Université Paris-Dauphine - PSL

Master of Quantitative Economic Analysis

Complex diagnosis, hidden factors - Analysing metabolic syndrome using patient data

---

Project report

Gaetan Le Floch, Julia Schmidt

Professor: Khalil El Mahrsi

15. December 2022

Contents

Abstract

Over the past decades, metabolic syndrome has been an emerging disease in line with increasing rates of obesity and diabetes. As symptoms linked to metabolic syndrome are highly intertwined and complex, this report analyses data on 2009 patients. The paper contributes to the current research by investigating the distribution of risk factors in both women and men, identifying the key drivers of metabolic syndrome and designing a predictive algorithm to classify patients suffering from the disease using a decision-tree. Based on the analysis, the report recommends three policies to improve the efficiency in diagnosing incoming patients.

Key words: correlation, logit regression, decision trees, metabolic syndrome

1 Why is metabolic syndrome important?

Metabolic syndrome (METSyn) is a combination of conditions that significantly raise the risk of a multitude of diseases, such as coronary heart diseases, diabetes or other illnesses that affect blood vessels (NHS 2022).

Over the past two decades, the number of people with METSyn worldwide has increased massively associated with the global epidemic of obesity and diabetes (Eckel et al. 2005). The syndrome can present itself in a large variety of symptoms, making it difficult to identify. In light of the recent COVID-19 pandemic, a renewed urgency to identify, treat and protect the patients at risk of METSyn emerged.

The concept of the METSyn was first described in the 1920s by Kylin, a Swedish physician, as the clustering of hypertension (elevated blood pressure), hyperglycemia (high blood sugar), and gout (swelling of joints). More recently, various institutions have defined the common symptoms of METSyn, albeit with slightly different thresholds and potential risk factors. The World Health Organization (WHO) for instance requires the presence of one out of four indicators relevant to diagnose METSyn next to an existing diabetes diagnosis, while the European Group for the Study of Insulin Resistance (EGIR) includes fasting plasma glucose (a measure of blood sugar), as one potential symptom next to the WHO indicators.

For our analysis, we follow the definition put forward by the American Heart Disease Association (American Heart Association 2022), as our patient sample shows similarities with the demography of the United States. According to their definition, common symptoms indicating that a patient has METSyn include: i) being overweight or having a large waist circle, ii) high levels of fat combined with low levels of cholesterol in the blood, iii) high blood pressure, and iv) an inability to control blood sugar levels, called insulin resistance.

2 Objective of the project

The aim of this project is to better understand, based on historical clinical data, the factors leading to METSyn in patients. In order to support hospitals in taking better decisions on how to identify patients with METSyn, the project identified the following three key objectives:

- Gain a general understanding of teh factors associated with the disease
- Identify the key drivers of METsyn
- Predict the probability of patients suffering from METSyn

3 Data

The data set contains 2009 observations, with no missing values. The information on the patient includes 14 features, which can be divided into five socioeconomic variables (age, sex, marital status, race and income) and nine biological variables. For the understanding of the analysis, the biological variables are briefly explained below.

- The **Body Mass Index (BMI)** measures the relationship between height and weight by dividing an adult's weight in kilograms by their height in metres squared. A BMI above 30 indicates obesity (Center for Disease Control 2022a).
- The **waist circle** defines the size of the waist and can indicate whether fat is overly distributed in the belly region. For men, more than 94 cm and for women more than 80 cm are a sign of obesity (Heart and Stroke Foundation 2022).
- The **stage of albuminuria** indicates whether a type of protein normally found in the blood is present in the urine. When it occurs in the urine, it is a strong indicator of kidney disease. The data set indicates stages of albuminuria: normal level (stage 0), microalbuminuria (stage 1) and macroalbuminuria (stage 2).
- The **Urine albumin to creatinine ratio (ACR)**, also known as urine microalbumin, helps identify kidney disease that can occur as a complication of diabetes. A very high ACR level indicates a more severe kidney disease (above 300). A slightly raised ACR level indicates early-stage kidney disease (between 30 - 300). A very low ACR value usually confirms healthy kidneys (below 300)(National Kidney Foundation, 2022)).
- **Uric Acid** is a waste product left over from normal chemical processes in the body and found in the urine and blood. Abnormal buildup of uric acid in the body may cause swelling of joints. Normal values are 1.5 to 6.0 milligrams/deciliter (mg/dL) for women and 2.5 to 7.0 mg/dL for men (Cleveland Clinic 2022b).
- **HDL-Cholesterol** absorbs cholesterol in the blood and carries it back to the liver. The liver then flushes it from the body. High levels of HDL cholesterol can lower the risk for heart disease and stroke. A good value for HDL is above 60 (Center for Disease Control 2022b).
- **Triglycerides** are a type of fat in your blood. Any value below 150 mg/dl is considered normal. The combination of high levels of triglycerides with low HDL cholesterol levels can increase the risk for health problems (Center for Disease Control 2022b).
- **Blood Glucose** measures the sugar levels circulating in your body. Values of 140 mg/dl are considered normal, while any value above 200 mg/dL is too high and an indicator of diabetes (Cleveland Clinic 2022a).

4 First insights into potential factors leading to metabolic syndrome

As METSyn is a complex phenomenon, it is imperative to first explore and understand the different variables in the data set and their relationship with each other. **Table 1** gives an overview over the occurrence of variables contained in the sample. The share of patients likely to be *affected by METSyn amounts to 35%*.

Variable	Count	Share	Variable	Count	Share
Sex			*Waist Circle (cm)*		
Male	987	49 %	[63 ; 94]	1175	58%
Female	1022	51%	Above 94	834	42%
Age			*Albuminuria*		
20-35	481	24%	No	1761	88%
36-60	890	44%	Micro	200	10%
60-80	638	32%	Macro	48	2%
Income			*BMI*		
[0 - 2000[	582	29%	Underweight	37	2%
[2000 - 5000[	722	36%	Healthy	594	30%
[5000 - 7000[	218	11%	Overweight	649	32%
[7000 - 9000]	487	24%	Obese	729	36%
Race			*Uric Acid*		
White	806	40%	Below 7mG.L	1712	85%
Black	462	23%	Above 7mG.L	297	15%
Asian	295	15%	*AC Ratio*		
Mexican-American	198	10%	Below 300	1961	98%
Hispanic	198	10%	Above 300	48	2%
Other	50	2%	*Blood glucose*		
Marital			Below 200	1946	97%
Married	1098	55%	Above 200	63	3%
Single	460	23%	*Triglycerides*		
Divorced	219	11%	Below 150	1509	75%
Widowed	144	7%	Above 150	500	25%
Separated	88	4%	*HDL*		
			Below 60	1420	71%
			Above 60	589	29%
			Metabolic Syndrome		
			Yes	712	35%
			No	1297	65%

Table 1: Count and share of each feature in the data set

The socioeconomic variables of the data set indicate that the hospital appears to cater to a specific strata of the population. While female (51%) and male (49%) are represented almost equally in the data set, the other variables are less balanced. Most people are married (55%). In terms of race, the majority of people are recorded as White (40%), followed by Black (34%) and Asian (15%). The median income in the sample is 3500 units/months. The median age in the sample is 49.

Most of the biological variables follow a bell shape distribution, with at times long tails on the right, pointing to patients with extreme values. Both the waist circle as well as the BMI median values are higher than recommended, at 97 and 27 respectively. Similarly, blood glucose levels (median = 100) as well as triglycerides (median = 103) show long tails on the right, indicating potential patients exhibiting risk factors. HDL values, too, appear to be rather low, with median values at 51 (below the threshold considered as healthy). On a positive note, Albuminuria, and the Urine albumin to Creatinine Ratio (ACR) are both showing low median values. Finally, Uric Acid appears to be within the normal range with the median value at 5.4.

How do different risk factors interact?

As many biological values require different thresholds for men and women, it is interesting to investigate whether the variables are distributed similarly in the two gender groups (male and female) (**Figure 1**). Overall, it appears as if slightly more women in the sample display healthy values for the biological values, however in both groups, patients at the risk of METSyn do exist.

For variables related to body weight, such as the waist circle and the BMI, the majority of females has lower values than men. This is in line with the average physique and indicates that the majority of women in the sample appear to have normal waist circle and BMI values. For all variables indicating potential kidney problems, most notably Albuminuria, ACR and Uric Acid, more men in the sample display higher values than females. The distribution of blood glucose as well as blood fat (triglycerides) shows very similar curves for both women and men, with a higher share of women showing healthy values. For the HDL, on the other side, the data show that men have lower values than women, indicating that their values are below the healthy threshold of 60 for this particular variable.

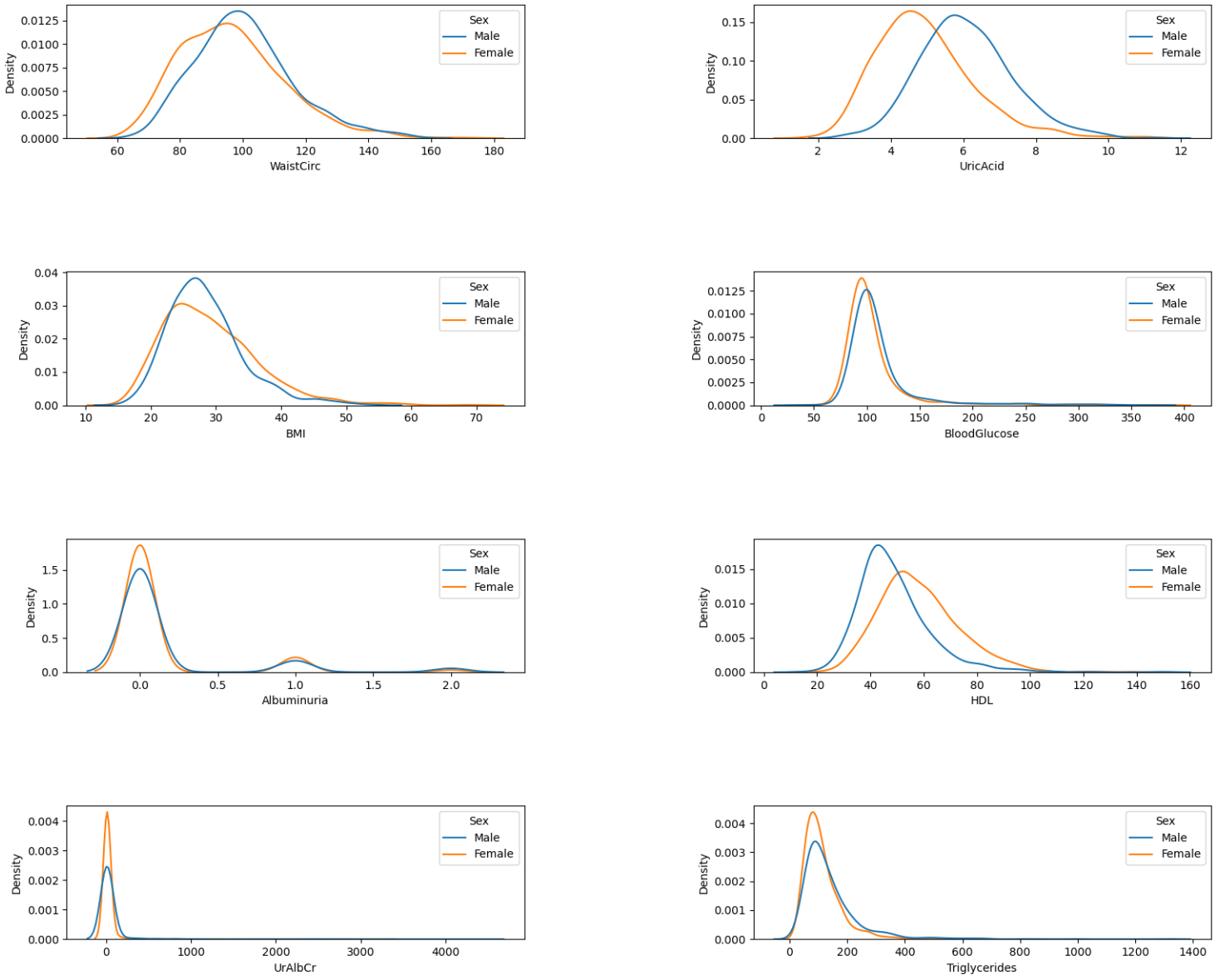

Figure 1: Distribution of biological variables by gender

5 Identifying the key drivers behind the syndrome

Before investigating potential causal drivers of METSyn, a correlation matrix gives important insights into which factors may be correlated and do condition each other **(Figure 2)**. The higher the correlation, the lighter the colour (white for maximum correlation).

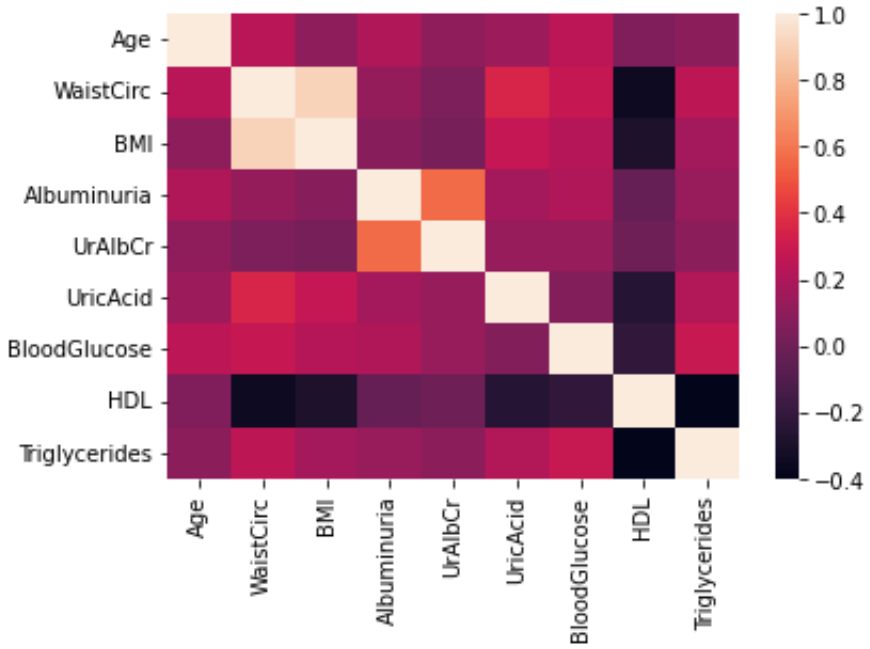

Figure 2: Correlation heatmap for all biological variables

The key correlations show that variables related to weight, most notably the waist circle and the BMI are highly correlated. This is expected as people with high (low) BMI usually tend to be heavier (lighter). This is also reflected in the waist circle. Similarly Albuminuria and the UrinAlbumin to Creatine ratio (ACR) are highly correlated, as both are indicators for kidney diseases. Conversely, HDL (good cholesterol) is negatively correlated with both waist circle and BMI, but also with blood glucose, triglycerides and Uric Acid. This is in line with medical thresholds, as the higher HDL values are, the more healthy the patient usually is - contrary to the other variables, where higher values indicate worsening conditions. Finally, higher correlations (between 0.2 and 0.4) can be observed for blood glucose and waist circle, BMI, albuminuria and triglycerides. These correlations are particularly interesting as they point to first components of the combination of symptoms potentially found in patients with METSyn.

With the aim to identify the key drivers of METSyn, a binary logistic regression is set up, with the following specification:

$$\mathbb{P}(Y = 1 \mid X) = \frac{1}{1 + e^{-X\beta}}$$

$\mathbb{P}(Y = 1 \mid X)$ is the probability to have METSyn given the characteristics of the patient. Logit models are commonly used to estimate effects on binary variables (Kleinbaum and Klein 2010). It is important to note that the data set is unbalanced (35% of the patients have MetSyn), which may limit the performance of the logit model. To account for patients' characteristics, we introduce fixed effects for marital status and race and compare the models with and without fixed effect within the scope of a robustness analysis.

The best model, based on our data and on the maximum log-likelihood is model 1 (**Table 2**), yet it is important to state that none of the models display a particularly high pseudo-R squared. Nevertheless, it allows us to derive some of the key drivers. Highly significant factors increasing the risk of METSyn appear to be age, sex, waist circle, blood glucose, and triglycerides level. HDL cholesterol is the only variable that has the ability to reduce the risk of METSyn.

- Each additional year of age increases the risk to have METSyn by 51%.
- Being a man increases the risk by 24%
- Each additional centimeter of waist circle multiplies the risk of having METSyn by 51.4%.
- High levels of blood glucose multiplies by 51% the risk of having METSyn.
- Higher levels of HDL decrease the risk of having such syndrome by 48% .
- Each additional unit of triglycerides multiplies by 50% the risk of having METSyn.

The BMI and the Albuminuria levels are only significant in the model without fixed effects (Model 4). Across all models, income remains insignificant and thus a bad predictor of MetSyn.

Dependant variable: Metabolic Syndrome				
Variable	(1)	(2)	(3)	(4)
Age	0.0394***	0.0348***	0.0395***	0.0335***
	(0.005)	(0.005)	(0.005)	(0.004)
Income	-0.00009	-0.00004	-0.00001	-0.00004
	(0.00002)	(0.00002)	(0.00003)	(0.00005)
Sex	-1.1325****	-1.0612***	-1.1056***	-0.9580***
	(0.178)	(0.174)	(0.157)	(0.158)
Waist circle	0.0629***	0.0612***	0.0593***	-0.0178*
	(0.012)	(0.012)	(0.012)	(0.01)
BMI	0.0092	0.0139	0.0179	0.1053***
	(0.028)	(0.028)	(0.027)	(0.025)
Albuminuria	0.1494	0.1192	0.1650	0.3602**
	(0.197)	(0.194)	(0.195)	(0.185)
Albumin on Creatinine	-0.0002	-0.0001	-0.0002	0.00002
	(0.000)	(0.000)	(0.000)	(0.0000)
Uric Acid	0.1124**	0.1014	0.1064*	-0.0704
	(0.056)	(0.055)	(0.055)	(0.050)
Blood Glucose	0.0247***	0.0249***	0.0247***	0.0111***
	(0.003)	(0.003)	(0.003)	(0.002)
HDL	-0.0469***	-0.0465***	-0.0473***	-0.0993***
	(0.007)	(0.007)	(0.007)	(0.006)
Triglycerides	0.0153***	0.0152***	0.0152***	0.0103***
	(0.001)	(0.001)	(0.001)	(0.001)
Fixed effects				
Marital status	True	False	True	False
Race	True	True	False	False
Performance				
Log-likelihood	-713.31	-717.09	-715.73	-829.75
Pseudo R^2	0.4539	0.4510	0.4520	0.3647

Table 2: Results of the logit model.

Note :, ** and *** denote significant results at the 10%, 5% and 1% level respectively. The socio-economic category comprises income and marital status. The socio-ethnic category comprises race.*

6 Predicting metabolic syndrome in patients

Finally, a predictive algorithm (decision tree classification) was set up to enable hospitals to identify future patients potentially at risk of metabolic syndrome.

Decision trees are non-parametric, supervised learning algorithms that have been widely used to predict both, categorical and continuous target variables. Classification trees predict the value of the target by learning simple 'if-then' decision rules from the data. They divide the sample of observations in sub-groups to minimise the classification error of the predicted variable. A prediction is made following a path in the tree and computing the average target value of the past observations that fall in the same leaf. Intuitively, classification trees may capture multiple interactions and thresholds effects, which make them particularly suitable for predicting the occurrence of a complex phenomenon, such as metabolic syndrome (De Ville 2013).

The data set is split into a test and a train sample, taking into account the imbalanced distribution of the target variable. First, all features were used to fit the tree and compute a cost-complexity function in order to identify the adequate range of pruning parameters (alpha) (**Figure 3**). The figure shows that the accuracy on the train and the test set jointly decline starting with an alpha higher than 0.05. To ensure we include the correct alpha in our grid search, we thus included alphas between 0 and 0.05 in our hyper parameter grid search.

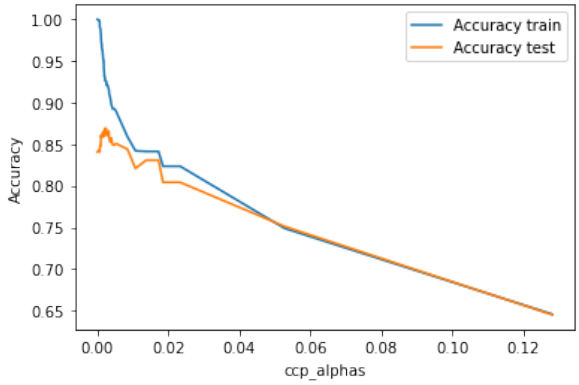

Figure 3: Pruning parameter alpha

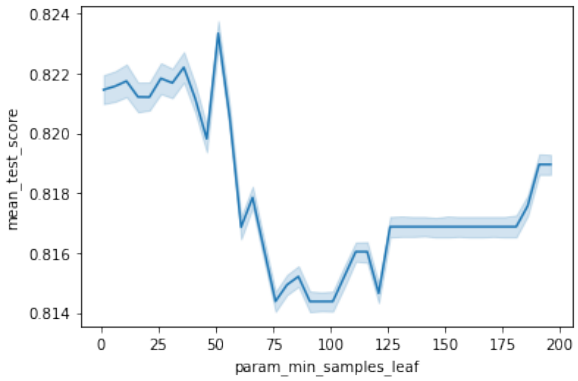

Figure 4: Minimum sample per leaf

In a second step, the tree was trained using stratified cross-validation and a hyper parameter grid search, including the previously identified alpha parameters. The best hyper parameters identified were alpha = 0, a maximum depth = 6, minimum sample size for a split = 22, and the minimum sample per leaf = 6 (**Figure 4**). In a final step, the target variable was predicted using the best parameters identified, with an overall accuracy of 84%, and a recall of 74% and a precision of 80% (**Table 3**). Overall, the decision tree confirms the results of the logistic regression, and adds additional information by illustrating the non-linearity and thus complexity of diagnosing the disease. The best tree prediction path is visualized in the Annex (**Figure 5, p.16**)

Model	Accuracy	Recall	Precision
	0.84	0.74	0.80

Table 3: Evaluation metrics for classification

Analysis of the prediction path: The prediction path of the decision tree illustrates which features are used to split the tree at each node, thereby effectively reducing the covariance within subgroups. The orange path in the tree indicates the best path to predict patients that do not have metabolic syndrome. The most important feature to split the data set appears to be blood glucose levels, dividing the data set in patients with glucose levels above 99.5 % and below. In a second step, the tree uses the BMI to group patients with blood glucose levels lower than 99.5 %. Patients with a BMI level below 24.8, are again split, analyzing their levels of cholesterol (HDL). These three nodes illustrate the non-linear interaction effect, the tree can model. In fourth step, the tree splits the patients below an age of 47 and then in a final step analyses the HDL levels one more time. This is a threshold effect, where an HDL level above 62 is considered as a good cut off, however, it yet again needs to be verified in case the patient is above 47. Overall, predictions with less accuracy on whether the patient has METSyn follow the left side of the tree (higher likelihood of patients without the syndrome), or the right side of the tree, where additional values on the health of the kidney are indicative of metabolic syndrome.

7 Recommendations

Based on the data analysed and the analysis conducted, this paper recommends the hospital to take specific steps to better identify and treat the incoming patients based on the data submitted. More specifically, based on the key drivers identified, the hospital could improve their pre-screening and treatment program:

- **Pre-screening of incoming patients:** Based on the classification model, the identification of healthy patients requires only five variables (blood glucose, triglycerides, HDL, BMI and age). The hospital could first filter out the healthy patients using the results of a blood test and a pre-screening tool, called 'MetSyn Watcher'. The METSyn watcher is a streamlit.app that was programmed based on the decision path of the classification model and can quickly analyse historical data. Once the patients receive the results of the blood test, they are asked to include their data in the 'MetSyn-Watcher'. Only the patients at risk proceed for additional urine tests to analyse their kidney values (especially to identify their Uric Acid levels). This measure can save cost and time in analysing the patients and increases the efficiency of the diagnosis.

- **Detailed monitoring of identified risk cases:** Pre-screened patients that are more complicated to classify as they show signs of METSyn, but their results do not fit within the suggested thresholds, require more detailed scanning and regular meetings with a doctor to closely monitor the thresholds and prevent a deterioration of the condition.

- **A forum of best practices and awareness:** Together with the METSyn Watcher, the hospital set up a Twitter channel to raise awareness on the metabolic syndrome. The Twitter channel also includes an anonymised tracker of the frequency with which the hospital pre-diagnoses patients with METSyn within the scope of the pre-screening. In addition, the channel could be used in the future to connect to other hospitals, and spread validated information on the risks of METSyn, especially within the local community (e.g. city/town).

References

American Heart Association. (2022). Metabolic syndrome. https://www.heart.org/en/health-topics/metabolic-syndrome/about-metabolic-syndrome

Center for Disease Control. (2022a). Defining adult overweight obesity. https://www.cdc.gov/obesity/basics/adult-defining.html#:~:text=If%20your%20BMI%20is%20less,falls%20within%20the%20obesity%20rang.

Center for Disease Control. (2022b). Ldl and hdl cholesterol and triglycerides. https://www.cdc.gov/cholesterol/ldl_hdl.htm#:~:text=HDL%20(high%2Ddensity%20lipoprotein),for%20heart%20disease%20and%20stroke.

Cleveland Clinic. (2022a). Blood glucose levels. https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test

Cleveland Clinic. (2022b). High uric acid levels. https://my.clevelandclinic.org/health/symptoms/17808-high-uric-acid-level

De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 448–455.

Eckel, R. H., Grundy, S. M., & Zimmet, P. Z. (2005). The metabolic syndrome. *The lancet*, *365*(9468), 1415–1428.

Heart and Stroke Foundation. (2022). Healthy weight and waist. https://www.heartandstroke.ca/healthy-living/healthy-weight/healthy-weight-and-waist

Kleinbaum, D. G., & Klein, M. (2010). Introduction to logistic regression. In *Logistic regression* (pp. 1–39). Springer.

National Kidney Foundation. (2022). Albuminuria. https://www.kidney.org/atoz/content/albuminuria

NHS. (2022). Metabolic syndrome. https://www.nhs.uk/conditions/metabolic-syndrome/

8 ANNEX

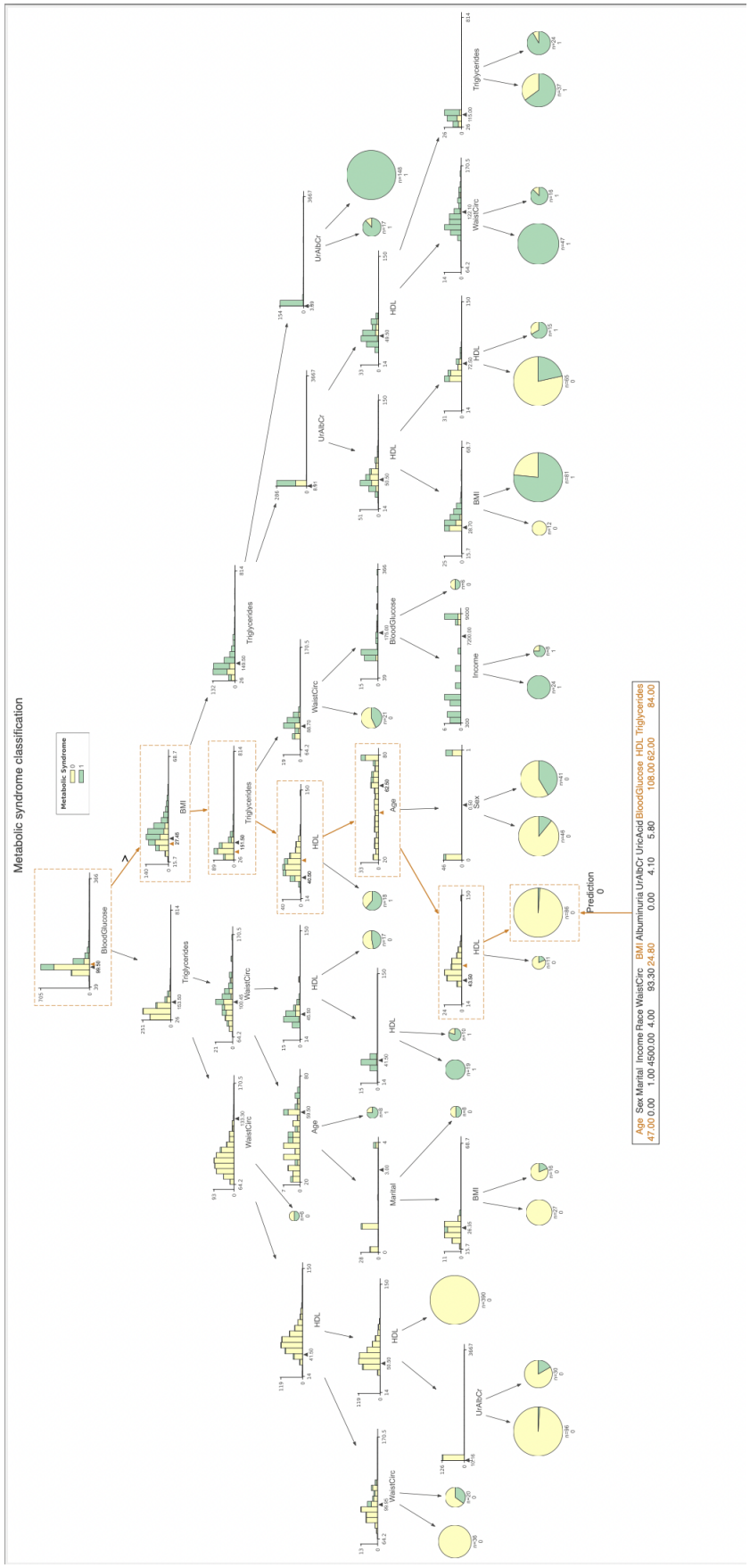

Figure 5: Best decision tree