**Machine Learning Unsupervised Learning 2**

**Batch 22 Term 2 – Group 17**

**Part 3: Project Final Report**

***Finding Factors Causing Diabetes and Suggesting Early Diagnosis & Effective Treatment based on Person health indicators, Lifestyles***

**Introduction**: Our project focuses on finding factors which causes diabetes, and our aim is to suggest early diagnosis and effective treatment based on the person’s health indicators and lifestyle.

**Problem Statement**: Diabetes is one of the most prevalent diseases in the world. The prevalence of diabetes is rising rapidly, highlighting the urgent need for effective interventions. Early diagnosis and personalized treatment strategies based on lifestyle modifications—such as diet and exercise—are crucial in preventing the onset of diabetes and mitigating its health consequences.

The goal of the project is to provide early indicators to people who have similar underlying factors or a sub-set of the underlying factors which, if not addressed could lead to them becoming diabetic in the future. Having this goal in mind, we can further dive into the below questions:

1. Understanding Health Indicators of individuals based on their Diabetes status
2. Finding Patterns & Associations between indicators causing Diabetes
3. Finding patterns in pre-conditions of people with Diabetes by comparing with people who are pre-diabetic or non-diabetic and suggest early lifestyle changes
4. Finding Anomalies of certain individuals based on the data and share these risk factors with health officials i.e., people who have more than one risk factors but have not yet been diagnosed to be pre/diabetic.
5. How to reduce economic burden on government for by providing early detection and suggesting corrective action.

Our objective is to address all the above questions, utilising the various unsupervised learning techniques.

A black background with white rectangles

Description automatically generated

**Data, Data Cleaning and Data Profiling**: The Behavioural Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the Center for Disease Control and Prevention (CDC). Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviours, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. For this project, we have used dataset available on [Kaggle](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download&select=diabetes_012_health_indicators_BRFSS2015.csv) for the year 2015. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

Since this data collection technique encompasses a very wide range of the population (400,000 Americans) and is voluntary, it is a highly clean dataset which can be taken as representative of the entire US population.

Based on the profiling report by pandas profiling, we see that this dataset contains 16 categorical and 6 numeric variables and has 0 missing cells. There are duplicate rows, but they are accepted as multiple definitions are assigned to the value 0 as per the attribute.A screenshot of a report

Description automatically generated

**EMPERICAL ANALYSIS**

**Understanding the data**: Glancing through the data, we see that this dataset has 253,680 survey responses and having 22 features (Indicators).

To find out relationships between the features, we ran a co-relation comparison among select few indicators (namely BMI, Age, Education, Income). We could not find any corelations among the selected indicators. Subsequently, we checked for potential outliers within the same selected indicators. *We observed that the BMI had the most outliers.*A group of graphs with numbers

Description automatically generated with medium confidencePlotting a graph showing the distribution of the people based on gender, we find that females are more represented (shown by 0) than males (shown by 1).A graph of a number of people with diabetes

Description automatically generatedWe can infer from this that females are healthier than males in the sample as the proportion of females with no diabetes are higher than the entire male sample.

**Basket Analysis (Association rules)**: The apriori algorithm is the technique we used here. Since this technique runs through the entire dataset multiple times and terminates when no associations are found within the item sets, we needed to divide the dataset into 3 baskets: People having no diabetes, people who are pre-diabetic and people who have diabetes. This helped reduce the processing time of the algorithm. For further division, we broke down the data further into males and females. We also narrowed down the attributes from 22 to 12 for even faster processing.

We also found a new feature in our dataset, ‘GenHlth,’ which is an indicator of the health of the individual. It is on a scale of 1 to 5, with 5 being the least healthy and the score of 1 being the healthiest. We broke down our dataset further based on this feature by stating that people with a score between 1 and 3 as being healthy, and people with score having 4 and 5 as unhealthy.

Having selected, cleaned, and filtered out our data, we then proceeded with doing basket analysis.

A screenshot of a computer

Description automatically generated*Snapshot of the General Health feature*

**BASKET ANALYSIS**

1. **People with no-diabetes basket**: For people with no-diabetes basket, we ran the apriori algorithm with a minimum support (the frequency of occurrence in the dataset) of 25%, we found that the most frequent attribute-pair is people getting a cholesterol check along with having no diabetes. This item-set (CholCheck, NoDiabetes) has a support of 95%. *This signifies that people with no diabetes also got their cholesterol checked in the last 5 years.*

The second most frequent attribute-pair is people doing a physical activity in the past 30 days having no diabetes. This attribute-pair (Gen Health, NoDiabetes) has a support of 86.94%. This signifies that *a large set of people who did a physical activity in the past 30 days had a greater chance of not developing diabetes.*

1. **People with diabetes basket:** For people with diabetes basket, we ran the apriori algorithm with a minimum support (the frequency of occurrence in the dataset) of 10%, we found that the most frequent attribute-pair is people getting a cholesterol check along with having diabetes. This item-set (CholCheck, Diabetes) has a support of 99%. *This signifies that almost everyone with diabetes also got their cholesterol checked in the last 5 years. This goes with intuition as people with diabetes would also get other health indicators like cholesterol checked to get a sense of their overall health profile at the time.*

The second most frequent attribute-pair is people having high blood pressure also having diabetes. This attribute-pair (Diabetes, Veggies) has a support of 75%. This signifies that *a large set of people who have diabetes consume more than 1 vegetable daily. This is not according to intuition, and we can look further into this. It is surprising that a lot of people who have diabetes consume more than 1 vegetable daily.*

Further, we can see a spread of confidence (the increase in occurrence of one attribute when compared to another attribute) against the support along with the lift coefficient in the association rules strength distribution visualisation.

A screen shot of a graph

Description automatically generatedThe chart above visualizes relationships between support (x-axis), confidence (y-axis), and lift (color scale). Data points are shaded based on lift values, with darker shades indicating higher lift. The chart highlights patterns in association rules strength across different support and confidence levels.

We can also see the distribution of the item sets by its support and confidence.

A diagram with blue dots

Description automatically generated

Preliminary analysis of the association rule mining: Using this technique, we answered the first three objectives we set out. We can find important indicators for diabetes; we found associations between indicators, and we also compared two baskets of non-diabetic and diabetic people. Here are some of the answers to the questions we formulated so far:

1) We saw that Gen Health status, Diabetes Check and Blood Pressure check were one of the primary check-ups that people in this dataset undertook. We see that most of the people (200,000+) have no diabetes.

2) Based on association rule mining, we saw that the users took a lot of preventive tests done (General Health Checkup, Blood Pressure, Cholesterol) as a pre-requisite. This is in line with intuition about how people react during annual health surveys.

3) Patterns were recognised which were in-line with the intuition. People having no diabetes were very actively looking after themselves by doing regular checkups and physical activity. In the case of diabetic people, we see that they get their preventive tests done but also consume more than 1 vegetable in a day. This can be representative of the sample as being pro-active in trying to maintain a healthy lifestyle. This basket of diabetic people can be explored further.

**NEXT STEPS**: Our remaining objectives need to be accomplished. They are:

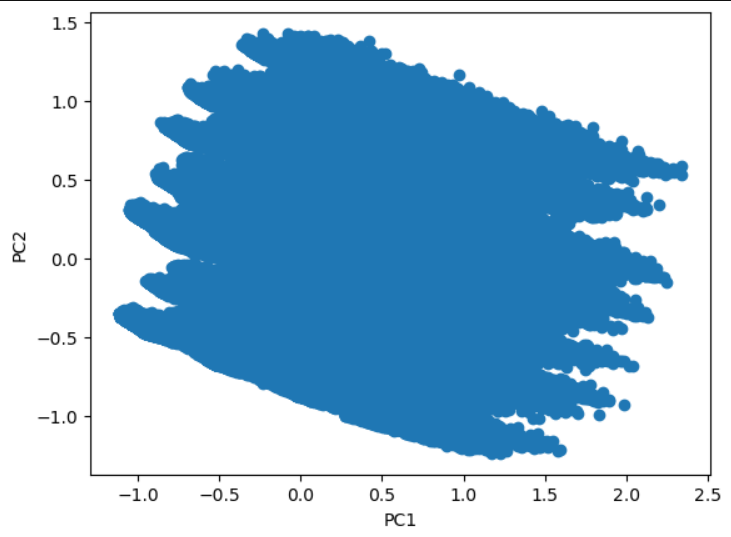
1. Finding Anomalies of certain individuals based on the data and share these risk factors with health officials i.e., people who have more than one risk factors but have not yet been diagnosed to be pre/diabetic.
2. How to reduce economic burden on government for by providing early detection and suggesting corrective action.

**ANOMALY DETECTION**

We took the original dataset and did scaling of all the features. We are using both Local Outlier Factor (LOF) method as well as PCA (Principal Component Analysis) to detect outliers. Our objective is to detect these outliers by both methods and do a comparison of how both techniques performed.

**Local Outlier Factor (LOF)**: In this method, we created an LOF model with the 10 nearest neighbours and decided to consider the rest of the neighbours as outliers (this is the definition of an LOF score). The distribution of the diabetic people with outlier score of more than is as follows.

A graph with a bar

Description automatically generated**Principal Component Analysis (PCA)**: We used PCA to reduce all the features into 2 principal components (PC1 and PC2). In this data frame, we added the LOF Outlier Score Indicator to do a comparison between the PCA and LOF methods.

A blue and orange dot diagram

Description automatically generated

Here, we observe that using PCA, we are not able to differentiate between the outliers as there are a lot of data points, however, when the LOF scores are superimposed, we can observe which points are considered outliers in the dataset. We see that LOF is more suitable for datasets where relationships between features are complex. PCA may not have a straightforward interpretation of the features, but LOF scores are detecting outliers effectively. If we expanded PCA into a third dimension or a fourth, we could have seen the outliers.

**RECOMMENDATION SYSTEM**

As part of the recommendation system, we will focus on finding similar people with the diabetic conditions based on their health characteristics. We first reduced the dimensions of the dataset using PCA, we then used co-sine similarity to find the nearest person. We then built a recommendation model in which the reduced dimensioned dataset was utilised. After building the recommendation system, we selected a random person with the person identifier. When we fed this identifier to the model, it gave out two persons with similar features as the test person.

For example, when we ran the code with person identifier 68105, it gave two similar persons with IDs 89909 and 171623. We observed that the suggested persons were non-diabetic. So, if we could recommend person 68105 to change his lifestyle and try to imitate either of the suggested person’s lifestyle, he/she can defeat diabetes.

A screenshot of a computer

Description automatically generated

Therefore, we can run this recommendation model to the authorities, and they can cluster similar patient profiles and run awareness campaigns to defeat diabetes.

**MANAGERIAL IMPLICATIONS**

* **Early Detection**: The findings underscore the importance of early diagnosis through regular health screenings (e.g., cholesterol checks). Organizations can implement programs promoting preventive healthcare to reduce future diabetes cases.
* **Targeted Health Campaigns/Awareness**: By understanding the lifestyle habits associated with diabetes risk, health authorities can design targeted awareness campaigns that encourage healthier behaviours among at-risk populations.
* **Resource Allocation**: Insights from the data can guide healthcare providers in allocating resources more effectively towards preventive measures rather than reactive treatments. It can also help insurance providers to better understand the risk profile of patients and curate better, human-centric insurance policies.

**CONCLUSION AND RECOMMENDATIONS**

This dataset highlights critical health parameters that can lead to diabetes detection among people. It emphasises on the need of early detection and lifestyle modifications for people who are already diabetic. One positive aspect of this dataset is that 84.2% of respondents are non-diabetic. This is a very pro-active sample of the population and might not be representing the exact picture of the United States. Hence, more scrutiny or data collection is required to give better insights/recommendations. Following is some of the recommendations from this dataset:

1. **Implement Regular Health Screenings**: Encourage routine cholesterol checks and other preventive tests among at-risk populations to facilitate early detection of diabetes.
2. **Promote Healthy Lifestyles**: Develop awareness programs that focus on physical activity and nutrition education, particularly emphasizing the consumption of vegetables and engaging in some sort of physical activity.
3. **Enhance Collaboration with Health Officials**: Share identified risk factors and anomalies with healthcare providers to enable timely interventions for individuals at considerable risk of developing diabetes.
4. **Expand Recommendation Systems**: Utilize the developed recommendation model to create personalized lifestyle change plans for individuals based on their health profiles, potentially integrating this into public health strategies.

By using these recommendations, governments/ public health organizations can mitigate the impact of diabetes in their community and facilitate better human focused policy making.