IS3221 Project-8
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 (SAC)**. The **data** to be used in this exercise will be the **Census-mod.xlsx** file.

Objectives

The census is a special, wide-range activity, which takes place once a decade in the entire country. Its purpose is to gather information about the general population, in order to present a full and reliable picture of the population in the country - its housing conditions and demographic, social and economic characteristics. The information collected includes data on age, gender, country of origin, year of immigration, marital status, housing conditions, marriage, number of children, education, employment, traveling habits, etc.

We use a census when we want accurate information for many subdivisions of the population. Such a survey usually requires a very large sample size and often a census offers the best solution.

With the exception of basic population counts probably the most interesting topics for census data users are income and poverty. People want to know how many people live in a place and they want to know something about how well those people are living. Income is generally used as a measure of the economic well-being of individuals and communities.

Local, state, tribal, and federal agencies use published income data to plan and fund programs that provide economic assistance for populations in need. Income data measure the economic well-being of the nation.

The results of the census help determine how hundreds of billions of dollars in federal funding, including grants and support to states, counties and communities are spent every year for the next decade. It helps communities get its fair share for schools, hospitals, roads, and public works.

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The prediction task is to determine whether a person makes over $50K a year.

The objective of the analysis is to understand any **general relationships between different income attributes** and the probability someone earning greater than $50K:

- **Objective 1**: Understand differences in the measurements recorded between the **group that earn an income >$50K** and the **group that earn an income below <=$50K**.
- **Objective 2**: Identify **associations** between the different factors that lead to the different income groups.
- **Objective 3**: **Develop a predictive model** to estimate whether a person will earn >$50K based on some attributes of the person.

The Adult database was extracted from the census bureau database in 1994 by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). It now being used to predict whether income exceeds USD 50K/yr based on census data.

A data set containing **48,870 observations** has been made available. It contains records describing **15** different attributes of an individual. Please take note there are 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.

This dataset contain attributes of individuals taken in a census done in the US. *Age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country and income* are attributes in dataset. The **target** (**Income** class) is **categorical** (> $50K earning income greater $50K, and <=$50K earning income less than $50K).

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).

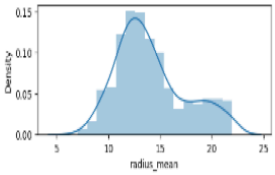
Perform a **preliminary analysis** of the data and **tabulate and identify** if there any **duplicate records** (not duplicate cells) in which case you need to remove them**.** Identify these duplicates in your report. Remove all the **data quality issues** you might spot. There are several in this dataset observe the spellings, replace attribute information so that they are **consistent** within each attribute e.g 's' and S so you change 's' to capital S. There are also **missing values or null values** issues that needs to be attended to and will depend on how you want to resolve them. 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).In this dataset missing values in some columns are represented by "**?**" and in some columns missing values are represented by "**99999**". You need to check for both and apply your imputation strategy on them.

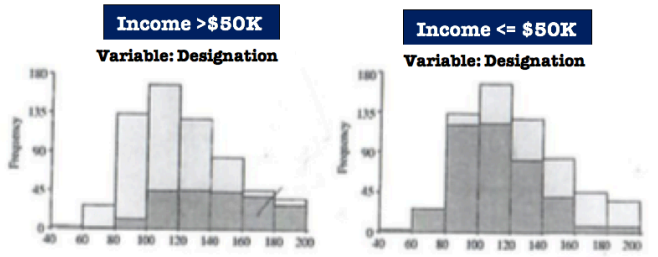
You will need plot a **scatter matrix plot** between all the variables in the data set and calculate the **r-correlation** coefficient. Conduct this investigation and decide on what needs to be done with correlated variables. Investigate if all variables are necessary in your analysis. You can ignore the ***fnlwgt* 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/density 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 **radius_mean** (not part of your data) is as shown: Observe these frequency distributions and give a general comment on each of them. By the way below is based on some dummy data and description may not make sense but image below is given to give an idea of what is expected.

Results	**Observations/Comments**
 --- Central Tendency --- Mean: 4.441412520064206 Median: 4.0 Mode: 1.0 --- Variation --- Variance: 9.957574850453929 Standard Deviation: 3.155562525201161 --- Shape --- Kurtosis: -0.1346301215353524 8 Skewness: 0.8253745177280933	The graph shows the distribution of the count of the number of pregnancies based on the Pregnant variable. 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.

Next you need to create the frequency distribution (**overlap frequency histograms**) for the following variables *age, gender, race, hours-per-week,* and *education* to understand **differences between the two groups** (**with** *Income >$50K* and *Income <=$50K*). You may want to plot **additional box plots** for these three variables for the two groups and compare them and comment on your findings. One example is as shown below for variable **Designation** (example shown is not part of your data).



Frequency distribution for Designation (grouped by Income)

The frequency distribution above shows the distribution of the variable **Designation** in **light gray** on both images. Observations belonging to the two *Income* groups are highlighted in **dark gray**. In the histogram on the left, the **dark gray** highlighted observations belong to the group that **earned >$50K**. The observations **dark gray** highlighted on the right histogram belong to group that eraned less than $50K**.** These histograms indicate that the distribution of variable **Designation** data between the groups is significantly different. Almost all the individuals with highest **Designation** values earn greater than $50K. Almost all the individuals with lowest **Designation** values earn less than $50K. (*Hint: Use Python draw this type of charts*).

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 look for opportunities to transform your data if there is a need. The following transformations can be considered: **normalization, discretization**, and **new calculated field**. Since most of the data are categorical there still room for further transformation.

One of the objective of this investigation is to **classify general associations** between the different attributes of the dataset. To this end, each variable is **binned** into a **small number of categories**. We will bin the following attributes as shown below:

- **Age-Grp**: **Young** (0-25), **Adult** (26-35), **Middle-aged** (36-45), **Senior** (46-65) and **Old** (>65)
- **Workclass**-grp: **Public-sector** (Federal-gov, Local-gov, State-gov), **Private-sector** (Private), **Self-employed** (Self-emp-inc, Self-emp-not-inc), **Unemployed** (Never-worked), **Others** (Without-pay)
- **Education-Grp**: **Preschool** (Preschool, 1st-4th); **Primary** (5th-6th , 7th -8th , 9th ,10th); **Secondary** (11th, 12th HS-grad, Some-college); **University** (Prof-school, Assoc-acdm, Assoc-voc, Bachelors); **Post-Grad** (Masters, Doctorate)
- **Education-num-Grp: Junior-yrs** (1 to 6), **Senior-yrs** (7 to 12) and **College-yrs** (13 to 16)

- **Hours-clocked-Grp**: hours-per-week cut into levels as **Part-time** (<40), **Full-time** (40 to 45), **Over-time** (>45)
- **Capital-gain-Grp:** each cut into levels **None** (0), **Low** (0 < median of the values greater zero < max) and **High** (>=max)
- **Capital-loss-Grp**: each cut into levels **None** (0), **Low** (0 < median of the values greater zero < max) and **High** (>=max)
- **Married-Status-Grp: Divorced** (Divorced, Separated), **Married** (Married-AF-spouse, Married-civ-spouse, Married-spouse-absent), **Single** (Never-married, Widowed)
- **Occupation-Grp: Tier-1-Occ** (Exec-managerial, Prof-specialty, Armed-Forces, Protective-serv), **Tier-2-Occ** (Tech-support, Sales, Craft-repair, Transport-moving), **Tier-3-Occ** (Adm-clerical, Machine-op-inspct, Farming-fishing, Handlers-cleaners, Other-service, Priv-house-serv)
- **Income-Grp: 1** (income >= $50K) and **0** (income <$50K)

You **can bin other attributes** if you can come up with your own categorisation to reduce the number categories. This will help in the analysis interpretation of the results.

Once the above is done please provide screen shot of the **binning process or discretisation** as appendix to your report. The results from this stage of the project is a cleaned and transformed data set ready for analysis.

You are 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. 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 will be selected. Using the following variables, and apply the **Apriori 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).

Look at the rules generated and look at both the Income groups 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 individuals into two categories: (1) **earning greater than 50K** and (2)

and earning less than 50K. Since the **response variable** (**Income**) 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 **Income column** and saved it somewhere for later comparison. Run the model with the test dataset and see if your classification model predict the income 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.

Apply the data with these suggested algorithms; **Logistic Regression, Naïve Bayes, Decision Tree Algorithm (J48), Random Forest, Support Vector Machine (SVM)** and **NNet Neural Network** from the *sklearn* **Python library** to build the models. Compare the results of the **confusion matrix** from SAC, and 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) how data analytics have solved or helped business solved problems and which industries are exploiting these techniques to their advanatge. (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
age	Age of an individual Min=17 and Max=90
workclass	General term to represent the employment status of an individual (eg. Private, Self-employed etc.)
fnlwgt	Number of people the census believes the entry represents
education	Highest level of education achieved by an individual
education-num	Number of years of study from 1 to 16
marital-status	Marital status of an individual. (eg. Married-civ-spouse corresponds to a civilian spouse while Married-AF-spouse is a spouse in the Armed Forces etc.)
occupation	General type of occupation of an individual
relationship	Represents what this individual is relative to others. (eg. Husband, Own-child etc.)
race	Descriptions of an individual's race (ethnicity)
sex	Biological gender of the individual (Male, Female)
capital-gain	Annual capital gains tax for an individual
capital-loss	Annual capital losses for an individual
hours-per-week	Number of hours worked per week Min=1 and Max=99
native-country	Country of origin for an individual
income	Whether or not an individual makes more than $50,000 annually

Additional Note(s): *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**.*

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*