the other hand the high sugar level which is above 240 mg, people with this level of blood sugar level should be highly diet conscious.

Fasting blood sugar level		Risk level and suggested action
0	50 mg/dl or under	Dangerously low: Seek medical attention
1	70–90 mg/dl	Possibly too low: Consume sugar upon experiencing symptoms of low blood sugar, or seek medical attention
2	90–120 mg/dl	Normal range
3	120–160 mg/dl	Medium: Seek medical attention
4	160–240 mg/dl	Too high: Work to bring down blood sugar levels
5	240–300 mg/dl	Much too high: This could be a sign of ineffective glucose management, so see a doctor
6	300 mg/dl or above	Very high: Seek immediate medical attention

Figure 3: Blood sugar level and risk level

<i>Blood Glucose Chart</i>			
<i>Mg/DL</i>	<i>Fasting</i>	<i>2-3 hours After Eating</i>	<i>After Eating</i>
<i>Normal</i>	<i>80-100</i>	<i>120-140</i>	<i>170-200</i>
<i>Impaired Glucose</i>	<i>101-125</i>	<i>140-160</i>	<i>190-230</i>
<i>Diabetic</i>	<i>126 +</i>	<i>200 +</i>	<i>220-300</i>

Figure 4: Blood sugar chart

4.2 Statistical Analysis

The statistical analysis in the project involves both morning and evening reading of the blood sugar level in the patients. The below table gives the mean, standard and maximum value of the blood sugar level.

	count	mean	std	min	25%	50%	75%	max
Morning Reading	246.0	226.203252	15.010179	189.0	215.0	225.5	235.0	270.0
Evening Reading	249.0	250.128514	16.725956	190.0	238.0	248.0	262.0	310.0

4.2.1 Analyzing how low spiciness influences blood sugar level

To analyze the spice in the food, this research analyzed the food intake and noted the spice composition in those food which is depicted in the below chart. The spice composition is portrayed on the bases of very good. Based on the below chart found that low spiciness food such as brown rice, vegetables, black lentils, yellow peas leads to high sugar level.

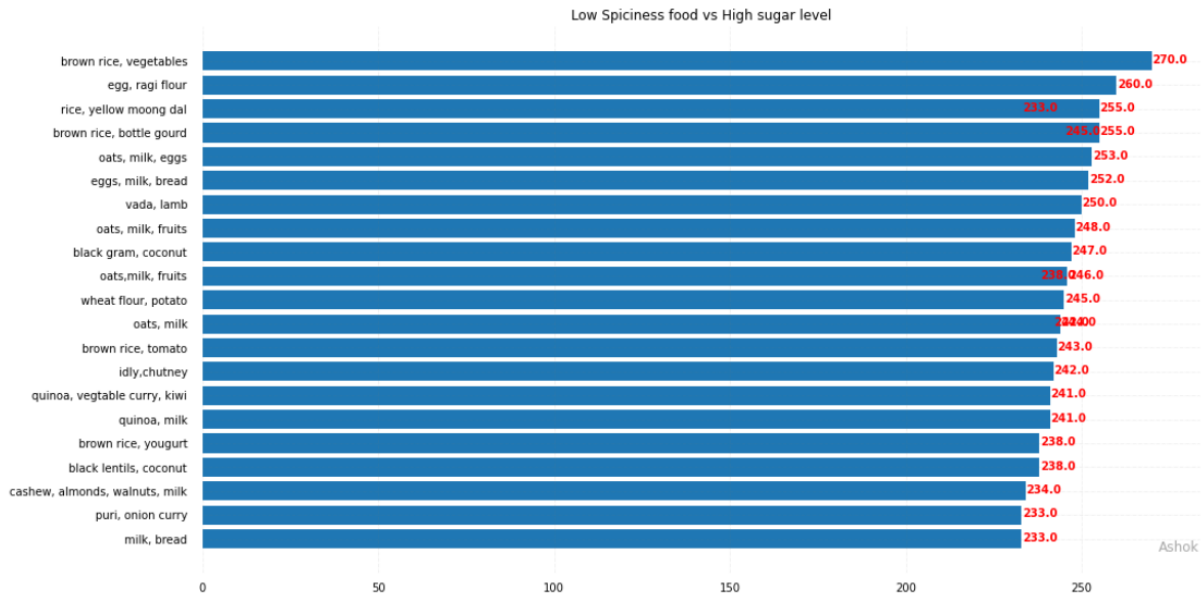


Figure 5: Low spiciness vs high sugar level

Like the above graph, it is necessary to calculate the low spiciness vs normal sugar level, in order to advice the people to consume the right food for their blood sugar level. Based on the above chart found that low spiciness food such as upma, idly, dosa, tomato rice would help to maintain normal sugar level as per the blood sugar chart shown above

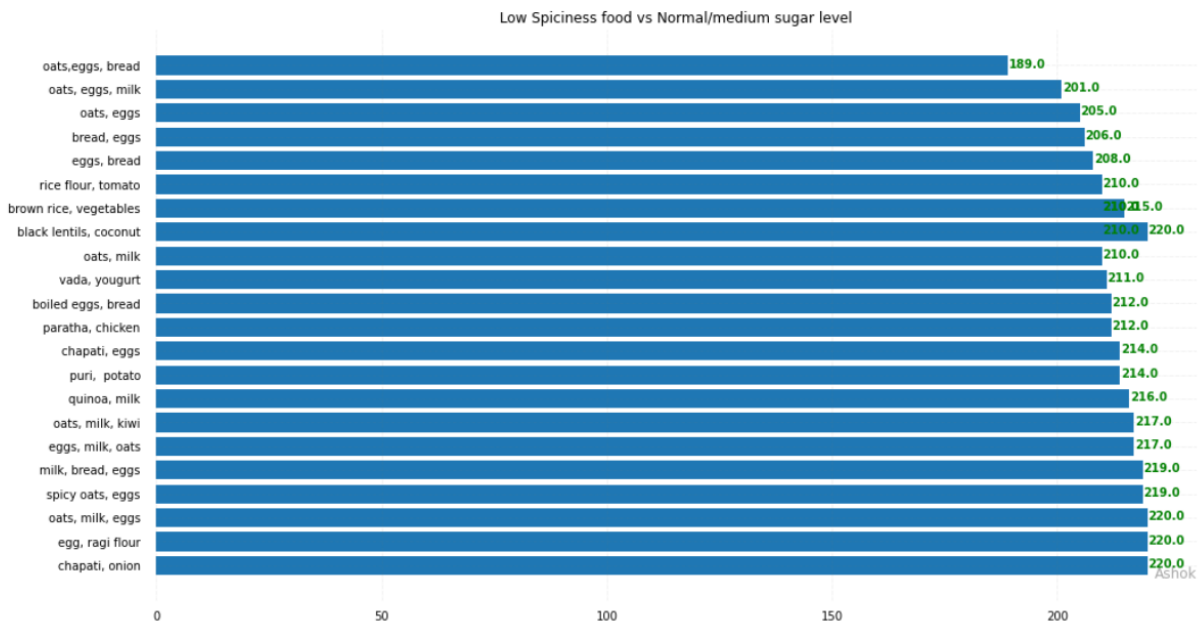


Figure 6: Low spiciness food vs Normal/medium sugar level

4.2.2 Analyzing how medium spiciness influences blood sugar level

The below chart represents the classification of medium spiciness vs high sugar level in the food intake and Based on the above chart found that black lentils,yellow peas,dosa ,chapathi rises sugar level to high

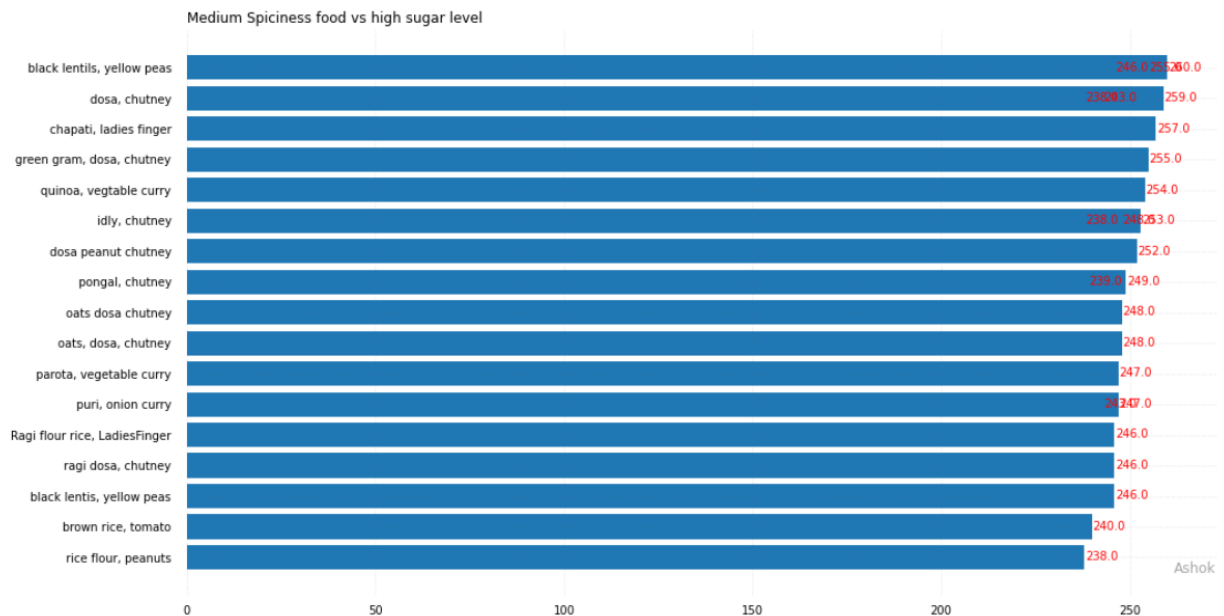


Figure 7: Medium spiciness food vs high sugar level

Like the above chart, I have predicted the medium spiciness vs normal sugar level present in the meal. Based on the above chart combination of (oats,egg,bread),(pongal sambar), (oats,egg) helps to maintain normal sugar level.

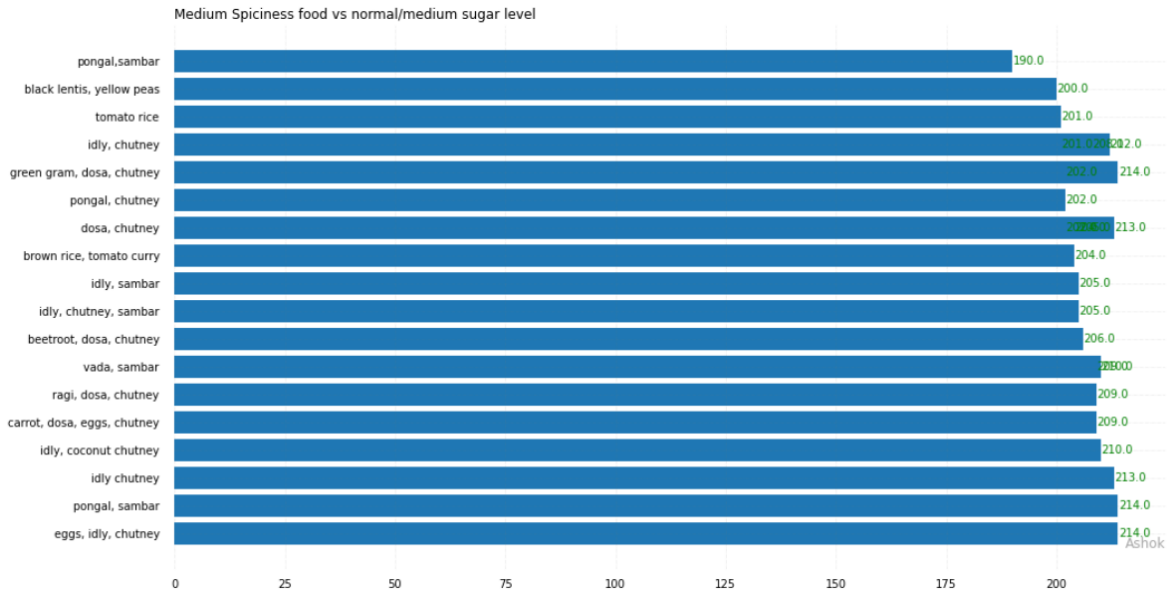
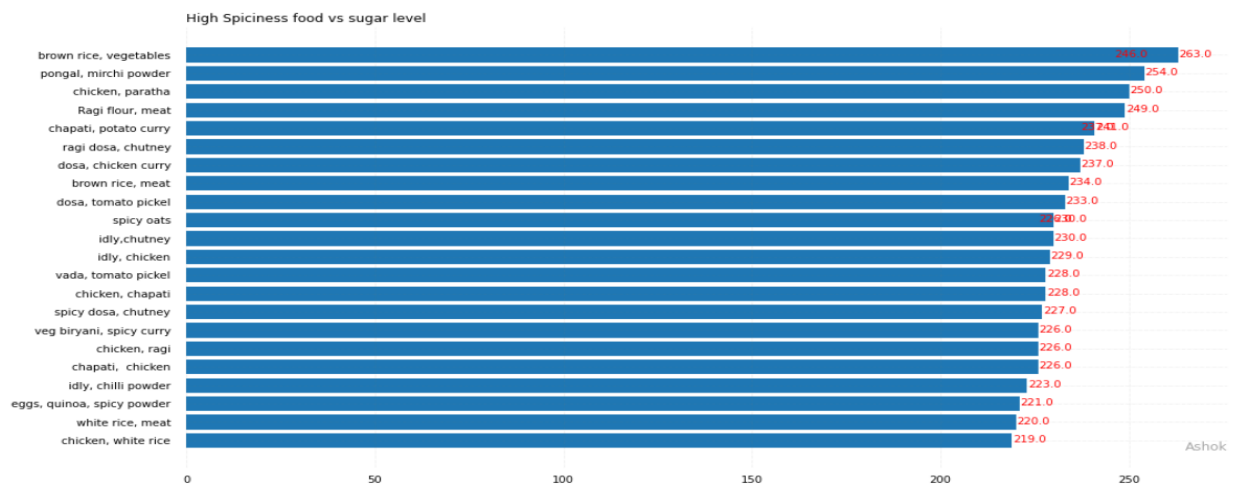


Figure 8: Medium spiciness food vs normal/medium sugar level

4.2.3 Analyzing how high spiciness influences blood sugar level

The below graphs depict the graph for high spiciness vs sugar level in the food. This is most prominent area compared to other two. As per various reports it is donated a high spice is directly proportional to the high sugar level. From observation and based on the below two charts found that high spiciness food lead to high sugar level



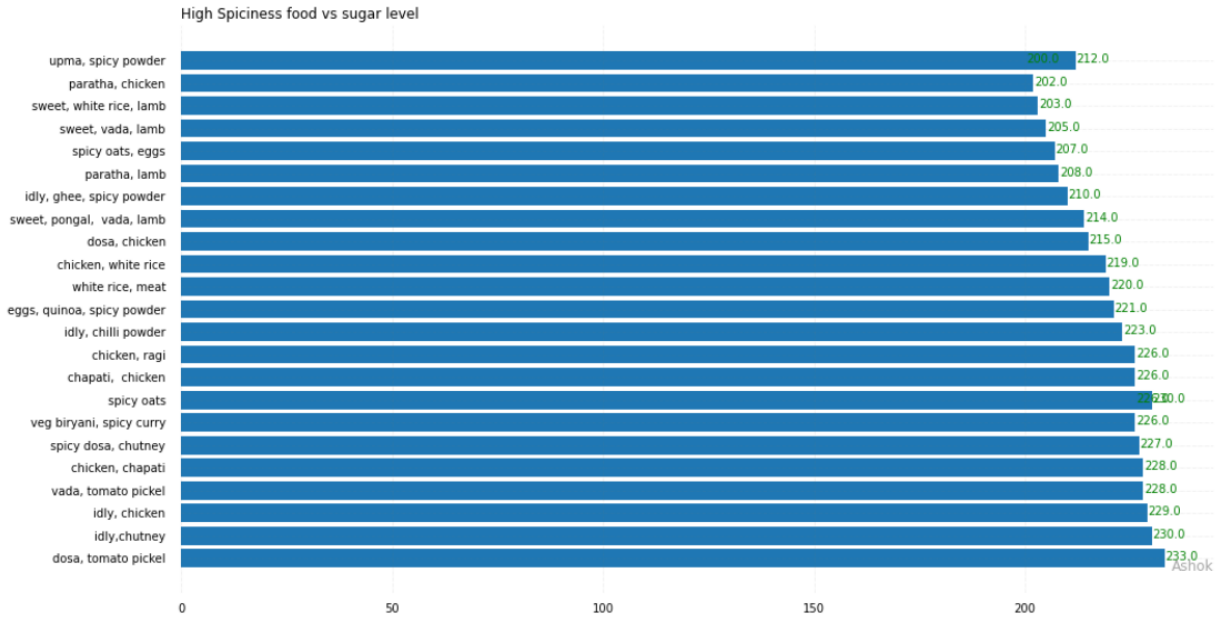


Figure 9: High spiciness vs sugar level

4.3 Food distribution based on spiciness

To recommend a proper diet to the patients, the amount of spice consumed by the patients now has to be identified. From consolidating the data from the dataset, In the following pie chart the values are depicted, which says that diabetic patient prefer medium and low spicy food.

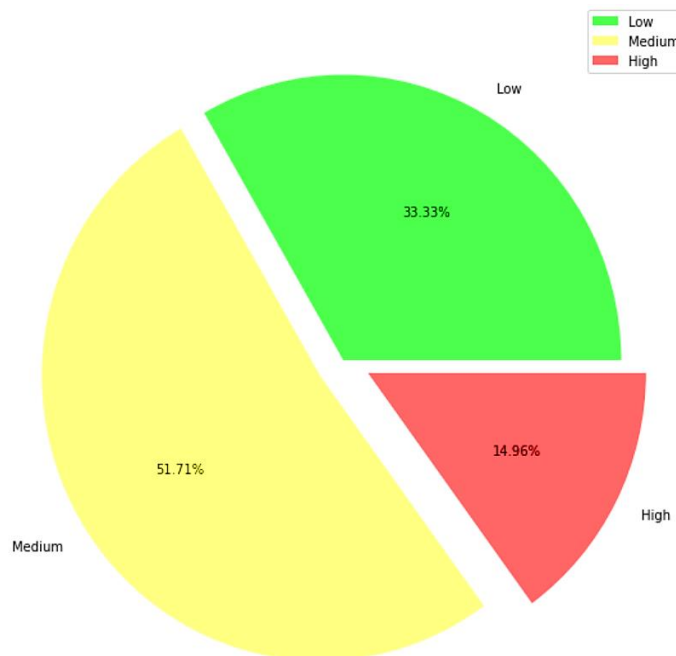


Figure 10: Spice consumption based on the levels

4.4 Food consumed during morning and evening

In this section I have segregated the food based on the time of consumption, like in the morning and evening. I analyzed type of food people intake in the morning and evening. This was helpful in identifying the sugar and spice intake. Generally, people are advised to take a good healthy meal in the morning and a light meal in the evening. But in reality, people do it in reverse action which eventually increases the blood sugar level and leads to various health composition.

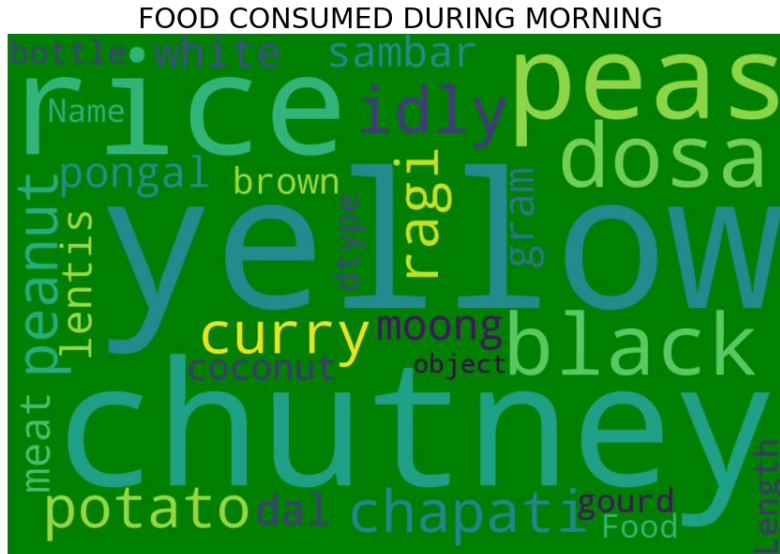


Figure 11: Food consumed during morning

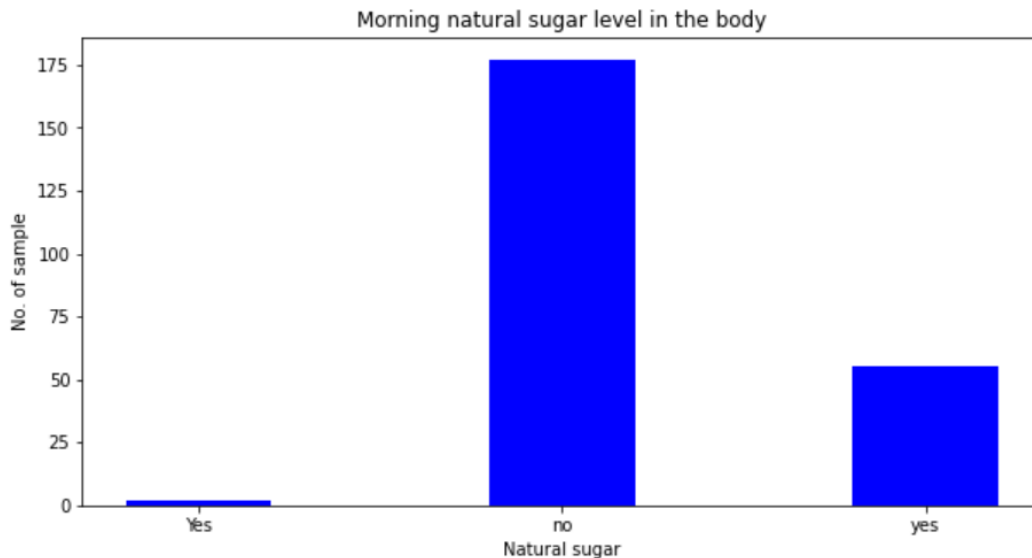


Figure 12: Sugar level in the morning food

Based on the above chart found that 85% of the blood sample doesn't have natural sugar level

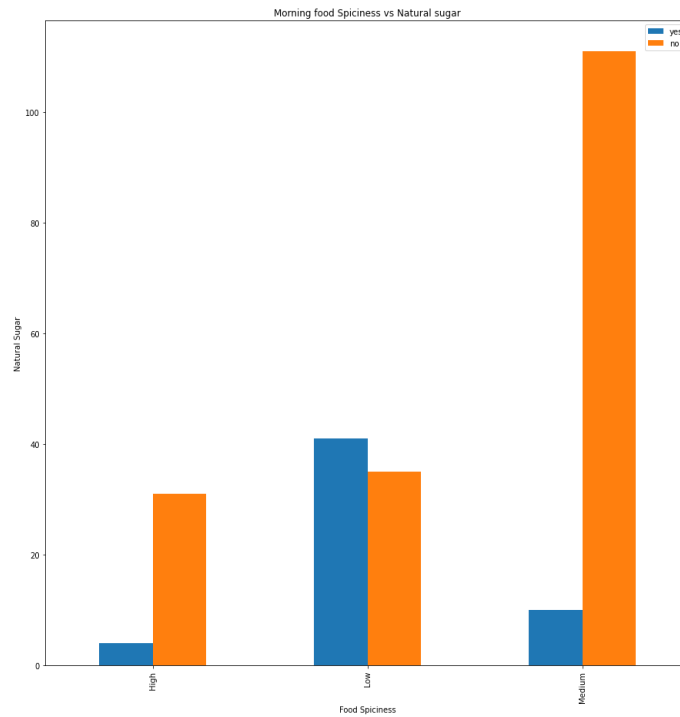


Figure 13: Morning spiciness vs natural sugar level

Food consumed in the evening

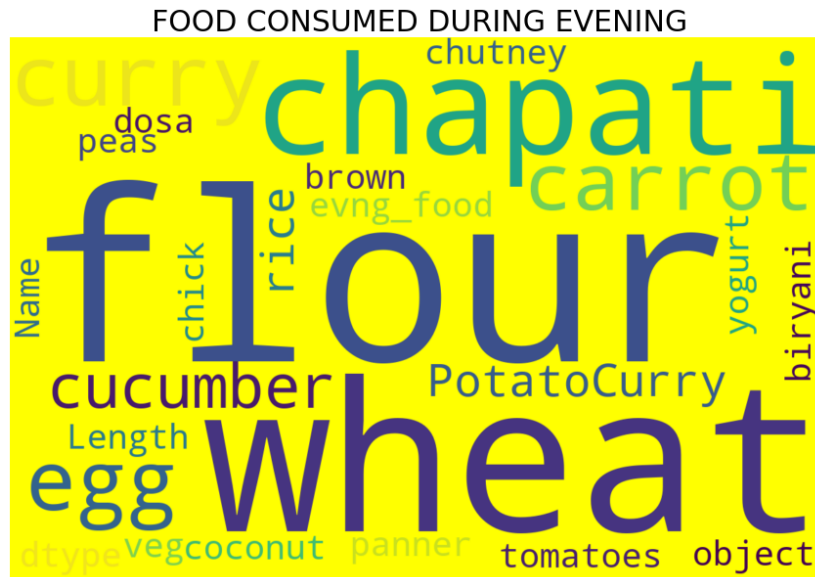


Figure 14: Food consumed during evening

Analyzing evening natural sugar level

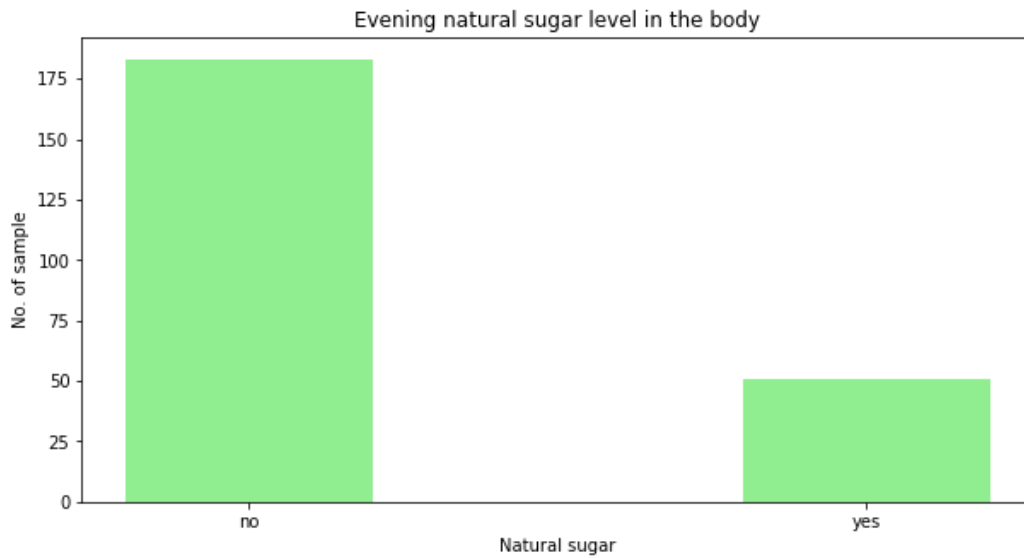


Figure 15: Natural sugar level in the evening

Based on the above chart found that 85% of the blood sample doesn't have natural sugar level. Based on natural sugar level chart it can be predicted that morning and evening follows similar pattern

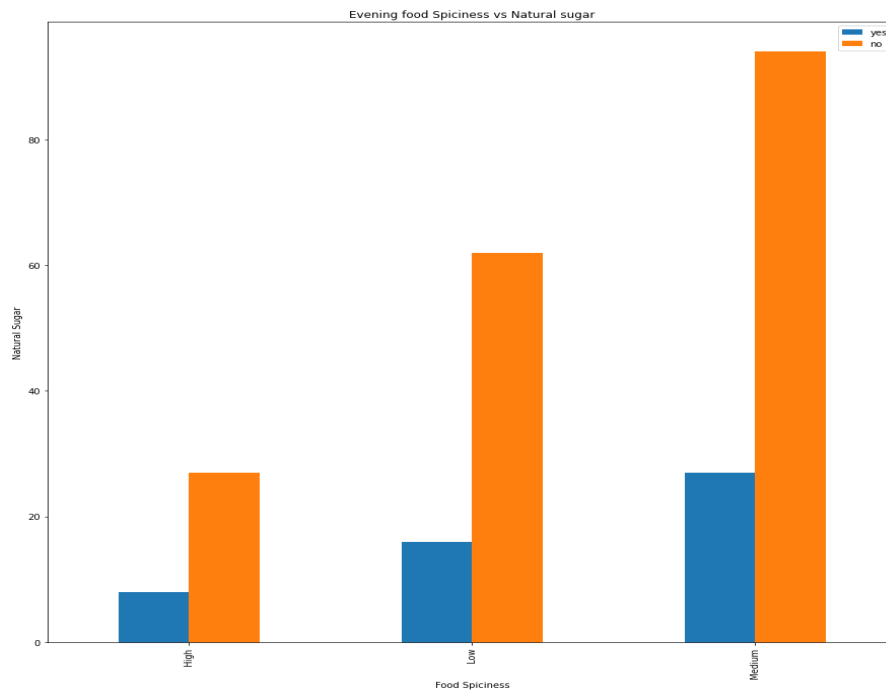


Figure 16: Food spiciness vs natural sugar in the evening food

4.5 Modelling

	Date	Sugar_Reading	Food	SpicyLevel	SugarLevel	NaturalSugars	ExternalSugars	measured_time
0	2020-02-01	222.0	idly, yellow peas chutney	medium	low	no	no	morning
1	2020-02-02	237.0	chapati, potato curry	high	high	yes	no	morning
2	2020-02-03	252.0	dosa peanut chutney	medium	low	no	no	morning
3	2020-02-04	238.0	ragi dosa, chutney	high	low	no	no	morning
4	2020-02-05	222.0	pongal, sambar	medium	high	yes	no	morning

Data cleaning

Row count before data cleaning 470

Row count after data cleaning 453

4.6 Finding outliers in the dataset

Based on the below chart find that blood sugar level less than 200 and sugar level greater than 280 are found to be outliers

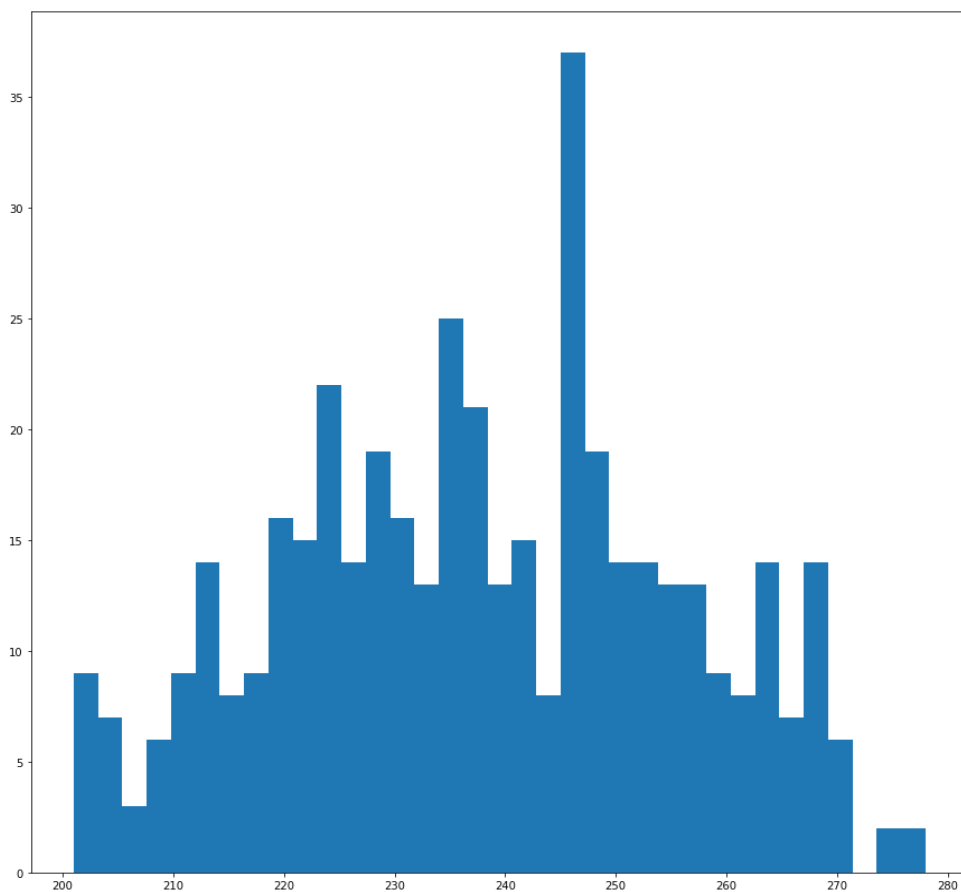


Figure 17: Data outliers of the dataset

Visualizing the data after removing outliers

```
count  453.000000
mean   238.796909
std    20.201710
min    189.000000
25%    224.000000
50%    238.000000
75%    252.000000
max    310.000000
```

Name: Sugar_Reading, dtype: float64

Splitting the dataset into train, test

```
<class 'pandas.core.frame.DataFrame'>
```

```
Int64Index: 302 entries, 391 to 217
```

```
Data columns (total 10 columns):
```

```
# Column          Non-Null Count  Dtype
---
```

```
0 Sugar_Reading    302 non-null  float64
1 SpicyLevel_high  302 non-null  uint8
2 SpicyLevel_low   302 non-null  uint8
3 SpicyLevel_medium 302 non-null  uint8
4 NaturalSugars_no 302 non-null  uint8
5 NaturalSugars_yes 302 non-null  uint8
6 ExternalSugars_no 302 non-null  uint8
7 ExternalSugars_yes 302 non-null  uint8
8 measured_time_morning 302 non-null  uint8
9 Clean_text       302 non-null  object
```

```
dtypes: float64(1), object(1), uint8(8)
```

```
memory usage: 9.4+ KB
```

4.6 Machine learning algorithms and accuracies

In order to perform this experiment, following models are effectively utilized KNN, random forest, and logistic regression and support vector mechanism. I have imported the dataset and processed with these machine learning algorithms and obtained different accuracies as below.

4.6.1 KNN

One performing the process K-nearest neighbor algorithm is accuracy which was obtained was 87%, which is the second best algorithm suitable for this process

```
for c in [3,5,7]:
```

```
    #defining knn model
```

```
    model_knn = KNeighborsClassifier(n_neighbors=3)
```

```
    model_knn.fit(X_train, y_train)
```

```

#predicting for test data
y_pred_knn = model_knn.predict(X_test)
#Checking accuracy for KNN model
print("Classification report for KNN- \n{}:\n{}\n".format(model_knn, classification_report(y_test, y_pred_knn)))
accuracy_knn = classification_report(y_test, y_pred_knn)

```

Output:

```

accuracy          0.87    136
macro avg         0.78    0.63    0.68    136
weighted avg      0.86    0.87    0.86    136

```

4.6.2 Logistic Regression

The logistic regression gave the least accuracy among all the other algorithm which is of 84%.

for c in [0.01, 0.05, 0.25, 0.5, 1]:

```

lr = LogisticRegression(C=c)
lr.fit(X_train, y_train)
predictions = lr.predict(X_test)
print("Classification report for Logistic regression- \n{}:\n{}\n".format(lr, classification_report(lr.predictions, y_test)))

```

4.6.3 Random Forest

The highest accuracy for this process is obtained by Random forest with 90% and considered as best model for diabetes sugar level prediction.

for estimator in [5, 10, 15, 20]:

```

model_rf = RandomForestClassifier(n_estimators=estimator)
model_rf.fit(X_train, y_train)
#predicting for test data
y_pred_rf = model_rf.predict(X_test)
#Checking accuracy for random forest model
print("Classification report for Random forest- \n{}:\n{}\n".format(model_rf, classification_report(y_test, y_pred_rf)))
accuracy_rf = classification_report(y_test, y_pred_rf)

```

Output:

```

accuracy          0.90    136
macro avg         0.81    0.77    0.79    136
weighted avg      0.89    0.90    0.90    136

```

4.6.4 SVM

Like the random forest model, support vector machine is also considered as the best model which showed 90% accuracy with this dataset.

```
for c in [0.01, 0.05, 0.25, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4]:
    svm = LinearSVC(C=c,max_iter=100)
    svm.fit(X_train, y_train)
    #predicting for test data
    y_pred_svm = svm.predict(X_test)
    print("Classification report for SVM- \n{}:\n{}\n".format(svm, classification_report(y_test, y_pred_svm)))
```

Output

accuracy		0.90		136
macro avg	0.84	0.72	0.77	136
weighted avg	0.89	0.90	0.89	136

5. Statistical Testing.

Statistical Testing have been conducted on the dataset using R language,

Null Hypothesis: *“Spice levels and Natural sugars in food doesn’t impact the blood glucose levels”*

Research Hypothesis: *“Spice levels and Natural sugars in food impact blood glucose levels”*

Chi-square test has been conducted as statistical testing and to observe the impact of two components (spice level and natural sugars) on blood glucose levels, with confidence level of 90%

The obtained chi-square test statistic value is greater than the critical value, hence the null hypothesis can be rejected.

6. Discussion

Discussion on Research Questions:

In this section the summary of the work done in this research is outlined, the research questions are evaluated, and the limitations of the work are explained.

Research Question One:

How spice levels can impact sugar levels in the food and what are the suitable method to identify it?

There are many factors in a food that is consumed can impact the sugar levels in the blood, one of the factor is spiciness of the food, so analysis has been made by identifying the spiciness from the food and categorizing into High, Medium, Low levels. In Explanatory data analysis it is identified that the food that has low spiciness has not increased the blood sugar levels

Research Question Two:

How natural sugars in the meal impact the blood sugar levels?

The Natural sugars of the food are one of the reasons that contributes the sugar levels, the natural sugars cannot be avoided, for an example a fruit has natural sweetness and when it is consumed the sugar level gets increased, the estimations of the natural sugars are done by understanding the diet information. And it is observed that natural sugars and external sugars follows a pattern in the diet. And the impact of them has been explained in pervious chapters.

Research Question Three:

How to choose a suitable algorithm from the comparison of results with different classification algorithms

The main objective of this research is to classify the sugar level present in the food, KNN, Logistic regression, Random forest, Support vector machine algorithms are used to evaluate performance, and it is observed that Support vector machine has performed well by training the model with count vectorizer, compared to SVM the accuracy of other models are less.

7. Limitations of work

There are few limitations in the systems and it is only a prototype version that can be enhanced. The data analysis has been conducted on the data that is collected from one person, and the dataset has very limited number of rows, the estimations of spiciness and natural sugars are done manually by understanding the diet specifications from the person. Due to this reason the predictions made by the model can only suggest the sugar levels for one person, and can't be generalized to everyone.

8. Conclusion and Future work

Diabetes is considered be one of the serious diseases which made numerous suffer across the world. According world health organization, people above the age of 60 seems to suffer more. Also, there is no cure found yet for this disease. Once someone is tested with low or high blood sugar level, the respective person should undergo lifelong medication and be on a strict diet. Diabetes also leads various other health complication, in the worst-case heart disease tend to occur. People lack awareness on the serious of this disease and also in the diet that needs to be followed. To address this issue, this research is conducted on one patient, the dataset is prepared by collecting the diet information from the patient.

With the help of the machine learning algorithms, the analysis has bought best possible prediction results of the sugar level in the meal. Spiciness is also considered in the analysis because the level of spiciness will impact the sugar level in the food. Random forest, KNN, logistic Regression, KNN and SVM algorithms are utilized in model development phase, and then compared their accuracies. Out of which Random forest and SVM showed highest accuracy with 90% and considered as the best models for this process. The present research could be extended to create a generalized model that can help to predict the diet chart information, and it can also be extended to provide a reminder to do proper exercise or proper meal for next intake if high sugar content food is consumed in a day.

9. Reference

- Péter Gyuk, István Vassányi, István Kósa, 2019 "Blood Glucose Level Prediction for Diabetics Based on Nutrition and Insulin Administration Logs Using Personalized Mathematical Models", *Journal of Healthcare Engineering*, vol. 2019, Article ID 8605206, 12 pages.<https://doi.org/10.1155/2019/8605206>
- Guariguata, L., Whiting, D.R., Hambleton, I., Beagley, J., Linnenkamp, U. and Shaw, J.E., 2014. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes research and clinical practice*, 103(2), pp.137-149.
- Florkowski, C., 2013. HbA1c as a diagnostic test for diabetes mellitus—reviewing the evidence. *The Clinical Biochemist Reviews*, 34(2), p.75.
- Nicholas, J., Charlton, J., Dregan, A. and Gulliford, M.C., 2013. Recent HbA1c values and mortality risk in type 2 diabetes. population-based case-control study. *PLoS One*, 8(7), p.e68008.
- American Diabetes Association, 2015. Standards of medical care in diabetes—2015 abridged for primary care providers. *Clinical diabetes: a publication of the American Diabetes Association*, 33(2), p.97.
- Eskaif, K., Ritchings, T. and Bedawy, O., 2014. Online prediction of blood glucose levels using genetic algorithm. In *Biologically-Inspired Techniques for Knowledge Discovery and Data Mining* (pp. 299-310). IGI Global.
- Ståhl, F. and Johansson, R., 2009. Diabetes mellitus modeling and short-term prediction based on blood glucose measurements *Mathematical biosciences*, 217(2), pp.101-117.
- Ståhl, F., Johansson, R. and Renard, E., 2010, August. Post-prandial plasma glucose prediction in type i diabetes based on impulse response models. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology* (pp. 1324-1327). IEEE.
- Clarke, W.L., Cox, D., Gonder-Frederick, L.A., Carter, W. and Pohl, S.L., 1987. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes care*, 10(5), pp.622-628.
- Shanthi, S. and Kumar, D., 2012. Prediction of blood glucose concentration ahead of time with feature based neural network. *Malaysian Journal of Computer Science*, 25(3), pp.136-148.
- Robertson, G., Lehmann, E.D., Sandham, W. and Hamilton, D., 2011. Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study. *Journal of Electrical and Computer Engineering*, 2011.
- Plis, K., Bunescu, R., Marling, C., Snoderook, J. and Schwartz, F., 2014, June. A machine learning approach to predicting blood glucose levels for diabetes management. In *Workshops at the Twenty-Eighth AAAI conference on artificial intelligence*.

Simon, D., 2006. Optimal state estimation: Kalman, H infinity, and nonlinear approaches. John Wiley & Sons.

Eskaf, K., Ritchings, T. and Bedawy, O., 2014. Online prediction of blood glucose levels using genetic algorithm. In *Biologically-Inspired Techniques for Knowledge Discovery and Data Mining* (pp. 299-310). IGI Global.

Chuah, Z.M., Paramesran, R., Thambiratnam, K. and Poh, S.C., 2010. A two-level partial least squares system for non-invasive blood glucose concentration prediction. *Chemometrics and Intelligent Laboratory Systems*, 104(2), pp.347-351.

Iyer, A., Jeyalatha, S. and Sumbaly, R., 2015. Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774.

Rajesh, K. and Sangeetha, V., 2012. Application of data mining methods and techniques for diabetes diagnosis. *International Journal of Engineering and Innovative Technology (IJEIT)*, 2(3), pp.224-229.

Kahramanli, H. and Allahverdi, N., 2008. Design of a hybrid system for the diabetes and heart diseases. *Expert systems with applications*, 35(1-2), pp.82-89.

Shanker, M.S., 1996. Using neural networks to predict the onset of diabetes mellitus. *Journal of chemical information and computer sciences*, 36(1), pp.35-41.

Sandham, W.A., Hamilton, D.J., Japp, A. and Patterson, K., 1998, November. Neural network and neuro-fuzzy systems for improving diabetes therapy. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286) (Vol. 3, pp. 1438-1441)*. IEEE.

El-Jabali, A.K., 2005. Neural network modeling and control of type 1 diabetes mellitus. *Bioprocess and biosystems engineering*, 27(2), pp.75-79.

Dazzi, D., Taddei, F., Gavarini, A., Uggeri, E., Negro, R. and Pezzarossa, A., 2001. The control of blood glucose in the critical diabetic patient: a neuro-fuzzy method. *Journal of Diabetes and its Complications*, 15(2), pp.80-87.

Pociot, F., Karlsen, A.E., Pedersen, C.B., Aalund, M., Nerup, J. and European Consortium for IDDM Genome Studies, 2004. Novel analytical methods applied to type 1 diabetes genome-scan data. *The American Journal of Human Genetics*, 74(4), pp.647-660.

Zorman, M., Masuda, G., Kokol, P., Yamamoto, R. and Stiglic, B., 2002, June. Mining diabetes database with decision trees and association rules. In *Proceedings of 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002)* (pp. 134-139). IEEE.

Hsu, W., Lee, M.L., Liu, B. and Ling, T.W., 2000, August. Exploration mining in diabetic patients databases: findings and conclusions. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 430-436).

Silverstein, C., Brin, S., Motwani, R. and Ullman, J., 2000. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2-3), pp.163-192.

Park, J. and Edington, D.W., 2001. A sequential neural network model for diabetes prediction. *Artificial intelligence in medicine*, 23(3), pp.277-293.

Haller, M.J., Wasserfall, C.H., McGrail, K.M., Cintron, M., Brusko, T.M., Wingard, J.R., Kelly, S.S., Shuster, J.J., Atkinson, M.A. and Schatz, D.A., 2009. Autologous umbilical cord blood transfusion in very young children with type 1 diabetes. *Diabetes Care*, 32(11), pp.2041-2046.

Chase, H.P., Cuthbertson, D.D., Dolan, L.M., Kaufman, F., Krischer, J.P., Schatz, D.A., White, N.H., Wilson, D.M., Wolfsdorf, J. and Diabetes Prevention Trial–Type 1 Study Group, 2001. First-phase insulin release during the intravenous glucose tolerance test as a risk factor for type 1 diabetes. *The Journal of pediatrics*, 138(2), pp.244-249.

Greenbaum, C.J., Cuthbertson, D. and Krischer, J.P., 2001. Type 1 diabetes manifested solely by 2-h oral glucose tolerance test criteria. *Diabetes*, 50(2), pp.470-476.

Vapnik, V., 2013. *The nature of statistical learning theory*. Springer science & business media.

Theodoridis, T., Solachidis, V., Dimitropoulos, K., Gymnopoulos, L. and Daras, P., 2019, June. A survey on AI nutrition recommender systems. In *Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments* (pp. 540-546).

Smola, A.J. and Schölkopf, B., 2004. A tutorial on support vector regression. *Statistics and computing*, 14(3), pp.199-222.

Appendix

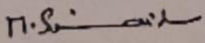
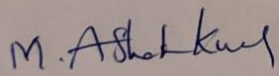
Appendix A

INFORMED CONSENT FORM

PROJECT TITLE:
DIET ANALYSIS FOR DIABETES USING MACHINE LEARNING.

PROJECT SUMMARY:
You are being asked to take part in the research study on diet analysis for diabetes using machine learning approach, this research would help in understanding the impact of food on blood sugar levels and how can intake of medicines like insulin can be reduced. This study aims to provide a model that predicts the sugar levels present in the food that falls under one of these categories (High, medium, low), to conduct the research the data collected from your day to day activities are used.

By signing below, you are agreeing that: (1) you have read and understood the Participant Information Sheet, (2) questions about your participation in this study have been answered satisfactorily, (3) you are aware of the potential risks (if any), and (4) you are taking part in this research study voluntarily (without coercion).

 (Participant's Signature)	M SIVAIAH (Participant's Name)
 (Student Signature)	ASHOK KUMAR MOVVA Student Name

11/01/2021

Date

Appendix B

Explanatory Data Analysis Code Snippets.

```
[1] #Visualisation packages
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.font_manager as fm
from matplotlib import figure
import pylab as pl
import matplotlib.figure
from wordcloud import WordCloud
pd.options.display.max_colwidth=500

from pathlib import Path
from google.colab import drive
drive.mount('/content/drive')

Mounted at /content/drive

15] df = pd.read_excel ('/content/drive/MyDrive/Projects/Machine Learning/diabetes prediction/originalData.xlsx', sheet_name='morning readings')
df.head()

[3] df.describe().T

[4] #removing null values
df = df.dropna(how='any',axis=0)
df['Mrng_SpicyLevel'] = df['Mrng_SpicyLevel'].str.lower()

[5] #Low spiciness infuleced sugar level.
df_low = df[df['Mrng_SpicyLevel'] == 'low']
df_low = df_low.sort_values(['Morning Reading'], ascending=[False])
df_low_high = df_low.head(25)
food = df_low_high['Food']
sugar_level = df_low_high['Morning Reading']
fig, ax = plt.subplots(figsize =(16, 9))
ax.barh(food, sugar_level)
ax.invert_yaxis()
for i in ax.patches:
    plt.text(i.get_width()+0.2, i.get_y()+0.5,
             str(round((i.get_width()), 2)),
             fontsize = 10, fontweight = 'bold',
             color = 'red')
ax.set_title('Low Spiciness food vs High sugar level',loc = 'center', )
plt.show()
```

```
[6] df_low = df_low.sort_values(['Morning Reading'], ascending=[True])
df_low_low = df_low.head(25)
food = df_low_low['Food']
sugar_level = df_low_low['Morning Reading']
fig, ax = plt.subplots(figsize=(16, 9))
ax.barh(food, sugar_level)
ax.invert_yaxis()
for i in ax.patches:
    plt.text(i.get_width()+0.2, i.get_y()+0.5,
             str(round((i.get_width()), 2)),
             fontsize = 10, fontweight = 'bold',
             color = 'green')
plt.show()

[7] #Medium Spiciness
df_medium = df[df['Mrng_SpicyLevel'] == 'medium']
df_medium = df_medium.sort_values(['Morning Reading'], ascending=[False])
df_medium_high = df_medium.head(25)
df_medium = df_medium.sort_values(['Morning Reading'], ascending=[True])
df_medium_low = df_medium.head(25)

[8] #medium spicy and high sugar level
food = df_medium_high['Food']
sugar_level = df_medium_high['Morning Reading']
fig, ax = plt.subplots(figsize=(16, 9))
ax.barh(food, sugar_level)
ax.invert_yaxis()
for i in ax.patches:
    plt.text(i.get_width()+0.2, i.get_y()+0.5,
             str(round((i.get_width()), 2)),
             fontsize = 10,
             color = 'red')
plt.show()
```

```

▶ #medium spicy and low sugar levels
food = df_medium_low['Food']
sugar_level = df_medium_low['Morning Reading']
fig, ax = plt.subplots(figsize =(16, 9))
ax.barh(food, sugar_level)
ax.invert_yaxis()
for i in ax.patches:
    plt.text(i.get_width()+0.2, i.get_y()+0.5,
             str(round((i.get_width()), 2)),
             fontsize = 10,
             color = 'green')
ax.set_title('Medium Spiciness food vs normal/medium sugar level',
            loc = 'left', )
plt.show()

```

```

[10] #Food consumed in the morning.
from wordcloud import WordCloud, STOPWORDS
wordcloud = WordCloud(background_color = 'grey', width = 1200, height = 800).generate(str(df['Food']))

plt.rcParams['figure.figsize'] = (10, 10)
plt.title('FOOD CONSUMED DURING MORNING', fontsize = 30)
plt.axis('off')
plt.imshow(wordcloud,interpolation = 'bilinear')
plt.show()

```

```

[11] #Food consumed in the evening.
wordcloud = WordCloud(background_color = 'orange', width = 1200, height = 800).generate(str(df['evng_food']))
plt.rcParams['figure.figsize'] = (10, 10)
plt.title('FOOD CONSUMED DURING EVENING', fontsize = 30)
plt.axis('off')
plt.imshow(wordcloud,interpolation = 'bilinear')
plt.show()

```

```

[12] #Natural Sugars in the morning.
natural_sugar = df.groupby(["Mrng_NaturalSugars"])["Date"].count().reset_index(name="count")
courses = list(natural_sugar['Mrng_NaturalSugars'])
values = list(natural_sugar['count'])
fig = plt.figure(figsize = (10, 5))

# creating the bar plot
plt.bar(courses, values, color = 'blue',
        width = 0.4)

plt.xlabel("Natural sugar")
plt.ylabel("No. of sample")
plt.title("Morning natural sugar level in the food")
plt.show()

```

```
[13] df['Mrng_SpicyLevel'] = df['Mrng_SpicyLevel'].str.lower()
natural_sugar = df.groupby(["Mrng_NaturalSugars", "Mrng_SpicyLevel"]).count()
df_no = df[df["Mrng_NaturalSugars"] == 'no']
df_yes = df[df["Mrng_NaturalSugars"] == 'yes']
df_no = df_no.groupby(["Mrng_NaturalSugars", "Mrng_SpicyLevel"])["Date"].count().reset_index(name="count")
df_yes = df_yes.groupby(["Mrng_NaturalSugars", "Mrng_SpicyLevel"])["Date"].count().reset_index(name="count")
plotdata = pd.DataFrame({
    "yes":df_yes['count'].to_list(),
    "no":df_no['count'].to_list()
}),
    index=["High", "Low", "Medium"]
)
plotdata.plot(kind="bar")
plt.title("Morning food Spiciness vs Natural sugar")
plt.xlabel("Food Spiciness")
plt.ylabel("Natural Sugar")
plt.figure(figsize = (5, 5))
```

```
▶ df['Evng_NaturalSugars'] = df['Evng_NaturalSugars'].str.lower()
natural_sugar = df.groupby(["Evng_NaturalSugars"])["Date"].count().reset_index(name="count")
courses = list(natural_sugar['Evng_NaturalSugars'])
values = list(natural_sugar['count'])
fig = plt.figure(figsize = (5, 5))

# creating the bar plot
plt.bar(courses, values, color = 'lightgreen',
        width = 0.4)

plt.xlabel("Natural sugar")
plt.ylabel("No. of sample")
plt.title("Evening natural sugar level in the food")
plt.show()
```

Machine Learning Modelling Code Snippets.

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV, train_test_split
from sklearn import svm
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.svm import SVC
from sklearn.svm import LinearSVC
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import classification_report
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
import warnings
warnings.filterwarnings('ignore')
```

```
[ ] from pathlib import Path
from google.colab import drive
drive.mount('/content/drive')
```

```
[ ] df = pd.read_excel('/content/drive/MyDrive/Projects/Machine Learning/diabetes prediction/originalData.xlsx', sheet_name='Sheet2')
df['SpicyLevel'] = df['SpicyLevel'].str.lower()
df['SugarLevel'] = df['SugarLevel'].str.lower()
df.head()
```

```
[ ] #removing null values
print('Row count before data cleaning',len(df))
df = df.dropna(how='any',axis=0)
print('Row count after data cleaning',len(df))
```

```
[ ] #outliers
features = df['Sugar_Reading'].to_list()
plt.hist(features, bins=35)
plt.show()
#removing outliers
df_no = df[(df["Sugar_Reading"] > 200) & (df["Sugar_Reading"] < 280)]
features = df_no['Sugar_Reading'].to_list()
plt.hist(features, bins=35)
plt.show()
```

```
[ ] cols_to_drop = ['Date']
df = df.drop(cols_to_drop, axis=1)
```

```
[ ] df['Sugar_Reading'].describe()
```

```
[ ] df.loc[df['SugarLevel'] == 'low', "type"] = int(0)
df.loc[df['SugarLevel'] == 'medium', "type"] = int(1)
df.loc[df['SugarLevel'] == 'high', "type"] = int(2)
df = df.drop(['SugarLevel'], axis=1)
df["type"] = df["type"].astype(int)
df.info()
```

```

[9] X_train, X_test, y_train, y_test = train_test_split(df['Food'], df['type'], stratify = df['type'], random_state=6, test_size = 0.3)
train = X_train; test = X_test

[10] vectorizer = CountVectorizer(token_pattern=r'\b\w+\b')
X_train = vectorizer.fit_transform(train)
X_test = vectorizer.transform(test)

[11] #Random Forest
for estimator in [5, 10, 15, 20]:
    model_rf = RandomForestClassifier(n_estimators=estimator)
    model_rf.fit(X_train, y_train)
    y_pred_rf = model_rf.predict(X_test)
    print("Classification report for Random forest- \n{}:\n{}\n".format(model_rf, classification_report(y_test, y_pred_rf)))
    accuracy_rf = classification_report(y_test, y_pred_rf)

[12] #SVM
for c in [0.01, 0.05, 0.25, 0.5, 1, 1.5, 2, 2.5, 3]:
    svm = LinearSVC(C=c, max_iter=100)
    svm.fit(X_train, y_train)
    y_pred_svm = svm.predict(X_test)
    print("Classification report for SVM- \n{}:\n{}\n".format(svm, classification_report(y_test, y_pred_svm)))

[13] #Logistic regression
for c in [0.01, 0.05, 0.25, 0.5, 1]:
    lr = LogisticRegression(C=c)
    lr.fit(X_train, y_train)
    lr_predictions = lr.predict(X_test)
    print("Classification report for Logistic regression- \n{}:\n{}\n".format(lr, classification_report(lr_predictions, y_test)))

▶ #KNN
for c in [3, 5, 7]:
    model_knn = KNeighborsClassifier(n_neighbors=c)
    model_knn.fit(X_train, y_train)
    y_pred_knn = model_knn.predict(X_test)
    print("Classification report for KNN- \n{}:\n{}\n".format(model_knn, classification_report(y_test, y_pred_knn)))
    accuracy_knn = classification_report(y_test, y_pred_knn)

```