Assignment Report: Big Data Analytics

DO NOT PUT YOUR NAME OR ANY OTHER IDENTIFIER ON THIS DOCUMENT

Section 1 — Upload final data file[pass/fail]

Before submitting your report, you should ensure you have uploaded the following:

- The final version of your data that you used for analysis, in ARFF format or a standard .csv format (with attribute names in the first row)
    - This may be provided as a single file (if appropriate) or as one file per research question (which may be necessary depending on how you have processed your data to support each question)

Notes:

- Failure to submit the above will result in a mark of zero for the whole assessment.
- Anything you submit may be used by us for testing.

Word counts must be adhered to and anything exceeding the limit will not be marked. Diagrams and/or screenshots may be used to support your discussion where applicable. These should be readable at an appropriate size and there should be no more than **5** in the whole document.

You should support your discussion with appropriate reference to relevant sources using correct citation and reference structure as indicated in the guide to the IEEE referencing system.

Section 2 — Business/research questions [10 marks]

1. **[10 marks] State the three business or research questions that you have attempted to answer through your analysis, and justify why they are interesting (300 words maximum)**

1) How can the smoking habit be predicted from marital status and general health for males aged 35 – 54?

In the USA, cigarette smoking is still a major public health problem as it was associated with lung cancer [1], stomach cancer [2], lifetime depression and anxiety [3], alcohol misuse [4] and many more health problems. Little research was done to explore the effect of psychosocial factors marital status [9] on smoking and smoking cessation. Expanding research on marital status and general health as factors that can predict smoking habits [9, 10] could assist in understanding which factors could be used to increase smoking cessation [5] and reduce the number of new smokers [6], which would help to avoid mentioned health problems and improve public health.

2) Is it possible that the employment status of a person can be a predictive factor for an individual's cholesterol level?

It is evident at this point that hypercholesterolemia is the most common cause of cardiovascular disease [7, 8]. Every 1 in 250 individuals can be affected by it globally [8]. At the same time, it can be observed that social and psychological circumstances can affect cholesterol levels and daily fat intake [11], for instance, employment status. Development on the research of social-psychological circumstances potentially can help creating predictive models for an individual's cholesterol level and even save lives by reducing rates of cardiovascular disease.

3) Can income level and physical activity be used to predict mental health problems?

In modern society, we notice growth in mental health problems with around 14% of the global disease burden caused by mental disorders [12]. Causes for mental health problems can often be attributed to behavioural and socioeconomic factors [12, 13, 14]. Expanding research on predictive factors for mental health can help decrease the scale of this problem.

Section 3 — Processing the data [20 marks]

Evidence for learning outcomes: LO2 Manipulate a data set to extract statistics and features

2. [10 marks] Describe how you explored the data, why you did it that way, and what conclusions you drew about it (300 words maximum)

The first two stages of CRISP-DM, business understanding and data understanding [15], were used in order to start exploration in this section. Business understanding was gained by analysing documentation that comes with the original data set [16] and data understanding was done by analysing the data dictionary [17] and exploring the data set itself using data analysis software WEKA and statistical data exploration methods.

Behavioural Risk Factor Surveillance System (BRFSS) is a telephone survey that gathers data statewide on behavioural factors and preventive health practices that affect certain medical conditions and traumas for people 18 years old and older [16]. Data set consists of 12338 instances with 414 attributes for each instance.

Firstly, analysis of attribute names and their values was done in WEKA using data breakdown and histograms to indicate distinct values for each attribute, their distribution, share of missing values [18]. It is clear that the majority of data provided is categorical and a smaller share of it is numerical. After the assessment of data dictionary and histograms, a group of attributes were selected that are relevant to the research questions. However, some numerical attributes, despite of appropriate deviation, were reconsidered after indicating a large share of missing values on dot plots [18].

Moreover, it was discovered via WEKA that certain attributes are having values that provide no weighting, for instance, "Refused", "Not asked", "Don't know", "Missing". It was confirmed via Python, using function “value_counts()”. Certain attributes had around 90% of this type of data and could not be used for analysis due to lack of weighting [19].

Basic statistics were helpful in comparing relevant attributes and selecting correct ones for research questions, findings were used to refine research questions. It is important to evaluate potential errors before cleaning data to make future predictive models more efficient [19].

3. [10 marks] Describe the cleaning/fixing you did on the data, and why (300 words)

For the data preparation step of CRISP-DM, including data cleaning and fixing, a Python programming language was used with Numpy and Pandas libraries, as they provide comprehensive tools for data processing and data analysis [20].

Firstly, data was imported into Python data frame by using function ".read_csv()". Data frame object can be thought of as a 2-dimentional table which makes it easier to work with data in Python [20]. As machine learning models can be confused by irrelevant attributes [19], next step was removing irrelevant attributes and keeping the ones that would need to be used for each research question. This was done by using ".drop()" and "columns.difference()" functions [20], which allowed removing all irrelevant attributes with names of relevant ones passed as parameters. Remaining attributes are:

Q1 - "SMOKDAY2", "MARITAL", "GENHLTH", "SEX", "AGE".

Q2 - "EMPLOY1", "_RFCHOL".

Q3 - "MENTHLTH", "EXERANY2", "INCOME2"

Secondly, breakdown of values of each attribute indicated values that do not carry any weighting, like "Refused", "Don´t Know/Not Sure", etc. Decision was made to remove these values via ".loc[]" function due to their small share and lack of weighting [21]. In order to handle missing values, instances with them were removed via ".dropna()" function to avoid confusing machine learning algorithms [21].

Extra steps were taken to filter data for Q1 via sex and age, this was done by via ".loc[]" and ".isin()" functions, after filtering these attributes were removed as they were intended only for filtering. Additionally, values in attributes for Q1 were simplified via mapping function ".map()" to simplify potential decision tree and make it more readable [21].

Finally, data was exported into files for each question via ".to_csv()". For future iterations of CRISP-DM, it is recommended to try computating synthetical data for missing values to see if it will improve machine learning models [19].

Section 4 — Data analysis [40 marks]

Evidence for learning outcomes: LO3 Critically evaluate and apply data mining techniques/tools to build a classifier or regression model, and predict values for new examples

4. [10 marks] Explain what analysis techniques you used to answer your business/research questions, and why (300 words maximum)

Several analysis techniques were considered for given research questions, such as regression, clustering and classification. Regression is a solid analysis technique for numerical data, especially with linear trends [19]. However, in this case all relevant attributes are nominal and expected output is not a number. Therefore, decision was made on not using regression modelling.

Clustering is a strong technique to classify data that does not have any labels on it, which means that it is a unsupervised learning algorithm and it is very efficient in dividing known data in yet unknown groups [22]. Nonetheless, in given research questions necessary labels are provided and the task is predicting output for future instances based on current ones, not identifying groups to begin with. Due to that, clustering was not used as an analysis technique for research questions.

As per Rob Alexander's lecture, in cases when it is needed to identify classes from labelled data, i.e. supervised learning, classification should be strongly considered [19]. Also, per the same lecture from Rob Alexander, feature interactions can potentially be expected, and decision trees can make them more evident [23]. That makes classification group of algorithms relevant to use with all given research questions, as they all have categorical attributes, they all have labelled data, and expected output is categorical data as well.

There are several types of classification algorithms available in WEKA, including:

- C4.5 tree (J48)

- Random tree

- Random forest

- Sequential minimal optimisation

- Naive Bayes

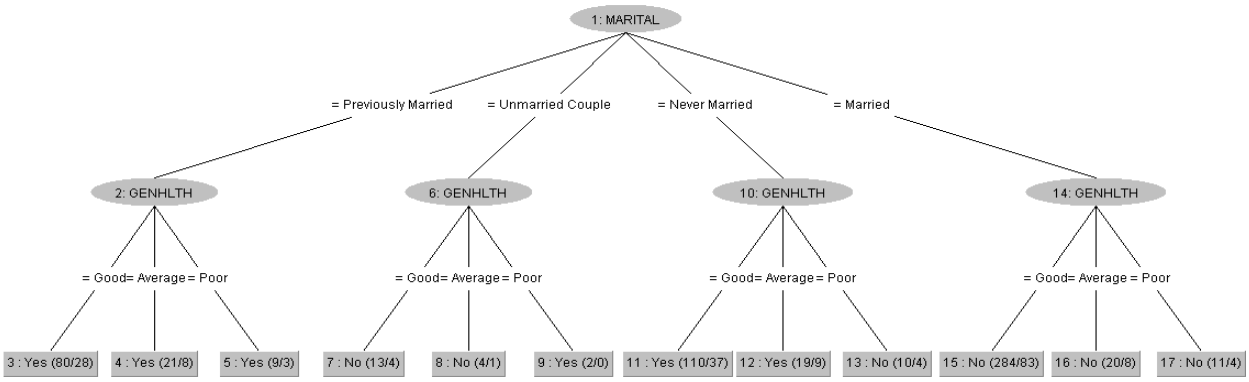
Above classifiers were used in analysis and modelling for all three research questions. Accuracy measurements and confusion matrices were used to compare results and identify best the most accurate model [19]. Decision trees were used for certain classifiers where relevant. Finally, cross-validation technique was used to ensure correct distribution between training and test instances [24].

5. [10 marks] Summarise the results of your analysis (300 words maximum)

As mentioned in the previous section, several classifiers were used in creating machine learning models in WEKA. Cross-validation was used for each of them with arbitrary 10 folds [24]. Note that only one variation of fold was used in order to keep this section within word count. Correctly classified instances were taken from confusion matrices for each question and classifier. TP rate, FP Rate, Precision, Recall and F-Measure all indicate accuracy of the created model. Generally for these characteristics of accuracy the bigger, the better, except for FP Rate, for it the smaller, the better [19].

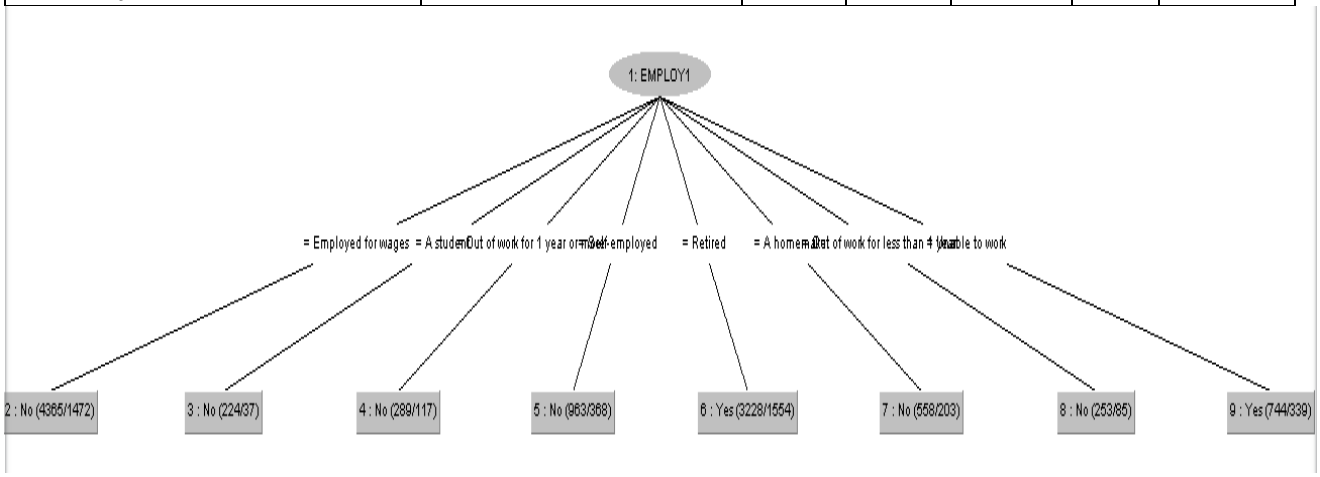
Q1) Below summarised results of analysis for question 1

Cross-validation 10 folds	Correctly Classified Instances	TP Rate	FP Rate	Precision	Recall	F-Measure
C4.5 decision tree (J48 in WEKA)	66.90%	0.669	0.343	0.668	0.669	0.668
Random tree	66.90%	0.669	0.346	0.667	0.669	0.667
Random forest	66.90%	0.669	0.349	0.667	0.669	0.667
Sequential minimal optimisation (SMO)	66.90%	0.669	0.343	0.668	0.669	0.668
Naive Bayes	66.72%	0.667	0.344	0.666	0.667	0.667



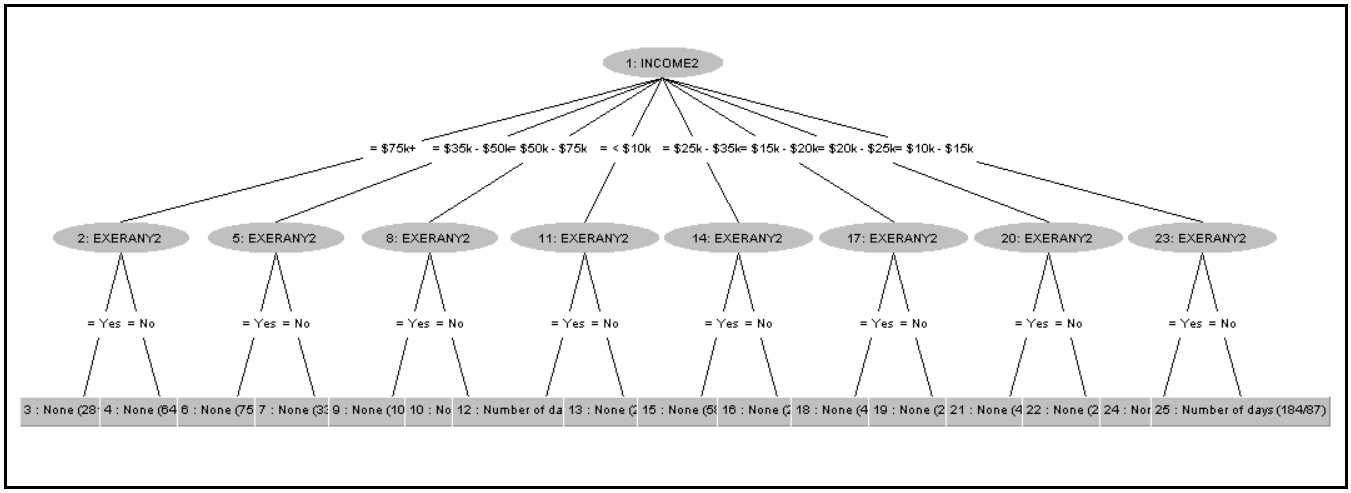
Q2) Below summarised results of analysis for question 2

Cross-validation 10 folds	Correctly Classified Instances	TP Rate	FP Rate	Precision	Recall	F-Measure
C4.5 decision tree (J48 in WEKA)	60.70%	0.607	0.433	0.602	0.607	0.604
Random tree	60.70%	0.607	0.433	0.602	0.607	0.604
Random forest	60.70%	0.607	0.433	0.602	0.607	0.604
Sequential minimal optimisation (SMO)	60.70%	0.607	0.433	0.602	0.607	0.604
Naive Bayes	60.70%	0.607	0.433	0.602	0.607	0.604



Q3) Below summarised results of analysis for question 3

Cross-validation 10 folds	Correctly Classified Instances	TP Rate	FP Rate	Precision	Recall	F-Measure
C4.5 decision tree (J48 in WEKA)	65.79%	0.658	0.658	0.610	0.658	0.566
Random tree	66.05%	0.661	0.618	0.619	0.661	0.566
Random forest	66.05%	0.661	0.618	0.619	0.661	0.566
Sequential minimal optimisation (SMO)	65.81%	0.658	0.617	0.611	0.658	0.567
Naive Bayes	65.85%	0.658	0.615	0.612	0.658	0.569


1: INCOME2
= $75k+
= $35k - $50k
= $50k - $75k
= < $10k
= $25k - $35k
= $15k - $20k
= $20k - $25k
= $10k - $15k
2: EXERANY2
5: EXERANY2
8: EXERANY2
11: EXERANY2
14: EXERANY2
17: EXERANY2
20: EXERANY2
23: EXERANY2
= Yes
= No
3 : None (28
4 : None (64
6 : None (75
7 : None (3
9 : None (10
10 : No
12 : Number of da
13 : None (2
15 : None (5
16 : None (2
18 : None (4
19 : None (2
21 : None (4
22 : None (2
24 : Nor
25 : Number of days (184/87)

6. [10 marks] What do the results say in answer to your business/research questions? (300 words maximum)

RQ1) After critically evaluating results, it is evident that the smoking can be predicted from marital status and general health for males aged 35–54 with 66.9% accuracy. All classifiers provided similar rate of correctly classified instances with the same weighted average TP, FP Rate, Precision, Recall, F-Measure. SMO and J48 are two classifiers that provided the best Kappa statistics results with 0.3273 for both of them, which indicated the quality of predictive model as "fair" [25, 26]. Decision tree indicated that generally members of married or unmarried couples are less likely to smoke, wheres previously married and never married people are more likely to be smokers. Couple of exceptions noticed in extended tree: people from unmarried couples with poor health likely to be smoking and never married people with poor health less likely to be smoking.

Q2) Results of analysis for this question confirmed that the employment status of a person can be predictive factor for an individual's cholesterol level with 60.7% accuracy. Decision tree indicated that people who are retired or unable to work have higher chance of high cholesterol. For remaining people, regardless if they are students, employed or employed, it is more probable to have normal cholesterol level. There is not classifier that clearly stands out with performance, thus, they all can be used to the same extent as a first version for this model.

Q3) Finally, income level and physical activity can be used to predict mental health problems with approximately 65.85% accuracy. Decision tree indicated that people with income < $10k who exercised and people with income $10k-$15k who did not exercised are more likely to have mental health problems. Random tree proved to be the best algorithm with kappa score 0.531, which makes predictive model of "fair to good" quality [27, 25, 26].

For future research, classification with percentage split and additional predictive attributes should be considered.

7. [10 marks] Describe the most salient threats to validity that remain in your analysis (300 words maximum)

RQ1) The first threat to validity in the analysis is the original quality of data. Specifically, it is right away evident that selected attributes missing some values. Usually, they would take the form of "Not asked or missing" record or similar. For instance, attribute "SMOKDAY2" missing 7269 values out of 12338 instances, which is approximately 59% of data for this attribute. As result, valuable observations potentially could be missed due to that problem.

Additionally, a share of data being lost when filtering by age and sex is being applied, which leaves a low quantity of instances for learning. Precisely, 583 instances left after removing missing values and filtering by age and gender. Classifiers can still build a solid model on this amount of instances, but it would have to be further trained on incoming instances to ensure its validity and to avoid

overfitting [15, 19]. Computing missing values could be considered as a way to provide more training instances [28].

Finally, certain data simplifications were made to make decision tree more readable, but they could affect accuracy of prediction. For example, in "GENHLTH" attribute "Excellent", "Good" and "Very good", were all combined in "Good". As a solution, a new model could be trained without modifying attributes and compared to the current model.

RQ2) Similar to the first research question, validity of research could be affected by 1575 missing values in "_RFCHOL" attribute. Besides that, this attribute contains computed values, and the quality of computation can affect the validity of the attribute and prediction model.

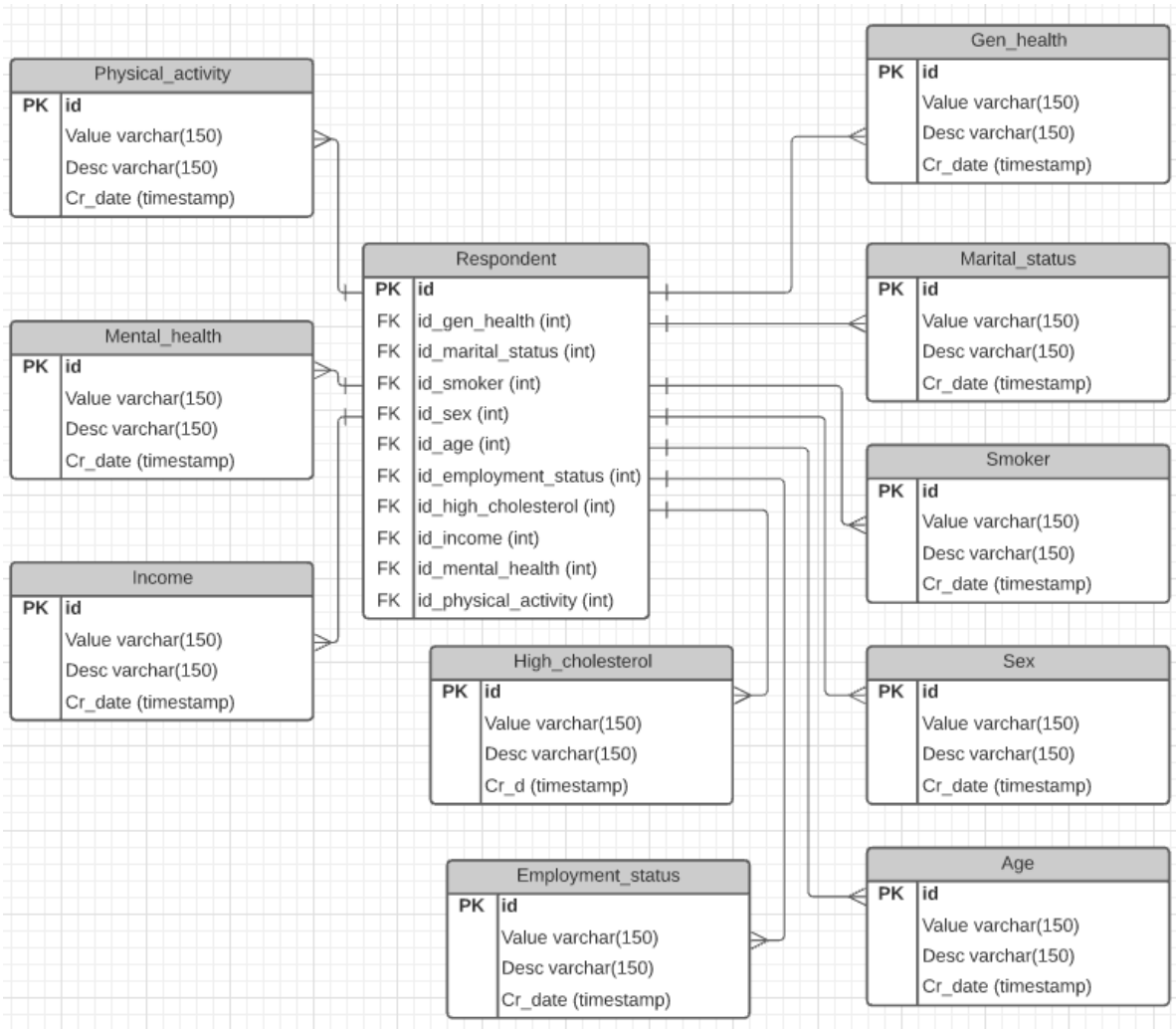
RQ3) Besides already mentioned above, "EXERANY2" attribute contains binary data on whether the person did any physical activity or exercise in the last 30 days. To build a more precise model, it could be beneficial to divide physical activity and exercise, and to consider their frequencies.

Section 5 — Dealing with large data sets [20 marks]

Evidence for learning outcomes: LO1 create a data set using modern database models and technology *and* LO4 Analyse and communicate issues with scaling up to large data sets, and use appropriate techniques to scale up the computation

8. **[10 marks] Describe how you could represent the data in a relational database — give a suitable schema, and describe a mechanism for converting it to a suitable input form for WEKA (300 words maximum)**

Below schema indicates how normalised data would look in RDBS MySQL.



There are several ways to prepare data set for appropriate format in WEKA, like compile into .csv file or more advanced technique connecting WEKA and SQL database to write queries directly from WEKA to import data in it. This can be done by using JDBC API. JDBC driver and MySQL connector would have to be installed, DatabaseUtils.props file customised [33, 34].

Further SQL queries to mine data and process it into WEKA suitable format can be seen. Due to word count restrictions, only the query for the most complex RQ1 provided. Queries for RQ2 and RQ3 would be simplified version of it.

select (CASE WHEN gh.value = 'Poor' THEN gh.value WHEN gh.value = 'Fair' THEN 'Average' WHEN gh.value = 'Excellent' OR gh.value = 'Good' OR gh.value = 'Very good' THEN 'Good' END IF) as Gen_health, (CASE WHEN ms.value = 'Married' THEN ms.value WHEN ms.value = 'Never married' THEN ms.value WHEN ms.value = 'Divorced' OR ms.value = 'Widowed' OR ms.value = 'Separated' THEN 'Previously Married' WHEN ms.value = 'A member of an unmarried couple' THEN 'Unmarried Couple' END IF) as Marital_status, (CASE WHEN s.value = 'Not at all' THEN 'No' WHEN s.value = 'Every day' THEN 'Yes' WHEN s.value = 'Some days' THEN 'Yes' END IF) as smoke_status, Sex.value, Age.value

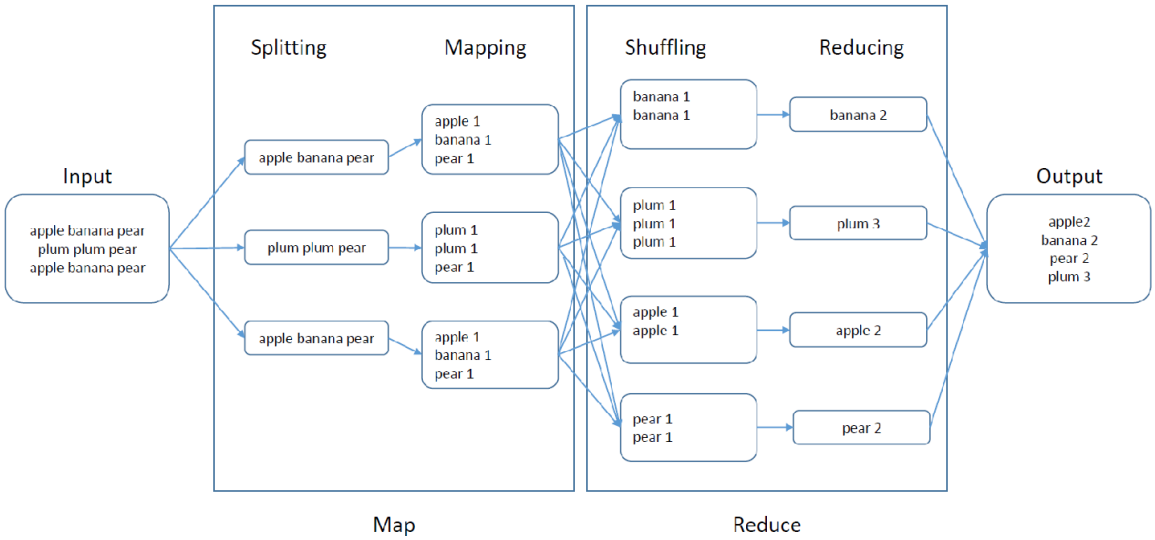
```
from respondent as resp

join Gen_health as gh a on gh.id = resp.id_gen_health

join Marital_status as ms on ms.id = resp.id_marital_status

join Smoker as s on s.id = resp.id_smoker

join Sex on Sex.id = resp.id_sex

join Age on Age.id = resp.id_age

where Age.value in ('Age 45 - 54', 'Age 35 – 44') and Sex.value = 'Male' and gh.value not in ('Refused', 'Don't know/Not Sure') and s.value not in ('Not asked or Missing', 'Refused', 'Don´t Know/Not Sure') and ms.value not in ('Refused')
```

9. **[10 marks] Imagine that you have to deal with a much larger version of this dataset, with gigabytes of data already held and tens of megabytes of new data arriving each day. Assume that rapid response to new data is required e.g. certain messages being sent immediately as soon as certain automated analysis results produce results of a certain value (e.g. a count of items of a particular type exceeds a predetermined threshold). Describe a way that you could use appropriate technologies to spread the load over multiple computers, and justify why this would be a good approach. (300 words maximum)**

In the situation when a traditional storage unit and processing unit cannot handle the volume of incoming data, the use of Hadoop could be justified as Hadoop is using a cluster of commodity hardware, i.e. several storage units and processing units altogether.

Hadoop distributed file system (HDFS) allows storing data in separate blocks across several computers and different clusters. HDFS also stores copies of the data across many systems to reduce data loss. This distributed data storage approach is reliable, easy to scale and cost-efficient [35].

To process large volumes of data efficiently, MapReduce component of Hadoop could be used [36]. The general idea behind MapReduce is splitting incoming data into small chunks, then data is being mapped depending on the word's frequency, after that similar words are getting shuffled and grouped, and finally, on the reduce phase, all grouped words are given a count. In order to provide output aggregating the results will take place. This approach of MapReduce saves a lot of time and makes load balancing more efficient. Hadoop can work rapidly due to its ability of processing and storing data on the same server whereas traditionally data would have to move from storage to processing server first [36]. The screenshot below visualises MapReduce process.



After data was mapped, YARN component can be used to further process it and to manage available resources efficiently [37]. Allocation of resources in Hadoop cluster is one of the primary functions of YARN. Besides that, YARN is responsible for scheduling tasks to be completed in various cluster nodes. Finally, YARN can extend Hadoop functionality with various modules [37].

For the situation when big incoming data volume should be processed quickly and output should be provided to users rapidly, Apache Spark module should be used for advanced, flexible and rapid distributed processing[38].

Section 6 — Privacy [10 marks]

Evidence for learning outcomes: LO5 Critically discuss the need for privacy, identify privacy risks in releasing information, and design techniques to mediate these risks

10. [10 marks] Imagine that this data will be used in support of a public-facing application. List the three most salient privacy issues related to this analysis and give strategies you could use to address each of them. (300 words maximum)

Even though provided data set anonymised, i.e.missing any names, certain individuals can still be identified from provided data, for instance from a rare medical condition, age and state can identify respondent, and with it, respondent's income, health conditions, marital status, etc. Auxiliary data can be the form of filter for de-anonymisation or re-identification process. This de-anonymised data potentially could be used for the hacking of bank accounts or other unethical purposes [30]. The most radical solution could be imposing strong restrictions on data access, but it is not always practical as data used for processing purposes and should be somehow shared with relevant stakeholders [29]. To protect respondents from the de-anonymisation risks, data could be slightly modified without changing the overall trends of the data. Also, the pseudonymisation

technique can be used in cases when data needs to be moved between entities of an organisation [31].

Another risk for data privacy is that data may be used for purposes other than those stated and intended. For instance, data could be collected to analysed to identify threats to public health, but was used to market commercial products relevant to respondents that may need this product [29]. To avoid this problem, respondents should clearly be informed about the purposes of data collection [31]. Also, organisations should define in advance a wide range of purposes for data collection and use that may arise in the future [29].

Finally, security breaches are a massive risk to data privacy that is hard to evaluate and fully predict, especially for healthcare data [32]. If a data breach will happen, data can be sold or used for further hacking purposes. Both using secure up-to-date systems and creating relevant business processes and training for employees can reduce the chance of both potential cyber attacks and social engineering [32].

Section 7 — Report references

11. Provide a correctly structured list of references to all the resources used for this development and report (no word limit)

[1] L. A. Loeb, V. L. Emster, K. E. Warner, J. Abbotts, and J. Laszlo, “Smoking and Lung Cancer: An Overview,” Cancer Res, vol. 44, no. 12 Part 1, pp. 5940–5958, Dec. 1984.
[2] A. Chao, M. J. Thun, S. J. Henley, E. J. Jacobs, M. L. McCullough, and E. E. Calle, “Cigarette smoking, use of other tobacco products and stomach cancer mortality in US adults: The Cancer Prevention Study II,” *International Journal of Cancer*, vol. 101, no. 4, pp. 380–389, 2002, doi: 10.1002/ijc.10614.
[3] “Associations between smoking cessation and anxiety and depression among U.S. adults,” Addictive Behaviors, vol. 34, no. 6–7, pp. 491–497, Jun. 2009, doi: 10.1016/j.addbeh.2009.01.005.
[4] S. A. McKee, T. Falba, S. S. O’Malley, J. Sindelar, and P. G. O’Connor, “Smoking Status as a Clinical Indicator for Alcohol Misuse in US Adults,” Archives of Internal Medicine, vol. 167, no. 7, pp. 716–721, Apr. 2007, doi: 10.1001/archinte.167.7.716.
[5] W. V. Ark, L. J. DiNardo, and D. S. Oliver, “Factors Affecting Smoking Cessation in Patients With Head and Neck Cancer,” The Laryngoscope, vol. 107, no. 7, pp. 888–892, 1997, doi: 10.1097/00005537-199707000-00010.
[6] “*Psychosocial approaches to smoking prevention: A review of findings.* https://psycnet.apa.org/record/1986-31012-001 (accessed Jul. 12, 2021).
[7] D. Steinberg, S. Parthasarathy, T. E. Carew, J. C. Khoo, and J. L. Witztum, “Beyond Cholesterol,” *http://dx.doi.org/10.1056/NEJM198904063201407*, Jan. 14, 2010. https://www.nejm.org/doi/pdf/10.1056/NEJM198904063201407 (accessed Jul. 12, 2021).

[8] “Canadian Cardiovascular Society Position Statement on Familial Hypercholesterolemia: Update 2018,” Canadian Journal of Cardiology, vol. 34, no. 12, pp. 1553–1563, Dec. 2018, doi: 10.1016/j.cjca.2018.09.005.
[9] “Marital status and smoking in Korea: The influence of gender and age,” Social Science & Medicine, vol. 66, no. 3, pp. 609–619, Feb. 2008, doi: 10.1016/j.socscimed.2007.10.005.
[10] “Meaning in life, anxiety, depression, and general health among smoking cessation patients,” Journal of Psychosomatic Research, vol. 67, no. 4, pp. 353–358, Oct. 2009, doi: 10.1016/j.jpsychores.2009.02.006.
[11] “A review of selected studies assessing social-psychological determinants of fat and cholesterol intake,” Food Quality and Preference, vol. 3, no. 4, pp. 183–200, Jan. 1991, doi: 10.1016/0950-3293(91)90033-B.
[12] “No health without mental health,” The Lancet, vol. 370, no. 9590, pp. 859–877, Sep. 2007, doi: 10.1016/S0140-6736(07)61238-0.
[13] R. Araya, G. Lewis, G. Rojas, and R. Fritsch, “Education and income: which is more important for mental health?,” Journal of Epidemiology & Community Health, vol. 57, no. 7, pp. 501–505, Jul. 2003, doi: 10.1136/jech.57.7.501.
[14] J. S. Raglin, “Exercise and Mental Health,” Sports Med, vol. 9, no. 6, pp. 323–329, Jun. 1990, doi: 10.2165/00007256-199009060-00001.
[15] J. D. Kelleher and B. Tierney, Data Science. MIT Press, 2018.
[16] *Behavioral Risk Factor Surveillance System 2015 OVERVIEW,* NYSDOH Division of Chronic Disease and Prevention., New York, USA, 2015.
[17] *Data Dictionary 2015 OVERVIEW,* NYSDOH Division of Chronic Disease and Prevention., New York, USA, 2015.
[18] C. Barr, D. M. Diez, and C. Rundel, OpenIntro Statistics. Opentextbookstore, 2016. Accessed: Jul. 12, 2021. [Online]. Available: https://lib.hpu.edu.vn/handle/123456789/21792
[19] F. Azuaje, “Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques 2nd edition,” BioMedical Engineering OnLine, vol. 5, no. 1, p. 51, Sep. 2006, doi: 10.1186/1475-925X-5-51.
[20] W. McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O’Reilly Media, Inc., 2012.
[21] D. Cielen and A. Meysman, Introducing Data Science: Big data, machine learning, and more, using Python tools. Simon and Schuster, 2016.
[22] “Pattern classification and clustering: A review of partially supervised learning approaches,” Pattern Recognition Letters, vol. 37, pp. 4–14, Feb. 2014, doi: 10.1016/j.patrec.2013.10.017.
[23] Vens, Struyf, and Schiegat, et al. *Decision trees for hierarchical multi-label classification*. Springer, 2007.
[24] “Cross-Validation Methods,” Journal of Mathematical Psychology, vol. 44, no. 1, pp. 108–132, Mar. 2000, doi: 10.1006/jmps.1999.1279.
[25] D. Cicchetti and S. Sara, “Developing Criteria for Establishing Interrater Reliability of Specific Items: Applications to Assessment of Adaptive Behavior.,” American Journal of Mental Deficiency, vol. 86, no. 2, pp. 127–137, 1981.
[26] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” Biometrics, vol. 33, no. 1, pp. 159–174, 1977, doi: 10.2307/2529310.
[27] J. L. Fleiss, B. Levin, and M. C. Paik, Statistical Methods for Rates and Proportions. John Wiley

& Sons, 2013.
[28] E. Acuña and C. Rodriguez, “The Treatment of Missing Values and its Effect on Classifier Accuracy,” in Classification, Clustering, and Data Mining Applications, Berlin, Heidelberg, 2004, pp. 639–647. doi: 10.1007/978-3-642-17103-1_60.
[29] M. Altman, A. Wood, D. R. O’Brien, and U. Gasser, “Practical approaches to big data privacy over time,” International Data Privacy Law, vol. 8, no. 1, pp. 29–51, Feb. 2018, doi: 10.1093/idpl/ipx027.
[30] J. Qian, X.-Y. Li, C. Zhang, L. Chen, T. Jung, and J. Han, “Social Network De-Anonymization and Privacy Inference with Knowledge Graph Model,” IEEE Transactions on Dependable and Secure Computing, vol. 16, no. 4, pp. 679–692, Jul. 2019, doi: 10.1109/TDSC.2017.2697854.
[31] M. Hintze and K. El Emam, “Comparing the benefits of pseudonymisation and anonymisation under the GDPR,” *Journal of Data Protection & Privacy*, vol. 2, no. 2, pp. 145–158, Dec. 2018.
[32] “Digging Deeper into Data Breaches: An Exploratory Data Analysis of Hacking Breaches Over Time,” *Procedia Computer Science*, vol. 151, pp. 1004–1009, Jan. 2019, doi: 10.1016/j.procs.2019.04.141.
[33] “How do i connect to a database - Weka Wiki.” https://waikato.github.io/weka-wiki/faqs/how_do_i_connect_to_a_database/ (accessed Jul. 12, 2021).
[34] “MySQL :: Download Connector/J.” https://dev.mysql.com/downloads/connector/j/ (accessed Jul. 12, 2021).
[35] *Hadoop and distributed computation*. [Online]. Available: https://www.york.ac.uk/eldt/canvas/canvas_landing_page.html. [Accessed: 10 Jul. 2021].
[36] K. Shvachko et al., ‘The Hadoop Distributed File System’, in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 2010, pp. 1–10 [Online]. Available: 10.1109/MSST.2010.5496972.
[37] T. White, *Hadoop: The Definitive Guide*. O’Reilly Media, Inc., 2012.
[38] X. Meng et al., ‘MLlib: Machine Learning in Apache Spark’, *Journal of Machine Learning Research*, vol. 17, no. 34, pp. 1–7, 2016.