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](https://www.york.ac.uk/students/studying/skills/integrity/referencing-styles/ieee/) [system.](https://www.york.ac.uk/students/studying/skills/integrity/referencing-styles/ieee/)

# 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. What factors contribute to the correlation between wealth and health?  Studies indicate that health rises with income level for populations [4]. A multitude of factors could influence this association, such as being unable to afford medication. This study could reveal which income-related health factors are most influential to physical health. Beyond income group, this question would explore whether poor health is most affected by not being to afford medicine, not seeing a doctor, or not having access to Medicare. This could highlight what areas of medical support we could focus on to best address the negative effects of the wealth + health gap.  2. Can level of physical activity be predicted using various health factors such as smoking and high blood pressure?  According to an article by Haskell et all [5], there is “substantial data supporting an inverse relationship between the amount of habitual physical activity performed and a variety of negative health outcomes throughout the lifespan.” Predicting higher physical activity using a classifier trained on a selection of responses on health condition and habits would indicate what negative health features are most influenced by changes in physical activity.  3. Do healthier people consume more fruits and vegetables?  It is known that eating fruits and vegetables over other types of food benefits health by providing nutrients and minerals [8]. The performance of a model trained on BRFSS data would reveal how relevant diet choices are to overall health. Fruit and vegetable consumption analysis can be compared to see which is more relevant to health. This question matters because meal choices are simple changes that can be made every day that could have a severe impact on human health. |

# Section 3 — Processing the data [20 marks]

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

1. **[10 marks] Describe how you explored the data, why you did it that way, and what conclusions you drew about it (300 words maximum)**

|  |
| --- |
| Information was gathered from DataDictionary, a document containing descriptions of each feature, to begin formulating questions. Although some fields can be contextually understood, many names are undescriptive and meaning cannot be inferred from data points (e.g. yes or no). The DataDictionary provides a brief description that clarifies each field’s meaning, allowing a better view of the data.  Features of interest and potential relationships between them are noted for further analysis. According to John D. Kelleher and Brendan Tierney in “Data Science” pg 47 [1]: “It is frequently the case that the real value of a data science project is the identification of one or more important derived attributes that provide insight into a problem.” Derived attributes in this context refers to attributes emerging from interactions between the given attributes. By finding derived attributes, meaningful insights can be gained, leading to meaningful questions. The dataset provides some of these derived attributes, such as “COMPUTED INCOME CATEGORIES,” used to easily train prediction models.  This is a dataset of survey responses. Cognitive biases may affect the truth of subjective survey responses as outlined by Marianne Bertrand and Sendhil Mullainathan in their paper on “Implications for Subjective Survey Data” [2]. A prompt from the database: “During the past 30 days, about how often did you feel restless or fidgety?” can influence other questions within the survey just from being presented in a certain order, as seen when a “question induced people to focus on one aspect of their life, an aspect that had undue effects on their subsequent answer,” [2]. Responses addressing mental health, such as: “During the past 3 days, about how often did you feel that everything was an effort?” are vulnerable to this. Research questions that address opinions and feelings ought to be structured in a way that acknowledges potential biases. |

1. **[10 marks] Describe the cleaning/fixing you did on the data, and why (300 words)**

|  |
| --- |
| Preprocessing is done in a Python Jupyter Notebook. There are a number of columns containing inputs: “Data do not meet the criteria for statistical reliability, data quality, or confidentiality (data are suppressed).” These are dropped from the working dataframe. Any rows missing one or more desired features for analysis are dropped to keep the data consistent for training. Responses such as “Don't know/Not Sure/Refused/Missing”, are sought out and similarly removed. While Naive Bayes can function in the presence of N/A values [9], N/A values are still dropped from question 2’s dataset in order to simplify conversion to integers.  There are several hundred columns that could contain non-useful information for a predictive model. On “Irrelevant Features and the Subset Selection Problem”: “Ideally, the induction algorithm should use only the subset of features that leads to the best performance.” [3]. For this reason, only columns relevant to the questions are retained while the rest are filtered.  Binary responses, such as “Fair or poor health” vs “Good or Better Health“, are altered to a discrete numeric form in order to properly feed data into the classifier model. The dataframe map() method is used to specify which string values are converted to which binary values. Some categories, such as “\_RFHLTH" are converted to a float by the map method, so these are then converted into int data type using astype(). (fig1)  The income group feature is mapped to a binary 0 or 1 to represent <50k (non-wealthy) and >50k (wealthy) income respectively. Wealth groups are reduced to binary form as this question is interested in the difference between two representations of wealth and it doesn’t address differences between specific income groups. 50K is chosen as it is the upper limit of the data provided in the data set. |

# **Figure 1**

# Section 4 — Data analysis [40 marks]

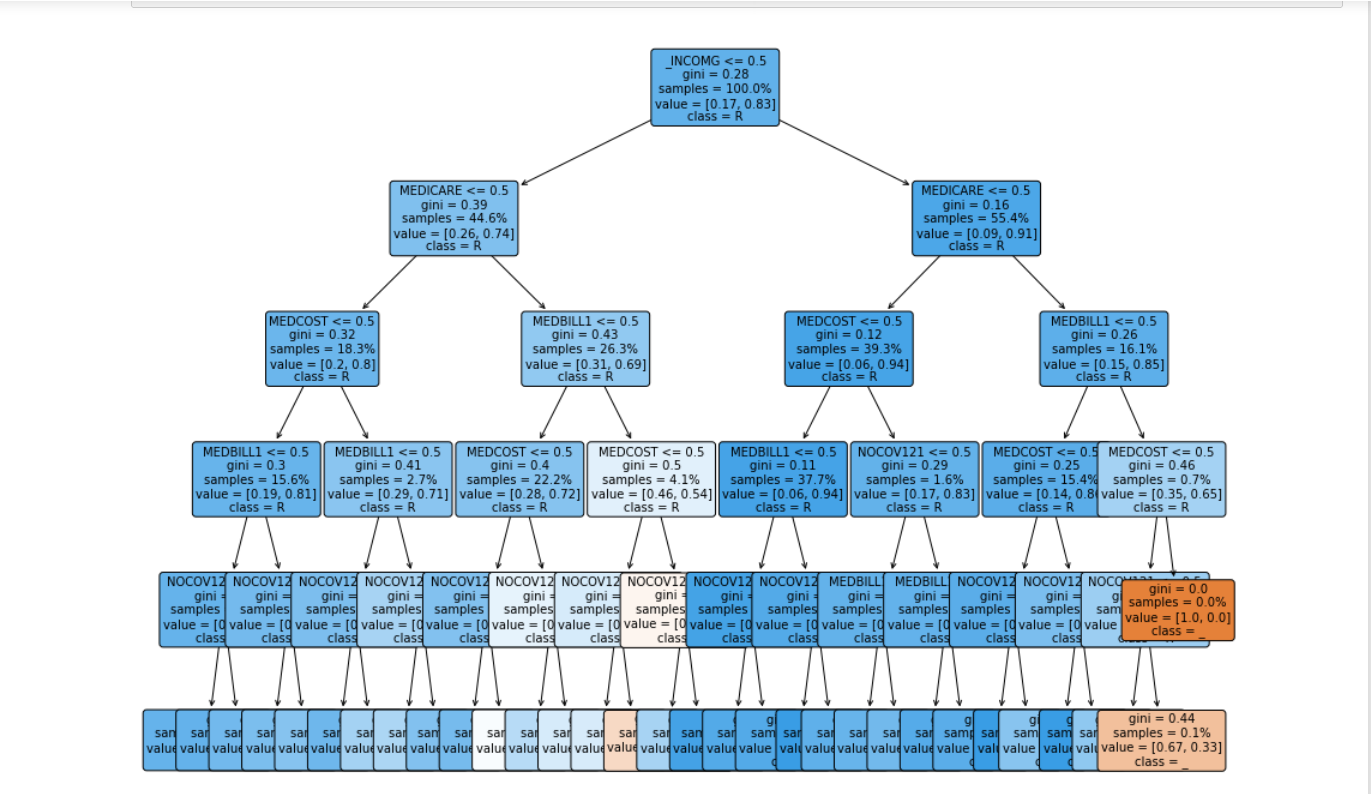
Evidence for learning outcomes: LO3 Critically evaluate and apply data mining techniques/tools to build a classifier or regression model, and predict values for new examples

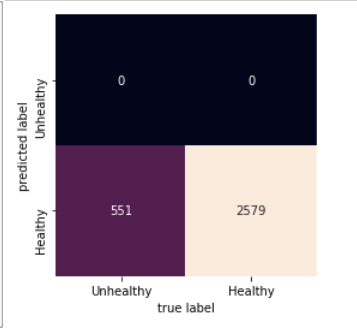
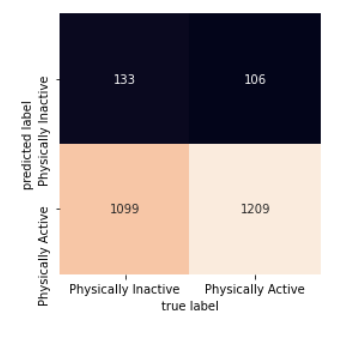
1. **[10 marks] Explain what analysis techniques you used to answer your business/research questions, and why (300 words maximum)**

|  |
| --- |
| 1. A decision tree is appropriate for classification using nominal inputs [1] such as the given data features. The decision tree can be optimized for income level prediction with the data, then broken down to observe which features were most important for accurate predictions. Correlation measurements can be compared with a confusion matrix visualization [8] represented by a coefficient. The greatest coefficients would indicate which of the features are most influential to decision making.  2. Naive Bayes is used as a classifier for this set of data, as it is ideal for probabilistic prediction based on discrete, nominal categories. “The naive Bayes classifier is optimal for any two-class concept with nominal features that assigns class 0 to exactly one example, and class 1 to the other examples,” according to “An empirical study of the naive Bayes classifier” theorem 1 [7]. Features such as high cholesterol and obesity contain yes/no responses that are mapped to 0 or 1. Independence between the variables is assumed as prerequisite to implementing Naive Bayes [7]. The negative health factors selected do not have clear dependence on each other, so Naive Bayes is appropriate.  3. “Support Vector Machines use linear models to implement nonlinear class boundaries” [9]. Fruit and vegetable consumption represent linear values that can be used to construct a classifier that predicts a non-linear binary health category. SVMs establish a maximum margin hyperplane that divides points into appropriate categories, healthy or unhealthy in this case. This method is chosen over Naive Bayes and a Decision Tree because an advantage to using SVMs is that over-fitting is reduced [9] as only training instances that are support vectors would make significant changes across the whole set. SVMs are commonly used to make classifications over linear data, which matches the profile of this data set. |

1. **[10 marks] Summarise the results of your analysis (300 words maximum)**

|  |
| --- |
| 1. The decision tree trained on income-related attributes was able to predict health category with an accuracy of ~84.15%. The decision tree diagram shows that decision splits are made first on income category, then medicare access, then on whether the participant currently is paying a medical bill if they do have medicare or whether the participant was able to afford medicine if they did not have medicare. Whether the participant had a period of time when they didn’t have health coverage is asked at the end before each leaf node. Income group had the highest importance score of ~0.6, Medicare: ~0.17, paying off medical bills: ~0.125, and so on. Whether the participant has any kind of health coverage had a feature importance of zero and is not present in the decision tree. (fig2.1)  2. The Multinomial Naive Bayes trained on the selection of health habits to predict physical activity found little correlation, resulting in a low accuracy score of ~52.69%. This is just over half of correct guesses on a binary category. Feature importance scores for each of the input features was averaged out to display. Importance scores were low, nearly zero for all features. A confusion matrix shows that the model over-predicts good health for the majority of the data set. ~100 ‘unhealthy’ predictions are made while ~1000 ‘healthy’ predictions are made, even though the amount of healthy and unhealthy respondents is nearly evenly split. (fig2.2)  3. The Support Vector Machine trained on fruit and vegetable consumption was able to predict the participant’s healthiness category with a score of ~82.4%. The confusion matrix representation of the results reveal that no ‘unhealthy’ predictions were made on the data set and that the model guessed ‘Healthy’ for all test data points. ‘Healthy’ responses made up ~82.4% of total responses (2579/3130). (fig2.3) |

**Figure 2.1**

****

**Figure 2.2 Figure 2.3**

1. **[10 marks] What do the results say in answer to your business/research questions? (300 words maximum)**

|  |
| --- |
| 1. Results show that of the selected features used for prediction, income group was the most important. This implies that wealth is most closely related with health, with deeper unseen factors than being able to access medicine or a doctor. Medicare is the second most important factor to healthiness, as it is the second to be factored into a decision after income group. Because ‘HLTHPLN1’ has no significance for the decision tree, having any kind of health coverage other than medicare is not a predictor for health. Medicare is particularly important for determining whether a participant belongs to the ‘healthy’ category.  2. Accuracy for the Naive Bayes is low, meaning that there is a low association between the selected features. The hypothesis of unhealthy attributes being closely correlated with reported physical activity level is thus disproved. Each of the features has little significance for the model’s predictions, further demonstrating that they are not correlated with physical activity. There are a disproportionate amount of physically active responses in comparison to inactive responses, which may not be representative of the general population. Furthermore, preprocessing had reduced the number of entries to just over 1300. More responses may be needed for a more accurate model.  3. This model was trained to make all ‘Healthy’ predictions which yielded a high accuracy score. In this case, the data was skewed enough such that the vast majority of responses, regardless of fruit and vegetable consumption, were ‘Healthy’. This result is not intended, but still insightful. A potential explanation for this result is that healthier individuals are more likely to respond to the question asking how many fruits and vegetables they consume. A healthier respondent may be more likely to keep track of their consumption and would have a response to the question. |

1. **[10 marks] Describe the most salient threats to validity that remain in your analysis (300 words maximum)**

|  |
| --- |
| 1. Construct validity is most salient with regards to classification of wealth group. Above and below 50k income is a clear division that does not capture the subtle factors of differences in wealth such as owned assets. Despite being a somewhat simplistic representation of wealth, income group (above or below 50k) was the most important feature for an accurate classifier. Threats to the validity of this important feature are significant to the whole analysis. As this definition of income group doesn’t capture the intricacies of wealth, the relation between income group and health demonstrated by this analysis may not accurately represent the true importance of specific factors included in the analysis such as access to Medicare or being able to afford medication.  2. Internal validity is possibly lacking as the built model is unable to use features to predict physical activity and none of the features are significantly relevant according to the analysis. Possible reasons for this result include an incorrectly chosen model or conflicting selected features due to selection bias, a potential bias that threatens internal validity [9]. A more deterministic selection of features or a thorough comparison of different models’ performance could reduce this bias and ensure internal validity with an accurate, telling model. For example, multiple models could be trained on the data to compare and select the most fitting.  3. The data contained a disproportionate amount of healthy participants providing fruit consumption data vs the amount of unhealthy participants (8679 healthy vs 1752), so much so that predicting ‘healthy’ for every participant was enough for the model to perform accurately. This threatens external validity as the results are idiosyncratic to this skewed data set and would not accurately perform on a more meaningful sample with an equal amount of healthy and unhealthy participants. |

# 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

1. **[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)**

|  |
| --- |
| 1. Schema:  CREATE TABLE ‘BRFSS’.dbo.‘IncomeHealth’ (  ‘participantId’ int NOT NULL AUTO\_INCREMENT,  ‘\_INCOMG’ varchar(50),  ‘NOCOV121’ varchar(50),  ‘MEDICARE’ varchar(50),  ‘MEDCOST’ varchar(50),  ‘MEDBILL1’ varchar(50),  ‘HLTHPLN1’ varchar(50),  ‘\_RFHLTH’ varchar(50),  PRIMARY KEY (‘participantId’));  2. Schema:  CREATE TABLE ‘BRFSS’.dbo.’Habits’ (  ‘participantId’ INT NOT NULL AUTO\_INCREMENT,  ‘\_\_PACAT1’ varchar(50),  ‘\_RFBING5’ varchar(50),  ‘\_RFBMI5’ varchar(50),  ‘\_RFCHOL’ varchar(50),  ‘\_RFHYPE5’ varchar(50),  ‘\_RFSMOK3’ varchar(50),  PRIMARY KEY (‘participantId’));  3. Schema:  CREATE TABLE ‘‘BRFSS’.dbo.’FruitsVeggies’ (  ‘participantId’ INT NOT NULL AUTO\_INCREMENT,  ‘\_RFHLTH’ varchar(50),  ‘\_FRUTSUM’ varchar(50),  ‘\_VEGESUM’ varchar(50),  PRIMARY KEY (‘participantId’));  A table to contain csv data, “dbo.BRFSS”, can be created using SQL Server Management Studio or similar tool. The csv file can be set as a source with schema auto generated from csv fields.  Data from dbo.BRFSS can be inserted into each of the question tables through a command such as:  INSERT INTO dbo.’FruitsVeggies’  SELECT ‘\_RFHLTH’, ‘\_FRUTSUM’, ‘\_VEGESUM’ FROM dbo.BRFSS  WHERE ‘\_RFHLTH’ != ‘Don't know/Not Sure/Refused/Missing’  Data can be further processed by dropping values:  DELETE FROM dbo.’FruitsVeggies’  WHERE ‘\_FRUTSUM’ IS NULL or ‘\_VEGESUM’ IS NULL  After installing WEKA 3.8.0 and the appropriate driver, the SQL database can be accessed in WEKA using the explorer and connecting to the hosted path/url.  Data can be retrieved with the command:  SELECT \* FROM ‘BRFSS’.dbo.’FruitsVeggies’  Data can be transformed in SQL with a command such as the following:  UPDATE PARTICIPANTS SET \_RFHLTH = REPLACE(\_RFHLTH, ‘Good or Better Health’, 1)  Which finds and replaces a string with a binary 0 or 1 that can then be typecast to an int using:  SELECT CAST(\_RFHLTH AS INT) FROM PARTICIPANTS  or typecast from a string of a number to a continuous float using:  SELECT CAST(\_FRUTSUM AS FLOAT) FROM PARTICIPANTS  Alternatively to preprocessing with SQL queries, “various data preprocessing tools, called “filters,”can be applied” [16] through WEKA using Java Database Connectivity. Numeric data can be made discrete, which is appropriate for training a classifier. |

1. **[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)**

|  |
| --- |
| This scenario requires rapid response to new data, in other words real-time data analysis, and is thus not an ideal use case for Hadoop as stated in the module lesson, “When to use Hadoop”. From my understanding of AWS cloud services as a Solutions Architect, I would use Kinesis [11] to set up real-time data streams to connect with Apache clusters for data analytics set up with EMR [10]. Results produced by the clusters can be forwarded to a serverless Lambda function [12] that would trigger an SNS [13] subpub messaging system to send out the appropriate message.  Hadoop can run large jobs using MapReduce architecture, splitting data into partitions run on a large number of machines, then merging the result [14]. A drawback of Hadoop is that jobs are run in batches, making reactions to influxes of new data slower than other distributed computing frameworks such as Apache Spark. “Spark copies the data into RAM (processing in-memory). This type of processing reduces the time needed to interact with physical servers and this makes Spark faster than Hadoop MapReduce” [14]. Spark is designed for fast queries on real-time data while handling large data jobs, well suited for handling the large quantity of already held and arriving data.  In the study by Aziz et all [14], a real-time big data analytics task, processing twitter data, is executed with Hadoop MapReduce and Apache Spark to compare performance. The runtime for processing was viewed through the Apache Spark web interface and through a Scala program for batches. It was found that Spark outperforms Hadoop because “Spark utilizes memory-based storage for RDDs and Hadoop MapReduce processes disk-based operations,” [14]. As Spark performs well on this example test case that fits the question’s scenario, using Apache Spark with cloud services is a good approach. |

# 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

1. **[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)**

|  |
| --- |
| “The GDPR provides a non-exhaustive list of identifiers, including: ...location data” [15]. Some participants included in the data set provide location information such as city and zip code. In combination with some specific health factors, this could be used to identify individuals. Thus it is vital that the program doesn’t show data to operators or users, or that only entries with location information filtered out are shown.  Participants have have already submitted responses to the brfss and should not be identified with their data. The organization conducting the survey should not give out any contact information to creators of the public-facing application. Therefore, the rights for participants to be informed, access their personal data, rectify the data, erase data, request restriction or suppression of their personal data, or object to applications’ use of data are severely limited for this scenario. To ensure individual privacy rights, brfss would have to be contacted in order to notify their participants if possible, or ensure consent by participants to have their data used for this application. The application must cite brfss data, with contact information or a request form to ask for data to be removed or hidden.  Data minimization practices help prevent complications and malpractice by limiting data to what is adequate, relevant, and necessary [15]. The data preprocessing steps for this project address this privacy issue adequately by removing unnecessary location data or personal data other than what is relevant for analyzing income, exercise, diet, etc. These data characteristics are not clearly defined, though a good bottom line for what features are appropriate is which features could not be an identifier. The purpose of these practices is that no harm is caused to any participants and that they are protected from any ill effects of using their information for a public application. |

# Section 7 — Report references

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

|  |
| --- |
| [1] Kelleher, John D., and Brendan Tierney. *Data science*. MIT Press, 2018.  [2] Bertrand, Marianne, and Sendhil Mullainathan. "Do people mean what they say? Implications for subjective survey data." *American Economic Review* 91.2 (2001): 67-72.  [3] Almuallim, Hussein, and Thomas G. Dietterich. "Learning boolean concepts in the presence of many irrelevant features." *Artificial intelligence* 69.1-2 (1994): 279-305.  [4] Semyonov, Moshe, Noah Lewin-Epstein, and Dina Maskileyson. "Where wealth matters more for health: The wealth–health gradient in 16 countries." *Social Science & Medicine* 81 (2013): 10-17.  [5] Haskell, William L., Steven N. Blair, and James O. Hill. "Physical activity: health outcomes and importance for public health policy." *Preventive medicine* 49.4 (2009): 280-282.  [6] Ohrnberger, Julius, Eleonora Fichera, and Matt Sutton. "The relationship between physical and mental health: A mediation analysis." *Social Science & Medicine* 195 (2017): 42-49.  [7] Rish, Irina. "An empirical study of the naive Bayes classifier." *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. No. 22. 2001.  [8] Bekena, Sisay Menji. "Using decision tree classifier to predict income levels." (2017).  [9] Witten, Ian H., et al. *Data Mining : Practical Machine Learning Tools and Techniques*, Elsevier Science & Technology, 2016. *ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/york-ebooks/detail.action?docID=4708912. Created from york-ebooks on 2021-06-07 05:28:34.  [8] Levander, Orville A. "Fruit and vegetable contributions to dietary mineral intake in human health and disease." *HortScience* 25.12 (1990): 1486-1488.  [9] Flannelly, Kevin J., Laura T. Flannelly, and Katherine RB Jankowski. "Threats to the internal validity of experimental and quasi-experimental research in healthcare." *Journal of health care chaplaincy* 24.3 (2018): 107-130.  [10] Gilmour, J. B., et al. “What is Amazon EMR?” *Amazon*, Amazon, 1986, docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html.  [11] Bender, Douglas. “What Is Amazon Kinesis Data Streams?” *Amazon*, Crabtree Publishing Company, 2022, docs.aws.amazon.com/streams/latest/dev/introduction.html.  [12] Hendrix, Roger W. “Lambda.” *Amazon*, Laboratory, 1983, docs.aws.amazon.com/lambda/?id=docs\_gateway.  [13] Treichler, Regula, and Christian Hardmeier. “What is Amazon SNS?” *Amazon*, SAB Verl., 2005, docs.aws.amazon.com/sns/latest/dg/welcome.html.  [14] Aziz, Khadija, Dounia Zaidouni, and Mostafa Bellafkih. "Real-time data analysis using Spark and Hadoop." *2018 4th international conference on optimization and applications (ICOA)*. IEEE, 2018.  [15] “Guide to the General Data Protection Regulation (GDPR).” 2 Aug. 2018.  [16] Frank, Eibe, et al. "Weka-a machine learning workbench for data mining." *Data mining and knowledge discovery handbook*. Springer, Boston, MA, 2009. 1269-1277. |