Metabolic Syndrome Risk Predictor

Patryk Slomka (2049498)
JM0150-M-6, Data Mining
JADS
's-Hertogenbosch, the Netherlands
p.slomka@tilburguniversity.edu

***Abstract*—**This study focuses on the predictive modeling of the Metabolic syndrome (MetS), the cluster of factors that lead to type 2 diabetes and heart disfunctions. The used data comes from National Health and Nutrition Examination Survey (NHANES) 2013-2014. The research involves machine learning and data mining techniques to accurately assess the MetS risk levels for individuals. The findings underline the need for machine learning in healthcare and highlights that the further refinement of MetS description would help in the personalized healthcare interventions.

Keywords—metabolic syndrome, machine learning, risk prediction, data mining, predictive modeling

I. Introduction

Health is becoming more and more important in today's world. The presence of fast-foods, alcohol, smoking and sedentary lifestyle leads humanity into the health problems trap. The Metabolic syndrome (MetS) is a clustered health disorder that can potentially develop into type 2 diabetes, heart diseases and stroke (Han & Lean, 2015). The prevailing trend indicates that this phenomenon is becoming more and more "popular" around the world, causing researchers to take closer look into its roots and probable causes. While there are many different definitions of MetS, as indicated by research of Eckel, Zimmet & Alberti (2005) and Samson & Garber (2014), they all point towards the same cluster of abnormalities. These include the higher fat body percentage, higher levels of glucose and triglycerides, and drinking and smoking. The causes are very multifactorial, but the physical inactivity and high fat diet are strongly linked to development of MetS (Yang et al. 2022). Given those serious health risks, it is crucial to identify the risk of each individual falling into MetS as early as possible to help overcome this disease. For that purpose, the use of Machine Learning (ML) comes in hand. By using the National Health and Nutrition Examination Survey data from 2013-2014, the paper builds on the creation and usage of the MetS risk prediction tool.

II. Problem statement

A. Real-world case and objectives

The researchers come into agreement that there is no complete definition of Metabolic syndrome (MetS). Although there is a consensus on the core components of MetS, it is still not widely spoken about. Early identification of the problem and the risk group an individual is in, may lead to better and timely medical intervention. By analyzing the key factor conditions that cluster into the MetS risk prediction, the Healthcare providers could personalize their approaches to individuals from different risk groups. This would include faster recognition of the state the user is in, which could lead to providing better help. By using data mining and Machine Learning techniques, this research seeks to develop the predictive models that could accurately assess the risk of an individual of MetS. Moreover, it aims to gain additional insights into the correlations between diet and lifestyle of the individuals, as these are one of the causing factors of MetS. The main research questions include:

Q1: "Are specific dietary patterns or behaviors associated with higher glucose levels and blood pressure?"

Q2: "How effective are machine learning models in predicting metabolic syndrome risk levels (low, moderate, high)?"

B. Motivation

The researchers underline that the varying definitions of MetS provide additional challenges in comparing the findings and developing more accurate guidelines for the healthcare providers. Genetics, lifestyle, and environmental influence in the form of smoking and alcohol, all add to the development of MetS. However, the studies often include the binary variables which might limit the accuracy of the simpler models already used (Tavares et al. 2022). Based on the importance of analyzing the MetS risk for the individuals, the researchers are looking for the more accurate predictive models to help them assess the user into groups. The research by Tavares et al. (2022) found out that the Neural Network (NN) performs better in analysis than the logistic regression. Other studies made use of checking which preliminary factors influence to the risk of MetS, including BMI, waist circumference, blood pressure and glucose levels (Kim, Mun, Lee, Jeong & Baek 2022). However, there are still challenges connected to imbalance of data, and missing data. To overcome those challenges, more sophisticated use of Machine Learning techniques would be useful to build a tool that can predict and identify the individuals who are at higher risk in terms of MetS.

C. Data Mining tasks

After stating the real-world case problem in the form of MetS, the relevant datasets had to be chosen. To ensure the provided data will be adequate for analysis, the first thing to undergo was to research about the data from the provided variables descriptions of every dataset. The choice was made to include 5 out of 6 datasets, leaving the medications one. The purposes of that process were, that the final five datasets included variables that were crucial for the MetS causes analysis and the medications might be more relevant in the context of overcoming MetS, which is a future action. After choosing the datasets, the first task to complete was to perform data cleaning. As National Health and Nutrition Examination Survey (NHANES) data included missing values, each dataset had to be analyzed and from that analysis, columns with more than 5000 missing values have been removed. The reason for that was that missing data could negatively impact the model performance. After that, the datasets were merged into one based on user-id variable

(SEQN). Based on the previous descriptions from the provided datasets, 26 features were chosen from the merged dataset. Those were based on the research by Han & Lean (2015) and Eckel, Zimmet & Alberti (2005). The next step involved applying feature engineering to create 2 new variables form the variables about blood pressure (diastolic and systolic). To receive more insights about the remaining data, the correlation heatmap was created, as well as the boxplots and pairplots of the most crucial variables for the analysis.

Following that, the data analysis also included two clustering methods, KMeans and Hierarchical clustering, to provide more information on the patterns in the data. To find even more insightful relations between the variables, Association rules were used, in the form of Apriori algorithm, which helped to find answers to two more questions related to MetS. Based on those actions, the data were ready to build the predictive models. The first ones to create were logarithmic regression and decision trees, which with fine-tuning gave impressive results in terms of accuracy and ROC AUC score. However, as we are dealing with health and risk factors, the more complex models are need for more accurate predictions. The choice went for the XGBoost model, which provided much better results. On top of that, the neural network was built to compare with the past models. Based on the evaluation metrics like accuracy, precision, recall, F1-score and ROC accuracy, the models could be compared and chosen for the best implementation within the prediction tool. The complete pipeline is presented in Figure 1.

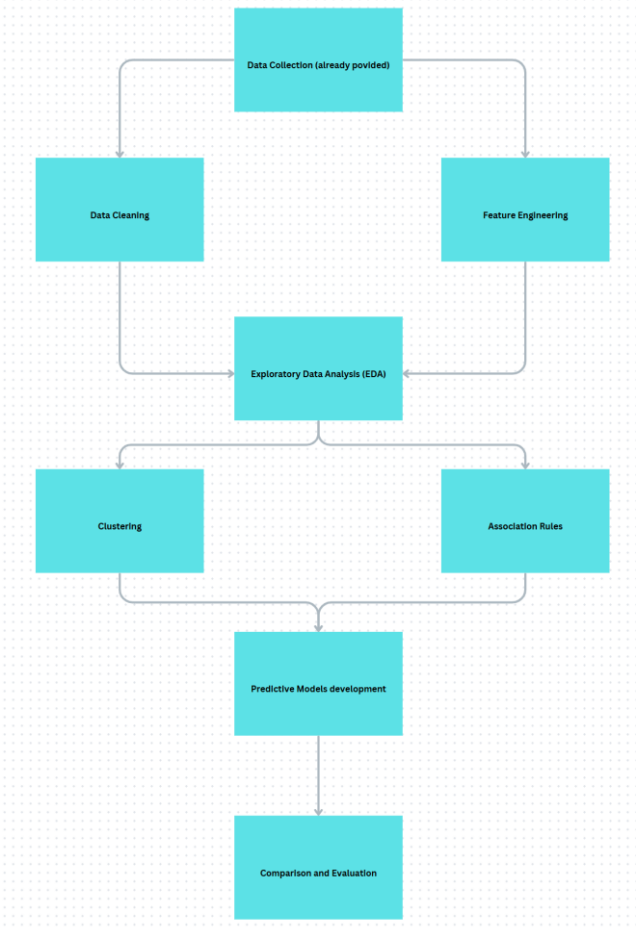

Figure 1 – *Data Architecture Pipeline*

III. Exploratory Data Analysis (EDA)

A. *Challenges and Limitations*

As stated earlier in the Data Mining tasks, the data were explored and analyzed to ensure the best performance of the models. The first challenge involved the missing values in the datasets. From the analysis, it occurred that there is a lot of columns with more than 5000 missing values, which had to be removed to ensure model's accuracy. Such approach was also beneficial when it comes to computational complexity, as with less amount of data, the machine could perform the tasks in lesser time. After choosing the crucial variables for the MetS analysis, the missing values were once again checked to avoid wrongly stated predictions by the model. As there were multiple variables for the blood pressure of the users, the new variables had been created out of those, by taking the averages. This included the diastolic and systolic blood pressure. The EDA also shown that there are outliers in some of the MetS features. However, there was several crucial features that could include outliers to better analyze the risk case of MetS. These included e.g. glucose levels and before-mentioned blood pressure. For the remaining variables with outliers, the method of winsorizing has been applied, where "the smallest and largest observations are not removed, but rather transformed by pulling them in" (Wilcox, 2003).

B. *Visualizations*

For the univariate analysis, the crucial features have been also summarized using descriptive statistics. Next to that, the multivariate analysis showed additional insights into the data. To see how the variables correlate with each other, the correlation heatmap has been created. It revealed strong correlations among certain variables, such as waist circumference ('BMXWAITS') and triglycerides ('LBXSTR'). This provides more insights into the impact on MetS risk.

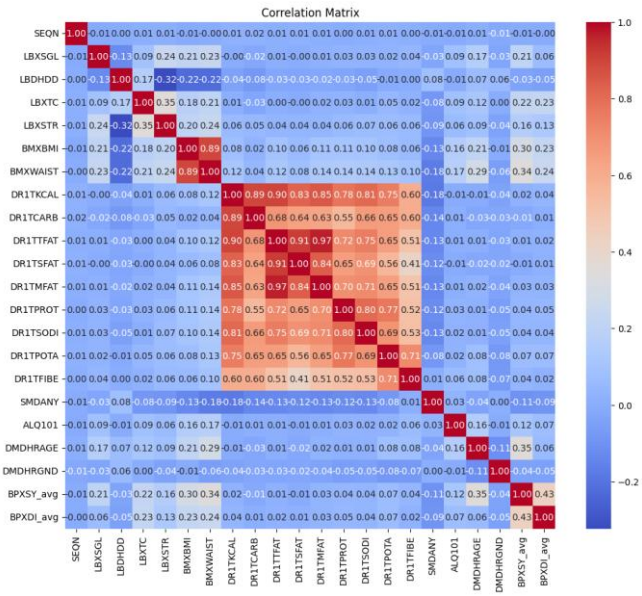

Figure 2 – *Correlation Heatmap*

On top of that, the boxplots of most important variables have been created to gain even more insights into the distribution of the data. They were chosen based on research

by Han & Lean (2015). Based on the distribution of the data, the glucose levels ('LBXSGL') showed lots of outliers above the upper whisker, which could include the individuals with hyperglycemia, which is one of the key risk factors for diabetes type 2 and MetS (Alexander, landsman & Grundy 2006).The cholesterol variables ('LBXTC' and 'LBDHDD') show that for the HDL cholesterol, most of the individuals are allocated below the threshold for the MetS, which could lead to heart diseases (Ali, Wonnerth, Huber & Wojta 2012). Another interesting factor to analyze was the waist circumference ('BMXWAIST'). As the data were collected in the United States of America, the measures for population for this region had to taken into account. Based on that, the data indicated that most of the individuals exceeded the region-specific thresholds for MetS, which could indicate the central obesity, one of the key risk factors.

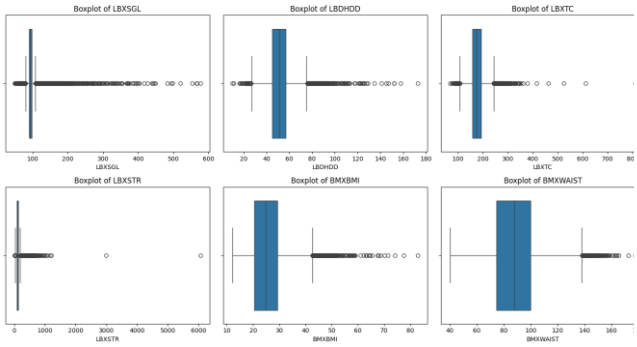

Figure 3 – *Boxplots of relevant features*

Next to the boxplots, the pairplots have also been created to highlight the relationships between the key selected variables. The obtained scatterplots showed that there are insightful trends in the data. An interesting example from the paiplots is the negative correlation between the HDL cholesterol and triglycerides, which are aligned with the known MetS risk factors. Also, another correlation that is important for the analysis is the fact that age trends show that with age increasing, the glucose and triglycerides levels also increase, but the BMI of the individual decreases.

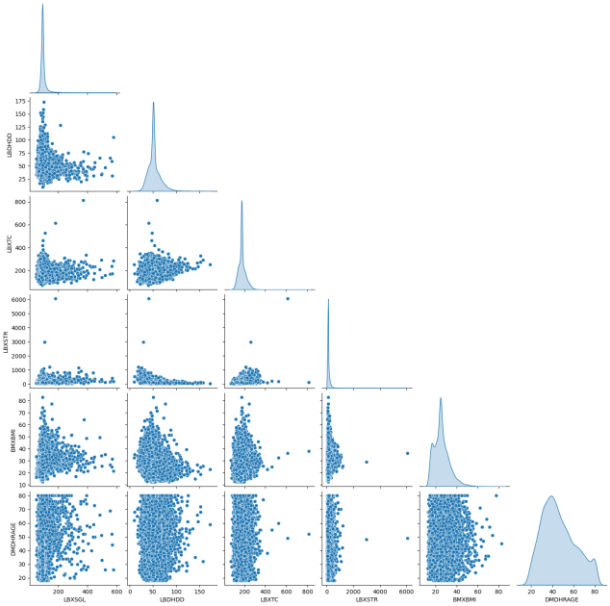

Figure 4 – *Pairplots of relevant features*

C. *Clustering*

The next part of data analysis included clustering methods. The above-mentioned plots showed which variables would be relevant for the cluster formation. The analysis used KMeans and Hierarchical clustering and calculated the silhouette score for each method to compare them. The features chosen for this analysis included e.g. glucose levels ('LBXSGL'), HDL cholesterol ('LBDHDD'), triglycerides ('LBXSTR'), waist circumference ('BMXWAIST'), and average blood pressure ('BPXSY_avg', 'BPXDI_avg'). From the KMeans clustering, three clusters were obtained that provided the best separation of the individuals. Cluster 1 represented the low-risk group of MetS, Cluster 2 the moderate-risk group and Cluster 3 the high-risk group. The silhouette score for the KMeans clustering was 0.27. However, for the hierarchical clustering method, which was created and performed using the Ward's linkage method, the dendrogram showed the same clusters as in the KMeans clustering. Moreover, it gave additional insights that Clusters 1 and 2 were closely related, and Cluster 3 was much more separated. The silhouette score in that case was 0.75. In both analyses the clusters indicated that the individuals with high-risk of MetS show the higher glucose, triglycerides and waist circumference factors, and lower HDL cholesterol. This insight is in line with the previously mentioned research of well-known risk factors for MetS.

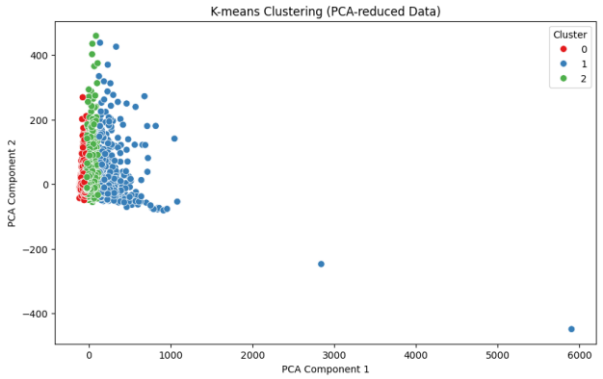

Figure 5 – *KMeans Clustering*

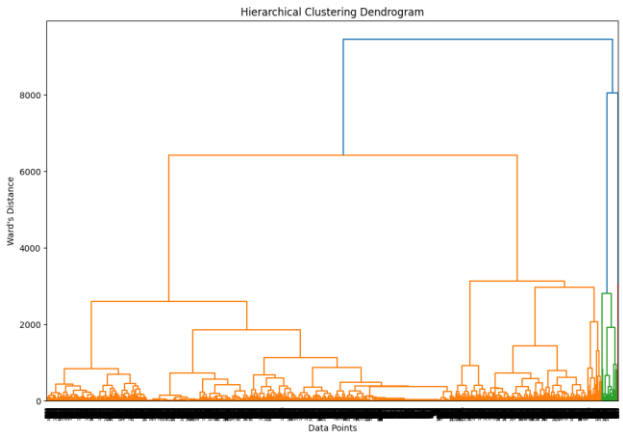

Figure 6 – *Hierarchical clustering dendrogram*

D. *Association rules*

For the research purposes, the paper addresses two questions that are based on the association rules. They were

used to identify the important relationship between certain features. The first one addresses the question:

Q1: "Do smokers who consume high sodium are likely to have high blood pressure?"

Which corresponds to the factors that increase risk of MetS for the individuals. To make use of association rules, the Apriori algorithm was used, which calculates the significant metrics to evaluate them. These include support, which measures how often does the itemset is present in the dataset; confidence, which states the probability of the itemset occurring together; and lift, which gives the evaluation of how strong the rule is. Therefore, the first rule was created for the purpose of finding whether there is an association between smoking, high sodium intake (from the dietary products) and high blood pressure. It appears that only one rule significant rule was found, with the extremely low support of 0.02. This indicated that such rule is not very frequent, however it also showed that smokers who have high sodium intake are 1.57 (lift) times more likely to have high blood pressure as compared to other non-smoking individuals.

	antecedents	consequents	support	confidence	lift
6	(Smoker, High_Sodium)	(High_BloodPressure)	0.020147	0.252153	1.567289

Figure 7 – *Q1 Association rule output*

The second question aimed to find more insights about the dietary factors and higher glucose levels of the individuals:

Q2: "Are there specific dietary factors (e.g., sodium, potassium, saturated fats) associated with higher glucose levels?"

From the analysis (using again Apriori rule), the first rule (Rule 14) has a support of 0.245, showing that the users with high sodium intake from their diet are 2.02 times more likely to also consume higher amounts of saturated fats and potassium, than the individuals that don't have that much sodium intake in their diet. Another interesting rule (Rule 11) showed a strong confidence of 0.848 the same lift as for Rule 14. This could suggest that there is a strong association between the dietary factors that include potassium, sodium and saturated fats.

	antecedents	consequents	support	confidence	lift
14	(High_Sodium)	(High_SaturatedFat, High_Potassium)	0.245307	0.585503	2.024293
11	(High_SaturatedFat, High_Potassium)	(High_Sodium)	0.245307	0.848114	2.024293
13	(High_SaturatedFat)	(High_Sodium, High_Potassium)	0.245307	0.585366	1.892026
12	(High_Sodium, High_Potassium)	(High_SaturatedFat)	0.245307	0.792884	1.892026
15	(High_Potassium)	(High_SaturatedFat, High_Sodium)	0.245307	0.585641	1.844858

Figure 8 – *Q2 Association rule output*

E. Insights and next steps

The insights from clustering and association rules could be used for the healthcare providers to personalize the interventions for the individuals and provide more tailored treatment for them in case of used medications. The different clusters could also receive other types of medical attention, e.g. Cluster 2 of moderate-risk group could be provided with dietary counselling to mitigate the risk to lower level. The association rules also revealed that dietary intake of certain features could lead to higher glucose levels, therefore providing insightful observations what to minimalize in the diets of individuals. That information aligns with the business objectives which were earlier mentioned in the paper. Understanding which risk factors are contributing more to the risk of MetS and helping to provide the more target intervention towards certain individuals. The next step of analysis focuses on building the predictive models, which with the use of Machine Learning can accurately analyze the individual cases and state to which risk group does it belong.

IV. MODELING AND DEPLOYMENT

The next step of research began with creation of the function to choose the certain thresholds of the relevant features for the models' purpose. The list of thresholds include:

- Gender ('DMDHRGND')
- Waist circumference ('BMXWAIST') – based on the research of Samson & Garber (2014), the threshold for the United States was chosen for men (≥102 cm0 and women (≥88 cm)
- HDL Cholesterol ('LBDHDD') – based on research by Yu et al. (2020), threshold was set for men (<40 mg/dL) and women (<50 mg/dL)
- Average blood pressure ('BPXSY_avg' and 'BPXDI_avg') – for systolic blood pressure the threshold was set same for both genders (≥130 mmHg) and for diastolic (≥85 mmHg), as indicated by Eckel, Zimmet & Alberti (2005)
- Triglycerides ('LBXSTR') - a threshold of ≥150 mg/dL is widely used to define hypertriglyceridemia in the context of MetS
- Glucose ('LBXSGL') – the indicated threshold by Ali, Wonnerth, Huber & Wojta (2012) is ≥100 mg/dL
- Dietary factors like calories intake, carbohydrates intake, total fat – chosen on the thresholds of United States individuals averages from Samson & Garber (2014), where the average daily calories intake is 2500, average carbohydrates intake is 300, and total fat of an individual is 80
- Smoking and alcohol ('SMDANY' and 'ALQ101') – indicating if the person smokes, and the second one if the person had at least 12 drinks I the past year, which doesn't necessarily show that the person is a heavy drinker, but indicates that drinking is a regularity and can be part of the lifestyle

Based on those factors, the low-risk of MetS would be stated if none of the conditions are met by an individual; moderate risk if one or two conditions are met; high-risk when more than two conditions are met. Based on this function, the dataset has a distribution as shown in Figure 9. However, as the classes are very imbalanced, this could lead to potential problems in the modeling and deployment of results, potentially harming the accuracy and trustworthiness of predictions. For that purpose, the analysis used Synthetic

Minority Oversampling Technique (SMOTE) to balance the dataset. Therefore, the final distribution was equal for all categories, as indicated in Figure 10.

```
MetS
Moderate    5193
High        4803
Low          179
Name: count, dtype: int64
```

Figure 9 – *Categories distribution before balancing*

```
Training set class distribution after SMOTE:
MetS
High        4174
Moderate    4174
Low         4174
Name: count, dtype: int64
```

Figure 10 – *Categories distribution after SMOTE*

A. Simple models

The first two predictive models that were created were Logistic regression and Decision Trees. Both of them were evaluated using multiple metrics to gain insights into weaknesses and strengths of each model. They included accuracy, precision, recall, F1-score, confusion matrix and ROC AUC score. The first one (Logistic Regression), achieved an accuracy of 87% and ROC AUC score of 0.96. after finetuning the model using cross-validation, the model did not improve its performance.

```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.87      0.89       960
           1       0.30      0.92      0.45        36
           2       0.89      0.86      0.88      1039

    accuracy                           0.87      2035
   macro avg       0.70      0.88      0.74      2035
weighted avg       0.89      0.87      0.88      2035


Confusion Matrix:
[[838  12 110]
 [  0  33   3]
 [ 76  65 898]]

ROC AUC Score: 0.9641
```

Figure 11 – Logistic regression model performance output

This is why the usage of the latter model was more crucial for the analysis. At first, the Decision Trees model achieved almost the same accuracy and ROC AUC scores as the Logistic Regression. However, after fine-tuning the model with Grid-search, the model performance increased significantly to the accuracy level of 95%. The model was more stable than the Logistic regression one.

```
Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.94      0.88       960
           1       0.49      1.00      0.65        36
           2       0.93      0.79      0.85      1039

    accuracy                           0.86      2035
   macro avg       0.75      0.91      0.80      2035
weighted avg       0.88      0.86      0.86      2035


Confusion Matrix:
[[902   1  57]
 [  0  36   0]
 [186  37 816]]

ROC AUC Score: 0.9608
```

Figure 12 – Decision Trees model performance output

```
Fine-Tuned Decision Tree Performance:
Accuracy: 0.9484
Precision: 0.9489
Recall: 0.9484
F1-Score: 0.9484
Confusion Matrix:
[[929   0  31]
 [  0  31   5]
 [ 65   4 970]]
```

Figure 13 – Decision Trees model performance after finetuning

B. Complex model

Afterward, the usage of more complex models came into action, with the XGBoost model. It immediately outperformed the previously mentioned models by achieving the stunning accuracy of 99.36% and ROC AUC score of 0.999. The confusion matrix of the model showed near-perfect classification across all categories. As the Logistic Regression and Decision trees provided solid results, the XGBoost emerged as the best model for the MetS risk prediction.

```
XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       960
           1       0.97      0.97      0.97        36
           2       1.00      0.99      0.99      1039

    accuracy                           0.99      2035
   macro avg       0.99      0.99      0.99      2035
weighted avg       0.99      0.99      0.99      2035


Confusion Matrix:
[[ 956    0    4]
 [   0   35    1]
 [   7    1 1031]]

ROC AUC Score: 0.9999
```

Figure 14 – *XGBoost model performance output*

C. Neural Network

As the last model in the analysis, the research used the Neural Network (NN) to check if the model would outperform already stunning results of XGBoost model. The baseline architecture of the NN included one hidden layer of 64 neurons and the ReLU activation function, and a single output layer which used softmax activation function. The suage of that function was made based on its performance in the multiclass classification problems, like the one in this MetS research. However, as the model did not perform very well with those features, it had to be fine-tuned and expanded. Therefore, the number of hidden layers were changed to two, each with 32 neurons. After checking the other activation functions like Tanh, the model stayed with them ReLU function. Moreover, the number of epochs was adjusted from 25 to 50, as the model performed better then. The batch size was also chosen to be 32, as this gave the right balance between accuracy and training time. The outcome of the NN after fine-tuning imposed accuracy results on the level of 92%. The model performed well, however not as great as the previous models, like XGBoost.

The final model used for the research purpose of predicting MetS risk was therefore the XGBoost model. The reason for that is not only the better accuracy of the model,

but its transparency and better explainability. The NN is often considered as black-box model due to its complex architecture, and harder explainability. The training time of NN is also significantly longer as compared to XGBoost and computationally more expensive, leaning the research direction even more into the XGBoost model.

```
Classification Report (Neural Network):
              precision    recall  f1-score   support

           0       0.92      0.94      0.93       960
           1       0.67      0.67      0.67        36
           2       0.93      0.91      0.92      1039

    accuracy                           0.92      2035
   macro avg       0.84      0.84      0.84      2035
weighted avg       0.92      0.92      0.92      2035

Confusion Matrix (Neural Network):
[[900   0  60]
 [  0  24  12]
 [ 80  12 947]]
```

Figure 15 – Neural Network model performance output

V. Limitations and Future Work

As the definition of MetS is not unified, the better refinement is needed to ensure the robustness of analysis and allow for comparison between research studies. For diagnostic criteria, this would be useful in building the next predictive models. Additionally, as the data was only focused o the population of US region, the need to more diverse data is needed to ensure the similar results in other parts of the world. Such action could give the healthcare providers more insights into how to tackle this issue. Based on the risk groups, the implementation of preventive strategies could be already done to ensure that each group receives more personalized approach which will be the most beneficial. By improving the risk predictive models and increasing awareness about the MetS, the researchers and healthcare providers can work together towards mitigating the risk of this syndrome and associated diseases.

References

[1] Han, T. S., & Lean, M. E. (2015). Metabolic syndrome. *Medicine*, *43*(2), 80-87.

[2] Eckel, R. H., Zimmet, P. Z., & Alberti, K. G. M. M. (2005). The metabolic syndrome. *The Lancet*, *365*(9468), 1415–1428. https://doi.org/10.1016/S0140-6736(05)66378-7

[3] Samson, S. L., & Garber, A. J. (2014). Metabolic syndrome. *Endocrinology and Metabolism Clinics*, *43*(1), 1-23.

[4] Yang, H., Yu, B., OUYang, P., Li, X., Lai, X., Zhang, G., & Zhang, H. (2022). Machine learning-aided risk prediction for metabolic syndrome based on 3 years study. *Scientific reports*, *12*(1), 2248.

[5] Tavares, L. D., Manoel, A., Donato, T. H. R., Cesena, F., Minanni, C. A., Kashiwagi, N. M., … Szlejf, C. (2022). Prediction of metabolic syndrome: A machine learning approach to help primary prevention. *Diabetes Research and Clinical Practice*, *191*, 110047. https://doi.org/10.1016/j.diabres.2022.110047

[6] Kim, J., Mun, S., Lee, S., Jeong, K., & Baek, Y. (2022). Prediction of metabolic and pre-metabolic syndromes using machine learning models with anthropometric, lifestyle, and biochemical factors from a middle-aged population in Korea. BMC Public Health, 22(1). https://doi.org/10.1186/s12889-022-13131-x

[7] Wilcox, R. R. (2003). Summarizing data. In Elsevier eBooks (pp. 55–91). https://doi.org/10.1016/b978-012751541-0/50024-9

[8] Alexander, C. M., Landsman, P. B., & Grundy, S. M. (2006). Metabolic syndrome and hyperglycemia: congruence and divergence. American Journal of Cardiology, 98(7), 982-985.

[9] Ali, K. M., Wonnerth, A., Huber, K., & Wojta, J. (2012). Cardiovascular disease risk reduction by raising HDL cholesterol–current therapies and future opportunities. British journal of pharmacology, 167(6), 1177-1194.

[10] Yu, C., Lin, Y., Lin, C., Wang, S., Lin, S., Lin, S. H., Wu, J. L., & Chang, S. (2020). Predicting metabolic syndrome with machine learning models using a decision tree algorithm: Retrospective cohort study. JMIR Medical Informatics, 8(3), e17110. https://doi.org/10.2196/17110