Biochemical Indicators and Lifestyle Diseases: A Data Mining Approach

Angel George
Masters of Data Science
Indiana University Bloomington
georgear@iu.edu

Brillia Benny
Masters of Data Science
Indiana University Bloomington
brbenny@iu.edu

Chaithra Lal Nair
Masters of Data Science
Indiana University Bloomington
cnair@iu.edu

project-georgear-brbenny-cnair

ABSTRACT

The emergence of data-mining technology in the field of medical research has marked a transformative shift in how healthcare data is utilized. With the increase in computational capabilities and the sheer amount of electronic health record data available, this industry has observed a shift in the information processed and insights generated. Researchers can now analyze patient data more thoroughly, thanks to the incorporation of data mining techniques, which have made it possible for them to forecast outcomes, identify risk factors, and customize interventions with exceptional accuracy and precision. The discovery of minor connections and trends that would have been missed in more manual, traditional analysis has been made easier by technological advances. Furthermore, data mining has sped up the process of identifying new biomarkers, refining treatment plans, and even forecasting the beginning of illnesses before the appearance of clinical signs. By leveraging data mining techniques on a diverse dataset sourced from laboratory tests, this study aims to identify key risk factors associated with lifestyle-related disorders such as cholesterol, blood pressure, etc. Through the application of classification, clustering, association rule mining, and feature selection, this project seeks to establish a comprehensive understanding of the predictive models for disease susceptibility, ultimately aiding in the development of proactive healthcare strategies.

Keywords: healthcare, outliers, decision tree, artificial neural networks, k-nearest neighbors, biochemical indicators, principal component analysis

1. DATA

Unlike the conventional approach where a cleaned and prepared dataset was selected directly from online sources such as Kaggle, a comprehensive dataset through the integration of diverse biochemical indicators associated with lifestyle diseases was created. The 2015-2016 Laboratory Data curated by National Health and Nutrition Examination Survey (NHANES), a part of Centers for Disease Control and Prevention (CDC) as a part of annual survey was chosen for the purpose of this analysis.

NHANES takes biological specimens (biospecimens) for laboratory analysis to offer precise information about the health and nutritional condition of participants. The gender and age of survey participants at the time of screening determine their eligibility for various laboratory tests. The biospecimens were collected in the mobile examination center (MEC). This includes blood, urine, and other types of specimen collection, processing, storage, and shipment. The MEC's controlled environment allowed laboratory measurements to be taken under identical conditions at each survey site.

The data was analyzed using Python and post handling outliers and missing values, techniques such as Principal Component Analysis (PCA) and neural networks were used for the assessment of risk factors of the lifestyle disease

2. BRIEF DESCRIPTION OF LABORATORY TESTS

For the analysis, the following laboratory tests were selected:

2.1. Indicator Variables:

2.1.1. Acrylamide & Glycidamide:

While the research is ongoing regarding the effects of acrylamide and glycidamide, there are studies which show a positive correlation between dietary acrylamide intake and higher total and LDL cholesterol levels and diabetes in adults. Information on exposure to these chemicals in the general population is needed to assess potential health effects associated with this exposure and to monitor changes in exposure over time. *(acry, gly_mide)*

2.1.2. Albumin & Creatinine - Urine:

Albumin and Creatinine levels in urine serve as valuable indicators of kidney function. Albumin, a protein, is typically retained in the bloodstream but can leak into urine when kidneys are not functioning optimally. Creatinine, a waste product from muscle activity, is measured alongside albumin to provide a standardized assessment of kidney function. *(albmn, creatin, uACR)*

2.1.3. Apolipoprotein B:

Apolipoprotein B is a protein associated with low-density lipoproteins (LDL) and very low-density lipoproteins (VLDL), playing a crucial role in lipid metabolism. Elevated Apolipoprotein B levels are linked to an increased risk of cardiovascular diseases. *(apolprtn)*

2.1.4. Complete Blood Count (CBC):

Arguably one of the most common lab tests routinely conducted, a CBC report can offer valuable insights into common lifestyle diseases. Elevated levels of white blood cells, platelets counts and red blood cell counts can indicate an elevation in diabetes, cholesterol and blood pressure. While CBC results cannot definitively diagnose the above conditions, the relationships are certainly present to delve deeper into. *(wbc_cnt, lymph_cnt, mono_cnt, neutro_cnt, eos_cnt,eso_cnt, rbc_cnt, hemo, hydctine, mcv, mch, mchc, rdw, plt_cnt)*

2.2. Response Variables:

2.2.1. Cholesterol - Total:

Total cholesterol levels, encompassing low-density lipoproteins (LDL), high-density lipoproteins (HDL), and very low-density lipoproteins (VLDL), are crucial indicators of cardiovascular health. Imbalances in cholesterol levels can significantly impact the risk of cardiovascular diseases. *(tot_chol)*

2.2.2. Glycohemoglobin:

Glycohemoglobin is a crucial marker used to measure average blood sugar levels over a few months. It plays a significant role in the diagnosis and management of diabetes mellitus. *(gly_hem)*

2.2.3. Systolic and Diastolic Blood Pressure

Systolic blood pressure represents the peak pressure in your arteries when your heart beats. Diastolic blood pressure represents the pressure in your arteries when your heart is at rest between beats. *(sys_bp, dia_bp)*

3. DATA PREPROCESSING

In any data mining problem, data preprocessing is the most important activity. The first step in using any algorithms is to clean the data. This entails dealing with a wide range of potential problems, including missing values, outliers, and different kinds of properties. We identify a variety of attribute types in the data we have been given, from continuous variables to categorical variables. Data preparation is more than just data cleaning; it includes important things like feature selection and data reduction. Data preparation is a complex procedure that prepares the ground for further analysis and creates the framework for reliable and informative data mining outcomes.

3.1 Missing Data Imputation

All the variables included in the analysis had missing values in their domain of varying degree. Out of 9544 instances, Acrylamide and Glycidamide had 7131 (74.7%) and 7277 (76.2%) missing values. Due to the very high missing values, the two tests were dropped from further analysis and imputing a large number of missing values would not lead to reliable results. 11 of the variables *(apolprtn, rbc_cnt, hemo, hydctine, mcv, mch, mchc, plt_cnt, tot_chol, dia_bp, sys_bp)* follow a fairly normal distribution. For these variables, median was imputed so as to not change the distribution. The above variables had less than 15% of the instances missing.

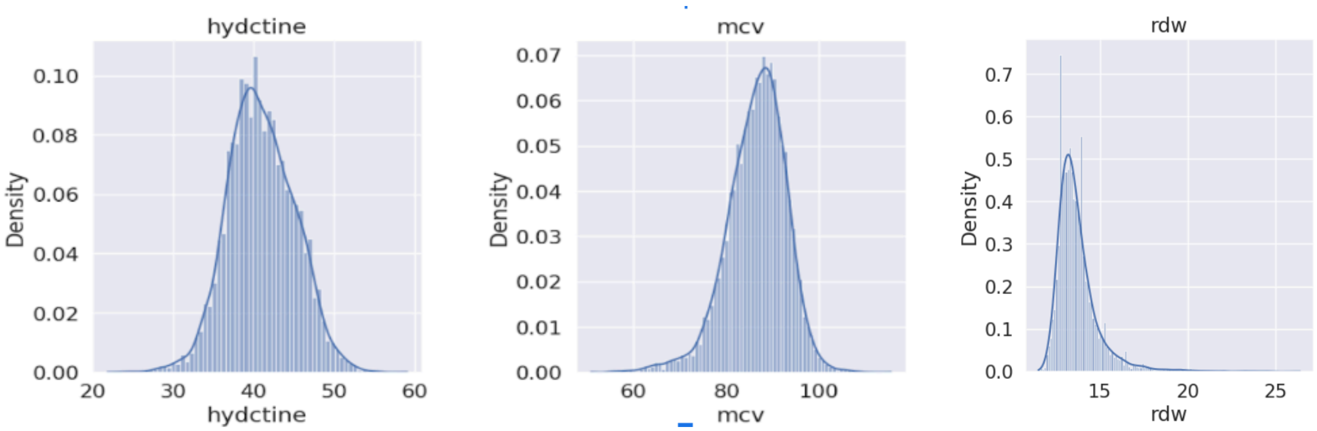

Figure 1

Distribution of hydctine, mcv, rdw

The target variables *(tot_chol, gly_hem, sys_bp, dia_bp)* have a higher percentage of missing values. Cholesterol had 2288 (24%) instances, glycohemoglobin 3218 (33.7%) and systolic and diastolic blood pressure had 2181 (22.8%) instances missing. The remaining indicator variables *(wbc_cnt, lymph_cnt, mono_cnt, neutro_cnt, eos_cnt, baso_cnt, rdw, hscrp)* have missing instances of 14-17% of the entire data. For the remaining variables, **K-Nearest Neighbours (KNN)** technique was used. KNN is a versatile machine learning technique that can be applied to various tasks, including imputing missing values. A k of 5 was selected and the missing values in the missing variables were imputed.

Certain variables such as *rdw* (as show above) and lymphocyte count (*lymph_cnt*) were highly skewed to the target variables. For missing values that were heavily skewed, KNN was used to impute them.

3.2 Outliers

An outlier is any observation that deviates from the sample's general pattern. Outliers in any data set have an impact on the training model, sometimes to an unexpected degree. Therefore, attention should be given to these outliers.

Outliers were determined based on the set of values that were possible for a person in the real-world setting. As such, outliers were observed in uACR, which is the ratio of albumin and creatinine, a key kidney function test. On further inspection, the instances which had abnormally high values of uACR had very abnormally high values of factors such as platelet count and lymphocyte count. This happens in the case of extremely sick patients (e.g. cancer). As this data was procured from a real world dataset, this seemed possible. To treat outliers, we have removed very extreme values (impossible) of uACR only.

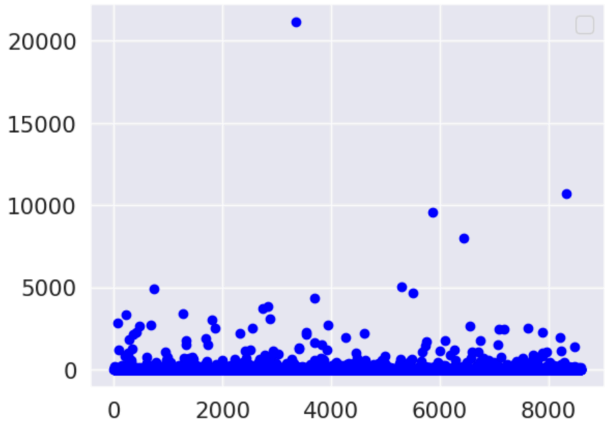

Figure 2.
Scatter Plot of uACR

3.3 Treatment of variables

For the analysis of the model, the target variables were converted to categorical by bucketing. Bucketing was done into three categories: Normal, Moderate and High. For Total Cholesterol, values less than 200 mg/dL were considered as normal, between 200 mg/dL and 240 mg/dL were considered moderate and above 240 mg/dL were considered high. For Diabetes, glycohemoglobin values less than 5.7% were considered normal, between 5.7% and 6.5% were considered prediabetic and above 6.5% were considered diabetic. For Blood Pressure, systolic values less than 120 and diastolic below 80 were considered normal, systolic between 120 and 130 and diastolic less than 80 were considered moderate and systolic greater than 130 and diastolic greater than 80 were considered high.

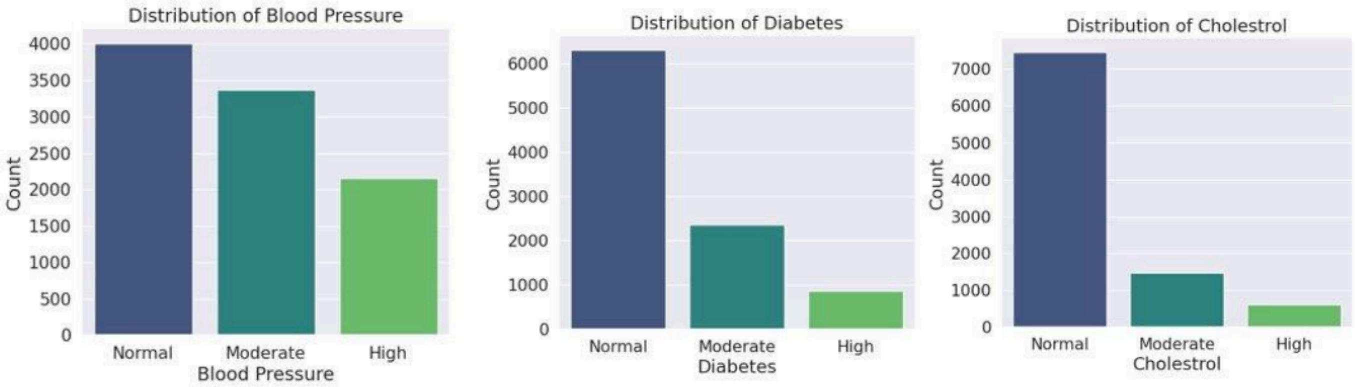

Figure 3
Histogram of target variables

3.4 Correlation Matrix

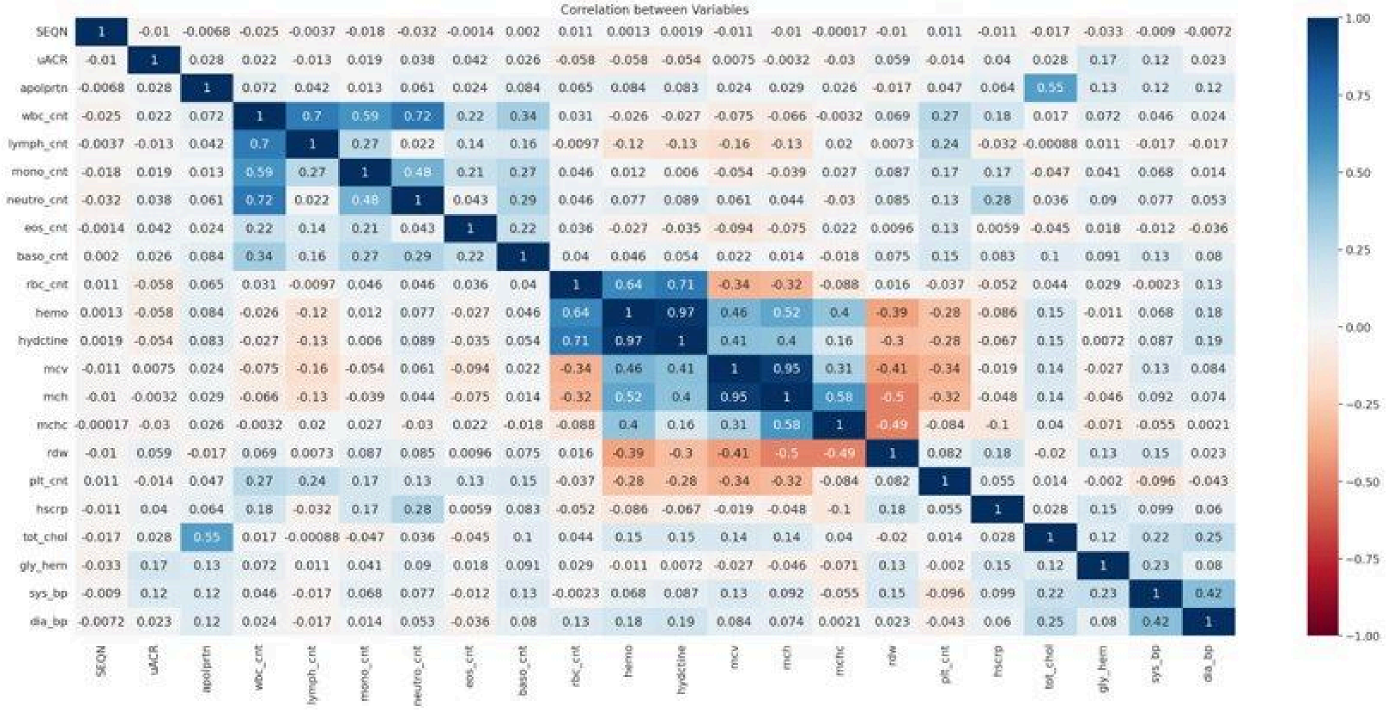

Figure 4
Correlation Plot

The correlation heatmap was meticulously constructed; nevertheless, it unveiled restrained or statistically insignificant correlations among the variables. This outcome can be attributed predominantly to the prevailing health status within the study population, wherein a substantial majority falls within the 'healthy' or 'normal' category, with only a minority exhibiting moderate to high severity levels.

4. ALGORITHM AND METHODOLOGY

The project mainly deals with using techniques like Principal Component Analysis (PCA), Neural Networks and Decision Tree.

4.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction in data analysis. It helps us understand and visualize complex data by compressing it into a smaller set of more relevant features and hence, improves the performance of other machine learning techniques

In the Principal Component Analysis (PCA) conducted on the dataset, the cumulative explained variance of the first two principal components, PC1 and PC2, was found to be 20.73% and 15.67%, respectively and for PC3 it was found to be 11.66%. As seen in the plots below PC1 and PC2 could not show any distinct clusters. Moreover, PC1, PC2 and PC3 together could only explain 48% of the variance in the total data. Thus, the choice to retain all variables to maintain the integrity of the dataset and preserve potentially important information encoded in the variables.

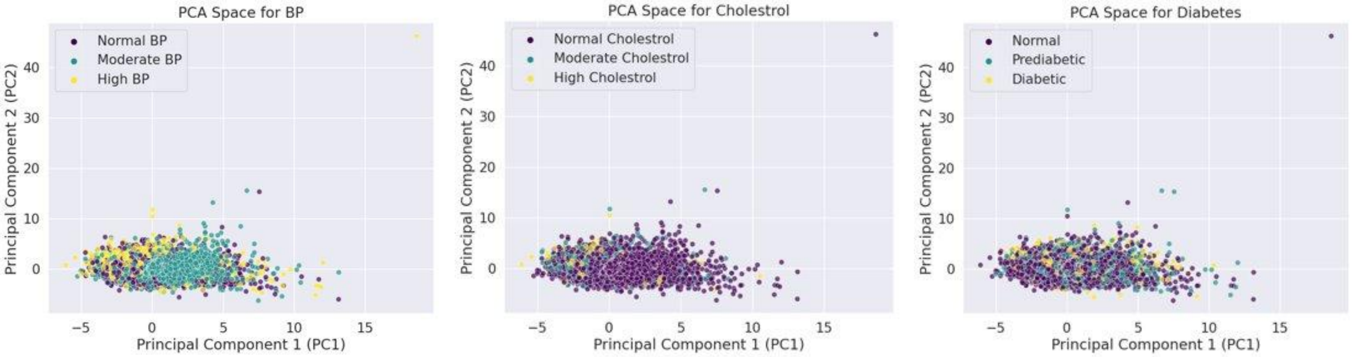

Figure 5
PCA Plot

4.2 Neural Networks

Artificial Neural Networks (ANNs) are computational models inspired by the structure and functioning of the human brain. They are a subset of machine learning algorithms that aim to mimic the way the human brain processes information and learns from data.

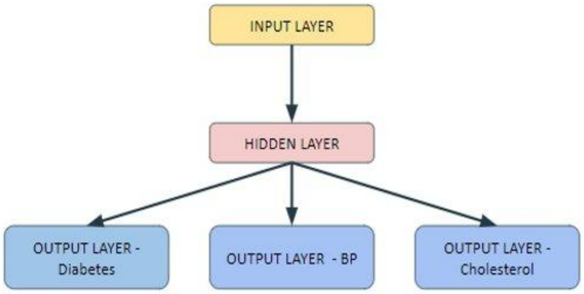

Figure 6
ANN Architecture

This framework allows for a comprehensive analysis of biochemical indicators related to specific lifestyle diseases, including cholesterol levels, blood pressure (BP), and diabetes. The multi-class (depending on Severity - Normal, Moderate, High) multi-output design accommodates various biochemical indicators, each serving as a distinct input feature. Dropout and regularization has also been implemented for this model. This inclusive approach ensures that the neural network can capture the nuanced relationships and interactions between different indicators, providing a holistic understanding of the biochemical landscape. For instance, inputs could encompass variables such as lipid profile, glucose levels, and other relevant markers. Simultaneously, the multi-output aspect addresses the unique characteristics of each lifestyle disease. By having separate outputs for cholesterol, BP, and diabetes, the neural network can simultaneously predict and delineate the influence of biochemical factors on each disease outcome.
The F1 score that we got for the neural network model is **0.6325** which, considering the complexity of the dataset, could be considered as a fairly good score.

4.3 Decision Trees

Decision Trees are a popular machine learning algorithm used for both classification and regression tasks. These trees use a hierarchical structure with nodes representing decision points and leaves denoting final outcomes. During the building process, decisions are made based on criteria like Gini impurity or entropy, and the tree is constructed recursively by splitting the dataset until specified conditions are met.

The accuracy of the decision tree w.r.t the different target diseases are as follows:

a. Blood Pressure - 0.4746
b. Diabetes - 0.5485
c. Cholesterol - 0.7271

The other metrics that were measured for the decision tree were precision, recall and F1 score. Given below are the description for the three different target diseases:

	Cholesterol			**Diabetes**			**Blood Pressure**		
class	**precision**	**recall**	**f1-score**	**precision**	**recall**	**f1-score**	**precision**	**recall**	**f1-score**
0 - Normal	0.86	0.83	0.85	0.71	0.68	0.7	0.51	0.50	0.50
1 - Moderate	0.29	0.33	0.31	0.29	0.33	0.31	0.51	0.53	0.52
2 - High	0.34	0.35	0.35	0.34	0.35	0.19	0.35	0.34	0.35

5. DISCUSSION

Since Decision Trees cannot handle multi-class multi-output design for multiple diseases, ANN would be a better choice. Also the decision tree didn't perform well with classes that didn't have a good representation in the data. For the ANN, the F1 score obtained on the test data was 0.6325 for cholesterol, blood pressure and diabetes in spite of the imbalanced dataset whereas the decision tree performed not up to mark for moderate and high classes of data. The decision tree gave a good F1 score for class 0 (Normal) due to the imbalance in the instances in that category.

We believe that the relatively low performance of the model occurred since the dataset was obtained from real-world sources.

6. USE CASE

Prediction of lifestyle diseases using common lab tests could be a crucial step in the early detection of diseases. While this is not comprehensive, this will enable patients to adopt measures that lead to healthier lifestyle and further measures. Such models would also be instrumental in creating personalized health plans based on various characteristics such as age, gender and economy. The usage of predictive models will help determine possible risk factors for those who are not aware of the various biochemical indicators and the role they play in influencing lifestyle diseases.

7. FUTURE STEPS

It is possible to increase the model performance by including more training instances and also through hyperparameter tuning. Furthermore, apriori algorithms can be utilized to discover relations between the various biochemical indicators and target variables. The usage of Gaussian processes can be used to obtain uncertainty estimations along with predictions. This uncertainty estimation is valuable, especially when dealing with medical diagnoses where the consequences of false predictions can be critical.

8. REFERENCES

Firoozeh Hosseini-Esfahani, Niloofar Beheshti, Amene Nematollahi, Glareh Koochakpoor, Soheil verij-Kazemi, Parvin Mirmiran & Fereidoon Azizi. The association between dietary acrylamide intake and the risk of type 2 diabetes incidence in the Tehran lipid and glucose study. *Scientific Reports*, 2023

Shi Gu, Anping Wang, Guang Ning, Linxi Zhang, Yiming Mu, Insulin resistance is associated with urinary albumin-creatinine ratio in normal weight individuals with hypertension and diabetes: The REACTION study. *Journal of Diabetes*, May 2020

Li Ming, Duan Wang and Yong Zhu. Association between urinary albumin-to-creatinine ratio within normal range and hypertension among adults in the United States: Data from the NHANES 2009–2018, *Clinical Cardiology*, June 2023

Maud Ahmad, Allan D. Sniderman, and Robert A. Hegele. Apolipoprotein B in cardiovascular risk assessment, *Canadian Medical Association Journal*, Aug 2023

Takanori Hasegawa, Rui Yamaguchi, Masanori Kakuta, Kaori Sawada, Kenichi Kawatani, Koichi Murashita, Shigeyuki Nakaji, Seiya Imoto. Prediction of blood test values under different lifestyle scenarios using time-series electronic health records. *PLOS One*, Mar 2020

Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, Machine Learning and Data Mining Methods in Diabetes Research, *Computational and Structural Biotechnology Journal,* Jan 2017