Exploring the Relationship Between Screen Time and the Mental Health of Young Adults

Dafydd Broom db16654 | Ben Hermans bh16979 | Sam James sj16765 | Marco Lewis ml16901 | Tim Roderick tr16259 | David Sharp ds16797

***Abstract*—In this report, we aim to see if there is any relationship between screen time and depression in young adults. We make use of a synthetic ALSPAC dataset to see if there is a way to model whether or not a young person suffers from depression from the length of time they spend looking at display screens. We firstly train models based on highly correlated variables to the depression diagnosis of a person (whether or not they have depression). Then we experiment with using data about the screen time of the data subjects to obtain a diagnosis. Finally, a third experiment attempts to use the screen time data to determine if a data subject may suffer from depressive symptoms and if this can be used to determine whether a person has depression. Our results show that there is no immediate correlation between screen time and depression. However, this is a reflection of the time that the data was collected and obtaining new, recent data would give a different outcome.**

Introduction

Technology plays such a central part in the day to day lives of young adults. Ofcom reports that 99% of 16-24 year olds use a mobile phone, 79% stream content online and 93% have a social media profile of some sort [19]. In fact, Ofcom state that only 1% of this age group doesn't use the internet at all. Although older Ofcom reports state similar figures for mobile phone usage, it is clear the way that phones are being used continually changes. The use of screens is seen as vital in all aspects of life: work, leisure, communication.

At the same time, there is a substantial increase in young adults suffering from poor mental health, especially regarding anxiety and depression [24]. Not only is this increase very noticeable but mental health has become one of the most pressing issues for young people amongst their social circles [17].

These statistics are unsettling, but it is imperative for research to take place into what the root causes are of the degradation of mental health and find ways to support medical professionals in identifying and, furthermore, treating those who are affected. It is widely discussed and generally divisive whether the increase of screen time, or even the types of content that young people consume, has an adverse effect on mental health.

We will, therefore, explore the data available on this issue and attempt to generate a way to diagnose depression or anxiety based on information of screen time, provided by the patient. We will explore whether this information plays a more or less important role in the diagnosis than features like sex which are well documented [9]. Thus we take a deeper approach than just statistical analysis of the data available and see whether the associations between variables are strong enough to build a reliable classifier upon.

If successful we would be able to conclude that there is a substantial enough link between screen time and depression and be able to provide medical professionals with a model which allows them to suggest restriction of certain types of screen time to patients in order to improve their mental health.

We will begin by exploring other related work in the area. This will be followed by analysing the dataset we will be using, how it was collected, the ethical implications, and its relevance. With the data collected we will discuss the decisions needed to clean the data such that it is usable for further modelling. After exploring the data and selecting important features, we will push the remaining data through several standard classification models. Finally, we will explore which models are the most relevant for this problem and evaluate their success using a metric that strongly weights against false negatives.

Related Work

There have been numerous papers trying to answer the question of whether screen time is associated with anxiety or depression in young adults. The most relevant of these, [18], analyses the same dataset we will be using, from the Avon Longitudinal Study of Parents And Children (ALSPAC) [16]. This paper multiple imputation to deal with missing data alongside ordinal regression to adjust for confounding variables, concluded that there were some associations between the two. These results were particularly strong for computer usage and the suggestion was that there is a small increase in risk by spending more time using screens which increased further, in the case of anxiety, when the subject spent more time alone.

Research into a population-based study in the US [23] suggests that young people who spent more time using screens had worse psychological well being. In particular, they were more likely to act in a less controlled way; arguing or losing their patience, struggling to make friends and failing to see tasks out to completion. A Chinese study found that high screen time was particularly connected with worse mental health when the subject did not take part in enough physical activity [12], the two of which are strongly linked. These papers have the benefit that they used much more modern data, both collected within

the last decade, but fall down because the data collection was less extensive than the ALSPAC specification.

Data Collection

The Dataset

The ALSPAC dataset is a multi-purpose collection of data sourced by tracking a large cohort of pregnant mothers due to give birth between 1991 and 1992 in the Bristol and Avon areas. These parents and their children were then monitored until adulthood, collecting data on environmental factors as well as usage of technology collected when participants were around 16 years old and a small number of usage metrics for the family in the first year.

From the original dataset a synthesised dataset was generated using the R package *SynthPop* by Arbon et al. [2] for use by multiple researchers to investigate any links between screen time and anxiety/depression diagnosis. As a result, none of the data corresponds to a specific real-world individual but the distribution of values (including missing data) has been maintained. In total the dataset contains 13,734 entries, of which 4,513 entries were tested for depression at age 18, with 389 of those entries being diagnosed with depression. We cannot draw any conclusions about the discrepancy between the total number of entries and the number that were tested. Although this figure could include many subjects who were deemed healthy enough to not be tested at all, it is not clear whether this would be the case for all of these entries. Rather, many of these missing values could simply be people who had dropped out of the study by that age. Because this knowledge of the data collection is not clear, we cannot infer anything from these missing data values.

Relevance of the dataset

It is important to note that technology in our lives has a rapidly changing scale and impact. It is inarguable that the way we interact with technology in our lives now e.g the pervasiveness of smartphones and the scale and number of social media platforms means that the definition of ”screen time” has drastically changed from 13 years ago when the screen time usage metrics were gathered in ALSPAC, in a study [14] on the hours of screen time per day for under six-year-olds between 1997 and 2014 found that screen time doubled over the period.

Ethics

Mothers were given the chance to opt-in to the ALSPAC project, which involved consenting to the use of the data by approved researchers. Before participants turned 18, data was gathered by parental consent for use of that data by approved researchers under ALSPAC. Consent for the use of this data could be revoked at any time at the participant's will.

As this dataset is fully synthetic, breach of privacy concerns are extremely limited, as the process utilised ensures that none of the data entries correspond to a real ALSPAC participant; ALSPAC themselves have also removed specific participants i.e triplets/quadruplets from the dataset since they do not have sufficient size to provide anonymity to those participants.

Data Exploration

To explore the surface level properties of the data, and create visualisations of data distributions, the data needed to be encoded into a consistent format. This process was not concerned with missing values, instead with translating each measured variable into a usable format. Analysis of the distribution of null values could, therefore, be done from the start.

The few continuous variables in the dataset (height, weight, age , etc.) were left untouched, for the most part, except for converting to a standard numerical type. The vast majority of the data, however, was in the form of ordinals. These were categorical variables where there was a distinct order in the responses such as the amount of time spent on a device, which was measured in increasing time intervals. To deal with all of these ordinal values, which had varying possible inputs, ranges, and formatting, we converted the original variables to a corresponding number (for example 1, 2, 3). The inherent order of the data was maintained, however, they were all in a consistent format where it was clear that 1 was the lowest value up to a maximum value (in this example 3).

It is important to note that the decision to encode the data in this way may not necessarily have been the best way to do it. Namely, if the intention is to use the data as categories (and thus disregard order) then this should work without drawback, however, in the case of a model which incorporates information on the order in the data (such as most linear models), we have made the substantial assumption that the categories are equidistant, which, in practise, may not be entirely correct. However, given the data on hand, there is no clear understanding of what the distance between the variables should be. Thus we felt this assumption is fair, it just needed to be considered with caution.

In this way, we removed all the complicated text strings and numerical ranges and encoded these as numbers which persisted in representing the categories in the correct order. In a few cases, it was not clear, despite the ALSPAC data dictionary listing the variable as ordinal, what the distinct order was and in these cases, we experimented with combining different variables that appeared to have the same ordinality, as well as introducing half-steps (such as 3.5) to represent an alternative to category 3 that did not necessarily imply a higher or lower relationship with the original category. Of course, inherently these half steps were encoded as numbers which did have a higher value than members of the same class, meaning that this encoding primarily works to introduce another category to the data and should no longer be treated as ordered, as an assumption has been made both on the order and the distance between the variables.

The vast majority of the data we were working with was ordinal data which we encoded and then treated as discrete

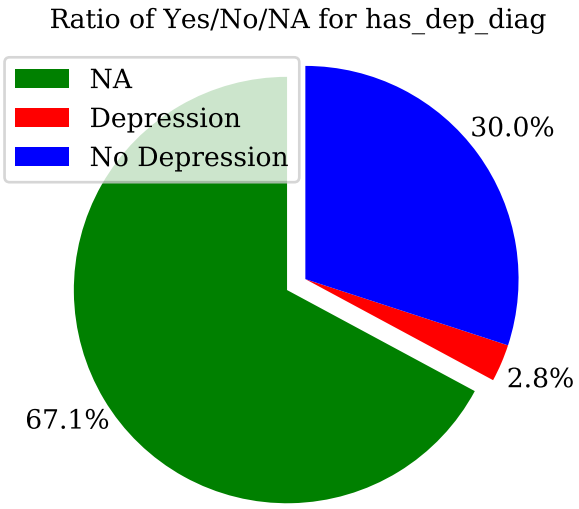

Fig. 1. Pie chart representing the proportion of positive and negative depression diagnoses in the data set. It also includes the proportion of subjects in the study with the relevant data missing.

categories to overcome the need for encoding the data in some machine-friendly way. Due to the resulting wealth of categorical data, there was little need for data cleaning as there was no way to determine if there were anomalous variables when they could only take a limited number of discrete values. On top of this, any necessary data cleaning was performed after data collection by the data handler before we received it.

With the encoding measures in place, it was possible to explore the initial properties of the data. To achieve this, we made it possible to plot the distributions of individual features in a relevant format. This gave us the ability to, for example, show the split of how many people in the study were diagnosed with or without depression, shown in Figure 1. Similarly, we could determine which categories in the ordinal data were the most common and by what proportion.

To compare the associations between two variables we implemented a way to calculate correlation metrics between any two variables in the data. The first metric used was Pearson's correlation which measures the linear correlation between the two where 1 is highly positively correlated and -1 is highly negatively correlated. Another metric we used was the odds ratio test, which evaluates the strength of correlation between the two variables by comparing the ratio of the presence of one without the other and the ratio of those happening in conjunction. This works only for the categorical data in the dataset, or if the ordinal data is treated as such. If the odds ratio is greater than 1 then we say that they are correlated. The final way in which we can explore the initial associations in the data is to simply plot the two against each other and see what patterns emerge. With these methods in place it was easier to test the associations between two variables, albeit naively, in order to aid data exploration, prior to the feature selection process.

In many cases, plotting categorical data against more categorical data provided a less than insightful visualisation of the data. Instead, it was far more beneficial to visualise the distribution of data across multiple classes by using stacked bar charts. With these, we could represent the spread of a variable in relation to another chosen variable. An example of this can be seen in Figure 2, where each bar shows the number of children who had a computer in their room at age 9 alongside whether they were diagnosed with depression later on or not.

The most noticeable property of the dataset, however, was the vast amount of missing data. As seen in Figure 1, only 32.8% of the members of the study received a diagnosis of any kind. As discussed in our data collection, with our knowledge of the dataset there is no inference we could make about the missing data as it could represent several things (dropping out of the study, no concern toward mental health, etc). This meant that there was no way to use the entire dataset for whichever model or technique we wished to apply. Instead, we would have to explore methods to either predict missing values or carefully decide which features are of interest so we didn't have to cut out too many records.

In this investigation about screen time, it is key that the data relating to screen time is recorded. Even considering the data subjects that have a depression diagnosis, not all of them have enough data in other fields to be considered useful. If data subjects are missing data on how much time they spend watching TV or texting, for example, then no deduction can be made about whether the amount of screen time they have is associated with their depression diagnosis. This reduces the

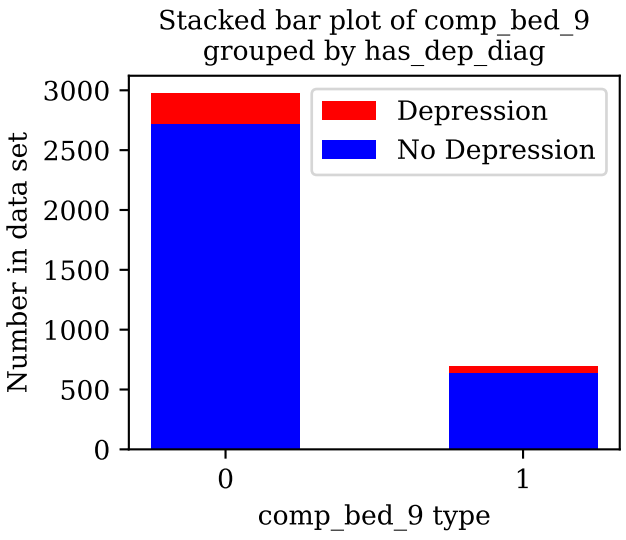

Fig. 2. Visualisation of the data we had on whether the child had a computer in their bedroom at age 9 given that they did or did not get diagnosed with depression later on

amount of useful data subjects that we had access to even more. Out of the 4,513 young adults who had received a test for depression at the culmination of the study, only 2,539 had information about screen time usage provided. Likewise, out of those 4,513 entries only 1,314 were accompanied with complete data for the features we found were most highly correlated with a depression diagnosis, which we determined using a Chi Square Test.

Chi Square Test

The Chi Square Test allowed us to find out which columns in the dataset were highly correlated to the depression diagnosis. This allowed us to get a sense of which columns in the dataset might be useful to use in our models. When the test was performed on each column with respect to the depression diagnosis and a statistical significance of 0.05%, the results yielded 8 columns (excluding the depression diagnosis) that had a high correlation. These resulting columns are what we refer to as the primary dependencies.

These primary dependencies are columns that referred to the medical data of the data subject, such as their panic score or depression score. Since we want to find a good model for predicting depression largely on screen time, then including these scores would give good predictions at the cost of not understanding whether or not screen time is a major influence.

Following on from this, we decided to see which columns were highly correlated to each of the primary dependencies. By performing the Chi Square Test on the rest of the columns and each of the primary dependencies, we found that there were 48 columns in total that the primary dependencies were highly correlated to (with a statistical significance of 0.5%). These columns are referred to as the secondary dependencies and they include a variety of columns from both medical data and screen time.

Mutual Information

Before preparing the data, we decided to investigate the mutual information between the primary and secondary dependencies. The mutual information is a measure of how dependent two columns are on each other. If the score produced is 0, then the columns are independent. Therefore a higher mutual information score will reveal that the columns are highly correlated.

When performed on the depression diagnosis, primary dependencies, and secondary dependencies, it revealed that a lot of the column pairs had a mutual information score close to 0. This meant that the data columns in the set are quite independent of one another. The columns that had a high mutual information score were related to the medical data about the data subject, e.g. primary diagnosis and depression score.

Data Preparation

Before we were able to start with building models for our data and determining whether screen time metrics would be useful at predicting a patient's diagnosis of depression, we needed to prepare the datasets we would be training with.

Missing Value Handling

Since the dataset is fully synthetic with missing values also included in the creation distribution it is difficult to make any inferences about whether data is Missing At Random (MAR) or Missing Not At Random (MNAR). However, any model that required a majority of fields from the dataset would inevitably require some form of handling missing data as no entry had complete data. As such we experimented with a couple of ways of dealing with missing values under the MAR assumption:

Single Feature Replacement

Initially, we simply replaced any missing values in model-relevant fields with the mean value from that field, if the MAR assumption holds then this maintains an unbiased sample estimate of the mean. However, this removes information on the spread of data and how different fields co-vary.

To maintain spread, we then changed our approach to using a discrete distribution from existing data and replaced missing data in this way.

Multiple Imputation

Once we had reduced the number of fields we wished to use in a model to those features that the depression diagnosis had a strong dependency on, we used multiple imputation methods to try and infer missing values not only from the distribution of the field but also how each fields co-varies with every other relevant field. In multiple imputation, single imputation is iteratively chained together in order to get multiple sets of plausible values for missing data [5], which are later pooled together into a single multiple imputation result. Multiple imputation was carried out on the parental and childhood covariates as determined by Khouja et al. [18].

Dropping Missing Data

For the final models, the method of simply dropping any participants with missing data for features of interest was used. This ended up being sufficient for our purposes as we had just over 1000 participants with complete data for both feature subsets, allowing us to best maintain the underlying distribution as it existed in the full dataset. It is possible that the performance of the final models could differ by using imputed data, however it is unlikely this would have changed the results so dramatically that the conclusion would be altered.

Feature Aggregation

To reduce the dimensionality of the data without discarding much information, we aggregated some features together to make a combined feature. This was done with the participant's height and weight measurements which were combined into a Body Mass Index (BMI) feature. This was also down with the maternal anxiety measurements which were measured at seven points after the child reached seven months, the new

feature was positive if any of the seven measurements were positive as described by Khouja et al. [18]. As neither of these new features presented a strong correlation with a depression diagnosis, as determined by the chi squared test, they were not ultimately used in our modelling. However, they could be used in a further, more complex, model which takes into account information like height and weight, which were some of the features will the least missing data.

Screen Time Features

Of the numerous features available in the dataset, 12 are selected that relate to screen time and form a separate dataset of inputs that can be used in models relating screen time and depression. The dataset is made up of 'comp_week', 'comp_wend', 'talk_mob_week', 'talk_mob_wend', 'phone_14_week', 'phone_14_wend', 'talk_phon_week', 'talk_phon_wend', 'tv_week', 'tv_wend', 'text_week' and 'text_wend' (see Appendix for explanations). Within this dataset any row that contains a N/A value is removed.

To check the impact of removing the rows containing N/A, the proportion of different feature values that appear in each dataset is calculated and compared. This provides a metric of similarity between the datasets for each feature, excluding those selected for the smaller dataset, which is between 0 and 1 will higher values signifying a higher similarity. The mean value is 0.653, which appears to be lower than would be wanted. However, the 'sex' and 'birth_number' features, neither of which contain N/A values, have similarity values of 0.913 and 0.998. The mean value is brought down by some features which contain multiple N/A values that get disproportionately removed in the screen time dataset, for example, because they were questions that occurred on the same questionnaire. As a result, it is found that the screen time dataset is representative of the whole dataset based on the 'sex' and 'birth_number' features.

Recursive Feature Elimination

As an alternative to dimensionality reduction, recursive feature elimination, a form of feature selection, was considered as an alternative method of reducing the number of input features. This method involves iteratively modelling the data, each time with different input features, and removing features that have the least impact on the model. The result is a ranking of all the inputs based on their impact on the model with a '1' signifying the optimal features to be selected, for a set number of inputs.

Recursive feature elimination was performed on the dataset of the 12 screen time-related features, with all N/A removed, finding the optimal model with 6 inputs to be 'comp_week', 'comp_wend', 'talk_mob_week', 'talk_phone_wend', 'tv_week', 'tv_wend' (for descriptions see the Appendix). The representation of each category was adjusted to reflect their bounds of time, first with the centre of the bound being chosen, next splitting the category into two features, the upper and the lower bound of each category, and finally one-hot encoding the categories. The first two adjustments resulted in a very similar selection of features with the centre bounds swapping one feature, the bounds method selecting an upper or lower bound of each of the same previously selected categories. The one-hot encode changed its selected features, however, the model produced always predicted no depression. As a result, the original category values were kept.

Despite the higher interpretability of a model using feature selection over dimensionality reduction to reduce the number of inputs, dimensionality reduction was chosen because it caused a smaller loss in information. This method was chosen as a result of the difficulty in developing a model that predicted any true depression diagnosis with fewer inputs.

Multiple Correspondence Analysis (MCA)

Related to principal component analysis (PCA), MCA is a method to reduce the dimensionality of data. Normally, it reduces the number of dimensions for nominal data, which have no ordering. However, MCA can accommodate for ordered data in bins (e.g. "less than 1 hour", "1-2 hours") [6]. This makes it useful for this project since we have a large number of categorical columns within our dataset that are in this bin data format.

MCA is an extension of Correspondence Analysis (CA), which works on two features, to multiple features. Here the process of CA is briefly described. Working on a contingency table/matrix, firstly row and column weights of the matrix are found and the matrix is normalised, denote this normalised matrix M. This normalised matrix is then decomposed using transformation matrices to obtain a decomposed matrix, Σ. One transformation matrix, V, normalises the matrix of column weights, W_c, to the identity matrix and the other, U, normalises the matrix of row weights, W_r. The computation of Σ and the transformation matrices is

$$M = U\Sigma V^*$$
$$U^* W_r U = I = V^* W_c V$$

Using the row weight matrix, row transformation and decomposed matrix, a row factor score matrix can be computed (similarly this can be done for the columns). The factor score of a data subject's row is the transformation of the original data into an orthogonal plane. This reduces the number of features and uses orthogonal factors.

Name	Gender
Alice	F
Bob	M

→

Name	Female	Male
Alice	1	0
Bob	0	1

Fig. 3. Transformation of the gender feature into indicator features

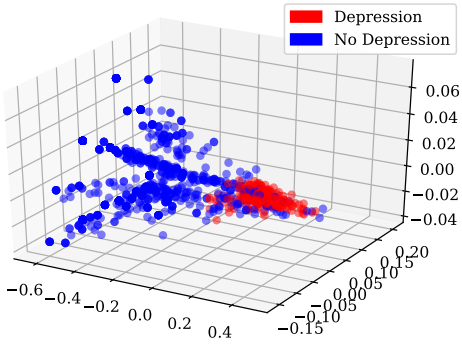

Fig. 4. The 3D output of MCA for the input of primary dependencies

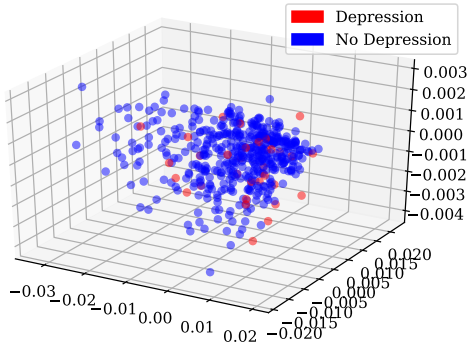

Fig. 5. The 3D output of MCA for the input of secondary dependencies

A full explanation of MCA is presented by Eric Beh [11]. For MCA, the data features are transformed into indicator features, then CA is performed and the factor scores are scaled. An example of a transformation into indicator features is shown in Figure 3.

Usually, this mapping reduces the number of dimensions to 2-3 factor scores. The number of factor scores is dependent on how much variance is wanted to be kept in the data. The distance between subjects' factor scores represents how similar the data of the subjects are. So two subjects that are far away from each other will have very different data.

For us, this will make it very useful to reduce the dimensions of groups of features. For example, the various screen time columns can be reduced to features that allow us to easily tell if someone spends a lot of time looking at screens or not. Also, we can use MCA to reduce the various depression scores/data to a few features to tell if someone is depressed or not.

Compare the graphs in Figures 4 and 5. They show that for primary dependencies there's a strong conical shape whereas for secondary dependencies it is a messy cloud. The first graph shows a good clustering for those who suffer from depression, whereas the second graph doesn't. It should be noted that the second graph is forming some grouping of similar data subjects, however, it isn't necessarily to do with whether or not they have depression.

Modelling

Having now identified the strong relationships in the data set, as well as the features of interest, we would now explore the models we could use to try and build our diagnosis classifier. For a given selection of features we would perform dimensionality reduction using MCA, as we treated all the data as categorical, and then provide this as input to all of the models described below.

We would compare then the ability of a classifier built upon some of the factors linked most closely with a depression diagnosis and a classifier which incorporates primarily information about a patient's tech usage. To compare the performance of different classifiers, all of which have well-documented usage in similar tasks, we would first need to design our evaluation procedure.

Evaluation Design

Medical data, by nature, is often completely imbalanced. That is, normally there are far more people without the condition than there are with the condition, and thus a data set that represents the population would have far more negative diagnoses than positive. Likewise, real-life medical data is also skewed due simply to the fact that many people will have a condition but never receive a formal diagnosis.

The task at hand for data-driven medical diagnosis is that we must endeavour to focus on the special cases to correctly diagnose a condition when the information points towards it despite the apparent rarity of the condition in the first place. It is therefore clear that a conventional notion of accuracy will not suffice here because the diagnosis model could simply classify everyone as without the condition and achieve an accuracy of well above 90%.

The evaluation metrics that we must use then need to include a notion of misclassification cost; weighting model performance to avoid the worst possible outcome of missing a clear diagnosis. Such methods rely on a confusion matrix, shown in Figure 6, which provides the counts of each possible type of prediction. What we are most interested in is reducing the number of false diagnoses but prioritising True Negatives which, in our implementation, represent the depression diagnosis for someone who actually suffers from depression.

Once calculated, there are multiple ways to use this confusion matrix to determine the *accuracy* of our model. The first would be to multiply the matrix by an associated cost matrix which weights the True Negatives far more highly than the rest. The total cost is given by this multiplication and then

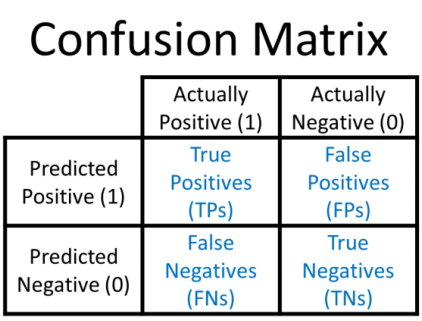

Fig. 6. Confusion Matrix for a classification task [3]

we can optimise over it. This cost matrix would need to be a decision made with the end-user of such a model.

The second way is to plot a Receiver Operating Characteristic (ROC) curve which displays the True Positive Rate (the proportion of the True Positives out of all True Positives and False Negatives) against the False Positive Rate (the proportion of False Positives out of all False Positives and True Negatives). We can use the area under this curve as an evaluation metric with the baseline result being 0.5 and the optimal being as close to 1 as possible.

The third way is to use *recall* which is the proportion of True Positives within the predicted results. How we measure the accuracy of our models is a slight extension on this, F1-score [4], which is derived from both the *recall* and the *precision*, which is the proportion of True Positives out of all positive results. The F1 score is formulated as the following.

$$2 \cdot w \cdot \frac{precision \cdot recall}{precision + recall}$$

Where $w = 100$ is the weighted conversion of the F1-score to be expressed as a percentage.

By using this as our evaluation metric we can understand the relevant performance of the models with which we experiment. In our naive case described above where we simply say everyone has not got depression to achieve a high initial accuracy, we would find a distinctly small F1-score as we have not correctly classified any of the actual depression cases - exactly what we need.

Naive Bayes

The first model which we chose to deploy for our classification is Naive Bayes, which is a probabilistic classifier based on using Bayes Theorem and assuming independence between all features. The classifier decides on the diagnosis of our patient by calculating the probability of a given classification based on all seen data, formulating a simple probability distribution for each feature as it encounters more data. Despite its naivety, there are still common uses for Naive Bayes in a range of classification problems. It is also still experimented within medical classification due to its simplicity, efficiency, and great scalability to large amounts of data [7].

Gaussian Process

Gaussian processes are another Bayesian method used in machine learning and classification. At their heart, Gaussian processes are a Bayesian approach to function approximation and nonlinear regression. We learn a function that approximates the black-box function that describes our classification problem in some space. Based on our training samples, the Laplace approximation is utilised to estimate the posterior distribution over our feature space to classify samples. A benefit to Gaussian processes for regression is that along with producing a function estimation, it also produces the uncertainty of this estimation in the form of covariance. Gaussian Process classification has also shown to be successful in medical diagnosis, especially with the research into Alzheimer's disease [13]. For this implementation, we used the default radial basis function (RBF) kernel and other default parameters.

SVM

Support-vector machines (SVM) are commonly applied to scientific classification problems. The general function of support vector machines is to generate a hyperplane that separates classes in the feature space of our samples. This is done usually by maximising the distance between elements that are closest to each other in potential classification, these samples are what is referred to as support vectors. As our number of considered training samples is relatively small (less than tens of thousands of samples), non-linear support-vector classification was feasible and appropriate for this problem. SVMs have been shown to perform well for cancer diagnosis [22], [25] and thus have recognised the potential for more general medical classification tasks. For this implementation, we also utilised the default RBF kernel and other default parameters.

Decision Tree

Another possible model we could use is a decision tree classifier. A decision tree is built up from observed data, with branches on the tree representing observed values and the leafs representing outcomes. In our case we are particularly interested in using a classification tree as the target variable, whether the subject has depression or anxiety, is discrete (in particular, binary). Decision trees are advantageous because they are simple to understand as they can be visualised. They also require much less data preparation and can handle the mixture of categorical and numerical data we are working with. They are efficient however their efficiency often leads to overly complex decision trees which can lead to overfitting to training data. Classification trees of many different forms have been shown to provide very accurate automated diagnoses in a similar medical domain [10].

Random Forest

One final classification method we could use, which builds directly upon the decision trees previously mentioned, is using Random Forests. Random forests are composed of a vast amount of individual classification trees, trained on various sub-samples of the data set. These all operate together as each tree in the forest will vote for a classification and the class with the most votes will be selected as the output. Random forests are powerful because on average a large number of uncorrelated classifiers operating together will outperform any of the individual members. Because they rely on decision trees many of the advantages and disadvantages of these carry through to random forests, just with improved predictive accuracy and less vulnerability to over-fitting. As with all the methods we have considered, random forests have been shown to perform well on medical classification tasks where the data often resembles the dataset we are working with [8].

EXPERIMENTS

To evaluate the relationship between screen time and mental health we composed three experiments to perform using the models described. Using the results of these experiments we would be able to formulate a justifiable conclusion about this relationship.

To reiterate, the selected features for each experiment would go through MCA dimensionality reduction to take categorical data (which we have disregarded any information of order from) and essentially convert this into a multi-dimensional continuous space. This would be the input for the models described.

Importantly, we choose to perform classification using all of the individual models specified in the previous section with the same training data and evaluation set. The reason for this is that we want to be comparative in our assessment of the efficacy of the models. Doing this provides us with the distinct advantage of greater interpretability of the models. If we chose to develop and configure a vastly complex model, that may or may not be able to classify people with depression with high accuracy, then we are not fulfilling the purpose of our modelling. We do not attempt to generate a high-accuracy classifier for depression diagnosis, but instead, try to demonstrate an interpretable link between depression and other features.

To evaluate the models we would need a train/test split of the complete data we would be using. To do this we utilise a simple 75/25 train/test split of the data.

Experiment 1

The first experiment we would perform is to train a classifier on the primary dependencies that we discovered in our data exploration earlier. We would train a classifier, one of each type of model that we introduced, using the train/test split. We would then evaluate each of these classifiers, using the metrics discussed, on the test portion of the dataset, and compare the results.

Experiment 2

The second experiment we would perform is to train the same set of classifiers on the data we have about screen time. With this new set of classifiers, we can once again evaluate the accuracy they produce on the test set. We can then compare the results between each classifier but also, more conclusively, compare the accuracy scores we can produce using this set of input data with the classifiers that were built upon data we had found to have a high correlation with depression diagnoses in our exploration.

If we find that there is a substantial drop-off in accuracy between the results of these two experiments then we would conclude that information about screen time is not considerable enough to build a model upon. Thus the relationship between screen time and mental health is too weak to be conclusively reasoned upon with the data we possess.

Experiment 3

The third experiment we would perform is to train the classifiers on a new binary column of depression symptoms. The depression symptoms are defined by the summation of the depression score and depression thoughts columns. These are used as its the closest measure of symptoms in the set while being generated in similar ways. A simple threshold for the summation leads to either a no depression diagnosis corresponding to 0 and has depression symptoms to 1. The results from this can then be compared to the previous experiments as a way to see if there is an increase in accuracy depending on screen time or the primary dependencies.

If the results from this show an increase in accuracy compared to the previous experiments then a conclusion that a correlation between the topics is stronger for symptoms rather than a full diagnosis can be drawn.

RESULTS

For evaluating our results, we mostly compare the results of experiment 1 and experiment 2 . This is because they are performing the same task, only with different features chosen. In the following subsections, we explore the results of these experiments.

Comparison of results on our evaluation set

First, we look at the results for experiment 1. In this experiment, we view the effectiveness of the models trained on the primary dependencies as found in the Chi Square Test section. We generate a train/test split as described in the Experiments section. We trained each of the models described in the modelling section on our training set and evaluated them on our evaluation set. The results of this evaluation can be found for experiment 1 in Figure 7.

The accuracy found during testing for experiment 1 highlights that regression has taken place as a result of our training.

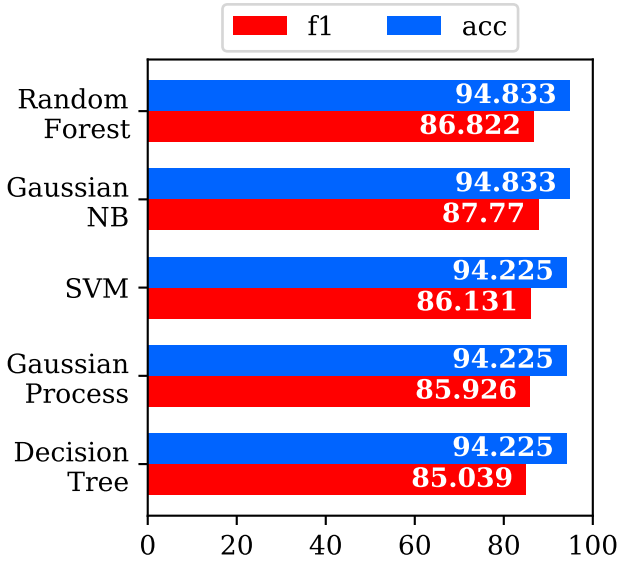

Fig. 7. Experimentation results for models trained on data of the primary dependencies of having a depression diagnosis. "acc" here refers to the test accuracy found during the evaluation. "f1" here refers to the F1-score achieved on our evaluation set. These values were plot for each model described in the modelling section.

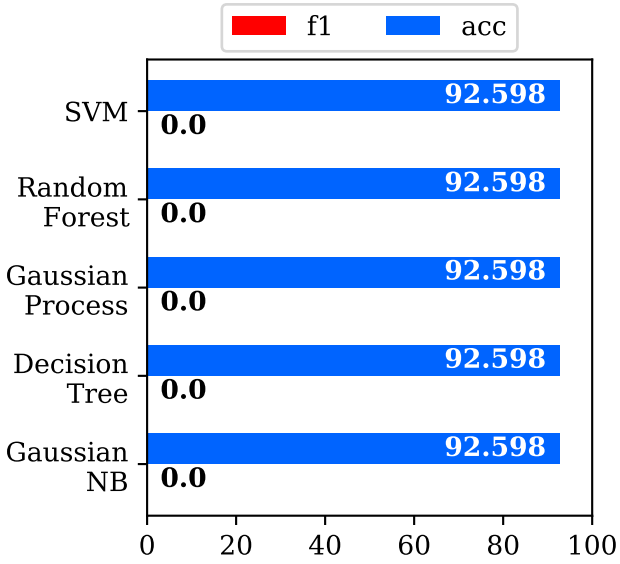

Fig. 9. Experimentation results for models trained on data about screen and technology usage for diagnosing depression. "acc" here refers to the test accuracy found during evaluation. "f1" here refers to the F1-score achieved on our evaluation set. These values were plot for each model described in the modelling section.

In terms of what this implies for our model's effectiveness in depression prediction is less significant. The training accuracy ignores critical information that is considered with other measures such as the F1-score.

For the F1-score, we found that the models produced very similar results, there is a 2% range from the smallest to largest F1-score. All of these models produced a high F1-score, where an F1-score above 80% is often considered good in non-trivial binary classification. We can also observe the ROC curve for these models as found in Figure 8. The efficiency in this graph is measured as the area under the curve. An area of 1 represents a perfect classifier, where an area of 0.5 is no better than random. A rough estimate of ROC curve performance classifications can be found at [15]. In this, we can view ROC areas as the following: (a) 0.90-1.00: excellent accuracy, (b) 0.80-0.90: good accuracy. From the curve, we can clearly see that all the models achieve good to excellent accuracy.

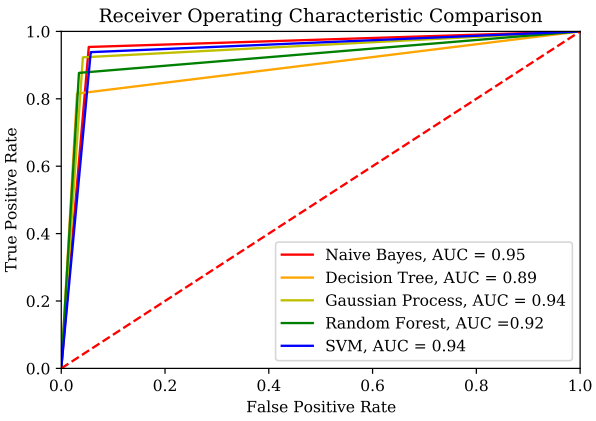

Fig. 8. The ROC curves for each model when trained on primary dependencies. The area under each curve is found to be above 0.8 meaning they can all be described as having good accuracy.

When observing the contents of the primary dependencies with depression however, this may not come as a surprise. The primary dependencies contain the features we would directly associate with someone who may have depression. For example, a subject's primary depression diagnosis and secondary depression diagnosis belong to the primary dependencies. This reduces this model's effectiveness as a depression predictor in a usable sense, but that is inherently not what we are attempting to achieve. What we want to illustrate here is the effectiveness of the models chosen, for depression classification, so that we have a benchmark comparison for experiment 2. Even before modelling, the separability of the features after performing MCA is visible as shown in 4. This when compared to the same figure for *secondary* dependencies in Figure 5 highlights this further.

With this, we can now explore the results of experiment 2 as found in Figure 9. In this experiment, we are observing the effectiveness of using the screen time-based features for the use of determining whether someone may or may not have depression. Immediately it is clear for us to see the high test accuracy found in the results in contrast to the F1-score/ This highlights the issue described before, as the

	Predicted Negative	Predicted Positive
True Negative	588	0
True Positive	47	0

Fig. 10. Here we describe the confusion matrix for one of the models trained for experiment 2 . We only describe one confusion matrix as they are all the same for each model. We can see that there are no positive classifications made by the model. This means that the "predicted positive" column is all 0. As for the F1-score we utilise $precision \cdot recall$, this highlights why it is 0 as if there are no positive classifications the $precision$ must be 0.

test accuracy doesn't capture the objectives we care about for in binary classification. This is further illustrated by the confusion matrix found in Figure 10. This shows that the model exclusively finds that no one has depression in the evaluation set. This is also why the F1-score is 0 as the F1-score is the harmonic mean of $precision$ and $recall$. As precision is the number of true-positive classifications out of the total number of positive classifications, this term is 0. This, in comparison to the primary dependency-trained model, is far less effective at demonstrating a link between the input features a depression diagnosis as it was not able to correctly classify any samples as having depression, unlike experiment 1.

Experiment 3 acts as a continuation of experiment 2 where the screen time features to determine whether someone has no depression or may have depression symptoms. Figure 11 shows the results for each of the models, the accuracy of each model shows a sharp decrease from that seen in experiments

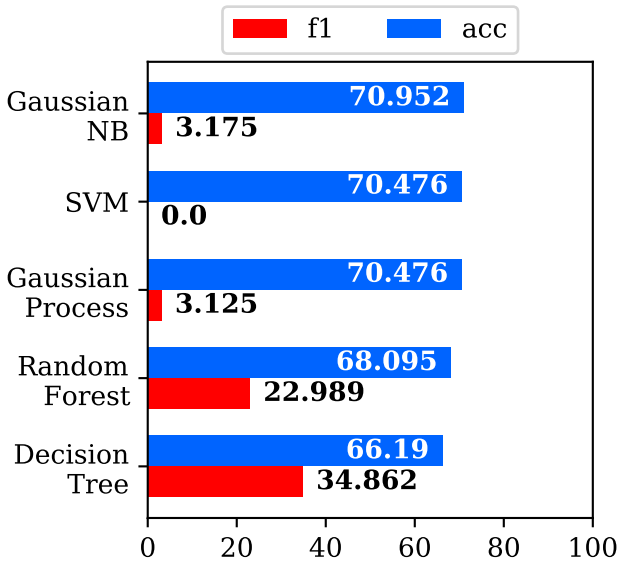

Fig. 11. Experimentation results for models trained on data about screen and technology usage for diagnosing depression symptoms. "acc" here refers to the test accuracy found during evaluation. "f1" here refers to the F1-score achieved on our evaluation set. These values were plot for each model described in the modelling section.

1 and 2. However, the increase in the f1 score is where the experiment differs significantly from experiment 2. The only consistency between the two is that the SVM method again predicts that every subject has no depression leading to the F1-score of 0. Although the models still fail to significantly increase the F1-score to the levels seen in experiment 1. While the decision tree and random forest methods can find some correlations between screen time features and depression symptoms the correlations, however, are still too weak.

Conclusion

The results from our experimentation are clear, with the data available and using simple (yet proven) models, it is not possible to use information of screen time to predict whether a young adult is suffering from depression or not. Whatever statistical associations there may be between the two, we cannot conclude that the relationship is strong enough to aid diagnosis.

We feel, however, that, regardless of our conclusion, the relevancy of the dataset is substantially lacking and thus the ALSPAC dataset should not be used (at least not on its own) to make claims about screen time and mental health in the present. It has been over a decade since the data was collected. This time period has seen not only a huge change in the way we use technology but our understanding and focus on mental health issues. This data was collected almost entirely before the advent of smartphone technology and its subsequent effect on the lives of young adults.

It is suspicious that the increase of technology usage in the UK has coincided with the increase of depression diagnoses, however we are unable to draw a black and white conclusion from the data available. There is no reason to assume a link between the two, let alone to assume a causal relationship between one or the other.

Future Work

There is a lot of scope to perform further investigations on this data to perhaps provide a more accurate or specialised conclusion abut the relationship between screen time and mental health.

Focusing particularly on the methodology we used, it is clear that one major decision we made was to essentially disregard the ordering of the variables as it was not clear how to encode this without also assuming the distance between them. Furthermore, the feature selection methods and models that we deployed would not take into account the order of discrete variables regardless, so we treated the ordinal data as categorical instead. It would be particularly interesting to look further into ways we can incorporate this information of order into the entire process to perhaps increase the predictive capacity of our models without needing more data, rather just using all the information encoded in the data given to us.

It is possible that building a more complex model, possibly using multiple imputation to gain access to more data,

would also provide a better classifier based primarily on the information of screen time that we possess. If no significant performance increase can be achieved by this avenue, however, we would suggest that using a newer, more relevant dataset should be the starting point for further research into the area. If significant quantities of more modern data could not be collected, it could be advantageous to combine this dataset with the selection of data that could. In this case, it would specifically be interesting to incorporate a weighting term to the data as to not abstract the problem from time, which, as we discussed, has an enormous impact on the data of two areas which are constantly changing.

Finally, one possible area of expansion is to use the theory of causality as composed by Judea Pearl [20] to make scientifically accurate claims about a causal influence of screen time on mental health (or indeed vice versa) as opposed to simply masquerading an inference of some association between the two as causality. Techniques of causal discovery [21] would allow us to input the data given and determine the causal structure of the data and conclude if, with the quantity of data available, we can make a substantial claim of a causal impact of screen time on mental health. After all, even if the two are statistically linked (and just not strong enough for us to build a model upon), there is nothing to say that poor mental health causes people to spend more time looking at screens as a form of escape. Similarly, it is possible that the entire relationship between the two is due to some other factor, such as a seemingly invisible cultural shift. It is clear that the vast quantity of content that we can now absorb with modern technology could, in fact, be used to encourage those who are struggling with their mental health to seek help - increasing the number of diagnoses. Therefore, research into this area needs to start with an open mindset, as opposed to to jumping to the conclusion that screen time is unhealthy and working from there.

REFERENCES

[1] Alspac data dictionary. http://www.bristol.ac.uk/alspac/researchers/our-data/. Accessed: 2020-05-10.

[2] MAPS: Mapping the Analytical Paths of a Crowdsourced Data Analysis — OSF Registries.

[3] Measuring performance: The confusion matrix. https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/. Accessed: 2020-05-02.

[4] Wikipedia entry for the f1-score. https://en.wikipedia.org/wiki/F1_score. Accessed: 2020-05-07.

[5] When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. *BMC Medical Research Methodology*, 17(1):162, dec 2017.

[6] Hervé Abdi and Dominique Valentin. Multiple correspondence analysis. *Encyclopedia of Measurement and Statistics*, 01 2007.

[7] KM Al-Aidaroos, AA Bakar, and Z Othman. Medical data classification with naive bayes approach. *Information Technology Journal*, 11(9):1166–1174, 2012.

[8] Md Zahangir Alam, M Saifur Rahman, and M Sohel Rahman. A random forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, 15:100180, 2019.

[9] Paul R Albert. Why is depression more prevalent in women? *Journal of psychiatry & neuroscience: JPN*, 40(4):219, 2015.

[10] Ahmad Taher Azar and Shereen M El-Metwally. Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23(7-8):2387–2403, 2013.

[11] Eric Beh. Simple correspondence analysis: A bibliographic review. *International Statistical Review*, 72, 08 2004.

[12] Hui Cao, Qingwen Qian, Tingting Weng, Changjiang Yuan, Ying Sun, Hui Wang, and Fangbiao Tao. Screen time, physical activity and mental health among urban adolescents in china. *Preventive medicine*, 53(4-5):316–320, 2011.

[13] Edward Challis, Peter Hurley, Laura Serra, Marco Bozzali, Seb Oliver, and Mara Cercignani. Gaussian process classification of alzheimer's disease and mild cognitive impairment from resting-state fmri. *NeuroImage*, 112:232–243, 2015.

[14] Weiwei Chen and Jessica L. Adler. Assessment of Screen Exposure in Young Children, 1997 to 2014. 173(4):391–393, apr 2019.

[15] Tape Thomas G. The area under an auc curve. http://gim.unmc.edu/dxtests/Default.htm. Accessed: 2020-05-10.

[16] Jean Golding, Marcus Pembrey, and Richard Jones. Alspac–the avon longitudinal study of parents and children. i. study methodology. *Paediatric and perinatal epidemiology*, 15(1):74–87, 2001.

[17] JM Horowitz and N Graf. Most us teens see anxiety and depression as a major problem among their peers, 2019.

[18] Jasmine N Khouja, Marcus R Munafò, Kate Tilling, Nicola J Wiles, Carol Joinson, Peter J Etchells, Ann John, Fiona M Hayes, Suzanne H Gage, and Rosie P Cornish. Is screen time associated with anxiety or depression in young people? results from a uk birth cohort. *BMC public health*, 19(1):1–11, 2019.

[19] Ofcom. Adults: Media use and attitudes report 2019. https://www.ofcom.org.uk/__data/assets/pdf_file/0021/149124/adults-media-use-and-attitudes-report.pdf, May 2019.

[20] Judea Pearl. An introduction to causal inference. *The international journal of biostatistics*, 6(2), 2010.

[21] Bernhard Schölkopf. Causality for machine learning, 2019.

[22] Nasser H Sweilam, AA Tharwat, and NK Abdel Moniem. Support vector machine for diagnosis cancer disease: A comparative study. *Egyptian Informatics Journal*, 11(2):81–92, 2010.

[23] Jean M Twenge and W Keith Campbell. Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study. *Preventive medicine reports*, 12:271–283, 2018.

[24] Jean M Twenge, A Bell Cooper, Thomas E Joiner, Mary E Duffy, and Sarah G Binau. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005–2017. *Journal of Abnormal Psychology*, 2019.

[25] Hui Wang and Gang Huang. Application of support vector machine in cancer diagnosis. *Medical oncology*, 28(1):613–618, 2011.

APPENDIX

List of Variables

The following table gives a comprehensive list of all the variables in the dataset alongisde a small explanation for what the variable represents. This information was given to us at the beginning of the project and resembles the ALSPAC Data Dictionary [1], [16]. We have reduced the table down to the relevant columns and have highlighted in light blue the 'screen time' features that we discuss throughout our report and in green the 'primary dependencies' we discover in our feature selection.

Variable Name	Variable Description
agg_score	Aggression score of partnership
alon_week	Average time child spent per day doing things by yourself on a typical weekday
alon_wend	Average time child spent per day doing things by yourself on a typical weekend day
anx_band_07	Child's has any anxiety disorder (DAWBA band prediction)
anx_band_10	Child's has any anxiety disorder (DAWBA band prediction)
anx_band_13	Child's has any anxiety disorder (DAWBA band prediction)
anx_band_15	Child's has any anxiety disorder (DAWBA band prediction)
birth_order	Order in which study child was born within pregnancy.
child_bull	Child has experienced bullying by another person since the age of 12
comp_bed_9	Child has computer in bedroom
comp_games	Child has computer games
comp_house	Computer without internet access is in the study child's house but not in the study child's room
comp_int_bed_16	Computer with internet access is more or less permanently study child's room
comp_noint_bed_16	Computer without internet access is more or less permanently in study child's room
comp_week	Average time child spent per day using a computer on a typical weekday
comp_wend	Average time child spent per day using a computer on a typical weekend day
creat_14	Frequency child does creative activities: art/acting/music/making things
dep_band_07	Child has depression (DAWBA band prediction)
dep_band_10	Child has depression (DAWBA band prediction)
dep_band_13	Child has depression (DAWBA band prediction)
dep_band_15	Child has depression (DAWBA band prediction)
dep_score	Child's depression score on CIS-R.
dep_thoughts	Child's number of depressive thoughts on CIS-R.
draw_week	Average time child spent per day drawing/making/constructing things on a typical weekday
draw_wend	Average time child spent per day drawing/making/constructing things on a typical weekend day
emot_cruel	Partner was emotionally cruel to mother since child was born
exercise	Frequency during the past year study child did exercise
fam_tv_aft	Frequency TV is on in the afternoons
fam_tv_eve	Frequency TV is on in the evenings
fam_tv_mor	Frequency TV is on in the mornings
has_dep_diag	Child has ICD-10 diagnosis of depression
height_16	Child's height (cm)
iq	Child's total IQ score on WISC-III
mat_age	Grouped age of mother at delivery
mat_anx_0m	Mother experienced anxiety or 'nerves' since study child was born
mat_anx_1	Mother experienced anxiety or 'nerves' in past year
mat_anx_18m	Mother experienced anxiety or 'nerves' since study child was 18 months old
mat_anx_8m	Mother experienced anxiety or 'nerves' since study child was 8 months old
mat_dep	Mother's Edinburgh Postnatal Depression Scale (EPDS) score.
mat_edu	Motherâ€™s highest educational qualification during pregnancy
mat_ses	Mother's social class during pregnancy.
musi_13	Child plays a musical instrument
musi_week	Average time child spent per day playing musical instruments on a typical weekday
musi_wend	Average time child spent per day playing musical instruments on a typical weekend day
num_home	Total number of people in study child's household

out_sum_week	Average time child spent per day out of doors in summer on a typical weekday
out_sum_wend	Average time child spent per day out of doors in summer on a typical weekend day
out_win_week	Average time child spent per day out of doors in winter on a typical weekday
out_win_wend	Average time child spent per day out of doors in winter on a typical weekend day
own_mob	Child has used their own mobile phone
panic_score	Child's panic score on CIS-R.
parity	Mother's number of previous pregnancies resulting in either a livebirth or a stillbirth.
pat_edu	Partner's highest educational qualification during pregnancy
pat_pres	Biological father lives with the study child
pat_pres_10	Biological father lives with study child
pat_pres_8	Biological father lives with study child
pat_ses	Partner's social class during pregnancy,
phone_14_week	Average time child spent on a school weekday talking on an ordinary phone
phone_14_wend	Average time child spent on a weekend day talking on an ordinary phone
phys_cruel	Partner was physically cruel to mother since child was born
play_week	Average time child spent per day with other young people on a typical weekday
play_wend	Average time child spent per day with other young people on a typical weekend day
prim_diag	Child's primary diagnosis in accordance with ICD-10 criteria.
read_week	Average time child spent per day reading books for pleasure on a typical weekday
read_wend	Average time child spent per day reading books for pleasure on a typical weekend day
secd_diag	Child's secondary diagnosis in accordance with ICD-10.
sex	Sex
talk_mob_week	Average time child spent per day talking on a mobile phone on a typical weekday
talk_mob_wend	Average time child spent per day talking on a mobile phone on a typical weekend day
talk_phon_week	Average time child spent per day talking on an ordinary phone on a typical weekday
talk_phon_wend	Average time child spent per day talking on an ordinary phone on a typical weekend day
text_week	Average time child spent per day texting on a typical weekday
text_wend	Average time child spent per day texting on a typical weekend day
tran_week	Average time child spent per day in a car/bus/other transport on a typical weekday
tran_wend	Average time child spent per day in a car/bus/other transport on a typical weekend day
tv_bed_16	Study child has a TV set more or less permanently in their room
tv_bed_9	Child has TV in bedroom
tv_week	Average time child spent per day watching TV on a typical weekday
tv_wend	Average time child spent per day watching TV on a typical weekend day
weight_16	Child's weight (kg)
work_week	Average time child spent per day doing school or college homework on a typical weekday
work_wend	Average time child spent per day doing school or college homework on a typical weekend day