

IS3221 Project-6

SAP Analytics Cloud (SAC) Project – Classification Analysis & Association Analysis

This assignment is designed to introduce you to some of the methods we can use to undertake **Classification Analysis** and **Association Analysis** using **SAP Analytics Cloud**. The **data** to be used in this exercise will be the **Cardio-mod.xlsx** file.

Objectives

Cardiovascular disease (**CVD**) is a general term that describes a disease of the heart or blood vessels. Blood flow to the heart, brain or body can be reduced because of a: blood clot (thrombosis). Cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. Over three quarters of CVD deaths take place in low- and middle-income countries.

Most cardiovascular diseases **can be prevented** by **addressing behavioural risk factors** such as **tobacco (smoking)** use, **unhealthy diet** and **obesity**, **physical inactivity** and harmful **use of alcohol**. It is important to detect cardiovascular disease as early as possible so that management with counselling and medicines can begin.

The objective of the analysis is to understand any **general relationships between different patient characteristics** and the propensity to develop CVD, specifically:

- **Objective 1:** Understand differences in the measurements recorded between the **group that have CVD** and the **group that does not have CVD**.
- **Objective 2:** Identify **associations between the different factors and the development of CVD** that could be used for education and intervention purposes. Any associations need to make use of general categories, such as high blood pressure etc, to be useful.
- **Objective 3: Develop a predictive model** to estimate whether a patient will develop CVD.

The population of this study consists of individuals from the age of **30 and 65**.

A data set containing **70,040 observations** has been made available. It contains patient records describing a number of attributes in addition to whether the patient went on to develop CVD. Please take note there many analysis done on this data on the internet. However, **word of caution** the data that is being provided here have been **modified** hence the conclusion and findings you will reach will be **VERY different** from what you read from other sources.

The **metadata** description is given in the last page of this document.

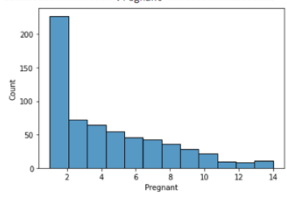
Before you start you will need to **prepare the dataset for analysis** (take note you might need to convert your data to numerical from categorical or vice versa in order to perform the analysis depending what you are trying to achieve).

Perform a preliminary analysis of the data and **tabulate and identify** if there any **duplicate records** in which case you need to remove them. You need to check if there any **null values** in the records as well. You are told that only the following columns (**height, weight, ap_hi, ap_lo, choles** and **gluc**) have **0** has the **missing values**. You need to decide how best to deal with the **missing values**. You are **not** allowed to remove any records with missing data. You must consider **imputation method** to replace the missing values and justify your approach taken (discuss with instructor in class on your decision).

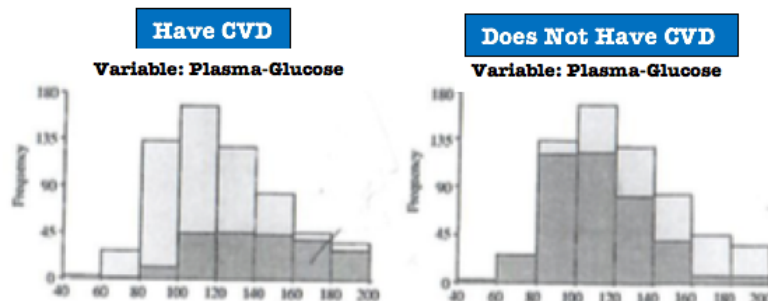
You will need plot a **scatter matrix plot** between all the variables in the data set and calculate the **r-correlation** coefficient. Conduct this investigation and decide on what needs to be done with correlated variables. Investigate if all variables are necessary in your analysis. You can ignore the **id column** in all your analysis.

Investigate also to see if there are any **outliers** in the data. Conduct this analysis for **all** the variables. If you decide to ignore any variables in your analysis, remove them with justifications. Plot **box-plots** to do this investigation.

For each variable, generate a **frequency distribution** and present alongside a series of descriptive statistics in order to characterize the variables. These descriptive stats should include **Central tendency** (Mode, Median, and Mean), **Variation** (Variance, Std Dev) and **Shape** (Skewness and Kutosis). One example for the variable **Pregnant** (not part of your data shown only as example) is as shown: Observe these frequency distributions and give a general comment on each of them.

Results	Observations/Comments
<p>***** Pregnant *****</p>  <p>--- Central Tendency --- Mean: 4.441412520064206 Median: 4.0 Mode: 1.0 --- Variation --- Variance: 9.957574850453929 Standard Deviation: 3.15562525201161 --- Shape --- Kurtosis: -0.1346301215353248 Skewness: 0.8253745177208933</p>	<p>The graph shows the distribution of the count of the number of pregnancies based on the Pregnant variable.</p> <p>As the skewness of the graph has a positive value of 0.825, indicating that the graph is skewed right whereas the kurtosis value of less than 3 suggests that the tails of the graph are shorter and thinner compared to a normal distribution.</p>

Next you need to create the frequency distribution (**overlap frequency histograms**) for **all** the variables to understand **differences between the two groups** (with CVD and no CVD). One example is as shown below for variable **Plasma-Glucose** (example shown is not part of your data).



Frequency distribution for Plasma-Glucose (grouped by CVD)

The frequency distribution above shows the distribution of the variable **Plasma-Glucose** in **light gray** on both images. Observations belonging to the two groups of CVD are highlighted in **dark gray**. In the histogram on the left, the **dark gray** highlighted observations belong to the group that **had CVD**. The observations **dark gray** highlighted on the right histogram are patients that **did not have CVD**. These histograms indicate that the distribution of variable **Plasma-Glucose** data between the groups is significantly different. Almost all the patients with highest **Plasma-Glucose** values went on to develop CVD. Almost all the patients with lowest **Plasma-Glucose** values did not go on to develop CVD. (*Hint: Use Python draw this type of charts*).

Once this is done observe and comment on **three** of the frequency distributions namely **Glucose**, **Cholesterol** and **Systolic blood pressure** for the two groups (CVD and no CVD). You may want to plot **additional box plots** for these three variables for the two groups and compare them and comment on your findings.

For **all the above task mentioned** you can easily accomplish using **Python**. Once you have completed the above you will be in the position to upload the data to SAC for further analysis. You will create a **data model** on your data and create a story and apply **smart assist** tools (eg **smart discovery**; **smart insights** and **search to insights**) to better understand your data.

Once you uploaded the data to **SAC** as a **model** you can now perform some **transformation** on the data if there is a need. The following transformations can be considered within this analysis: **normalization**, **discretization**, and **new calculated field**. You only need to do discretization and calculate a new field call **BMI**.

One of the objective of this investigation is to **classify general associations** between classes of variables, such as **high blood pressure**, and **CVD**. To this end, each variable is **binned** into a small number of categories. To calculate BMI

apply the following formula **kg/m²** – where kg is a person’s weight in kilograms and m² is their height in meters squared. The following summarizes the cut-off values (shown in parentheses) you can use along with the names of the bins for the variables:

- **Age-(group):** **young-adulthood** (25 to 39), **middle-adulthood** (40 to 59), **elderly** (> 60)
- **BMI-(group):** **underweight** (< 18.5), **healthy-weight** (18.5 to < 25), **overweight** (25 to < 30), **obese** (30 to <35), severely-obese (>35)
- **Gender-(group):** **women** (1) and **men** (0)
- **DiastolicBP-(group):** **normal** (< 80), **normal-to-high** (80-89), **high** (> 90)
- **SystolicBP-(group):** **normal** (< 120), **normal-to-high** (120-139), **high** (> 140)
- **Cholesterol-(group):** **c-normal** (1), **c-abv-normal** (2), **c-high-normal** (3)
- **Glucose-(group):** **g-normal** (1), **g-abv-normal** (2), **g-high-normal** (3)
- **Smoker-(group):** **non-smoker** (0), **smoker** (1)
- **Physical Activity-(group):** **non-active** (0), **active** (1)
- **Cardio-(group):** **no-CVD** (0), **yes-CVD** (1)

Once the above is done please provide screen shot of the **binning process or discretisation** as appendix to your report. Please provide the screen shot of the calculated BMI filed in your appendix. The results from this stage of the project is a cleaned and transformed data set ready for analysis. The **new columns (headers)** in your dataset will have names as shown above in black letters.

You can now ready to create any charts, tables graphs to analyse your data in **SAC Story**. Apply **smart assist** to learn more about the data before moving on to the investigate the next objective. Investigate what the main **variable influencers** in your dataset are when you run with smart discovery. **Take note** of them and apply this information when you perform the third objective later.

The **second objective** was to identify **general associations** in the data to understand the relationship between the measured fields and whether the patient goes on to develop CVD. Since the analysis will make use of **categorical data**, requires the identification of associations, and must be easy to interpret, the associative rule grouping approach was selected. Using the following variables, and apply the **A-priori algorithm** technique to come with the association rules.

To accomplish this you need to **export your model dataset** into a csv file and then **re-import to SAC** but **this time import as data file**. Once this is done you can then perform your data **association analysis** (refer to **Tutorial 6b** as a guide) based on the following columns.

- **Age-(group):**
- **BMI-(group):**
- **Gender-(group):**
- **DiastolicBP-(group):**
- **SystolicBP-(group):**
- **Cholesterol-(group):**
- **Glucose-(group):**
- **Smoker-(group):**
- **Physical Activity-(group):**
- **Cardio-(group):**

Look at the rules generated and look at both the groups (CVD and no CVD) and interpret the results with respect to their **support, confidence** and **lift values**. Identify the top 5 rules in this analysis and give your insights if the results makes sense. Plot any necessary charts or tables for your findings of this investigation.

The **third objective** of this exercise was to develop a **predictive model** to classify patients into two categories: (1) **patients that have CVD** and (2) **patients that do not have CVD**. Since the **response variable (Cardio)** is categorical, we must develop a **classification model**. To be useful the model must have **sensitivity and specificity** values greater than **60% to 70%**.

For this **third objective** you can use SAP **Smart Predict** to accomplish this. Use the same data file you create above. Since you are performing a **classification model** we also need some **test dataset** besides the **training dataset**. You need to export you dataset used in Association analysis above to a csv file. Then select **10%** of your dataset as you **test dataset** and remaining **90%** will be used as you **training dataset**. You need to apply a **random sampling distribution** to select this 10%. This you can once again use Python to accomplish this. Once, separated you once again upload the 90% training dataset as a **data file** to SAC.

Once uploaded perform a **smart predict analysis** with **classification** as the option and using the 90% data as your **training dataset**. Use the **Tutorial 7** as your guide. Look at the results generated, interpret the **confusion matrix** generated, plot any necessary graphs or charts to explain your findings. Then use this classification model and run it against the 10% test dataset you saved earlier. Before you run with the test dataset remove the **class column** and saved it somewhere for later comparison. Run the model with the test dataset and see if your classification model predict the class correctly for the 10% test dataset. Compare the predicted results with the actual and comment on your findings.

Next, you may want to remove some of the variables and only focus on the main influencers the system found when you did you initial analysis earlier. Compare the results of this new model with the previous model.

There are many alternative classification algorithms that you could consider. The one used by **Smart Predict** is done by the SAP system and this not known to us which classification algorithm was used in the analysis. The next step for you is to select the data for the best model results you have so far and train the data with other classification algorithms.

You can select any two of these suggested algorithms; **k-Nearest Neighbors (kNN)**, **Random Forest** and **NNet Neural Network** from the **Python library** to build the models. Compare the results of the **confusion matrix** from SAC, and any two of the above algorithms using the following KPIs **Accuracy**, **Precision**, **Sensitivity**, **Specificity**, and **F1 score** and comment on your findings. Take note some of these algorithms will accept numeric data and some categorical data. Apply the necessary data conversions where necessary.

Software: Use SAP Analytics Cloud (SAC) & Python

Deliverables:

1. Create a **graphic display** on the complete steps you took to reach the final results. You can come up with your own design for the graphics display.
2. Describe *briefly* in your report the steps you took to complete the project based on your graphic display you created above. From the pre-modelling stage to the final model. Ensure you provide the reason and justifications for all your decisions you took along the way. Any **Python code** used to create a table or chart must be **displayed with full comments** next or after the table or chart at the appropriate place in your report.
3. Provide all **relevant screenshots** of all the visualization you created. Each visualization should have a brief description on what it is representing.
4. Explain *briefly* how **performance of classification models are measured** and what are the various metrics used. Using the test dataset metrics and final model metrics do a comparison analysis based on the metrics you mentioned.
5. Create a **role matrix** table with 4 columns. First column insert the **names** second column **the person role(s)** played by each group member in the whole process. Then the third column titled “**Predictive Analytics Process**” you insert the step/steps completed by that individual and fourth column insert a very **brief description** of the step.
6. You can have one person in your group to execute the predictive analytics or you can have all executing in your individual laptops and then share your results for final reporting. All team members must assist or play the role that you have been assigned by your group leader in ensuring the completion of the whole project. It is important for each group member to contribute substantially to final submitted work. You can research the internet for any process steps that you may not be clear on how to proceed.
7. **General Question:** Explain *briefly* with diagram(s) the Cross Industry Standard Process for Data Mining – **CRISP-DM** (Total word count for this

question not more than 2000 words with citations). Please **insert** the exact word count at the end of the article.

8. Explain *briefly* what is **overfitting of data** and what is the common way to avoid overfitting of data.
9. **Submit a report** in word document with all the above information and upload **one softcopy** to the submission folder (will be created later) in Canvas and submit **one hardcopy** report to me. The report should have one **cover sheet** with all the details of the **title of the project, project number, group-id** and **group members names**. No need to submit in any fancy ring binders just the word document will be enough. Submission will be due during week 11/12. Tentative date of submission **Monday, 6th Oct 2023**.
10. Please submit this **project description** document together with your final report. This document should be placed just after your cover sheet with your name details and project title.
11. You must also **submit all the consolidated Python codes** used in your project as Appendix of the report with **full comments**. Please do **NOT paste an image** it must be in **text** so that it can be easily be copied and used directly on the **Python** editor.
12. You should also prepare a power-point **presentation deck** of this project for presentation during week 12/13. Detail of date and venue and format of presentation will be made known at a later date.

Take **note final report** will be subject to **Turnitin** analysis.

Metadata description of the dataset:

Attribute Name	Attribute Description	Column Name	Data interpretation
ID	ID number	id	id
Age	Age of patient in days	age	age
Height	Height of patient in cm	height	height
Weight	Weight of patient in kg	weight	Weight
Gender	Patient gender	gender	gender =1 (women) age=2 (men)
Systolic Blood Pressure	Systolic blood pressure (mm Hg)	ap_hi	systolic pressure
Diastolic Blood Pressure	Diastolic blood pressure (mm Hg)	ap_lo	diastolic pressure
Cholesterol	Cholesterol range	choles	choles=1 (o-normal); 2 (o-abv-normal) and 3 (o-high-normal)
Glucose	Sugar level	gluc	gluc=1 (g-normal); 2 (g-abv-normal) and 3 (g-high-normal)
Smoking	Smoker or non-Smoker	smoke	non-smoker=0; smoker=1
Alcohol Intake	Alcohol consumption	alco	non-alco=0; alco=1
Physical Activity	Active lifestyle	active	non-active=0; act=1
Class/Target	Target or class variable	cardio	cardio=0 - absence of cardiovascular disease (CVD) and cardio=1 - presence of cardiovascular disease (CVD)

Additional Note: Generally, a value of r (correlation coefficient) greater than or equal to 0.7 is considered a **strong correlation**. Anything between 0.5 and less than 0.7 is a **moderate correlation**, and anything less than 0.4 is considered a **weak or no correlation**.

It's the **Cholestreol value** (Cholesterol is a type of fat found in your blood) of your blood. In Adults, **200 mg/dL is desired** with 200 and 239 mg/dL as **Boderline High**. In Children, 170 mg/dL is desired with 170 and 199 mg/dL as Boderline High

gluc => It's the Glucose Level. They're **less than 100 mg/dL** after not eating (fasting) for at least 8 hours. And they're less than **140 mg/dL 2 hours after eating**. For most people without diabetes, blood sugar levels before meals hover around 70 to 80 mg/dL

ap_hi => It's the **Systolic blood pressure** i.e. Pressure exerted when Blood is ejected in arteries. **Normal value : 120mmhg or Below**

ap_low => It's the **Diastolic blood pressure** i.e. Pressure exerted when Blood exerts between arteries and heartbeats. **Normal Value : 80mmhg or Below**

Some of the useful **Python libraries** that you may want to use are as *pandas*; *matplotlib.pyplot*; *seaborn*; *numpy*; *scipy.stats* ; *os*; *re* and *sklearn*