Determination of Diabetes Predictors and the Likelihood of Diabetes Based on Dietary Habits

Menghan Jiao, Akshaymani Sornalingam, Shruti Ramesh

1. Introduction

This project aims to identify critical predictors of diabetes and explore the associations between the incidence of diabetes and certain dietary habits through a comprehensive data-driven approach [1]. Leveraging machine learning methodologies such as logistic regression , SVM (supervised learning techniques) and K-Means clustering (unsupervised learning techniques), the data is first annotated to facilitate analysis. Then the correlation between important disease predictors to determine the most appropriate parameters is investigated. The results of the analysis are considered, along with the consumption of various food types, to predict with a degree of certainty, whether an individual is likely to have diabetes or not using an SVM model. This project could lead to enhanced predictive models to understand the interplay between health indicators and dietary patterns, leading to improved strategies for disease management and prediction.

2. Background and Motivation

Current statistical data reveal a significant increase in diabetes worldwide [2], making it a critical public health challenge. Studies have also indicated that certain dietary components may influence the onset and management of diabetes [3], yet comprehensive models integrating these factors with clinical predictors are lacking.

This project aims to bridge this gap through machine learning techniques. This project not only contributes to existing research by providing a nuanced understanding of dietary factors associated with diabetes but also offers practical insights that could reform dietary recommendations and public health policies aimed at diabetes prevention. Moreover, the model developed in this project can be extrapolated to diseases other than diabetes, to offer holistic disease management and prevention.

3. Dataset Description

This project utilizes two distinct datasets to explore different facets of diabetes data. The first dataset, acquired from the UC Irvine Machine Learning Repository [5], originates from the 1994 AAAI Spring Symposium on Artificial Intelligence in Medicine. This dataset includes 70 data entries related to diabetes patients. Each entry provides a time series of numeric values concerning glucose levels, insulin readings, and various other lifestyle indicators recorded during a designated period of time.

The second dataset is derived from a nutrition study associated with the article "You Can’t Trust What You Read About Nutrition" [6] and consists of 54 comprehensive survey responses. These responses detail individual dietary habits and frequency of consumption across diverse food types, along with other pertinent variables. For the purpose of our analysis, descriptive data regarding respondents will be omitted to maintain focus on dietary patterns and their implications.

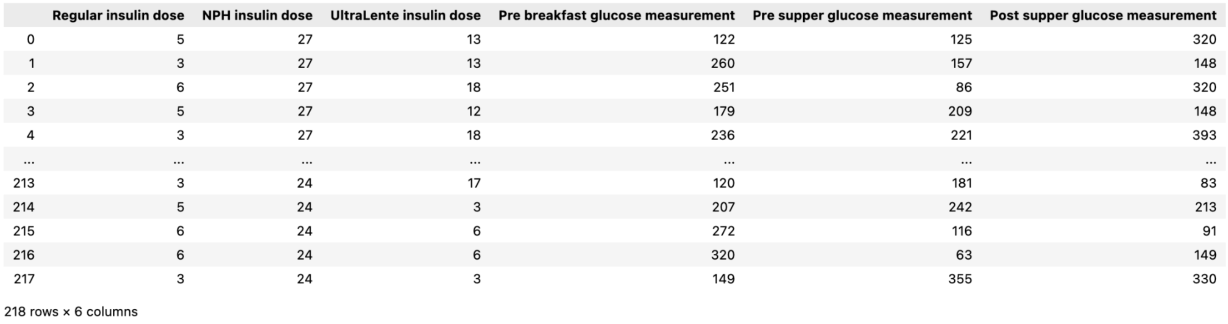
4. Methodology

Given that the project utilizes two different datasets, henceforth, the dataset sourced from the UCI Machine Learning Repository will be referred to as "dataset 1" or " records dataset," and the dataset associated with the article will be referred to as "dataset 2" or " nutrition dataset."

i. Data Pre-processing

The preprocessing of dataset 1 involves several key steps aimed at preparing the data for further analysis. This dataset is recorded in a way that each patient’s multivariate time series data is stored in different files, so the initial stages include loading each dataset into a data frame and performing data cleaning, such as removing the time stamp, stacking and grouping all similar values, handling missing values and erroneous entries. Once the data is cleaned and structured appropriately, more operations are executed, such as feature scaling and normalization to ensure that the model inputs are homogenized, thereby enhancing the algorithm's performance. The dataset also undergoes a transformation process where key features are engineered and selected based on their predictive power and relevance to the outcome variable, which in this case includes various metrics of diabetes diagnosis and management.

Based on the understanding of diabetes, the following six variables were selected for future analyses.



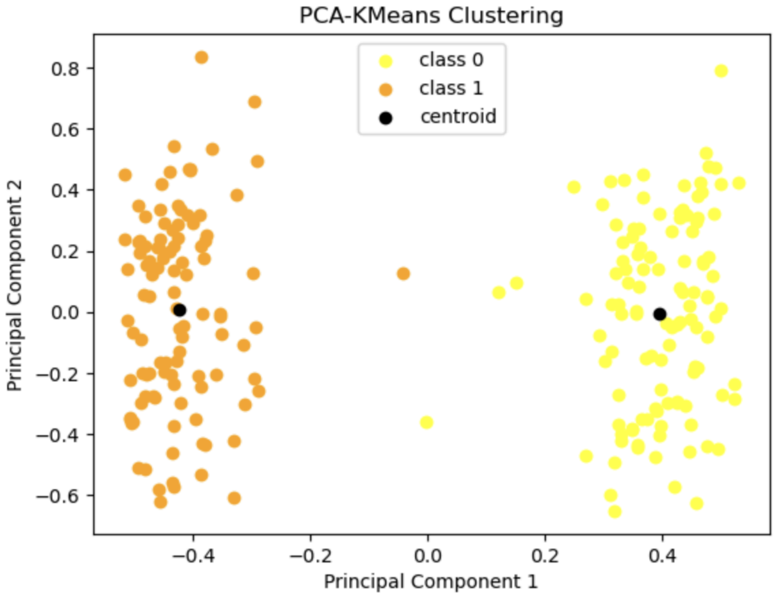
	Regular insulin dose	NPH insulin dose	UltraLente insulin dose	Pre breakfast glucose measurement	Pre supper glucose measurement	Post supper glucose measurement
0	5	27	13	122	125	320
1	3	27	13	260	157	148
2	6	27	18	251	86	320
3	5	27	12	179	209	148
4	3	27	18	236	221	393
...	...	...	...	...	...	...
213	3	24	17	120	181	83
214	5	24	3	207	242	213
215	6	24	6	272	116	91
216	6	24	6	320	63	149
217	3	24	3	149	355	330

218 rows × 6 columns

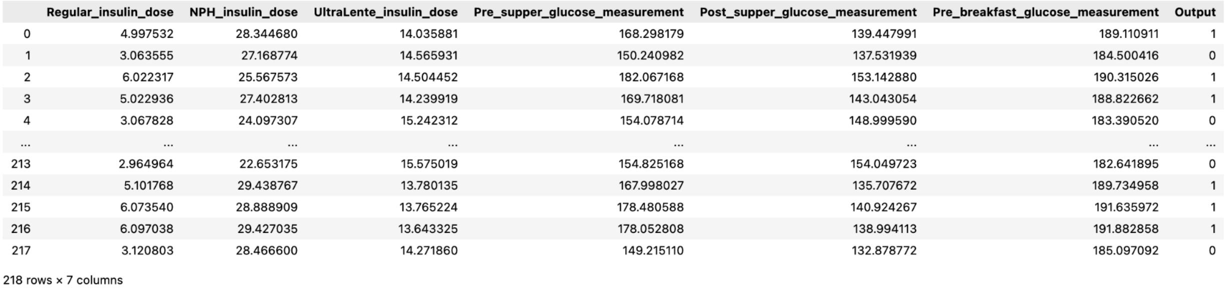
Dataset 2 consisted of meticulously calculated data about individual food intake and the nutritional values associated with it. Preprocessing this dataset required domain knowledge to manually cluster food items into broader food groups to facilitate better analysis. Data imputation was employed to handle missing data.

ii. K-Means Clustering

K-Means clustering was then performed on the cleaned diabetes dataset with two clusters, each representing the possibility of diabetes. Before K-Means clustering, Principal Component Analysis (PCA) was used. PCA is a machine learning technique designed to condense a large-scale dataset into a smaller dataset while preserving essential patterns and trends, thereby reducing the dimensionality of the data.

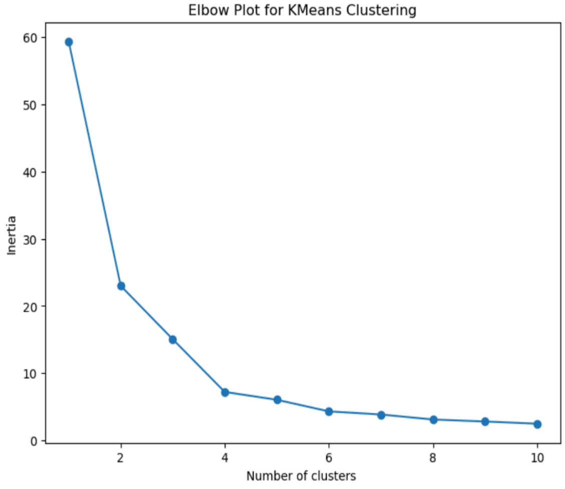


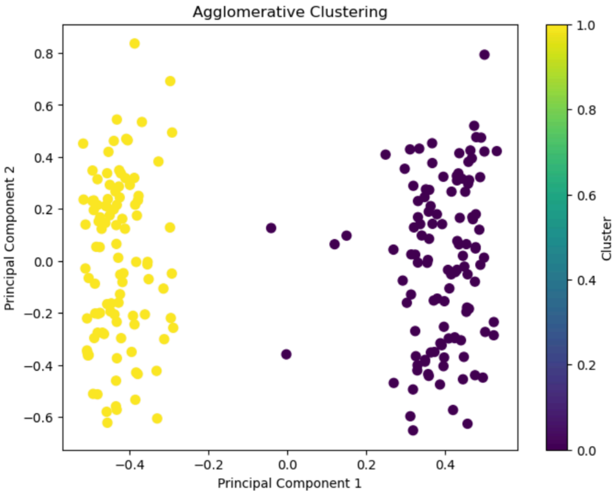
The output from the previous step was reverted to the original dataset using inverse PCA and inverse scaling techniques. With these adjustments, the dataset is now prepared for subsequent regression analyses.



	Regular_insulin_dose	NPH_insulin_dose	UltraLente_insulin_dose	Pre_supper_glucose_measurement	Post_supper_glucose_measurement	Pre_breakfast_glucose_measurement	Output
0	4.997532	28.344680	14.035881	168.298179	139.447991	189.110911	1
1	3.063555	27.168774	14.565931	150.240982	137.531939	184.500416	0
2	6.022317	25.567573	14.504452	182.067168	153.142880	190.315026	1
3	5.022936	27.402813	14.239919	169.718081	143.043054	188.822662	1
4	3.067828	24.097307	15.242312	154.078714	148.999590	183.390520	0
...	...	...	...	...	...	...	...
213	2.964964	22.653175	15.575019	154.825168	154.049723	182.641895	0
214	5.101768	29.438767	13.780135	167.998027	135.707672	189.734958	1
215	6.073540	28.888909	13.765224	178.480588	140.924267	191.635972	1
216	6.097038	29.427035	13.643325	178.052808	138.994113	191.882858	1
217	3.120803	28.466600	14.271860	149.215110	132.878772	185.097092	0

218 rows × 7 columns





iii. Regression Analysis

Then a logistic regression model was developed using ‘statsmodels.api’ to choose the parameters by applying the L1 penalty, a variable selection property, and then classify the data point as input by the user. Firstly, GridSearchCV was employed to find the most suitable degree for the logistic regression model and the optimal penalty parameter C based on the 'F1-scoring' metric. The parameters thus identified were integrated into a pipeline comprising a polynomial degree function, a standard scaler, and a logistic regression mode which operates based on a straightforward liblinear solver with the penalty C. The polynomial function adjusts predictor values to the determined optimal degree. And the standard scaler, or more specifically a min-max scaler was used to normalize values to a uniform scale. Following the model's training with the regression dataset, the list of predictors that were selected by the L1-regularized logistic regression model was identified.

Then the user was prompted to input specific parameters. All the user input data were collected and transformed into a data frame, wherein the probability of the data point belonging to a particular class was calculated. The decision threshold for class assignment was established at 0.5.

iv. Dietary Pattern Prediction

To predict the likelihood of diabetes based on dietary habits, a Support Vector Machine (SVM) model was used. Feature selection involved removing identifiers and the target variable from the nutrition dataset, leaving only the dietary attributes as predictors. These features were then standardized using StandardScaler to ensure that the SVM algorithm, which is sensitive to the scale of input features, performs optimally. Then the linear kernel was selected due to its simplicity and efficiency on linearly separable data. The model was trained on 80% of the data, which was randomly split from the dataset, with the remaining 20% reserved for testing and model evaluation. Model performance was assessed using standard classification metrics, including precision, recall, and the F1-score.

5. Results and Observations

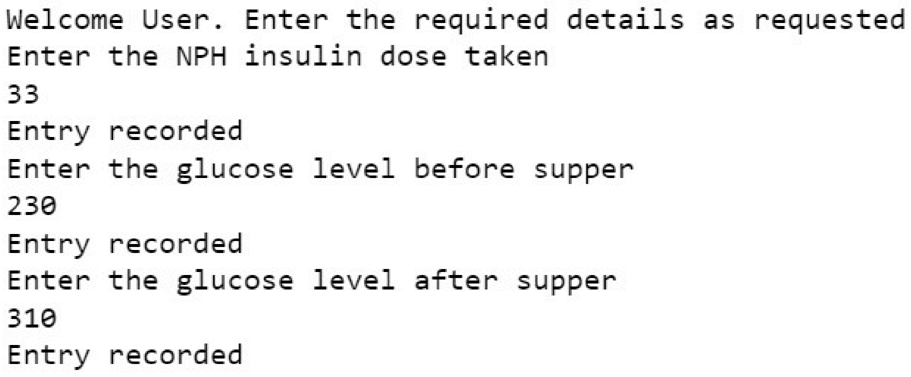
i. K-means Clustering

Standard clustering was selected over agglomerative clustering for the creation of the logistic regression dataset because the CH-index for standard clustering (341.54) was higher than that for agglomerative clustering (340). These evaluation metrics effectively assess both within-cluster variation and between-cluster variation.

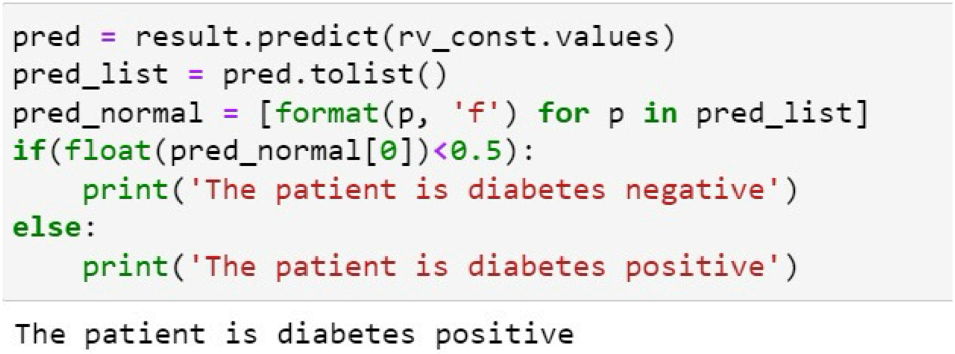
Using existing scientific knowledge about diabetes, it can be concluded that Cluster 0 is indicative of a diabetes-negative status, attributed to lower insulin doses, whereas Cluster 1 suggests a diabetes-positive status due to higher insulin doses and elevated glucose levels, particularly evident in pre-breakfast scenarios. The centroid for Cluster 0 is represented by the values [9.50847873, 5.29842949, 11.62588758, 199.36890395, 162.65034574, 180.17548561], and for Cluster 1, it is [10.55754194, 33.7550235, 11.4788067, 191.0220367, 164.90010411, 193.84923929]. The first three values in each list denote insulin dose levels, and the last three represent glucose levels. This information substantiates the rationale behind the labeling of the clusters.

ii. Logistic Regression

For regression analysis, the data indicated that the optimal polynomial degree is 1, and the optimal penalty parameter is 10. Upon implementing L1 penalization, the model selected NPH insulin dose, pre-supper glucose, and post-supper glucose as predictors. The coefficient associated with the NPH insulin predictor is notably high (+0.4045), suggesting that the amount of NPH insulin significantly aids in the diagnosis or prediction of diabetes. The rationale for selecting these predictors is based on the role of NPH insulin in stabilizing blood sugar levels between evening meals. Given that both pre-supper and post-supper glucose levels are also chosen, the model's selections are deemed appropriate.



```
Welcome User. Enter the required details as requested
Enter the NPH insulin dose taken
33
Entry recorded
Enter the glucose level before supper
230
Entry recorded
Enter the glucose level after supper
310
Entry recorded
```

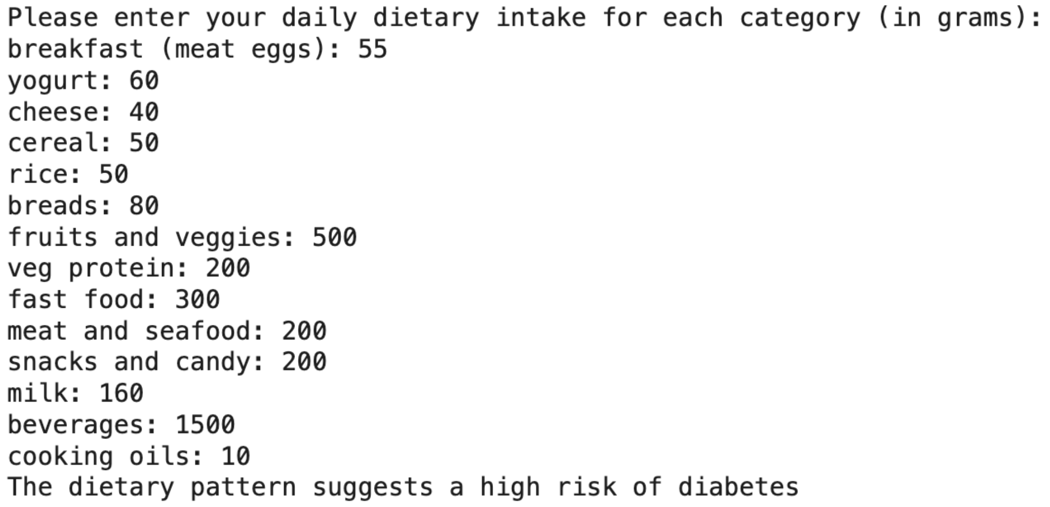


```
pred = result.predict(rv_const.values)
pred_list = pred.tolist()
pred_normal = [format(p, 'f') for p in pred_list]
if(float(pred_normal[0])<0.5):
    print('The patient is diabetes negative')
else:
    print('The patient is diabetes positive')
```

```
The patient is diabetes positive
```

iii. Dietary Pattern Prediction

Utilizing the SVM model trained on the nutrition dataset, an interactive module was created to enable users to input their dietary intake values and promptly obtain predictive outcomes. For example, the following entered daily dietary intake values indicate that if an individual maintains this dietary pattern, there is a heightened likelihood of developing diabetes. The primary objective of this model is to provide users with immediate feedback, thereby allowing them the opportunity to modify their dietary habits by the model's recommendations to mitigate their risk of diabetes.



```
Please enter your daily dietary intake for each category (in grams):
breakfast (meat eggs): 55
yogurt: 60
cheese: 40
cereal: 50
rice: 50
breads: 80
fruits and veggies: 500
veg protein: 200
fast food: 300
meat and seafood: 200
snacks and candy: 200
milk: 160
beverages: 1500
cooking oils: 10
The dietary pattern suggests a high risk of diabetes
```

6. Conclusion

In this project, it was found that pre and post-supper glucose levels and NPH insulin levels serve as good predictors of diabetes and this aligns with existing scientific understanding of the disease. Notably, the predictor coefficient value for NPH insulin is higher, indicating its greater relevance in the predictive model. Additionally, dietary data was analyzed to examine the consumption patterns of various food groups among individuals with and without diabetes. The SVM (Support Vector Machine) model developed from this data successfully classified the user as diabetes positive or negative based on their dietary habits (i.e. amount of food substances that they input). This project not only highlights critical predictors of diabetes but also demonstrates the potential of the model to analyze and predict other diseases. This approach could significantly enhance preventative strategies and personalized dietary recommendations in clinical settings.

References

[1] A. Mujumdar and V. Vaidehi, “Diabetes Prediction using Machine Learning Algorithms,” Procedia Computer Science, vol. 165, pp. 292–299, 2019, doi: https://doi.org/10.1016/j.procs.2020.01.047.

[2] Saeedi P, Petersohn I, Salpea P, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res Clin Pract. 2019;157:107843. doi:10.1016/j.diabres.2019.107843

[3] Hu FB. Globalization of diabetes: the role of diet, lifestyle, and genes. Diabetes Care. 2011;34(6):1249-1257. doi:10.2337/dc11-0442

[4] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, “A Machine Learning Methodology for Diagnosing Chronic Kidney Disease,” IEEE Access, vol. 8, pp. 20991–21002, 2020, doi: https://doi.org/10.1109/ACCESS.2019.2963053.

[5] M. Kahn, “UCI Machine Learning Repository,” archive.ics.uci.edu. https://archive.ics.uci.edu/dataset/34/diabetes

[6] A. M. Barry-Jester, “You Can’t Trust What You Read About Nutrition,” FiveThirtyEight, Jan. 06, 2016. https://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition/

[7] Md. K. Hasan, Md. A. Alam, D. Das, E. Hossain, and M. Hasan, “Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers,” IEEE Access, vol. 8, pp. 76516–76531, 2020, doi: https://doi.org/10.1109/access.2020.2989857.