Does screen use at age 16 predict depression at age 18?

Muntazim Ali ma16932@bristol.ac.uk
Kai Hulme kh16747@bristol.ac.uk
Kaihuang Huang jv18268@bristol.ac.uk
Zhisen Fu zf17507@bristol.ac.uk
Elvis Chen yc16483@bristol.ac.uk

***Abstract*—The link between mental health and digital technology is contested. We investigate the link between mental health in teenagers and the time spent using digital screens by analysing a synthetic dataset derived from the Avon Longitudinal Study of Parents and Children. An exploratory data analysis reveals large amounts of missing data. We address the missing data by evaluating several types of imputation. Evaluation is performed by comparing the F2 score of an XGBoost classifier trained on the imputed datasets. We find that an XGBoost based imputation method performs best. We proceed to train a random forest classifier and an XGBoost classifier on the best imputed dataset. From the classifiers we extract the feature importances and find that features corresponding to screen use have little to no importance. We further look at the correlation between features associated with screens and depression, finding little to no correlation. We conclude that the link between screen time at age 16 and depression at age 18 is very weak.**

I. Introduction

A. Background

With the prevalence of digital technology increasing, along with an increase in the amount of time we engage with said technology, it is important to understand the impact it may have upon mental health. A 2017 report suggests that children in Britain between the ages of 5 and 15 spend 1.5 hours more per week online rather than watching TV. This is in opposition to their findings from 2007, where children were spending approximately 5 hours more per week watching TV rather than online [1]. The effects of the Covid-19 pandemic have lead to another increase in screen time, with participants from the UK having an average of 7.2 hours of screen time overall per day [2]. Along with this, it has been reported that the incidence of mental health disorders such as anxiety has seen an upward trend [3]. It is therefore reasonable to assess whether there is a link between different types of screen time and this increase in mental health disorders. Studies as recent as May 2021 addressing the topic find that certain types of interactions, such as social media use, have less impact than previously thought [4].

The Avon Longitudinal Study of Parents and Children (ALSPAC) was a study proposed as part of an investigation into the modifiable influences on child health and development [5]. The study collected data on 14,775 children whose estimated delivery date fell between 1 April 1991 and 31 December 1992 inclusive. Data was collected through questionnaires distributed to the mothers of the children, the children themselves, and teachers, with the data collected covering a time period from birth to 18 years of age. A more in depth description of the cohort can be found at [5], [6].

MAPS (mapping the analytical paths of a crowdsourced data analysis) is a project proposed to study the impacts of the different choices researchers may make when analysing a dataset [7]. As part of the MAPS project the same dataset was distributed to a variety of researchers, each of which created their own analysis. The dataset provided was synthesised from a subset of the original ALSPAC data, with the synthesised set replicating the distribution of the original. From these analyses, the goal is to see how researchers' prior beliefs on the topic and choices in analysis affect the results produced. Along with this it is possible to contrast whether the analyses provide different results for the synthetic dataset and the original dataset. To test this, the MAPS researchers will run a multiverse analysis by running each individual analysis on the original ALSPAC dataset and comparing the results. The question asked as part of MAPS was "Based on this dataset: does screen use at age 16 predict depression at age 18?".

B. Project Goal

The goal of this project is to provide our own analysis on the link between screen time and mental health, using the synthesised dataset provided as part of the MAPS project. The original ALSPAC dataset required imputing values because of missing data [8]. Due to the synthetic dataset being based on the ALSPAC data we suppose that imputation will be a major part of our project.

C. Ethics statement

ALSPAC is a study that tracks a vulnerable group of participants over a period of time. As such, there are ethical concerns surrounding the gathering and handling of data from this study. The ALSPAC Ethics and Law Committee (ALEC) was founded to handle ethical decisions regarding research using the data, as well as any follow up studies [9] [10].

Ethical approval was granted by ALEC for the MAPS project to use the data for its study. Because of the ethical concerns with widely sharing this dataset, the MAPS researchers elected to synthesise a dataset for widespread circulation.

II. Methods

A. Exploratory Data Analysis

The description of the synthesis of the dataset is provided in detail at the project page on the Centre for Open Science's Open Science Framework [11].

The data were initially provided as a CSV file, which we investigated using Python's data analysis library *pandas*. There were 83 columns, with one being a flag marked "Synthetic". As this flag only indicates the fact that all the data are synthetic we can safely remove it without affecting our results. This results in a dataframe with 82 columns and 13734 rows. There are a mix of categorical, ordinal and numerical features. For the purposes of using standard statistical techniques we implemented a feature mapping dictionary to convert each feature into a numerical feature, while still retaining the ordinal or categorical nature of features. Binary categorical features are mapped to the values 0 and 1, while ordinal features are mapped to ascending numbers corresponding to the order of the values for the feature. Full details of the feature mapping dictionary can be found in the GitHub repository for this project [12].

We proceeded to explore the correlation between features to gain an insight into which features may be important to our analysis. We produced heatmaps to show the pairwise feature correlations over all features. Full size images for all figures can be found in the project GitHub repository [12].

Figures 1 to 3 show the heatmaps produced using three different correlation coefficients. In all three figures we have that as the gradient tends towards red there is positive correlation, and as the gradient tends towards blue we have negative correlation. No correlation is indicated by grey, and white represents missing correlation values. Numerically the correlation values range between -1 and 1, with 1 corresponding to positive correlation and -1 to the inverse.

Figure 1 illustrates the results of using the Pearson correlation coefficient, r_{xy}, which is given by the following formula [13]:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where:

n is the sample size

x_i, y_i are individual sample points indexed by i

$\bar{x}$ is the sample mean, calculated by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\bar{y}$ is calculated analogously to $\bar{x}$

The Pearson coefficient is used to assess the degree of relationship between two linearly related variables. This poses a limitation as there is no guarantee that our features are linearly related. Because of this we opted to produce the heatmaps seen in figures 2 and 3, using Spearman's coefficient and the Kendall coefficient respectively. Both the Spearman and Kendall coefficients are more robust with regards to outliers [13]. By doing so we aimed to see if there were any major differences produced by the different measures. We also note that although we did not remove any extremal values, if they were present they would have an effect on the result of the Pearson coefficient. This is important to note for the MAPS project. Since our data contains many ordinal features, both

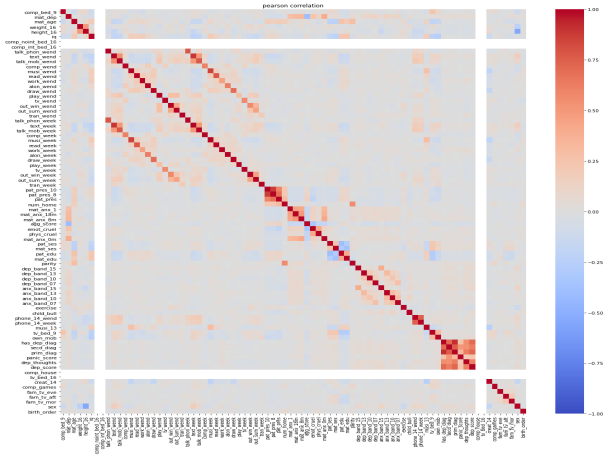

Fig. 1: Pairwise feature correlations using the Pearson correlation coefficient

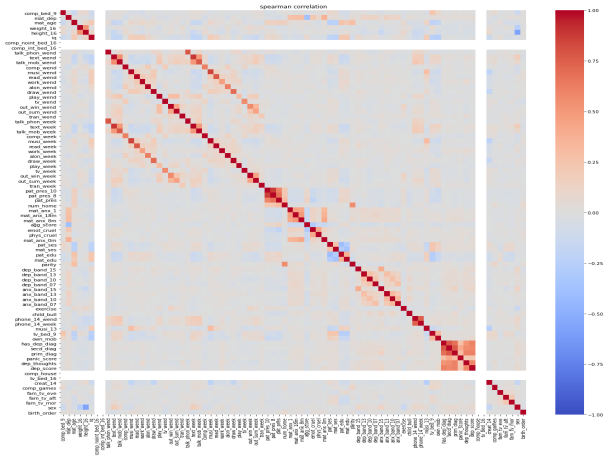

Fig. 2: Pairwise feature correlations using Spearman's rank correlation coefficient

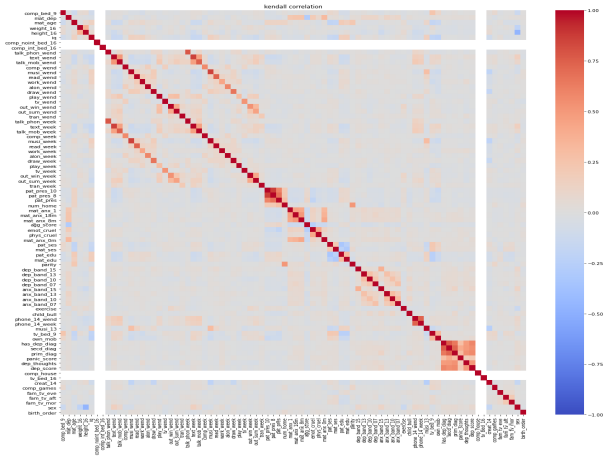

Fig. 3: Pairwise feature correlations using the Kendall rank correlation coefficient

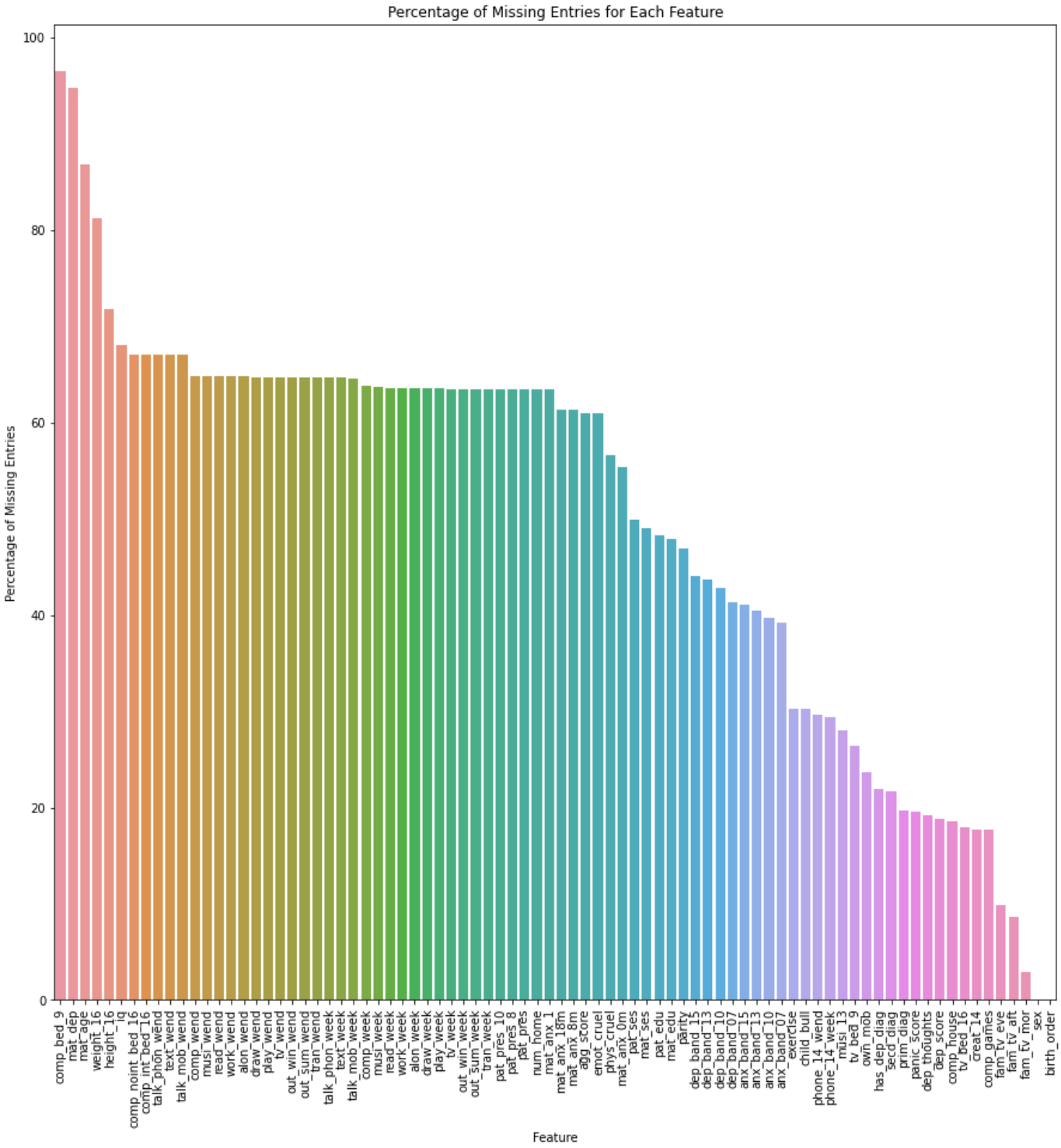

Fig. 4: Percentage of missing values for each feature

Spearman's and the Kendall coefficient are more suited to our analysis [14].

Spearman's rank correlation coefficient, r_s, is given by computing the Pearson correlation coefficient between the rank variables [15]. When we mapped our ordinal and categorical features to numerical representations, we implicitly defined this ranking in our dictionary. If these rank variables are designated as rg_{x_i} and rg_{y_i}, then we compute r_s using the following formula:

$$r_s = \rho_{rg_X rg_Y} (2)$$

Where ρ is the Pearson correlation coefficient.

The Kendall correlation coefficient, τ, is defined as follows [13]:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j) \quad (3)$$

Where sgn is the sign function, defined as follows:

$$sgn(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases} \quad (4)$$

The heatmaps are broadly similar, with some minor differences in clusters of features such as in the bottom right hand corner. There are some mild correlations between features such as whether they have a depression diagnosis and time spent texting, however the reliability of these correlations to make inferences at this stage is not very high, due to large numbers of missing values. Out of a total of 1,126,188 values, 563,352 were not-a-number (NaN) values, approximately 50%. We produce a bar graph to show the percentage of missing values for each feature, seen in figure 4.

There are only two features with no entries missing, sex and birth order. Around 40% of the features have between 60% and 70% of the entries missing, with some features having near 90% of the entries missing. Furthermore, we see there are no complete rows where every feature has a value. Therefore it does not make sense to drop rows containing NaNs.

Further analysis of the missing data produces the heatmap in figure 5, showing the correlation between NaNs for each feature. A high correlation between many features containing NaNs suggests to us that the data is missing not at random. As such, we decided to focus on developing a robust imputation method to replace the missing data.

B. Data Wrangling and Imputation Methods

Before attempting any statistical imputation methods we first see if there are any values we can deduce using a logical substitution. We found that there were four binary features which only took the value "Yes" (mapped to 1 by our feature mapping dictionary) or were a NaN.

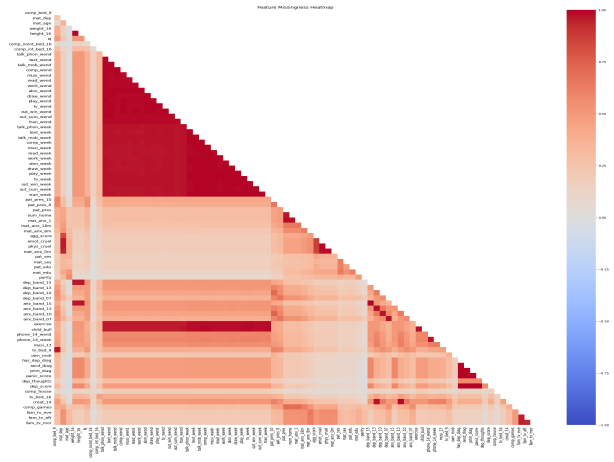

Fig. 5: Heatmap showing correlation between number of missing values

TABLE I: Binary categorical features with only one non-NaN value

Variable	Variable Descriptor
comp_house	Computer without internet access is in the house but not in the study child's room
comp_int_bed_16	Computer with internet access is more or less permanently in study child's room
comp_noint_bed_16	Computer without internet access is more or less permanently in study child's room
tv_bed_16	Study child has a TV set more or less permanently in their room

Looking at the original survey used to collect the ALSPAC dataset, there are questions that only allowed participants to answer "Yes". It seems reasonable then that some would choose to give a blank answer to represent a "No". As the synthesised data reflects the distribution of the ALSPAC data, it follows that we would have features where there are only "Yes" answers. If we were to use other imputation methods with these features as they are, we would end up with an over representation of the value "Yes". Using this reasoning, and given the variable descriptions in table 1, some logical assumptions we can make are:

- If comp_noint_bed_16 is Yes, then comp_int_bed_16 should be No
- If comp_int_bed_16 is Yes, then comp_noint_bed_16 should be No
- If comp_noint_bed_16 is Yes, then comp_house should be No
- If comp_house is Yes, then comp_noint_bed_16 should be No

We also use information from four other non-binary categorical features to further deduce that:

- If comp_week and comp_wend are "Not at all" (mapped to 0) then comp_house should be No
- If tv_week and tv_wend are "Not at all" (mapped to 0) then tv_bed_16 should be No

After imputing using these logical substitutions, we can see how many values were imputed as "No".

TABLE II: Count of each value for binary features after logical imputation

	Count of each value		
Variable	Yes	No	NaN
comp_noint_bed_16	471	2898	10365
comp_int_bed_16	2446	471	10817
comp_house	567	975	12192
tv_bed_16	3648	462	9624

Furthermore, we look at are the depression and anxiety band features. These features are ordinal categorical features, and take values from 0 to 5, which map to bands from the Development and Well-Being Assessment (DAWBA) for child mental health [16]. The exploratory data analysis shows that these features have a high correlation with each other. This is expected as the literature supports the idea that anxiety and depression are closely linked [17]. In the data we find that the feature anx_band_07, which is the DAWBA band for anxiety at age 7, has no values in the category "0". Unlike the binary features from before, there are no obvious ways to logically impute missing values here. However, there is only one category which is missing. Instead of imputing values for this category we combine the DAWBA bands into three categories: "low" (0-1), "medium" (2-3) and "high" (4-5). In this way there are no categories with zero entries.

We proceed to implement a variety of statistical imputations, and evaluate them by looking at performance metrics for a classifier trained on the imputed data.

Univariate imputation methods impute missing values based on calculations using current values for the feature. A simple univariate method would be to fill all missing values with the average for that feature. Numerical features would usually use the mean, however if the feature is highly skewed it may make more sense to use the median value. For categorical features this would use the mode of the feature. Random Hot-Deck imputation is a univariate method, that replaces missing values with a randomly chosen value from the current data [18].

Model-based imputation methods aim to use multiple features to infer missing values. They model the feature as a function of other features, allowing a more accurate imputation compared to univariate methods. Linear and logistic regression are examples of this. For numerical variables we use linear regression, which assumes that the relationship between the target feature we are modelling and the dependent variables is linear. A linear regression model is as follows:

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_t X_t \tag{5}$$

where:
Y is the target variable we are trying to predict
a is the intercept of the line
X_is are the regressor variables we use to predict Y
β_is are regression coefficients

We use Python's *sklearn* built in linear regression method, which itself uses ordinary least squares to estimate the β_is [19]. Ordinary least squares minimises the sum of the squares of the differences between the observations and the predictions.

For modelling categorical features we use multinomial logistic regression for non-binary features, and binomial logistic regression otherwise. A linear predictor function is used for logistic regression as in linear regression. However we now model the logarithm of the probability of a feature being a particular class as the linear predictor function. Using this model, a general equation for the probability of a particular class is [20]:

$$Pr(Y_i = c) = \frac{e^{\beta_c X_i}}{\sum_{k=1}^{K} e^{\beta_k X_i}} \tag{6}$$

Where:
Y_i is the target variable
c is a particular class
β_ks are the regression coefficients from the linear predictor function, as in linear regression
X_is are regressor variables
k indexes the number of possible classes

We use *sklearn*'s built in logistic regression method, which has parameters for choosing whether to use binomial or multinomial regression. For training both our linear and logistic regression models we initialise them using our mean/mode imputed dataset. The predicted feature values then replace the mean/mode imputed values to produce a new dataset.

Stochastic regression imputation follows the same steps as linear and logistic regression, but has the addition of randomness. The sklearn implementation we use normally outputs predictions that lie on the regression line. With stochastic regression, when the model predicts a value, the output is randomly adjusted to reflect the variance of real data. For our stochastic model we constrained the error such that ordinal categorical features have a 10% chance of differing by 1 category from the prediction. Binary categorical features have a 10% chance to swap. Numerical features output a value from a normal distribution centered around the prediction, and values should fall within 10% of the mean either way.

One more type of regression imputation we use on our data is a random forest regressor. Random forests are an example of ensemble learning methods, which are methods using multiple models to reduce overfitting [21]. In the case of random forests we construct multiple decision trees when training. Decision trees represent groups of observations as branches to reach conclusions about a target value represented as leaves of the tree. The output is then the class that is the mode of the classes for categorical features, or the mean for numerical features. Random forests also leverage the use of bootstrap aggregating to increase the performance of the model. Given a training set, bootstrap aggregating selects a random sample with replacement from the set, and fits the trees

to this random sample. In doing so the variance of the model is reduced without increasing the bias [22]. Random forests are particularly well suited for high-dimensional and non-linear data, making it a suitable choice for our dataset. As with linear and logistic regression, we use sklearn's method for random forest regression. We train on the dataset imputed using the mean/mode, and then replace values with the new predictions. To further increase the performance, we take an iterative approach to our random forest imputation. After the initial imputation produces a new dataset, we train another random forest regressor on this new dataset, taking the outputted predictions as our latest data. We repeat this process until there is no change in the predictions.

The final method of imputation we employed was to use gradient boosting. Gradient boosting is another machine learning technique that can be used for both regression and classification. Like random forests it is an ensemble method, using multiple prediction models to increase performance [23]. The idea behind gradient boosting is to find some function to approximate the target variable from the values of the inputs. Formally, we are trying to minimise a loss function of the difference between the estimated and true values [23].

$$\hat{F}(x) = \underset{F}{\operatorname{argmin}}(\mathbb{E}_{x,y}[L(y, F(x))]) \tag{7}$$

Where:
y is our target variable
x represents our regressor variables
L is the loss function to minimise

The form our approximation will take is a weighted sum of functions $h_i(x)$ which come from a class of weak learners [23]. A weak learner is simply defined as a classifier that is slightly correlated with the true classification [24].

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const.} \tag{8}$$

For our data we used the open-source software library XGBoost for Python [25]. XGBoost is a high-performance scalable implementation of gradient boosting, and is able to predict missing values from datasets [26]. Notably, we can impute values without using the mean/mode imputation as a training set. As we see in the results, this leads to better performance for our model.

C. Classification

To evaluate the performance of each imputation method we choose to see how an XGBoost classifier performs when trained on each imputed dataset. We are specifically classifying on the feature has_dep_diag which is a binary categorical feature for whether the subject has medically diagnosed depression. We create a test set and remove the features which are directly related to depression, as they will be correlated with our class labels and retaining them may result in inaccurate classification. A full list of the removed features can be found in the source code [12]. The XGBoost classifier is then trained using the modified imputed datasets. We output five key-metrics for evaluating the performance: accuracy, precision, recall, F1-score and F2-score.

TABLE III

		Actual Value	
		Positive	Negative
Predicted Value	Positive	True Positives	False Positives
	Negative	False Negatives	True Negatives

Table III shows a confusion matrix that illustrates how we can calculate these metrics. For convenience we us the following abbreviations for the entries in the matrix: TP, FP, FN, TN.

$$\textbf{Accuracy} = \frac{\text{TP + TN}}{\text{TP + FP + FN + TN}} \tag{9}$$

$$\textbf{Precision} = \frac{\text{TP}}{\text{TP + FP}} \tag{10}$$

$$\textbf{Recall} = \frac{\text{TP}}{\text{TP + FN}} \tag{11}$$

$$\textbf{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}} \tag{12}$$

$$\textbf{F2-Score} = 5 \cdot \frac{\text{Precision} \cdot \text{Recall}}{(4 \cdot \text{Precision}) + \text{Recall}} \tag{13}$$

Classifiers implemented with random forest and XGBoost methods are able to compute the importance of features for classification. We implement a random forest classifier, though we only use it to classify on the best performing imputed dataset. The metric we use to choose our best performing imputation is defined in the results secion. We then output the feature importances from the classifier for analysis. For robustness, we also compare the feature importances of the XGBoost classifier. To further improve performance we perform hyperparameter optimisation on the XGBoost classifier. To do so we use sklearn's Grid Search method. A grid search is an exhaustive sweep through a manually specified subset of the hyperparameters [27]. As the literature [17] supports a link between other mental health issues and depression, we trained our classifiers on reduced feature sets, one where medical features were removed. To specifically focus on the effects of screen time, we trained the XGBoost classifier on a feature set with only screen time related features.

III. Results

All images and code can be found in the project GitHub [12]. Figure 6 shows the distribution of the two classes for the feature has_dep_diag. Figure 7 is a heatmap visualising the results in table IV. Figures 8 and 9 show the importance of features for the classifiers, with one plot showing a reduced feature set where depression related features are removed and the other a feature set of only screen time features. Figure 10 shows the Spearman correlation coefficients overlayed on a

heatmap for screen related features with has_dep_diag. Tables V and VI are the performance metrics for the XGBoost classifier with the best imputed data, both before and after hyperparameter optimisation.

TABLE IV: Performance metrics for imputation methods to 2 significant figures

Imputation Method	Accuracy	Precision	Recall	F1-Score	F2-Score
No imputation	0.92	0.67	0.027	0.053	0.034
Mean/Mode	0.95	0.88	0.49	0.63	0.54
Hot-Deck	0.89	0.32	0.36	0.34	0.35
Linear/Logistic Reg.	0.96	1.0	0.53	0.70	0.59
Stochastic Reg.	0.89	0.32	0.33	0.32	0.33
Random Forest	0.95	0.94	0.45	0.61	0.50
Iterative RF	0.95	0.97	0.45	0.62	0.51
XGBoost	0.95	0.65	0.78	0.71	0.75

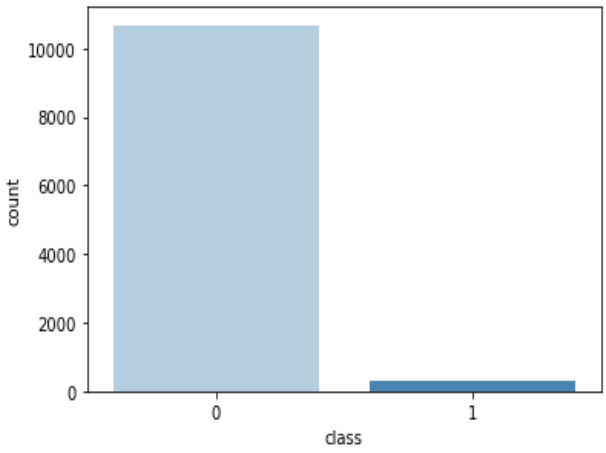

Fig. 6: Distribution of classes for has_dep_diag

TABLE V: Performance metrics (2 s.f.) for XGBoost classifier using different subsets of features

Feature Set	Accuracy	Precision	Recall	F1-Score	F2-Score
All	0.99	0.95	0.96	0.95	0.96
Non-Medical	0.96	0.89	0.56	0.70	0.62
Screen time	0.85	0.30	0.68	0.43	0.55

TABLE VI: Performance metrics (2 s.f.) for XGBoost classifier after hyperparameter optimisation

Feature Set	Accuracy	Precision	Recall	F1-Score	F2-Score
All	0.99	0.97	0.96	0.97	0.96
Non-Medical	0.97	0.96	0.71	0.82	0.75
Screen time	0.91	0.45	0.71	0.55	0.64

IV. Discussion

A. Imputation Performance Analysis

When evaluating which of the imputation methods leads to the best classifier performance we must understand what each metric represents. Accuracy is simply the percentage of correct responses and treats all results as having the same worth. However, for highly imbalanced datasets it can lead to inaccurate conclusions. We can see in figure 6 that our classes are highly imbalanced, so relying only on accuracy would not be ideal. Precision tells us what proportion of the positively predicted class were correctly predicted. This is a good measure if predicting false positives is costly. In our case a false positive is the case that we predict someone as being diagnosed with depression despite this not being the case. If we consider the false negative case, incorrectly diagnosing somebody who should have a depression diagnosis as healthy, then the false negative case has a higher cost. Recall is the metric which captures the number of true positives correctly predicted, and hence is best for when there is a high cost for false negatives. The F1 and F2 scores are metrics which take both the precision and recall into account. The difference is that while F1 score weights both precision and recall equally, the F2 score weights the recall as being twice as important. As mentioned, our dataset is both highly imbalanced and has a high cost associated with false negatives. Therefore F2-Score is the best of our metrics for evaluating performance.

We see in table IV that based upon F2-Score the XGBoost imputation produced the best classification performance, with a score of 0.75. Contrasting this with the classifier trained on the original dataset with no imputation, we see a stark increase in performance. Both Hot-Deck and Stochastic regression imputation have an element of randomness, however we can see that neither performs well. This is likely due to the method of randomness being sub-optimal. Along with this, Hot-Deck imputation is univariate and does not take into account the relationships between features. Another issue with Stochastic regression is that it assumes linearity between features, which is likely not the case. Similarly, linear and logistic regression also makes this assumption, and hence unsurprisingly does not perform particularly well. We do note that the non-stochastic regression performed better the stochastic method. This is again likely due to sub-optimal methods for introducing randomness for the stochastic regression. Random forest based imputation does not perform as well as we expected, and when implemented iteratively we see a slight increase from 0.50 to 0.51 for F2-Score. The performance of the random forest method is possibly due to using the mean/mode imputed dataset for the initial training. Imputation using the mean or mode can lead to over representation of one feature and a decrease in variance. Since we train using this, it is possible that these weaknesses are represented in the model. Supporting this idea is the fact the our XGBoost model performed best, as it does not need to train on the mean/mode imputations. Surprisingly, imputation using the mean and mode performed much better than expected. We suppose that this is possibly a result of the logical imputations made beforehand.

B. Hyperparameter Optimisation

Before hyperparameter optimisation, we see in table IV that the performance on reduced feature sets is reasonable. We note that training on all features results in very high performance.

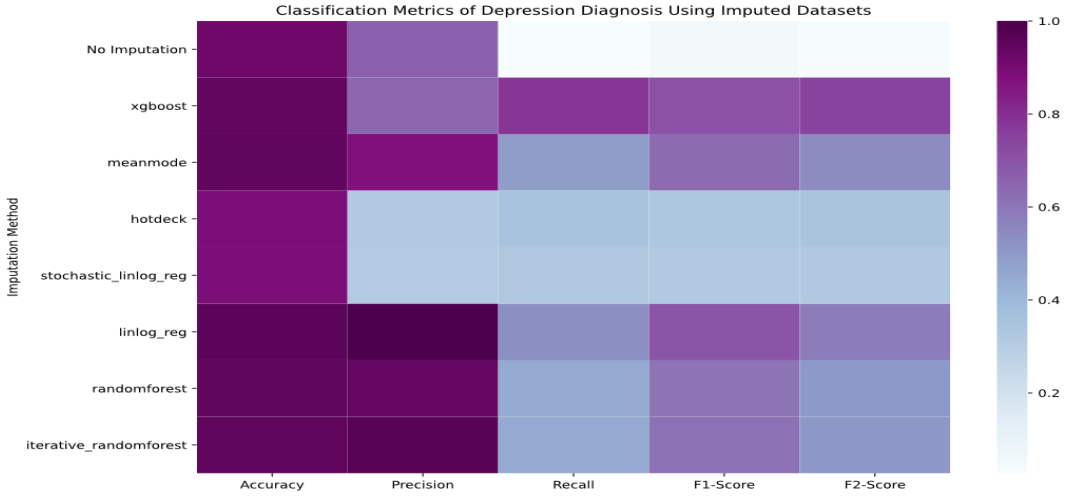

Fig. 7: Heatmap of Performance metrics

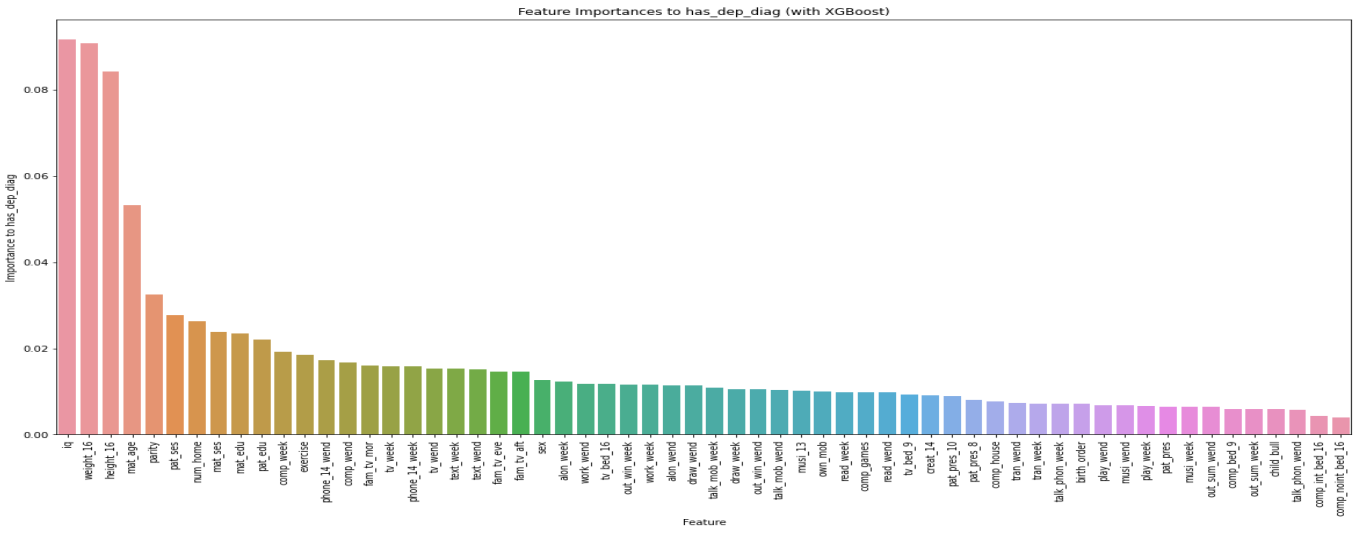

Fig. 8: Bar plot of feature importances on a reduced feature set for random forest classifier

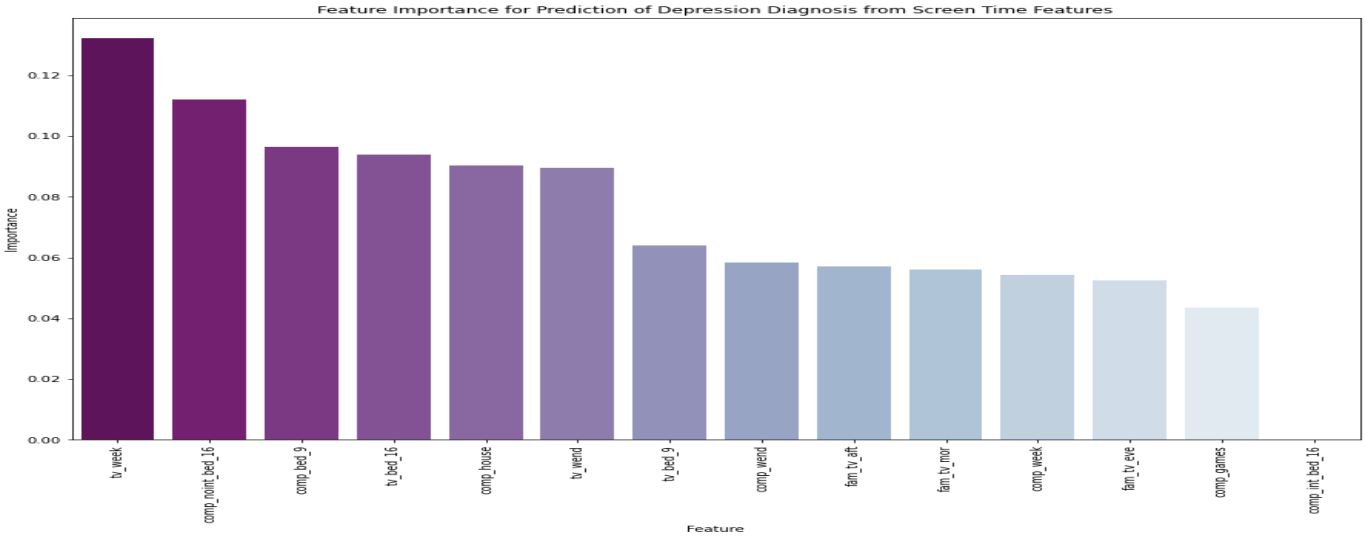

Fig. 9: Bar plot of feature importances for XGBoost classifier, with only screen time related features

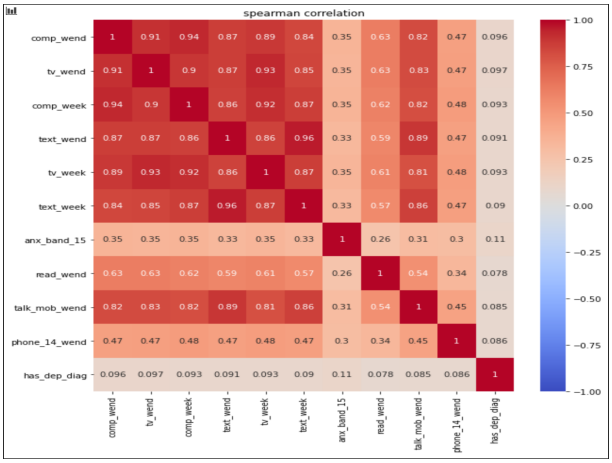

Fig. 10: Spearman correlation coefficients for screen related features and has_dep_diag

As mentioned previously, medical features such as depression score and anxiety score are highly linked with a depression diagnosis. Therefore we focused on the performance when these features are removed. Table V shows us that our grid search optimisation leads to an increase in performance for all metrics corresponding to the reduced feature sets.

C. Feature importance and screen time correlations

Taking the XGBoost imputed dataset, we use this to train a random forest classifier and output the importance of each feature to classifying whether a subject has depression. Figure 8 shows the feature importances after omission of depression related features. The three most important features for this classifier are IQ, and height and weight at 16 years old. We note that out of the screen related features the most important is comp_week, the amount of time spent during weekdays using a computer. However the importance is low, at less than 0.02. Figure 9 shows the feature importances for the XGBoost classifier, but trained on only features pertaining to screen time. Using this classifier observe an increase in all screen related features compared to the previous classifier, however they are still low, with the most important feature being the amount of time spent watching TV during weekdays. The importance of this feature is only about 0.13. This suggests to us that there is a very weak link between screen use at age 16 and depression at age 18. To further support this figure 9 shows that the correlation between the feature has_dep_diag and other screen related features is very weak. For each of the features, the correlation coefficient is never more than 0.1, supporting our previous conclusion. In the literature the link between depression and screen time is contested, with some suggesting a weak or no link [8], while others suggest a strong link [28].

D. Limitations and future work

Given that a large amount of our data were missing, this poses a potential limitation on the reliability of our results. As noted earlier, data seems to be missing not at random and there is potential for our imputation to introduce bias. Though we evaluated multiple methods for imputing missing values to attempt to mitigate this, we acknowledge that there are still possible improvements to make. Future studies could investigate the impact of using other imputation methods, notably multiple imputation. There is evidence that multiple imputation by chained equations is suitable for longitudinal data such as ours [29].

Commonly, studies addressing the relation between mental health and screen time are cross-sectional. This has been criticised before [30]. The ALSPAC data is longitudinal, and as our data is synthesised from the ALSPAC dataset, this is a strength of our study. Nonetheless, it is possible that there were important confounding features not measured by ALSPAC. Along with this is the potential for imperfect measures of the features that were included. As mentioned previously, there were features where it was only possible for participants to answer ”Yes”, introducing a large amount of what we suspect were incorrectly labelled NaN values. Future studies should aim to address this by having a more robust survey for gathering initial data.

One more limitation is that the nature of screen use has changed over time [31]. ALSPAC originally gathered the data for their study before the widespread use of smart phones, tablets and related technology. More modern technology and developments in areas such as social media allow interaction with digital screens in a fundamentally different way. The accessibility and relevance of technology to modern life has meant that high screen time no longer necessarily has to mean sedentary behaviour. There are even studies suggesting links between technology and social capital [32], and others showing a link between video games and physical activity [33]. This suggests the link between mental health and screen time may be more complicated than sometimes suggested. To further investigate this link it is important to factor in the nature of engagement with screens. This is already being addressed by some studies, though further work needs to be done [4].

V. Conclusion

Our aim was to investigate the link between mental health and screen time, using a dataset synthesised from a longitudinal study. Our results suggest that screen time does not have much, if any, impact on the mental health of an adolescent with regards to depression. We do however acknowledge several limitations in our study and suggest further improvements for future studies. As mentioned before, there is contention over the link, and one of the aims of MAPS is to see how different choices made in an analysis lead to different conclusions. We are also interested in seeing whether our model would give the same results for the original ALSPAC dataset. Should our model show similar results this could potentially impact the way medical practitioners look at diagnosing and treating depression, leading to a move away from discouraging technological use for adolescents with mental health issues.

REFERENCES

[1] Ofcom. (2017) Children and parents: media use and attitudes report. in: Chidren's media literacy. [Online]. Available: https://www.ofcom.org.uk/__data/assets/pdf_file/0020/108182/children-parents-media-use-attitudes-2017.pdf [Accessed: 2021-04-15]

[2] L. Smith, L. Jacob, M. Trott, A. Yakkundi, L. Butler, Y. Barnett, N. C. Armstrong, D. McDermott, F. Schuch, J. Meyer, R. López-Bueno, G. F. L. Sánchez, D. Bradley, and M. A. Tully, "The association between screen time and mental health during covid-19: A cross sectional study," *Psychiatry Research*, vol. 292, p. 113333, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165178120315663

[3] A. John, A. Marchant, J. McGregor, J. Tan, H. Hutchings, V. Kovess, S. Choppin, J. Macleod, M. Dennis, and K. Lloyd, "Recent trends in the incidence of anxiety and prescription of anxiolytics and hypnotics in children and young people: An e-cohort study," *Journal of Affective Disorders*, vol. 183, pp. 134–141, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165032715002943

[4] M. Vuorre, A. Orben, and A. K. Przybylski, "There is no evidence that associations between adolescents' digital technology engagement and mental health problems have increased," *Clinical Psychological Science*, vol. 0, no. 0, p. 2167702621994549, 05 2021. [Online]. Available: https://doi.org/10.1177/2167702621994549

[5] A. Boyd, J. Golding, J. Macleod, D. A. Lawlor, A. Fraser, J. Henderson, L. Molloy, A. Ness, S. Ring, and G. Davey Smith, "Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children," *International Journal of Epidemiology*, vol. 42, no. 1, pp. 111–127, 04 2012. [Online]. Available: https://doi.org/10.1093/ije/dys064

[6] A. Fraser, C. Macdonald-Wallis, K. Tilling, A. Boyd, J. Golding, G. Davey Smith, J. Henderson, J. Macleod, L. Molloy, A. Ness, S. Ring, S. M. Nelson, and D. A. Lawlor, "Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort," *International Journal of Epidemiology*, vol. 42, no. 1, pp. 97–110, 04 2012. [Online]. Available: https://doi.org/10.1093/ije/dys066

[7] M. Munafo, N. Timpson, K. Robson Brown, K. Northstone, A. Kwong, N. Thurlby, K. Drax, and R. Arbon. Maps project overview. [Online]. Available: https://jean-golding-institute.github.io/maps/overview/ [Accessed: 2021-04-15]

[8] J. Khouja, M. Munafò, K. Tilling, N. J. Wiles, C. Joinson, P. J. Etchells, A. John, F. M. Hayes, S. H. Gage, and R. P. Cornish, "Is screen time associated with anxiety or depression in young people? results from a uk birth cohort," *BMC Public Health*, vol. 19, 2019. [Online]. Available: https://doi.org/10.1186/s12889-018-6321-9

[9] S. E. Mumford, "Children of the 90s: ethical guidance for a longitudinal study," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 81, no. 2, pp. F146–F151, 1999. [Online]. Available: https://fn.bmj.com/content/81/2/F146

[10] ——, "Children of the 90s ii: challenges for the ethics and law committee," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 81, no. 3, pp. F228–F231, 1999. [Online]. Available: https://fn.bmj.com/content/81/3/F228

[11] R. Arbon. (2019) Synthetic data description. [Online]. Available: https://osf.io/785sx/ [Accessed: 2021-04-15]

[12] M. Ali, K. Hulme, K. Huang, Z. Fu, and E. Chen, "Mental health," https://github.com/ZF-u/Mental-Health, 2021.

[13] C. Croux and C. Dehon, "Influence functions of the spearman and kendall correlation measures," *Tilburg University, Center for Economic Research, Discussion Paper*, vol. 19, 01 2010. [Online]. Available: https://doi.org/10.1007/s10260-010-0142-z

[14] A. Lehman, *JMP for basic univariate and multivariate statistics: a step-by-step guide*. SAS Press, 2005, p. 123.

[15] J. L. Myers and A. Well, *Research design and statistical analysis*. Lawrence Erlbaum Associates, 2003, p. 508.

[16] R. Goodman, T. Ford, H. Richards, R. Gatward, and H. Meltzer, "The Development and Well-Being Assessment: description and initial validation of an integrated assessment of child and adolescent psychopathology," *Journal of child psychology and psychiatry, and allied disciplines*, vol. 41, no. 5, pp. 645–655, 2000. [Online]. Available: https://doi.org/10.1111/j.1469-7610.2000.tb02345.x

[17] E. Frank, M. K. Shear, P. Rucci, J. M. Cyranowski, J. Endicott, A. Fagiolini, P. Grochocinski, V Jand Houck, D. J. Kupfer, J. D. Maser, and G. B. Cassano, "Influence of panic-agoraphobic spectrum symptoms on treatment response in patients with recurrent major depression," *The American journal of psychiatry*, vol. 157, no. 7, pp. 1101–1107, 7 2000. [Online]. Available: https://doi.org/10.1176/appi.ajp.157.7.1101

[18] R. R. Andridge and R. J. A. Little, "A Review of Hot Deck Imputation for Survey Non-response," *International Statistical Review*, vol. 78, no. 1, pp. 40–64, 2010. [Online]. Available: https://doi.org/10.1111/j.1751-5823.2010.00103.x

[19] sklearn.linear_model.LinearRegression. [Online]. Available: https:scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html [Accessed: 2021-05-19]

[20] A. El-Habil, "An application on multinomial logistic regression model," *Pak.j.stat.oper.res.*, vol. 8, 03 2012. [Online]. Available: https://doi.org/10.18187/pjsor.v8i2.234

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[22] ——, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996. [Online]. Available: https://doi.org/10.1007/BF00058655

[23] T. Hastie, J. Friedman, and R. Tisbshirani, *The Elements of statistical learning: data mining, inference, and prediction*. Springer, 2017, p. 337–388.

[24] M. Kearns and L. G. Valiant, "Crytographic limitations on learning boolean formulae and finite automata," in *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, ser. STOC '89. New York, NY, USA: Association for Computing Machinery, 1989, p. 433–444. [Online]. Available: https://doi.org/10.1145/73007.73049

[25] Python XGBoost Documentation. [Online]. Available: https://xgboost.readthedocs.io/en/latest/python/python_intro.html [Accessed: 2021-05-19]

[26] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[27] C.-w. Hsu, C.-c. Chang, and C.-J. Lin, "A practical guide to support vector classification chih-wei hsu, chih-chung chang, and chih-jen lin," 11 2003.

[28] M. Liu, L. Wu, and S. Yao, "Dose–response association of screen time-based sedentary behaviour in children and adolescents and depression: a meta-analysis of observational studies," *British Journal of Sports Medicine*, vol. 50, no. 20, pp. 1252–1258, 2016.

[29] K. J. Lee, G. Roberts, L. W. Doyle, P. J. Anderson, and J. B. Carlin, "Multiple imputation for missing data in a longitudinal cohort study: a tutorial based on a detailed case study involving imputation of missing outcome data," *International Journal of Social Research Methodology*, vol. 19, no. 5, pp. 575–591, 2016. [Online]. Available: https://doi.org/10.1080/13645579.2015.1126486

[30] V. Suchert, R. Hanewinkel, and B. Isensee, "Sedentary behavior and indicators of mental health in school-aged children and adolescents: A systematic review," *Preventive Medicine*, vol. 76, pp. 48–57, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0091743515001073

[31] A. S. Lopes, K. S. Silva, V. C. Barbosa Filho, J. Bezerra, E. S. de Oliveira, and M. V. Nahas, "Trends in screen time on week and weekend days in a representative sample of Southern Brazil students," *Journal of Public Health*, vol. 36, no. 4, pp. 608–614, 02 2014. [Online]. Available: https://doi.org/10.1093/pubmed/fdt133

[32] M. Hooghe and J. Oser, "Internet, television and social capital: the effect of 'screen time' on social capital," *Information, Communication & Society*, vol. 18, no. 10, pp. 1175–1199, 2015. [Online]. Available: https://doi.org/10.1080/1369118X.2015.1022568

[33] T. Althoff, R. W. White, and E. Horvitz, "Influence of pokémon go on physical activity: Study and implications," *J Med Internet Res*, vol. 18, no. 12, p. e315, Dec 2016. [Online]. Available: http://www.jmir.org/2016/12/e315/