Data Mining and Decision Systems
Assigned Coursework Report

Name: Chinedu Kristian Usifoh

Student ID: 201912549

Date: 3rd November 2020

1 Methodology

According to (Sarkar et al.), CRISP-DM, which stands for Cross-Industry Standard Process for data mining, provides a guideline to data mining. The framework consists of six-phase with arrows indicating critical supported phases. The sequence can move back and forth between phases as necessary. The first phase of the CRISP-DM is the business phase which provides the flexibility to understand the business problem, goals and success criteria. Data understanding which is the next phase of the CRISP-DM lifecycle involves understanding the provided data. This phase involves exploring analysis to avoid unexcepted problems in the next phase- data preparation. The data preparation phase is one of the most critical stages and usually the longest. It is estimated to take 50-70% of project time and effort, which involves data cleaning, formatting, removing noisy data(non-informative data) and encoding data. Modelling phase involves conducting multiple iterations with several models using specific parameters to optimize the model performance. A model with low precision and recall will usually lead to returning to the data preparation stage. See section 1.2(Evaluation and hyperparameters) of the jupyter notebook for Precision and Recall explanation. Successful completion of the modelling phase will lead to the evaluation phase, which analysis the business success criteria established at the business understanding phase. Lastly, the deployment phase is the formal integration of the model that are read into a data warehouse. Which usually leads to a change in decision and policy in an organization.

2 CRISP-DM Lifecycle

2.1 Business Understanding

The main objective aims to develop a machine learning project with machine learning pipelines using the CRISP-DM methodology to ensure planning, guidelines, and documentation to meet project success criteria. The primary focus is to provide an algorithm that predicts if a patient is at Risk or not at Risk depending on specific medical conditions. In (section 3.1(o)) in the notebook shows a small proportion of positive cases which poses an unbalance nature of the data. The imbalance between precision and recall in the dataset is a challenge in developing a prediction model. (See notebook section 1.2 Evaluation and Hyperparameters) for precision and recall explanation. This can make prediction difficult for machine learning models because of the fewer positive outcome(patient is at Risk) to learn from. Due to these challenges, a high recall can be obtained at the expenses of precision and vice versa. This project's success criteria will be measured by an algorithm that can predict positive and negative outcomes with as few as possible "False positive" and "False Negative." i.e. precision and recall close to 100%. (See notebook section 1.2 Evaluation and Hyperparameters) for more details about "False positive" and "False Negative." Each model will be evaluated to look at the F1 and accuracy score, which considers precision and recall.

The historical dataset provided by the University of Hull contains 1520 data record and 11 factors. These factors include the following:

Random: is mainly used for sorting a patient's record.

ID: is mainly used for the unique identification of a patient.

Indication: The indication is made up of 4 values in our legacy dataset. This indicates what type of cardiovascular event can trigger hospitalized. The values are a-f which stands for atrial fibrillation(A-F). According to (Odutayo et al., 2016), A-F is an irregular heartbeat that can form blood clots and lead to stroke or heart failure.

According to (American Heart Association, 2017) cardiovascular accident (CVA) is a general term for conditions affecting the heart or blood vessels.

According to (Simantirakis et al., 2017), ASx which stands for Aspartic acid, is a non-essential human amino acid with a negative charge and plays a major role in the synthesis of other amino acids in the processes of citric acid and urea.

Transient ischaemic attack (also referred to as a TIA or "mini-stroke") is similar, but only temporarily disrupts blood supply to the brain. This can cause speech and visual disturbance, weakness in the face and leg. However, TIA's effect is usually resolved within 24hours. source:(NHS Choices, 2019b)

Atrial-Fibrillation(A-F) is a heart condition which causes irregular and abnormal heart rate. This increases the rate of stroke, heart failure and other heart-related conditions. Atrial-Fibrillation forms blood clots within the upper chambers of the heart. source:(NHS Choices, 2019a), (Mayo Clinic, 2018)

Diabetes is a long-term disorder that causes cholesterol levels to become too high for an individual. The insulin moves sugar from the blood into the cell, which can be used for energy or stored. However, with diabetes, the body does not make enough use of insulin. An untreated diabetic patient can be exposed to damages of the nerves, eyes, kidneys and other vital organs. source:(Watson, 2020)

IHD, which is also called coronary artery disease(CAD), is when the heart is blocked or interrupted from the blood supply by a build-up of fatty substances in the coronary arteries. source:(NHS Choices, 2019b)

Hypertension also knows as high blood pressure, is a health condition where the blood vessels raise above the standard level. Hypertension is a severe health condition which increases the risk of other diseases such as heart, brain and kidney diseases. source:(World Health Organisation, 2020)

Arrhythmia occurs when the electrical impulses that coordinate the heart fail work properly resulting in irregular heartbeats. source:(Mayo Clinic, 2017)

IPSI: cerebral ischemic lesions occur when insufficient blood flow to the brain meets metabolic demand. This leads to limited oxygen supply and can lead to stroke, brain tissue conditions and cerebral infarction. From this definition, we can link a high Ipsi value can be detrimental to a patient with heart conditions. source:(Columbia Neurosurgery, 2018)

Contra: Cerebral ischemia is a condition in which oxygen restriction through the blockage in an artery results in brain tissue damage. Cerebral ischemia can also lead to brain cell damage and brain cell death. Cerebral ischemia can cause a temporary loss of brain

function, which is also linked to a transient ischemic attack(TIA). Also, from observation, Contra and TIA positively negatively correlated

2.2 Data Understanding

2.2.1 Data Retriever

In this phase, the Pandas library was used to retrieve the data from the CSV file. This is the most efficient library in python for handling CSV and JSON flat file. (See notebook section 1.3(a), In 10).

2.2.2 Data Exploration

The dataset was explored to see the size, datatype and shape and graphical representation of the data.

- There are 1520 cells and 11 columns by the statement df.info in the notebook (section1.3(c)).
- Listed by df.dtypes (in section 1.3(f)) of the notebook, it can be seen that the resulting data type of ipsi and Contra does not correspond to the information the data dictionary.
- In the notebook section 1.3(g) it can be seen that 17 null values are found in the dataset.
- The indication attributes have inconsistent values that need to be mapped in the correct order. In notebook section 3.1(k); it can be seen that the indication attribute have five values in contrary to the legacy data dictionary.
- According to the legacy dictionary, the id and random columns aim to sort and identify a patient's record. These independent variable do not have a significant effect on the dependent variable Risk.
- The notebook (section 3.1(j), ii) shows that 90% of diabetes patients are at Risk. However, patients with a medical record are not at Risk.
- In (notebook section 3.1(o)), The dependent variable has inconsistent values that should be mapped into Yes or No.
- In the notebook section(3.1(n), In 40, 41,) shows arrhythmia is highly significant to hypertension. 22% of arrhythmia and hypertensive patients are at Risk.
- In the notebook section (3.1(m), iii), 52% of hypertensive patients are at Risk with no medical record.
- Attributes such as IPSI, Contra, AF, Asx, Diabetes, IHD, Hypertension, Arrhythmia are highly significant to the dependent variable Risk (in notebook section 7.1(a)).

- The notebook section (7.1(c), In 114, In 118), the Contra attribute is uniformly distributed while the IPSI attribute is left-skewed with the presence of outliers.
- In notebook (section 6.1), it can be seen that the use of boxplot to detect outliers in IPSI and statistical explanation of why it should not be dropped in notebook (section 7.1).
- The visualization plot in notebook section 7.1(In 118), shows how attributes are distributed in the data. This helped detect IPSI is left-skewed, and Contra is distributed uniformly.

2.2 Data Preparation

2.2.1 Selecting Data

Based on the previous CRISP-DM phase, relevant attributes were identified for the data mining goal. In this phase, the univariate selection technique was applied to noisy(non-informative) in the dataset and Support Vector Machines(SVM) was used to access the p-values for each univariate feature weight In notebook section 7.1(e). Features with the best nine p-values were selected with an accuracy score of 97%. Considering the entities in this universe are typically interconnected, this appears to be the most viable and practical solution.

2.2.2 Cleaning Data

The cleaning process was used to address the analysis of the univariate feature technique. The history and CVA attributes were dropped due to the low p-value and in heatmap correlation with the dependent variable Risk. See notebook section (7.1(a), ii). There are 17 nulls values, two anonymous data and one empty string among 1520 data entries. This makes about 20/1520 * 100 = 1.3% of the patients with null/inconsistent values. The percentage of missing values is tiny. In statistical language, according to (Madley-Dowd et al., 2019) if the number of missing entries is less than 5% of the sample, it is advisable to drop it because a missing rate of 5% is inconsequential. Also, 5% of missingness has been suggested as a lower threshold below the MI provides benefits. In contrast, the maximum threshold for a large dataset is 5%. Statistical guidance articles have also stated that analysis will likely be more bias if the meaningness is more than 10%. Result with meaningness data above 40% should be considered as a hypothesis. Considering the dataset with just 1.3%

meaningness, dropping all null and inconsistent values in line with the statistical guideline was applied in the notebook sections(4.1(a), In 48,), (4.1(d), In 54), and(5.1, In 61).

Previously, the indication attribute's visualization shows data inconsistency with five values contradicting the indication values in the legacy dictionary. According to (Simantirakis et al., 2017) Asx which stands for Aspartic acid is a non-essential human amino acid that has a negative charge and plays a significant role in the synthesis of other amino acids in the processes of citric acid and urea. Also, with the above definition, the correct syntax is ASx. Converting these values and renaming them for easier understanding and consistency. Pandas.str.lower() function convert these laters from upper case to lower case in notebook section(5.1(b), In 64).

The data dictionary shows that the categorical attributes are nominal. Nominal values consist of discrete attributes with no sense of order within them. The idea here is to translate these attributes into a numerical format that is more reflective and can be easily interpreted by downstream code and pipelines. The independent variables, pandas.get_dummies will be used to transform their values into 0 and 1 that our machine learning can understand where 0 means "No" and 1 signifies "Yes". See notebook section (5.1(c), In 66). The notebook section(5.1(c) In 68) dropped specific columns to avoid dummy trap. According to (www.algosome.com, n.d.) dummy variable trap is the process when one variable can use to predict the other variable. i.e. they are highly correlated. The sum of the dummy variable is equal to their intercept value which means the multi-collinearity is perfect. When there is n number of attributes, n-1 should be considered in our model. the left out values represent the changes from this reference.

Lastly, standardization is an essential part of data processing for our machine learning estimators to implement sklearn efficiently. They might be inaccurate if the individual features do not more or less appear in a standard naturally distrusted data. This means rescaling the features such that they have the properties of a standard normal distribution. Looking at the original dataset, we can see that the numeric attributes are in normalized distribution from a range of 1-100. However, MinMax scaler helps convert the numeric data attributes for our machine learning model understanding by scaling feature between a

giving minimum and maximum value. See notebook section (7.1(f), In 161, 162, 163,) for data transformation.

2.3 Modelling

In this Modelling phase, the machine learning algorithm to make predictions based on the training data provided was developed. Also, data will be split into train and test set. The machine learning algorithm will learn about the train set and predict the test set(unseen data). Lastly, the various matrix will be implemented to improve classification accuracy, such as hyperparameter.

2.3.1 Splitting Data Into Train and Test Set

Data partitioning is an essential part of machine learning. This involves splitting the data into features and target. Train-test split is a technique used in evaluating model performance. When building a machine learning algorithm, it is essential to split the dataset into two subsets. The first subset is used to fit the model, which is also called model training. This allows the model to learn about the data's patterns. This is used in model evaluation by predicting the test dataset. The reason for splitting these data into subset is to prevent the machine learning algorithm from learning useful mapping of inputs to output by providing fewer data. The training and test data are usually split into 70% training set and 30% as the test set. This implies training the model with 70% of the data and testing the model's accuracy, with 30% of the data. See notebook section 8.1(In 166, 167, 168) data partitioning.

2.3.2 Model Selection Technique and Result

Analysis from the data exploration phase, outliers were detected in the IPSI attribute. The notebook section 8.1(b), 8.1(c) and 8.1(c) determine which classifier is robust to outliers and performs better for each model. Cross-validation score helper function evaluates model accuracy by splitting the data into equal pieces, trains on each combination to understand the train set patterns and give back the accuracy score. Logistic regression, decision trees classifier, k neighbours classifier, linear discriminant analysis cause, naive Bayes classifier, SVC and random forest classifier algorithms for the model building were initialized one by one. Then loop through all the model dictionary, run KFold validation, and append the result

to an array to show each classifier's mean and standard deviation. Lastly, a graphical format for model comparison of results is displayed. The result shows that random forest, decision, SVC, and logistic regression have a score above 95%. See notebook section 1.2 for KFold explanation.

2.3.3 Model Building

Based on the model performance in report section 2.3.2, RandomForestClassifier, DecisionTreeClassifier, Support vector classifier SVC(), LogisticRegression. using GridSearchCV to implement Hyperparameter. GridSearchCv helps select the model with the best performance, which is also known as hyperparameter tuning. The performances of a model depend on the set values for the hyperparameters. This function helps to loop through predefined hyperparameters. Also, fit the model on the training set and use the best parameters from the listed hyperparameters to predict the test set.

2.4 Evaluation

As a result of reviewing the initial data mining project, Random Forest and support vector Classifier was the most successful when ranked by recall, precision and accuracy score. See notebook section (1.2) for the explanation. The accuracy score for random forest varies from 0.9822 to 0.9867; SVC is always constant with an accuracy score of 0.9844. On this basics, SVC is selected for implementation because of its consistency in predicting the outcome. This is important for the domain as the domain requires an accurate prediction of a patients life (Risk or Not at Risk). As anticipated, the logistic regression performed the poorest with an accuracy score of 0.9756.

Other challenges during the development phase, data understanding was not explored in details. This is because of inconsistent data that needs data cleaning. However, the CRISP-DM does not support back and forth movement within the business understand and data preparation. An alternative methodology is the SEMMA process with five stages for the process, which is more flexible to data understanding and preprocessing. The SEMMA stands for Sample, Explore, Modify and Assess.

2.5 Deployment

The notebook section 8.1(e, In 180), SVC supervised learning model can be prepared for deployment to the server or system. In doing this, the scaler object used in scaling the numerical features were saved. The will allow easy integration to the server or system to predict new patients record. However, an analyst dedicated to maintaining and updating records will be necessary as the scope of the project expands. Also, for optimal user experience, using more capable hardware and database to store data will be required. Lastly, more appropriate measures to make users feel safe in providing their data should be considered. This will be crucial for further data analysis while dealing with missing values.

3 Results

Model	**Optimized Model Accuracy Score**	**Unoptimized Model Accuracy Score**	**Optimized Model Recall Score**	**Precision Score**	**TP**	**TN**	**FP**	**FN**
Random Forest Classifier	0.9778	0.9822	0.98	0.98	297	145	2	8
SVC	0.9844	0.9689	0.98	0.98	297	146	2	5
Decision Tree Classifier	0.9733	0.9756	0.97	0.97	296	144	3	7
Logistic Regression	0.9667	0.9467	0.96	0.97	295	144	4	7

4 Evaluation & Discussion

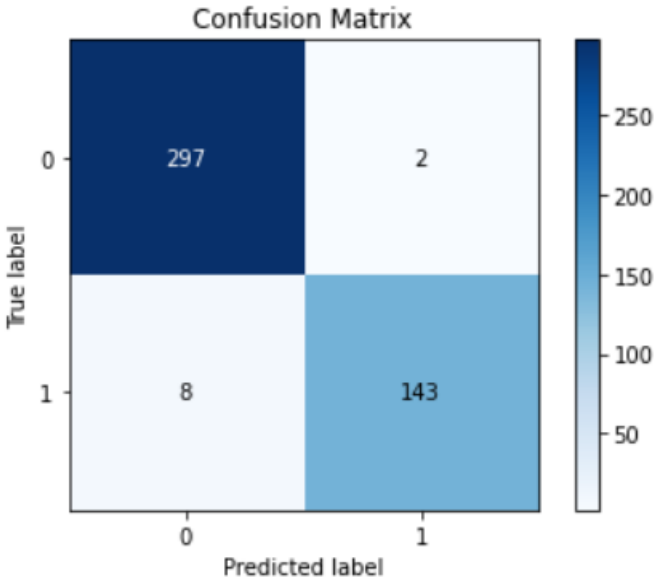
This section critically analyses report section 3. One of the key criteria of this project is for the model to emphasize recall score. Not capturing a patient at risk could result in death. The support vector and random forest classifier performed the best in this project with a recall score of 98%. This means the proportion of patients at risk accurately predicted. However, SVC has a better accuracy score compared to Random Forest. The hyperparameter(optimized model) in the notebook section (1.2, In 2) was essential in optimizing the SVC model predictions. Comparing their confusion matrix, SVC had a better prediction for True positive (At Risk) and True negative (Not at Risk). See report section 5 (Appendix A, 5.1 and .3) confusion matrix. The logistic regression was the least performing model with 0.96% for the optimized model. Finally, the selected model for implementation will be the support vector classifier. The model prediction consistency after multiple testing, better precision, recall and accuracy score was crucial in selecting the model.

Also, the CRISP-DM framework was crucial throughout the end to end development phase. The business understanding and data understanding phase were crucial for understanding the domain and the data. This provided important information about inconsistent data, wrong data types and null values. The data preparation phase supported the data cleaning, transformation and encoding for the CRISP-DM modelling phase. I could do the evaluation and if needed, re-iterate the process. One of CRIPS-DM's draw-back is its inability to move back and forth the data understanding and data preparation phase, which in the dataset given, most attributes needed data cleaning before exploring the data through visualization.

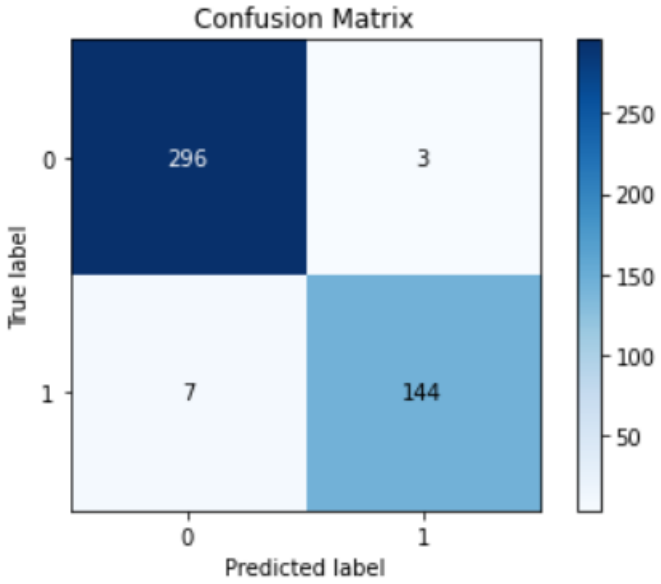
Lastly, see (Appendix A section 5.5) for foundational methodology. The foundational methodology supports the back and forth iteration between data collection and data understand and can iterate back from data preparation to data collection. The methodology would have been a flexible approach in this project compared to CRISP-DM.

5 Appendix A

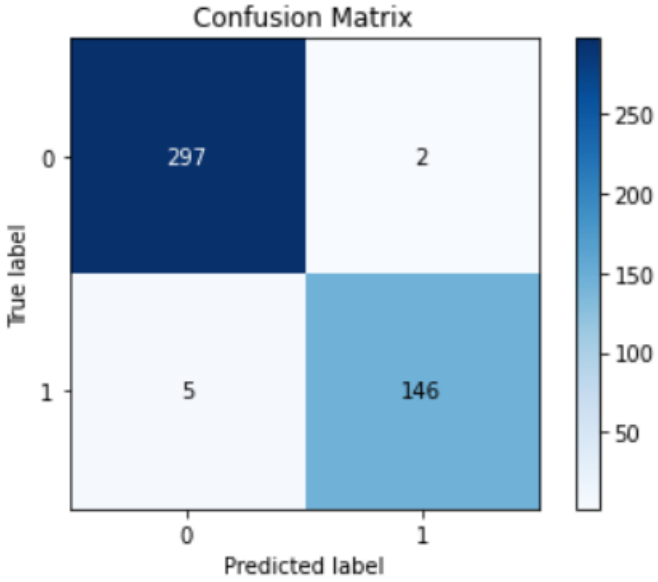
5.1 Random Forest Confusion Matrix



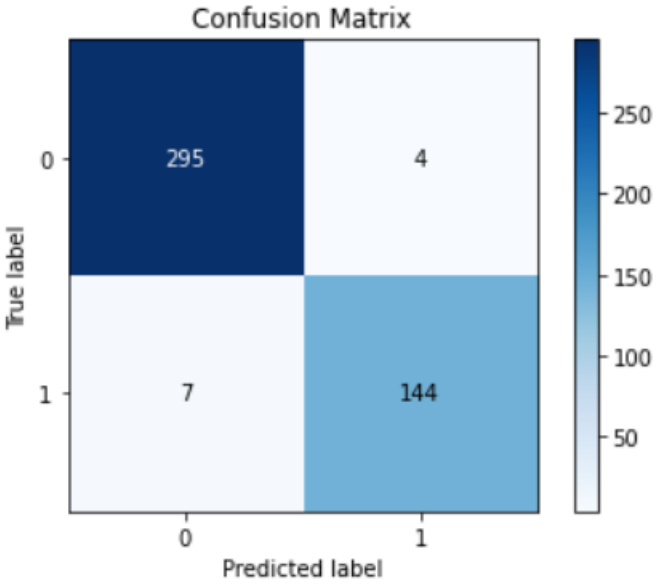
5.2 Decision Tree Confusion Matrix



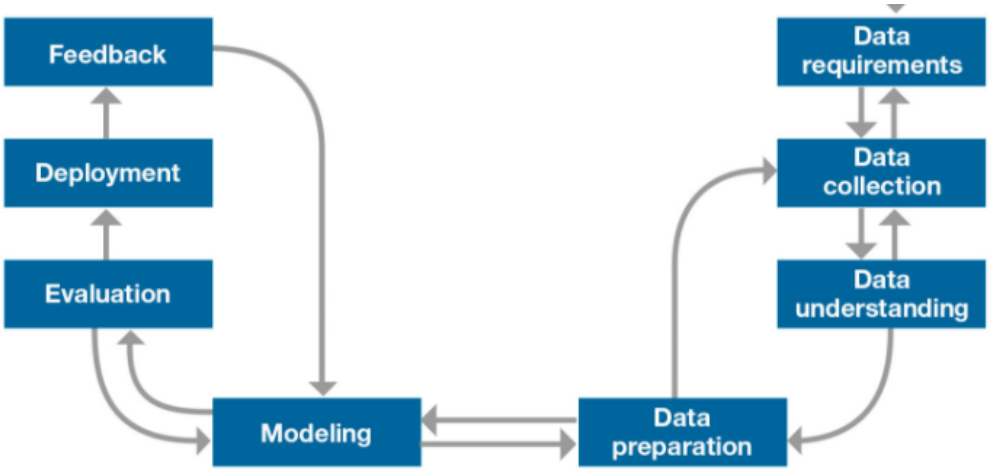
5.3 Support Vector Classifier Confusion Matrix



5.4 Logistic Regression Confusion Matrix



5.5 Foundational Methodology



Source: ("Why We Need a Methodology for Data Science")

6 Reference List

American Heart Association (2017). *What is Cardiovascular Disease?* [online] www.heart.org. Available at: https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease [Accessed 28 Jan. 2021].

Columbia Neurosurgery. (2018). *Cerebral Ischemia - Symptoms, Treatment, Recovery Columbia Neurosurgery*. [online] Available at: https://www.columbianeurosurgery.org/conditions/cerebral-ischemia/ [Accessed 28 Jan. 2021].

IBM Big Data & Analytics Hub. (2015). *Why we need a methodology for data science*. [online] Available at: https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science [Accessed 28 Jan. 2021].

Madley-Dowd, P., Hughes, R., Tilling, K. and Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110, pp.63–73.

Mayo Clinic (2017). *Heart arrhythmia - Symptoms and causes*. [online] Mayo Clinic. Available at: https://www.mayoclinic.org/diseases-conditions/heart-arrhythmia/symptoms-causes/syc-20350668 [Accessed 28 Jan. 2021].

Mayo Clinic (2018). *Atrial fibrillation - Symptoms and causes*. [online] Mayo Clinic. Available at: https://www.mayoclinic.org/diseases-conditions/atrial-fibrillation/symptoms-causes/syc-20350624 [Accessed 28 Jan. 2021].

NHS Choices (2019a). *Overview - Coronary heart disease*. [online] NHS. Available at: https://www.nhs.uk/conditions/coronary-heart-disease/ [Accessed 28 Jan. 2021].

NHS Choices (2019b). *Overview - Transient ischaemic attack (TIA)*. [online] NHS. Available at: https://www.nhs.uk/conditions/transient-ischaemic-attack-tia/ [Accessed 28 Jan. 2021].

Sarkar, D., Bali, R. and Sharma, T. (n.d.). *Practical Machine Learning with Python A Problem-Solver's Guide to Building Real-World Intelligent Systems*. [online] . Available at:

https://library.kre.dp.ua/Books/2-4%20kurs/%D0%9F%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D1%83%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F%20%2B%20%D0%BC%D0%BE%D0%B2%D0%B8%20%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D1%83%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F/Python/practical-machine-learning-python-problem-solvers.pdf [Accessed 28 Jan. 2021].

Simantirakis, E.N., Papakonstantinou, P.E., Chlouverakis, G.I., Kanoupakis, E.M., Mavrakis, H.E., Kallergis, E.M., Arkolaki, E.G. and Vardas, P.E. (2017). Asymptomatic versus symptomatic episodes in patients with paroxysmal atrial fibrillation via long-term monitoring with implantable loop recorders. *International Journal of Cardiology*, [online] 231, pp.125–130. Available at: https://www.internationaljournalofcardiology.com/article/S0167-5273(16)33814-1/abstract.

Watson, S. (2020). *Diabetes: Symptoms, Causes, Treatment, Prevention, and More*. [online] Healthline. Available at: https://www.healthline.com/health/diabetes#:~:text=Diabetes%20mellitus%2C%20commonly%20known%20as [Accessed 28 Jan. 2021].

www.algosome.com. (n.d.). *Dummy Variable Trap in Regression Models*. [online] Available at: https://www.algosome.com/articles/dummy-variable-trap-regression.html [Accessed 28 Jan. 2021].