**I. RESEARCH QUESTIONS:**

|  |
| --- |
| **Question1**: Can we predict whether a person has a mental health issue based on his/her general health, age and BMI?  In England, 1 in 6 people experienced a common mental health problem in any given week [4], however, mental illness often goes unnoticed or are only treated when advanced [5]. It was shown that around 30% of people with a long-term condition have a mental illness [6]. Nearly half of adult age 55+ experienced mental health problems [7] andpeople with obesity had a 55% increased risk of developing depression over time [8].  This prediction would allow improving awareness of mental health problems from early stages and enable personalised treatments*.*  **Question2:** Is there any significant association between per-day consumptions of vegetable and fruit with the risk of having diabetes?  Management of the rising prevalence of diabetes is one of the main challenges facing health-care systems worldwide. In the UK, there are around 4 million people diagnosed with diabetes in 2019 [9].  Some previous study has shown that intake of fruit and vegetable reduced the risk of having diabetes [10], while others reported no significant association between fruit and vegetable consumption with type 2 diabetes [11].  The answer tothis research question would contribute on settling this disagreement.  **Question3**: Can we predict the risk of being overweight/obese based on chronic health conditions (high blood pressure, diabetes), mental health, physical activities and age?  The Health Survey 2019 estimated around two thirds of adults in England was overweight or obese [12]. Factors affecting BMI include activeness and age [12]. Recent studies showed people with mental health problems had a 58% increased risk of obesity [13]. Moreover, obesity is associated with numerous chronic health conditions, including diabetes and hypertension [14].  The prediction of this risk would allow design of plans for obesity reduction. |

# 

**II. DATA EXPLORATION**

|  |
| --- |
| The first step was to research about the nature and purpose of the collected dataset, based on which research questions were described.  The next step consists in checking the types of the data, and then separating nominal and numeric attributes. This is important because the data type of an attribute affects the methods we can use to analyse and understand it, including basic statistics and complex algorithms used to identify relationships between attributes [2, p. 44].  As the result, I figured out the dataset is a report from a surveillance system that monitors modifiable risk behaviours and other factors contributing to the causes of morbidity and mortality of New York state population in 2015 [15]. The dataset was constructed with 12338 observations and 414 attributes, in which 313 attributes were nominal and 101 attributes were numeric.  After that, the prevalence of missing data in the dataset was explored. A list of missing data was built to find out the percentage of missing data for all attributes. Some attributes have a large fraction of missing data, particularly EXACTOT1 - 99%, EXACTOT2 – 98%. There are 26 attributes with more than 50% missing data.  This exploration enabled me to refine my research questions. In particular, I decided to exclude some attributes because using attributes with a large amount of missing data can have a negative effect on the performance of machine learning algorithms and accuracy of research findings [16].  Next, the focus was given to attributes that could be used to answer my proposed research questions. After carrying out descriptive statistics, histograms, bar charts and box plots were created to visualize the distribution of the data, and the possible existence of outliers. These analyses were helpful for understanding and cleaning data, and for building machine learning models [2, p. 44-46]. |

**III. DATA CLEANING**

|  |
| --- |
| First, all the attributes of interest were extracted to new csv files, where each csv file corresponds to a particular research question. As such, I only need to focus on attributes relevant to the research questions, simplifying the data cleaning process and improving performance of machine learning algorithms [1, p.288].  The next step was to clean data from missing values. At the first glance*,* the percentage of missing values for all chosen attributes seemed small (e.g., MENTHLTH 3%, \_AGE\_G 0%, GENHLTH 1%).  However, there was an issue in the dataset where many observations are categorized as “Not asked or Missing”, “Don't know/Not sure”, or “Refused”. For all nominal attributes, as the total percentage of all those categories is small, I decided to categorize them as missing values and remove them from the dataset. This was done taking into account the size of the dataset and the meaning of these values.  For numeric attributes \_FRUTSUM and \_VEGESUM, histograms were built to visualize their distributions. Both distributions were skewed to the right. Missing data was replaced by the median of the attribute because it is more robust to extreme observations [3, p.50]. Next, box plots were made to confirm the prevalence of outliers in the dataset. To reduce the effect of outliers and skewness of data, extreme outliers were dropped, and the rest was kept unchanged.  For DIABETE3 attribute, a simplifying transformation was performed based on the meaning of the values. For example, values "No, pre-diabetes or borderline diabetes", "Yes, but female told only during pregnancy", "Refused", "Don't know/Not sure" were mapped to “No”, given the domain knowledge about diabetes. Next, data was converted from nominal to binary in order to perform regression algorithms.  Grouping was applied to the attributes \_AGE\_G and GENHLTH, reducing their categories to a manageable size. |

**IV. MODEL BUILDING AND EVALUATION**

|  |
| --- |
| **Questions 1&3:** Decision tree J48 algorithm was used first as classifier because the goal is to predict the category of an attribute based on values of others, given that all attributes are nominal [2, p. 136]. J48 is also fast and produces quite accurate output [1, p. 209].  First, J48 was performed on the full training set to build a classifier, which was then evaluated by using 10 folds, 7 folds, 5 folds cross-validation and hold-out approach 66% split and 50% split to separate test data from training dataset [1, p.163 - 169]. The obtained results were then compared using accuracy and confusion matrices. Then, the difference between evaluation measures for all test options, including Weight Average of TP Rate, FP Rate, Precision, Recall and F-measure, was studied.  Next, other classifiers, namely Naïve Bayer and SMO, were applied to the full training dataset, 10 folds cross-validation and 66% split (question1)/ 50%split (question3). Their accuracy and evaluation measures were then compared to those obtained by J48.  **Question 2:** A prediction model was built using Linear Regression because \_FRUITSUM and \_VEGESUM (total fruits/vegetables consumed per day) are numeric attributes [2, p. 114]. Linear Regression was performed on the full training set to build a classifier. The obtained model was then evaluated by using 10 folds, 7 folds, 5 folds cross-validation and 66% split and 50% split test options. Then, I compared the coefficient corelation, Mean Absolute Square of different test options and used Mean-squared Error to evaluate the learnt model [1, p. 194-197].  After that, other regression algorithms, namely, Simple Linear Regression and SMO Regression - with 10 folds cross-validation and 66% split, were applied. The performance of Linear Regression was compared against Simple Linear Regression and SMO Regression according to their corelation coefficient and error scores [1, p. 194-197]. |

**V. RESULTS**

|  |
| --- |
| CC = *corelation coefficient*  LR = *linear regression*  C-V = *cross-validation*  **Question1:**  J48 performance: the same accuracy, TP, FP, Precision, Recall and F-measure were found with full training dataset, 10 folds, 7 folds and 5 folds C-V. J48 worked slightly worse on 50% split than 66% split.  Text  Description automatically generated with low confidence  *Image 1. J48(left) and Naïve Bayes(right) classifiers (10 folds cross-validation)*  When comparing J48 to alternative algorithms, J48 and Naïve Bayes gave the same values of evaluation measures, however accuracy of J48 was slightly better (68.59% vs 68.54%). SMO gave the lowest accuracy (67.56%). Conclusion: J48 gave the best performance with this dataset.  **Question2:**  Linear Regression performance: the same CC (0.074) and errors was found with 10 folds and 5 folds C-V, 7 folds gave slightly smaller CC (0.073). 50% split gave smallest CC (0.058) and highest error rates for all options.  Graphical user interface, text, application  Description automatically generated  *Image2. Linear Regression(left) and SMOreg (right) models (10 folds cross-validation)*  When comparing LR to alternative algorithms, SMOreg and LR gave the same CC, however, SMOreg gave smaller MAE and smaller RAE. Simple LR performed slightly worse than these two.  **Question3:**  Diagram  Description automatically generated  *Image3. J48 visualized tree*  Performance of J48: the best accuracy was found with 50% split (64.21%), then full training set (64.18%) and 5 folds C-V (63.91%). J48 performed slightly worse with 10 folds, 7 folds C-V and 66% split. TP, FP, Precision, Recall and F-measure are not much different between test options.  Therefore, full training dataset, 5 folds C-V, and 50% split were used to compare J48 , Naïve Bayes and SMO.  There is not much difference in performance of three classifiers. J48 gave the best performance with 50% tests options and Naïve Bayes worked better with C-V folds. SMO gave slightly worse performance than J48 and Naïve Bayes for all options.  Table  Description automatically generated with medium confidence  *Image 4. J48(left) and Naïve Bayes(right), 50% split* |

**VI. ANALYZING RESULT**

|  |
| --- |
| **Question 1:**  The model shows mental health disorders are more prevalent in people with poor and fair general health, while people with good general health is less likely to experience mental health problem. Age is another significant factor that affects mental health (as shown by other research). However, according to learnt model, overweight/obese is not found to be a significant feature to identify whether one has mental health. Removing attribute overweight/obese did not significantly impact the accuracy of the classifier, which is 68.6%. The high TP and low TN mean that the model classified most people without mental health correctly but misclassified many people with mental health problems. It might be because of noise in the data, or because of attribute selection.  **Question 2:**  The very low correlation coefficient and the large values of error measures in regression models indicated that there is hardly any association between the total amount of fruit and vegetable per day and the risk of having diabetes.  **Question 3:**  The high TP and low TN mean the model had a good performance on classifying people with overweight/obese but misclassified many without overweight/obese. The results also showed that diabetes, high blood pressure, and age are important attributes to use when classifying people with or without overweight/obese. The model again confirmed that there is no significant relationship between mental health problems and the risk of being overweight/obese. |