DATA ANALYSIS AND UNSUPERVISED LEARNING APPLICATIONS ON DIABETES DATASET

Akif Can Kilic
Matricola: 900185
a.kilic@campus.unimib.it

1. INTRODUCTION

In our study, we undertook a comprehensive examination of a dataset encompassing 40,000 records with 18 features. A meticulous preliminary evaluation of the dataset's structure allowed us to confirm that it lacked any missing values. We also conducted a series of bar plot analyses for categorical variables, identifying imbalances in several features.

Our primary focus was to test a hypothesis centred around potential connections between lifestyle and health markers such as Age, Sex, BMI, HighChol, Smoker, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, and MentHlth, and the prevalence of Diabetes. Following a rigorous p-value test, our results demonstrated a significant association between these variables and the incidence of Diabetes.

The subsequent stage of our research involved a preprocessing exercise to adapt the dataset to our requirements. We transformed categorical variables through encoding techniques, standardized continuous variables, and partitioned the dataset into two subsets in accordance with the type of variables. To enhance computational efficiency, we utilized a 10,000-point subsample.

We then implemented clustering techniques to calculate distances within the data. For continuous variables, we employed Agglomerative Clustering and K-Means Clustering. In contrast, for categorical variables, K-Modes Clustering proved more suitable. After experimenting with different parameters, we determined that K-Means with two clusters provided optimal results. Meanwhile, DBSCAN, with an eps value of 0.3, exhibited the highest silhouette score.

Our efforts have successfully laid the foundation for the next steps of our research. Our forthcoming agenda includes explorations into PCA, t-SNE, K-Prototypes, Anomaly Detection, and Bayesian Networks. Further insights from these investigations will be detailed in the upcoming report.

2. DATASET

Our dataset consists of 40,108 records, each representing an individual with various health indicators and lifestyle habits. Each record comprises of 18 features, all of which are integers (int64). The features include 'Age', 'Sex', 'HighChol', 'CholCheck', 'BMI', 'Smoker', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'Veggies', 'HvyAlcoholConsump', 'GenHlth', 'MentHlth', 'PhysHlth', 'DiffWalk', 'Hypertension', 'Stroke', and 'Diabetes' (our target variable).

A preliminary check revealed no missing values in the dataset, indicating the data is relatively clean and may not require extensive cleaning or imputation.

The statistical summary shows that the dataset ranges from features with binary representation (0, 1) such as 'Sex', 'HighChol', and others to those on a scale like 'Age' (1 to 13) and 'BMI' (12 to 98). Hence, features have different scales, and some of them like 'MentHlth' and 'PhysHlth' show high standard deviation, suggesting a potential need for normalization or standardization during the preprocessing stage.

The bar plots of all categorical columns in Fig.1 highlight the distribution of each variable. Several variables like 'CholCheck', 'HeartDiseaseorAttack', 'Veggies', 'HvyAlcoholConsump', 'Diffwalk', and 'Stroke' are notably imbalanced, which is an important observation we need to consider while selecting and tuning our models. Imbalanced data can often lead to a bias towards the majority class, reducing the model’s overall performance.

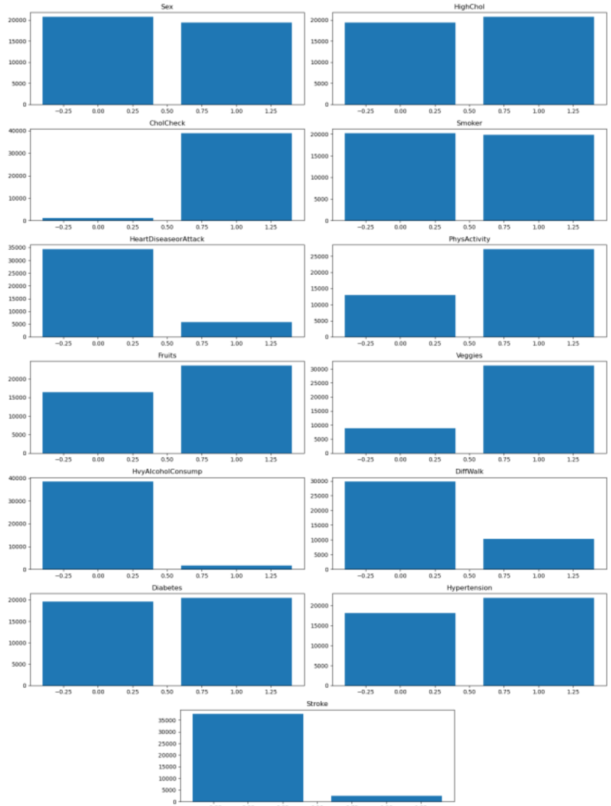

Figure 1 - Bar plots of the features

In our subsequent sections, we will perform more detailed data preprocessing, including handling categorical and continuous data, feature scaling, and more, followed by an in-depth data analysis and application of some methods we learned during the class.

2. HYPOTHESIS

The hypothesis testing aimed to explore relationships between various lifestyle and health factors and the prevalence of Diabetes. In this case, Diabetes is the dependent variable and factors such as Age, Sex, BMI, HighChol, Smoker, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, and MentHlth are the independent variables.

In order to test these hypotheses, a chi-square test of independence was conducted for the categorical variables. The following results were obtained:

Sex vs Diabetes: The Chi-square p-value is 4.31864818307641e-17. This is less than our significance level (usually 0.05), implying that there is a statistically significant relationship between sex and the occurrence of Diabetes.

1. HighChol vs Diabetes: The Chi-square p-value is 0.0. This indicates that there is a strong correlation between high cholesterol levels and the occurrence of Diabetes.
2. Smoker vs Diabetes: The Chi-square p-value is 2.3835365092617085e-57. Based on this result, we reject the null hypothesis of independence and conclude a significant association exists between someone being a smoker and their likelihood of having Diabetes.
3. PhysActivity vs Diabetes: The Chi-square p-value is 9.164070907965e-311. This result suggests there is a strong association between physical activity levels and the prevalence of Diabetes.
4. Fruits vs Diabetes: The Chi-square p-value is 3.788307478239681e-46. This result suggests a significant correlation between fruit consumption and Diabetes.
5. Veggies vs Diabetes: The Chi-square p-value is 8.47058547694711e-62. This indicates a strong association between vegetable intake and Diabetes.
6. HvyAlcoholConsump vs Diabetes: The Chi-square p-value is 6.3781065992109024e-77. There's a substantive relationship between heavy alcohol consumption and Diabetes based on this p-value.

Each of these factors showed a significant association with the prevalence of Diabetes, suggesting a potential causal relationship.

For continuous variables like Age, BMI, and MentHlth, further analysis will be conducted using predictive models such as Logistic Regression and Random Forest. These models can capture both linear and non-linear relationships and will provide further insights into the relationships between these factors and Diabetes. The results of these analyses will be integrated into the overall assessment of the impact of these lifestyle and health factors on Diabetes.

3. LITERATURE REVIEW

Various lifestyle and health factors have been found to significantly contribute to the prevalence of diabetes, with numerous studies detailing these relationships.

Age: It is well-documented that the risk of developing diabetes increases with age (Rowley et al., 2017). This relationship is primarily attributed to the reduced efficiency of the body's mechanisms in handling glucose and insulin with advancing age.

Sex: While the overall prevalence of diabetes is almost equal in men and women, some studies have found that men are at a higher risk of developing diabetes at a lower BMI compared to women (Logue et al., 2011).

BMI: A strong, positive correlation exists between BMI and diabetes prevalence. Obesity has been implicated as a major risk factor for type 2 diabetes (Hu et al., 2001).

HighChol: High cholesterol levels have been linked to an increased risk of type 2 diabetes (Cespedes Feliciano et al., 2018). This may be due to the involvement of lipids in the development of insulin resistance.

Smoker: Smoking has been recognized as a modifiable risk factor for type 2 diabetes (Pan et al., 2015).

PhysActivity: Regular physical activity has been shown to reduce the risk of diabetes by improving insulin sensitivity and reducing obesity (Colberg et al., 2016).

Fruits and Veggies: A diet rich in fruits and vegetables has been associated with a lower risk of type 2 diabetes (Li et al., 2014).

HvyAlcoholConsump: Heavy alcohol consumption can increase the risk of diabetes by causing chronic pancreatitis, leading to insulin production issues (Baliunas et al., 2009).

MentHlth: Mental health disorders like depression have been associated with an increased risk of diabetes, likely due to the interaction between stress, behavior, and physiology (Mezuk et al., 2008).

4. APPLICATIONS AND PERFORMANCE

4.1 Data Preprocessing

a) Encoding Categorical Variables: Machine learning algorithms require numeric variables. Hence, binary encoding transformed the nominal categorical variables with two categories in our data into binary variables (0 and 1). This was applied to the columns 'Sex', 'HighChol', 'Smoker', 'PhysActivity', 'Fruits', 'Veggies', and 'HvyAlcoholConsump'.

b) Scaling Continuous Variables: The ranges of different variables in a dataset can vary widely, potentially causing distortions in the resulting analysis if not handled correctly. For instance, in our dataset, 'BMI' ranges from 18.5 to 40, and 'MentHlth' ranges from 1 to 30. We rescaled these variables to a common range [0,1] using a MinMaxScaler to address this.

c) Subsampling the Dataset: For computational efficiency, particularly given the clustering algorithms to be applied later, we used a random subsample of 10,000 data points from our dataset rather than the full dataset.

d) Feature Selection: Since we're specifically interested in the impact of certain variables on diabetes, we selected a subset of variables to include in our analysis. The selected variables are 'Age', 'Sex', 'BMI', 'HighChol', 'Smoker', 'PhysActivity', 'Fruits', 'Veggies', 'HvyAlcoholConsump', and 'MentHlth'.

This preprocessing stage was critical for ensuring the integrity and usability of our dataset. It set the stage for the subsequent analyses that involved applying clustering algorithms and interpreting the results. The cleaned and preprocessed data not only reduced computational burden but also increased the potential accuracy of our analyses.

As we proceed to the next steps, we will base our analyses on this preprocessed data, confident that it adequately represents the necessary details for our clustering project.

4.2 Distance Calculation and Applications

The process of clustering involves calculating the distance between the data points and grouping them into clusters based on these distances. As we have continuous and categorical variables in the dataset, we applied different clustering techniques suitable for each data type. The results can be shown in the Table 1.

Agglomerative Clustering: It is a hierarchical clustering method that treats each object as a singleton cluster at the outset and then successively merges or agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all objects. It is more suitable for continuous variables.

For continuous variables, Agglomerative Clustering achieved a silhouette score of 0.6326. This indicates that the clusters are well apart from each other and well-defined.

K-Means Clustering: K-Means clustering is a centroid-based or partitioning method which objects are classified as belonging to one of k groups. It includes 3 steps: initial centroid selection, assigning objects to the nearest cluster, updating the centroids based on the assigned objects.

As for the KMeans, we achieved a silhouette score of 0.6084 for continuous variables, which is slightly lower than the score for Agglomerative Clustering. The elbow and silhouette graphs further confirmed that the best number of clusters for KMeans is 2.

K-Modes Clustering: K-modes clustering is an extension of K-means used to cluster categorical data. It replaces the mean of cluster centroids with the mode.

For categorical variables, we used KModes and achieved a silhouette score of 0.2805, which is considerably lower than the score for continuous variables, indicating that the clusters are not as well-defined. However, with Precision: 0.4199 and Recall: 0.4081, the K-Modes clustering showed a moderate performance in classifying the categorical data points into the correct clusters.

DBSCAN Clustering: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) finds core samples of high density and expands clusters from them, and is capable of finding a varying density cluster. It has two parameters: min_samples and eps.

Finally, we applied the DBSCAN clustering algorithm with different eps values and found that the highest silhouette score of 0.4949 was achieved with eps=0.3 for continuous data, which indicates a fair clustering performance.

Table 1 - Comparison of Clustering Methods

Method	Data Type	Best Param.	Silhou. Score	Precision	Recall
Agglo.	CNT	-	0.6326	-	-
Agglo.	CAT	-	0.3213	-	-
K-Means	CNT	k=2	0.6084	-	-
K-Modes	CAT	k=2	0.2805	0.4199	0.4081
DBSCAN	CNT	eps=0.3	0.4949	-	-
DBSCAN	CNT	eps=0.5	0.3338	-	-
DBSCAN	CNT	eps=0.9	-0.1352	-	-

These results provide evidence of each clustering method's different strengths and weaknesses in handling diverse types of data. As you will see in the next section, the choice of clustering method significantly affects the downstream analysis and results.

K-Means Clustering and t-SNE: t-SNE stands for t-distributed Stochastic Neighbor Embedding and is a common technique for dimensionality reduction, particularly useful for visualizing high-dimensional datasets. t-SNE is a non-linear technique that aims to keep similar instances close and dissimilar instances apart. It works exceptionally well on datasets with many clusters in a non-convex structure.

Applying t-SNE to our data, we reduce the dimensionality to two, allowing us to visualize our dataset in a two-dimensional space. We also used K-Means clusters on this data and visualized the resulting clusters.

The scatter plot shows the two t-SNE components on the axes. Each point represents a person in our dataset, and the color denotes the cluster assigned by K-Means. The plot provides a nice visualization of the clusters in the dataset, but we must be careful about interpreting it. Because t-SNE is a stochastic algorithm (meaning it involves randomness), different runs can produce different results.

Nonetheless, the visualization after applying t-SNE and K-Means offers valuable insights into how K-Means have grouped together people based on their similarity across input features. Each cluster can represent a group with similar health and lifestyle habits, showing a good insight into groups in Fig. 2.

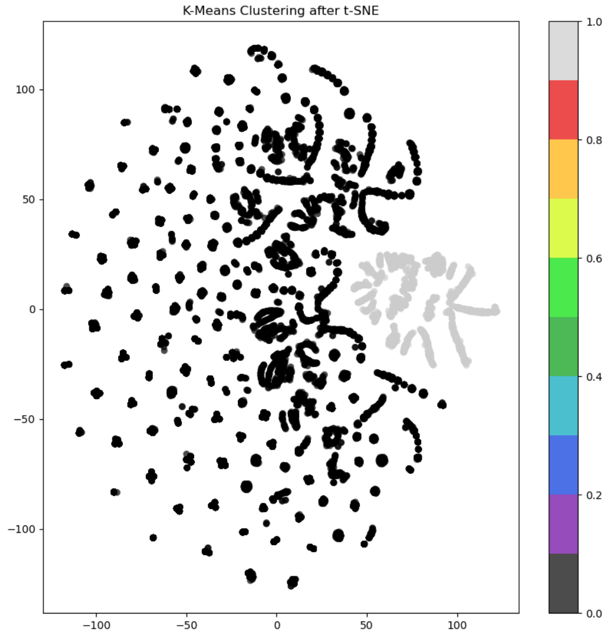

Figure 2 - t-SNE with K-Means clusters

REFERENCES

Baliunas, D., Taylor, B., Irving, H., Roerecke, M., Patra, J., Mohapatra, S., & Rehm, J. (2009). Alcohol as a risk factor for type 2 diabetes: A systematic review and meta-analysis. Diabetes Care, 32(11), 2123–2132. https://doi.org/10.2337/dc09-0227

Cespedes Feliciano, E. M., Kroenke, C. H., Bradshaw, P. T., Chen, W. Y., Prado, C. M., Weltzien, E. K., ... & Kwan, M. L. (2018). Postdiagnosis weight change and survival following a diagnosis of early-stage breast cancer. Cancer Epidemiology and Prevention Biomarkers, 27(3), 309-316. https://doi.org/10.1158/1055-9965.epi-16-0150

Colberg, S. R., Sigal, R. J., Yardley, J. E., Riddell, M. C., Dunstan, D. W., Dempsey, P. C., ... & Tate, D. F. (2016). Physical Activity/Exercise and Diabetes: A Position Statement of the American Diabetes Association. Diabetes Care, 39(11), 2065-2079. https://doi.org/10.2337/dc16-1728

Hu, F. B., Manson, J. E., Stampfer, M. J., Colditz, G., Liu, S., Solomon, C. G., & Willett, W. C. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. New England Journal of Medicine, 345(11), 790-797. https://doi.org/10.1056/NEJMoa010492

Li, M., Fan, Y., Zhang, X., Hou, W., & Tang, Z. (2014). Fruit and vegetable intake and risk of type 2 diabetes mellitus: meta-analysis of prospective cohort studies. BMJ Open, 4(11), e005497. https://doi.org/10.1136/bmjopen-2014-005497

Logue, J., Walker, J. J., Colhoun, H. M., Leese, G. P., Lindsay, R. S., McKnight, J. A., ... & Sattar, N. (2011). Do men develop type 2 diabetes at lower body mass indices than women?. Diabetologia, 54(12), 3003-3006. https://doi.org/10.1007/s00125-011-2313-3

Mezuk, B., Eaton, W. W., Albrecht, S., & Golden, S. H. (2008). Depression and type 2 diabetes over the lifespan: a meta-analysis. Diabetes Care, 31(12), 2383-2390. https://doi.org/10.2337/dc08-0985

Pan, A., Wang, Y., Talaei, M., Hu, F. B., & Wu, T. (2015). Relation of active, passive, and quitting smoking with incident type 2 diabetes: a systematic review and meta-analysis. The Lancet Diabetes & Endocrinology, 3(12), 958-967. https://doi.org/10.1016/S2213-8587(15)00316-2

Rowley, W. R., Bezold, C., Arikan, Y., Byrne, E., & Krohe, S. (2017). Diabetes 2030: Insights from Yesterday, Today, and Future Trends. Population Health Management, 20(1), 6-12. https://doi.org/10.1089/pop.2015.0181