Ðauphine | PSL
UNIVERSITÉ PARIS

Metabolic Syndrome Analysis

Final Project – Python for Data Science

Master´s in Quantitative Economic Analysis

Author: Felipe Fernando Calvo de Freitas

Date: Novermber 2025

1. Introduction

1.1. Purpose of the project

The purpose of this project is to show how certain biological (physiological or biochemical), demographic and socioeconomic indicators are related to the Metabolic Syndrome.

Our aim is to identify potential risk factors, patterns, and correlations that help us explain which characteristics are most strongly associated with the syndrome. Understanding these relationships can contribute to early detection, targeted prevention and a better understanding of how Metabolic Syndrome develops across different population groups.

1.2. Definition and prevalence of Metabolic Syndrome

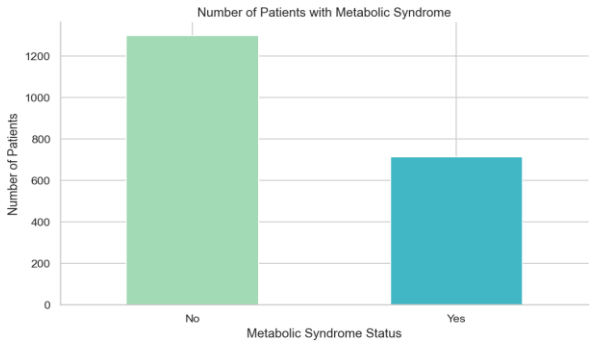
Metabolic Syndrome (MetSyn) is a group of conditions that increase the risk of heart disease, stroke and type 2 diabetes. According to a report of the National Cholesterol Education Program[1], if 3 of 5 of the following characteristics are present in a patient, a diagnosis of the syndrome can be made. These are[2]:

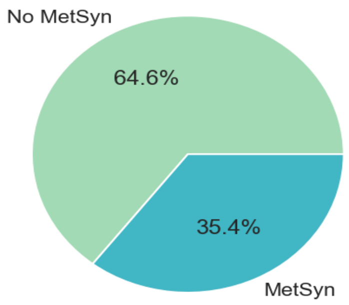
- Abdominal obesity: if Waist circumference is greater than 102cm for men or 88 cm for women.
- Elevated Triglycerides: if this indicator is greater or equal than 150 mg/dL.
- Low HDL cholesterol: if this variable is lower than 50 mg/dL for women or 40 for men.
- Elevated fasting glucose: if Blood Glucose is higher or equal than 100 mg/dL, which would count as impaired glucose regulation.
- Elevated blood pressure: if it is greater or equal than 130/85 mm Hg.

From our dataset and the 2009 observations, there´s a majority of patients that do not suffer from this disorder, almost 65% of them.

[1] Grundy, S. M., Brewer, H. B., Cleeman, J. I., Smith, S. C., & Lenfant, C. (2004).
Definition of metabolic syndrome: Report of the National Cholesterol Education Program (NCEP) Adult Treatment Panel III. Circulation, 109(3), 433–438.
https://doi.org/10.1161/01.CIR.0000111245.75752.C6

[2] All definitions and thresholds of the Metabolic Syndrome follow the NCEP report indications.





2. Distribution of variables

For better data analysis and interpretation, we will divide variables into numerical and categorical.

Numerical variables

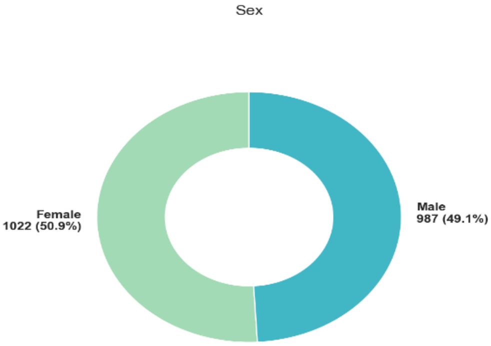
The numerical variables in the dataset are Age, Income, WaistCirc (Waist circumference), BMI (Body Mass Index), UricAcid (Uric acid level), BloodGlucose, HDL (High-Density Lipoprotein cholesterol), Triglycerides, and UrAlbCr (Urine albumin–creatinine ratio), which show diverse distributional shapes. Firstly, Age is relatively uniform across groups, whereas Income is right-skewed, as most individuals are lower income levels. For some biological measures (Triglycerides and BloodGlucose) there´s also a strong presence of right-skewness, which may indicate that patients lie in common health-levels (low and moderate levels) with a small subset of them having extreme magnitudes of those. Moreover, other variables show more symmetrical distributions. However, both by visualisation and numerical checks, we can highlight UrAlbCr (Urine albumin–creatinine ratio), which seems to be highly affected by extreme values. That´s why we have decided, for the sake of a better understanding of it, to log-transform it and plot an extra graph.

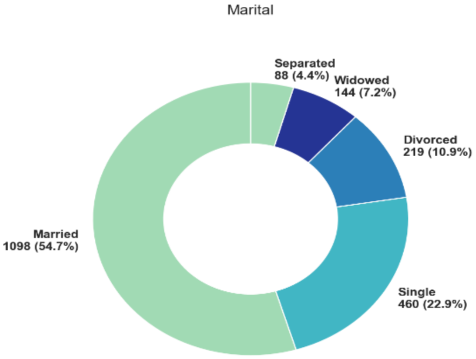
Numerical descriptive table

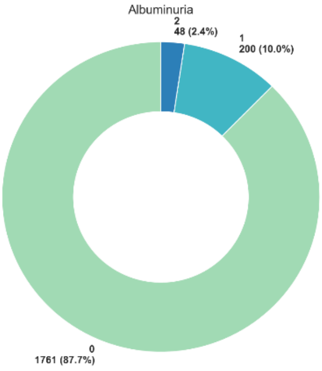
Variable	Mean	SD	Min	25%	50%	75%	Max
Age	49.26	17.42	20.00	35.00	49.00	63.00	80.00
Income	4147.19	2984.60	300.00	1600.00	3500.00	6200.00	9000.00
WaistCirc	98.52	16.31	63.10	86.90	97.10	107.80	170.50
BMI	28.73	6.58	15.70	24.10	27.70	32.10	68.70
UrAlbCr	42.25	241.42	1.40	4.46	6.96	13.49	4462.81
UricAcid	5.49	1.43	1.80	4.50	5.40	6.40	11.30
BloodGlucose	108.01	33.64	39.00	92.00	100.00	110.00	382.00
HDL	53.55	15.01	14.00	43.00	51.00	62.00	150.00
Triglycerides	126.89	89.82	26.00	75.00	103.00	149.00	1131.00

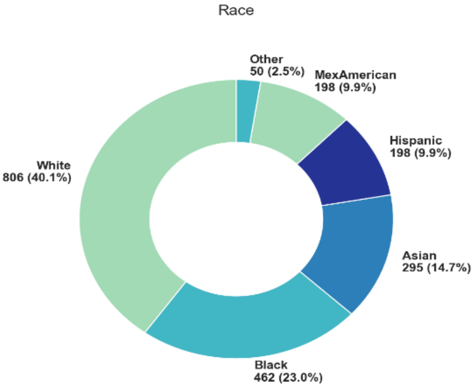
Categorical variables

Categorical variables include Sex, Marital, Race and Albuminuria stage[3]. Gender distributions show almost an identical spread, whereas for the rest, proportions are quite different. Marital status is dominated by married individuals (54.7%) with separated being its lowest group (4.4%). Race shows heterogeneity, as white ethnicity accounts for 40% in contrast with the category others, with a 2.5%. Finally, is the most unequal, with almost 90% of patients in the normal category.









3. Interaction of variables

In this section, we will present two main interactions of variables. Firstly, we will assess gender differences in biological variables, and then we will compute correlations between biological variables.

Biological variables are: WaistCirc, BMI, UricAcid, BloodGlucose, HDL, Triglycerides, UrAlbCr and Albuminuria.

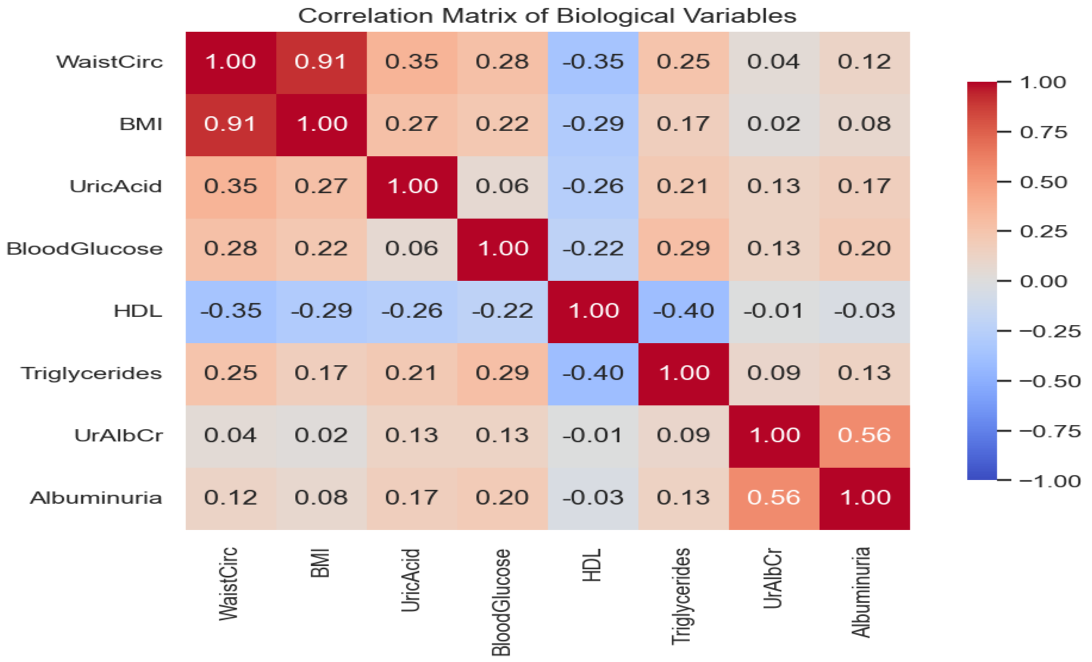
[3] Metabolic Syndrome is categorical, but we have decided to analyse it separately as it is the target variable (in part 1).

3.1. Gender differences in biological variables.

To start with, we will compute the differences in numerical biological levels between females and males. Female´s mean was slightly lower for Waist circumference, Blood glucose and Uric acid levels, whereas men possess smaller values for BMI and Triglycerides. HDL levels were found to be much higher for women than men (58.3 versus 48.7). For Urine albumin–creatinine ratio (UrAlbCr), we can observe after computing its a log-transformation, that levels for females are slightly higher than for male. Finally, Albuminuria percentages (the only categorical biological variable) are almost identical between both, even thought women presence on low and moderate levels are marginally higher.

3.2. Correlation between biological variables

Moving forward, we will assess the relationship that biological variables have with themselves, showed by a correlation matrix.



The strongest correlations are:

- **BMI and Waist Circumference (r = 0.91)**

This is the strongest correlation found and it is quite intuitive, as both measure body fat and obesity. While WaistCirc is directly linked to abdominal obesity (and also one of the five requisites of MetSyn), BMI also reflects body obesity as it is an indicator of overall body mass.

- **Urine albumin–creatinine ratio and Albuminuria (r = 0.56).**

Understanding this relationship required checking clinical sources, as it is not commonly known outside medical contexts. According to the Cleveland Clinic (2025), the Urine Albumin–Creatinine Ratio (UrAlbCr) measures how much albumin (a protein that healthy kidneys normally keep in the blood) is being lost into the urine, relative to creatinine (a waste product filtered by the kidneys). Elevated UrAlbCr levels may indicate impaired filtration of the kidney[4]. With this definition, the correlation now seems reasonable: higher values of this ratio correspond to more severe Albuminuria stages (which represents increasing levels of kidney damage, especially at level 2 (≥300 mg/g)) [5], hence reflecting significant renal dysfunction in both cases.

- **HDL and Triglycerides (r = -0.40)**

This moderate negative correlation reflects the following pattern: as Triglyceride levels increase, HDL ("good" cholesterol ones) decreases. It is interesting to highlight that both are key diagnostic components of MetSyn, so their inverse relationship is medically meaningful. Individuals having high levels of the former tend to have lower of the latter, a combination known as atherogenic dyslipidemia, which is a fundamental indicator of cardiovascular risk[6].

4. Factors most linked to Metabolic Syndrome

This section examines the relationship between Metabolic Syndrome and the biological, socioeconomic, and demographic indicators included in our study. Two complementary approaches are used. First, we analyse differences in means for numerical variables and distributional patterns for categorical variables. Then, we quantify the strength of these associations using Pearson correlations for numerical variables and Cramér´s V for categorical ones.

[4] Cleveland Clinic. (2025, January). Urine Albumin-Creatinine Ratio (uACR). Cleveland Clinic Health Library. https://my.clevelandclinic.org/health/diagnostics/urine-albumin-creatinine-ratio

[5] National Kidney Foundation. (n.d.). *Albuminuria (proteinuria): What it means and what you can do*. https://www.kidney.org/kidney-topics/albuminuria-proteinuria

[6] Manjunath, C. N., Rawal, J. R., Irani, P. M., & Madhu, K. (2013). Atherogenic dyslipidemia. Indian Journal of Endocrinology and Metabolism, 17(6), 969–976. https://doi.org/10.4103/2230-8210.122600

Numerical mean-difference table

Variables	Mean-difference	Explanation
Triglycerides	+ 83	Reflects dyslipidaemia (insulin resistance and impaired fat metabolism)
UrAlbCr	+ 37	Early kidney dysfunction
BloodGlucose	+ 26	Insulin resistance (also a market of Type 2 diabetes)
WaistCirc	+ 17	Abdominal obesity (body fat contributes to insulin resistance)
Age	+ 10	Aging increases cardiovascular risk factors
BMI	+ 6	Excesses in body weight reinforce patterns of obesity (remember correlation with WaitCirc)
UricAcid	+ 1	Insulin resistance (linked to mild hyperuricemia)
HDL	-12	Impaired lipid transport, contributes to cardiovascular risk
Income	-616	Access to preventive healthcare or to healthier diets

By comparing the mean levels of patients with and without a diagnose of the syndrome, we observe that the highest difference of all comes from Triglycerides (body fat) and elevated fasting glucose (measured by Blood Glucose), both being two main indicators for the metabolic diagnosis. Mean levels of Triglycerides from patients with MetSyn goes up to 180.5, compared to 97.5 from non-MetSyn, whereas it is 125 compared to 98.6 for Blood Glucose. Urine albumin-creatine ratio is also quite different across groups, with a difference of almost 37 points between them. This is interesting to observe as the variable was the most difficult ones to interpret due to its complex levels for plotting and analysis. Most of the very high levels of Urine albumin–creatinine ratio are found in risky patients (with abnormal levels), which explains why the mean is so different.

On the other side, HDL mean difference are negative, coherent with the pattern found with Triglycerides, but also with clinical investigations, in which lowest levels of it are considered healthier. Moreover, lower HDL is found together with high Triglycerides, corroborating our earlier correlation results.

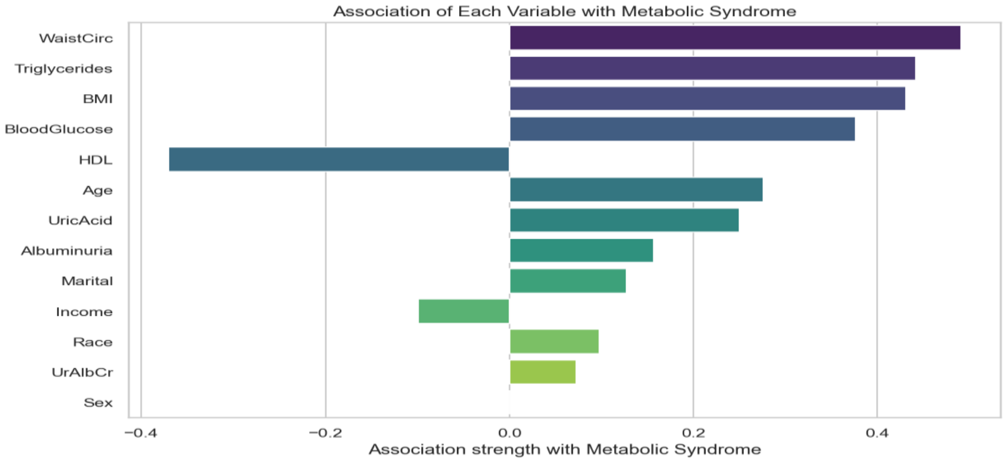
Categorical distributional table

Variables		% MetSyn	% no MetSyn
Sex	*Female*	*34.3*	*65.7*
	Male	*36.6*	*63.4*
Marital	Divorced	42.0	58.0
	Married	37.4	62.6
	Separated	36.4	63.6
	Single	24.3	75.7
	Widowed	45.1	54.9
Race	Asian	25.8	74.2
	Black	32.9	67.1
	Hispanic	42.4	57.6
	Mex-American	42.9	57.1
	Other	30.0	70.0
	White	37.2	62.8
Albuminuria	0 – normal level	32.6	67.4
	1 – moderate level	54.0	46.0
	2 – extreme level	62.5	37.5

We observe substantial distribution differences across categories. For example, patients with extreme Albuminuria levels show a 29.9-percentage-point (pp) higher prevalence of Metabolic Syndrome compared with those with normal Albuminuria levels (like UrAlbCr, abnormal levels lead to higher risk states). Similarly, in marital status, widowed individuals reveal a prevalence 20.8 pp higher than single individuals, but that might be affected by age differences. Moving to ethnicity, the Mexican-American cohort presents a 42.9% of its people with this syndrome, in contrast with the Asian community that exhibits 17.2 pp less than them, possibly because of health and diet decisions. Finally, Male individuals manifest slightly higher values of the syndrome by 2.2 percentage points.

Pearson and Cramér´s correlation tests

To quantify the association between Metabolic Syndrome and the numerical indicators, we compute a Pearson (point-biserial) correlation, measuring the strength of linear relationships between these continuous variables and a binary outcome (MetSyn). For categorical indicators, we cannot use this test, so we will use Cramér´s V chi-square statistic test instead, that captures the strength between two categorical variables.



The results reveal a consistent pattern linking the syndrome mainly to key biological factors. The strongest values are Waist circumference, Triglycerides, BMI and Blood glucose, reflecting the vital importance of abdominal obesity, dyslipidemia and insulin resistance in the profiles of patients with the syndrome. Besides, HDL showed a pronounced negative strong association, consistent with its negative pattern. Routinary tests of these factors and more focus on the prevention of high levels of them can lead to better health outcomes for society.

Moderate associations were observed for Age, Uric acid and Albuminuria, which indicate that older individuals and those with early signs of renal stress are more likely to be diagnosed with the syndrome. Correlations of socioeconomic and demographic variables are low, suggesting no high importance of these.

5. Random Forest predictive model

In this final section, we complement the descriptive and correlational analysis with machine learning tools that can help us develop a model of Metabolic Syndrome likelihood prediction based on the full dataset given. The model improves our analysis by capturing non-linear interactions across indicators and provides a robust validation of the patterns identified.

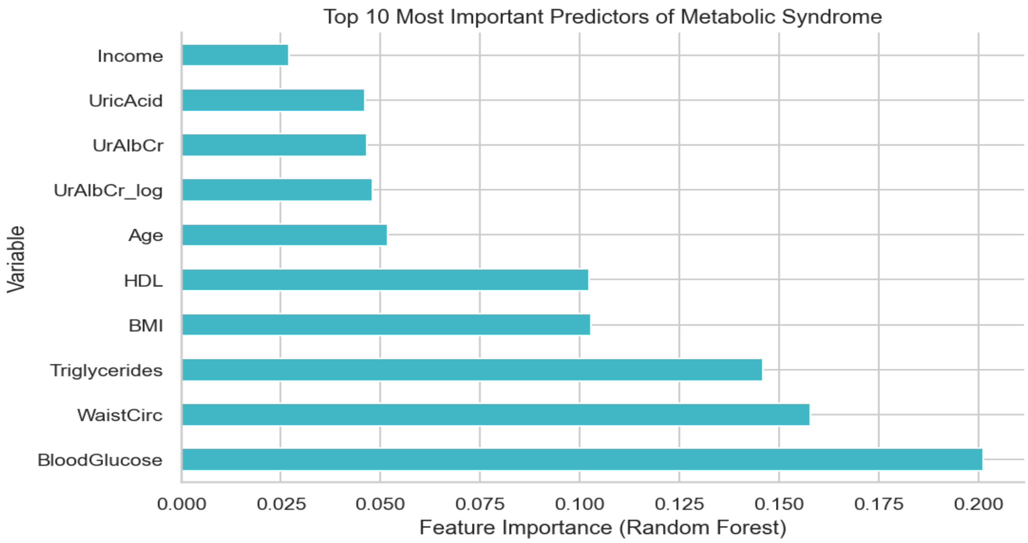
We selected the Random Forest algorithm because it is especially well suited for this type of dataset, where variable relationships often show non-linear and complex interactions. Besides, they are also robust to multicollinearity, provide interpretable measures and can handle both categorical and numerical data. Hence, this algorithm is a reliable choice for predictive modelling in this context.

Classification table

	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.87	0.92	0.90	389
CLASS 1	0.84	0.76	0.80	214
ACCURACY			0.86	603
MACRO AVG	0.86	0.84	0.85	603
WEIGHTED AVG	0.86	0.86	0.86	603

The model is more confident and accurate for detecting people without the syndrome (class 0), performing slightly worse for detecting it, but it still shows an acceptable recall and F1 scores (0.76 and 0.80).

Random Forest – Predictor importance plot



Corroborating our results from correlation and pattern tests, Blood glucose, Waist circumference, Triglycerides, BMI and HDL are found to be the strongest contributors of Metabolic Syndrome patients. Moreover, they align with the clinical definition of the syndrome, reflecting impaired glucose regulation, abdominal obesity and dyslipidemia as central drivers of metabolic risk.