# Prediction of Academic Development of Children in Ireland based on Screen-time: Longitudinal Study of Infant and Child Cohorts

Saumya Gautam
Student Id: 19211275
MCM, Department of Computing
Dublin City University
Dublin, Ireland
Saumya.gautam3@mail.dcu.ie

*Abstract*— **Screen-time has been identified as a major concern by the health organizations and parents with regards to development, physical and mental health, and academic performance of children. Many surveys and studies have been done in different countries to see the effects of the increasing screen-time, due to easy accessibility to devices and internet, on the development and health of children and adolescents.**

**In this paper I have tried to predict the academic performance of children in Ireland based on screen-time using Growing up in Ireland (GUI) data. The motivation for this study is that all the research done on this topic have focused on the effects of screen-time on the health and well-being of the children but effects on academic performance have not been studied yet and the studies were mostly statistical in nature or classification tasks while this is a prediction task. I have extended this study to infant cohort as well to examine effects of screen-time which has not been done on this dataset. Machine learning models like Decision tree, Random Forest and SVR have been applied to predict that academic performance of children. The models are then evaluated using RMSE, MAE and cross-validation r-squared values. The conclusion drawn from this study is that academic performance of the children can be predicted considerably, and that the performance of the infants is adversely affected by the screen-time.**

***Keywords— Growing up in Ireland, GUI, Child Cohort, Infant Cohort, Wave, Decision Tree algorithm, Random Forest Algorithm, SVR, Regression, R-Squared, Adjusted R-Squared***

## I. INTRODUCTION

Earlier television and its effect on children used to be a major concern for parents. But now that has shifted to screen-time, which is the term coined for the overall time spent interacting with devices having screens: TVs, mobiles, computers, laptops, and video games. This topic has attracted a lot of attention from WHO (World Health Organization), health care organizations, researchers, and parents across the world. Studies have shown that screen-time has association with the child's development, physical and mental health, academic performance, and sleep patterns. While usually screen-time is attributed to have negative effects on the children, some studies show screen-time have positive [1] impacts as well. Another study[2] suggested that screen-time alters the brain of children who use smartphones, video games, and tablets more than seven hours a day and children who had more than two hours of screen-time scored lower on thinking and language tests.

These studies are more important than ever because infants and children are getting exposed to the screens more than they used to earlier due to lifestyle changes. The brain development expedites in infancy and adolescence and these and these studies can be helpful in giving insights about the health and educational achievements of children as adults and this is the motivation behind this paper.

In this study I have tried to predict academic development of children and infant based on screen-time using Growing Up in Ireland (GUI) [3] dataset. The GUI is a longitudinal study done on two cohorts of children, Infant and Child, in Ireland. In this study the same set of participants for respective cohorts were interviewed over the years starting first in 2006. As of now 6 waves(phases) for infant cohort interviews have been completed out of which data for 5 waves is available. Child cohort has gone through 4 waves of interviews with data for 3 waves available for use. The novelty of this paper was that previous work[1] done on GUI data involves effects of screen-time on health, well-being, and physical activity of Irish Children, but as per GUI publications, there is no study to see effects of screen-time on academic achievements. From academic performance point of view there have been two studies, one considering mobile ownership and another home computer use, but their scope is only concerned with device usage. Also, most of the studies related to screen-time, some of which are discussed below, are based on adolescents and children above 9 years. Since there are fewer evidence of studies on infant cohort on this subject, I have extended the experiment to infant cohort as well. GUI data has not been used yet for studying the correlation between screen-time of infants and their health, growth and academic or cognitive performance. Another new aspect covered in this study is the

prediction of academic performance based on screen-time using machine learning, while other studies have focused on either classification, regression, or correlation tasks.

I have made use of the Decision Tree regressor, Random Forest regressor, and Support vector regressor for this prediction. The models have been evaluated using multiple regression metrics like R-squared, Adjusted R-squared, RMSE (Root Mean Squared Errors), MAE (Mean Absolute error). The results show that there is considerable predictability of academic development from screen-time and socio-economic factors. I have fitted the models on 4 different datasets i.e. Child wave 1, Child wave 2, Child wave 3, and Infant wave 5. Since the infant's waves do not have the academic test scores available except for the wave 5, I further checked the correlation between screen-time and socio-economic factors and picture recognition test scores using datasets for Infant wave 2 and Infant wave 3.

## II. Related Work

Dempsey S. et al. (2019) conducted a study [4] on GUI dataset to examine mobile phone ownership and its association with academic development of Irish Children. The study used the data on the 9-year old (wave 1) and 13 years old (Wave 2) children. Logistic regression and Cross-sectional OLS methodology was applied to see the association between owning a phone at age of nine and Drumcondra Reading and Mathematics scores, while using socio economic status as it was a string predictor for the mobile phone ownership and academic scores. The results showed that that there is negative association in mobile ownership and academic scores. The study also highlights that the negative association of mobile phone and academic performance in wave 2 is cumulative effect of the association observed in wave 1.

Casey A. et al. (2012) studied the link between children's home computing and their academic performance in the areas of reading and mathematics. For this study [5] they chose the 9-year-old cohort from Growing up in Ireland survey. In addition to academic performance they investigated effects of various applications on computer. They applied multiple controlling determinants such as income, socio economic status etc. also. As per the survey the most popular activity was searching for information on internet rather than communicating. OLS Regression model was used to conclude that computer use is positively and significantly associated with reading and math scores. Apart from this, the purpose of use also had it effects on the academic scores, wherein activities like emailing, surfing internet for fun and information, and projects had positive impact while instant messaging, downloading music or watching movies are negatively associated with the reading and mathematics scores.

Peiro´ -Velert C et al. (2014) described the interrelations between screen-media usage, sleep-time and academic performance in adolescents in their paper [6]. The study took sample comprising of 3095 Spanish adolescents, aged 12 to 18 years. Self-organizing maps, which is a competitive non-supervised neuronal network algorithms, was applied to this study to establish the non-linear relationship between the variables and the identified behavioral patterns in subsequent cluster analysis. Cluster 1 consisted of boys who spent more than 5.5 hours on screen with low academic performance with average of 8 hours of sleep. And the cluster 2 comprised of the girls with excellent academic performance, who gave less time to sedentary screen usage and slept around 9 hours a day. The analysis shows that the academic performance is directly related to sleep time and inversely related to overall sedentary Screen Media Usage among the Canadian students who participated in this study. It also demonstrated that boys used more passive games and computers for playing than the girls, who mostly used mobile for communicating. The youngest adolescents slept more, had less sedentary screen usage and performed highest in academics.

Sharif, I. et al (2006) conducted a cross-sectional survey on adolescents in the US which focused on weekday and weekend Television and video game screen-time, parental and content restrictions to see their effects on school performance. The study[7] used ordinal logistic-regression analysis to calculate adjusted odds ratios (ORs) and 95% CIs for the relationship between school performance and each of the media use variables while adjusting for parenting style, child personality, demographics, and clustering by school. With an ordered dependent variable, these models had the advantage of retaining information that would be lost by combining the data into 2 arbitrary groups, as one does when using logistic regression. It was concluded that there are strong independent relations between the screen-time and content of exposure and school performance.

Chandra, M. et al (2016) tried something new as compared to previous studies discussed earlier. They performed this study [8] on the infants of 18 months of age in Australia. The aim was to discern the total screen-time of infants in 18 months of age and establish reasons contributing to excessive screen usage. The study used contingency tables, X2 tests and multivariable logistic regression model to conduct analysis. The study presented evidence that there is adverse effect on cognitive development and health of infants due to excess screen-time and most of the infants had screen-time >2 hours per day which is beyond the recommended time for infants. Factors that contribute to excess screen-time in infants were also identified.

Pagani, L. et al (2010) performed a longitudinal study on the children of Canada to estimate the influence of early childhood television exposure on academic, mental and physical well-being of fourth grade children. For the study [9] the parents reported the weekly hours the children spent watching television at age of 29 months and 53 months (approx. 2 years and 4 years). The academic, psychological and health factors were then reported by parents and teachers for same children at the age of 10 years. A statistical analysis was conducted using a series of ordinary least-squares regressions by linearly regressing the academic, psychological and lifestyle characteristics on early life television exposure. The results obtained from the study

showed that every additional hour of television exposure, when adjusted against family and individual factors, contributed to 7%and 6% unit decrease in classroom engagement and maths achievement, respectively; 13% decrease in weekend physical activity and a higher consumption of snacks and soft drinks. Conclusion drawn was that higher levels of early exposure and the long-term risk associated with it may lead towards unhealthy dispositions in adolescence.

## III. Dataset and Data Pre-Processing

### *3.1 About the data*

The data source used for this prediction task is the Growing up In Ireland (GUI) survey data which is a National longitudinal study of infants and children in Ireland. Fig.1 describes the process of data collection for GUI survey. The survey is being carried out by the Economic and Social Research Institute (ESRI) and Trinity College Dublin (TCD). Data is provided by ISSDA (Irish Social Science Data Archive) in AMF (Anonymized Microdata File) version. The survey started in 2006 adopting a two-age cohort design, one cohort being the infant cohort (aged 9 months) with 11,134 infants and other one being the child cohort (9 years old) with 8,568 children. A random sample of participants for infant cohort was selected from the records of Child Benefit Register as it is believed to have all the necessary characteristics for use as sampling frame. For child cohort, sample was taken from the National School System which offered analytical benefits over other sampling frame such as Child Benefit Register.

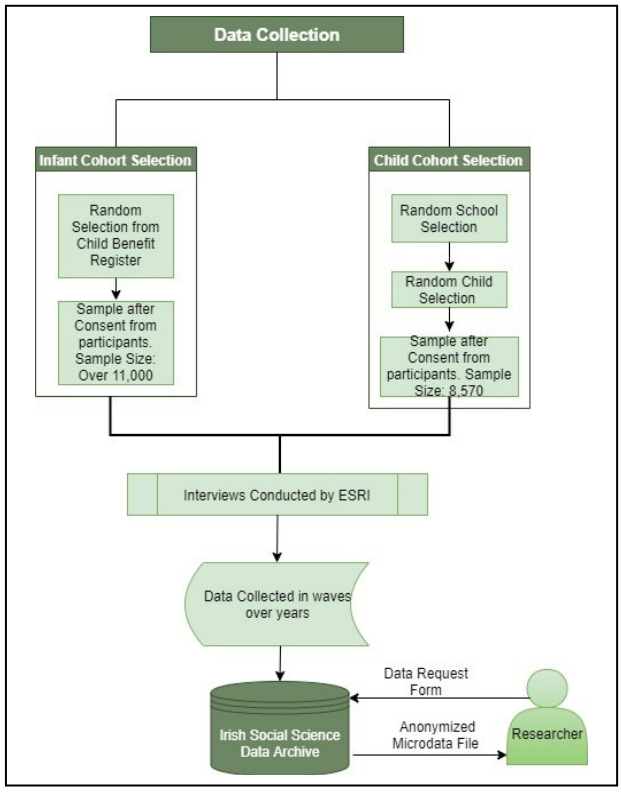


Fig.1. Data Collection Process

Due to longitudinal nature of the survey, participants in both cohorts are interviewed on several occasions over the years. Phases in which participants are interviewed are called as waves as described in Fig. 2, Infant cohort Wave 1 was recorded for infants at age of 9 months, then the same participants were interviewed at the age 3 years for wave 2, at 5 years for wave 3 and subsequently at age 7 years and 9 years for wave 4 and wave 5, respectively. Similarly, Child cohort wave 1 was interviewed at age 9 years and subsequently at age 13 years and 17 years for wave 2 and wave 3, respectively.

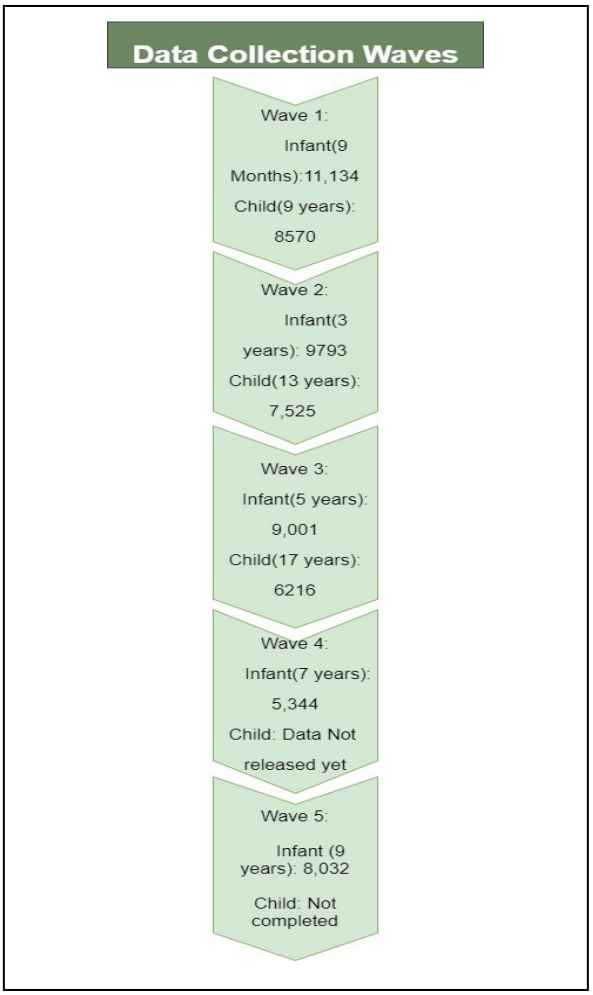


Fig. 2. Waves of Data Collection through interviews

### *3.2 Varaibles*

#### 3.2.1 Dependent Variable

Dependent Variable: Educational test scores are taken as the target variables. Educational Research Centre has worked on a scoring system that measures academic performance of the students based on Drumcondra Primary Reading Test and Drumcondra Primary Mathematics Test. These tests are designed for the students to be taken over the years to generate reports that provide a summary of pupil's performance. These tests are indicative of academic achievement and performance. The Drumcondra Test Scores for Mathematics and Cognitive Math Score have been chosen for the Child Cohort and

Drumcondra Reading Test scores were selected as target variable for Infant Cohort wave 5.

| Target Variable | Count | Mean | STD | Min | 1st Quintile | 2nd Quintile | 3rd Quintile | Max |
|---|---|---|---|---|---|---|---|---|
| Drumcondra Maths Test: Child Wave 1 | 8449 | 56.4297 | 20.994 | 0 | 40 | 56.67 | 72 | 100 |
| Drumcondra Maths Test: Child Wave 2 | 7148 | 53.5275 | 22.897 | 0 | 35 | 55 | 70 | 100 |
| Cognitive Maths Score: Child Wave 3 | 6165 | 2.4738 | 1.2105 | 0 | 2 | 3 | 3 | 4 |
| Drumcondra Reading Test: Infant Wave 5 | 7751 | 77.0625 | 17.958 | 5.56 | 67.5 | 82.5 | 90 | 100 |

Table 1. Statistical Summary of Target Variables

#### 3.2.2 Independent Variables

The survey data consisted of various questions covering multiple aspects of the life of children. The data files for the different waves of data consisted of variables ranging from 95 to 1222 depending on the questionnaire for interview. For this study, I have chosen variables related to Screen-time: Average hours spent watching TV, using computer, playing video games, on mobile phones, or other screen related activities. Questions implying device ownership like the study child having TV, gaming console or computer in their bedroom and if they own mobile phone are included as well.

Apart from screen-time and device ownership, variables for purpose of use were taken into consideration. These variables If the child used the device for communication, research, entertainment, playing games or listening to music.

While screen-time features are the main predictors for this study, socio-economic features cannot be ignored as they contribute to the overall device ownership, and academic performance. It has been seen that academic performance depends greatly on the factors like family income, area of residing (urban or rural), parental education and gender of child etc. and many previous studies have considered them in their analysis. For this purpose, I have selected Annual family income, Highest education received by the Primary Caregiver of the child, Region, gender of child and household type as the socio-economic controllers.

### *3.3 Data Pre-processing*

Variables were reviewed for their datatypes to see if they were any mismatches. There were white spaces in the data which were replaced with NaN using regex.

#### 3.3.1 Handling Missing Values

There are two types of missing values in each wave. First kind is the blank cells and the second kind is cells with "Don't know values" , wherever participants replied with Don't Know for any question, and are represented using different values (8 / 9 / 99 / 98 / 9999 / 666 etc.) in different columns. So, to handle these missing values, I first imputed "Don't know" values with mean for numerical columns where the value was continuous in nature, for example: annual income. For columns with categorical data, mode of each column was used as the columns only accepted certain values like 1,2,3,4 where each number indicates a response from participant. Similarly, the blank values were replaced using column mean in numerical columns and with mode in categorical columns.

#### 3.3.2 One Hot Encoding

One hot encoding [10] is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

The Dataset contains both numerical and categorical variables. Before feeding data into ML model for training I pre-processed categorical columns using one hot encoding. Let us say there is a variable Color which can take 3 different values Red, Green, Blue. One hot encoder will replace Column Color with three new binary columns 'Is_Red', 'Is_Green', 'Is_Blue'. I have used get_dummies function from pandas for one hot encoding.

#### 3.3.3 Data Normalization

Before running Support Vector regressor, I normalized numerical columns using standard scalar method [11], whereas one hot encoding is used to process categorical features. Data normalization can be done using min-max scalar also. But standard scalar is generally preferred over min-max scalar [12]. Standard scalar works by subtracting mean value and dividing by standard deviation of the column which replaces X by its z score. Whereas Min max scalar works by subtracting minimum and dividing by (maximum - minimum). So, in case there is a single outlier in a column, min-max scalar gets impacted badly which can result in poorer results. That is why I have selected standard scalar instead of min-max scalar.

#### 3.3.4 Feature Engineering by removing correlated features using VIF score (Variance Inflation Factor)

In some of the waves number of features are too high which may result in reducing the adjusted r-squared of the model. So, I have removed correlated features using Variance inflation factor (VIF score) [13]. Basically, when training a machine learning algorithm, multicollinearity in the data should be removed. I have used correlation matrix to find out which features are highly correlated with each other. But, even after find pairs of correlated features, we need to decide which one to drop. To make this decision automatic, I made use of VIF score. Let us understand how this method works. Let us say initially there are n features. We train n distinct regression models where in each model we use one of the n features as target variable and remaining (n - 1) features as predictors. So, there are n models, we get r-squared for each model. And VIF score for each model is calculated as (1)

$$VIF = 1 / (1 - r^2) \quad (1)$$

So, here if r-squared is more than 10, it means very high multicollinearity, $4 < VIF < 10$ means considerable multicollinearity, and multicollinearity and VIF less than 4 mean very less multicollinearity. So, if R-squared of each all models are less than 4, then I did not drop any features, whereas if VIF > 4 then I dropped the feature corresponding to the model with target variable having maximum VIF score. And features were dropped until I got VIF less than 4 for all models.

## IV. METHODOLOGY

For this study I have applied the Decision Tree regressor, Random Forest regressor, and Support vector regressor for the prediction. The models have been evaluated using multiple regression metrics like R-squared, Adjusted R-squared, RMSE (Root Mean Squared Errors), MAE (Mean Absolute error). I have fitted the models on 4 different datasets i.e. Child wave 1, Child wave 2, Child wave 3, and Infant wave 5. Since the infant's waves do not have the academic test scores available except for the wave 5, I checked the correlation between screen-time and socio-economic factors and picture recognition test scores using datasets for Infant wave 2 and Infant wave 3.

### *4.1 Decision Tree*

Decision Trees (DTs)[14] are a non-parametric supervised learning method used for classification and regression. It works by creating a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Some advantages of using decision tree are:

- it is the easiest to interpret
- does not require feature scaling
- computationally less expensive than other methods

### *4.2 Random Forest*

A Random Forest [15] is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.
Advantages of Random Forrest are:

- Random Forest increases predictive power of the algorithm and helps prevent overfitting.
- Ability to handle multiple input features without need for feature deletion.
- Prediction is based on input features considered important for classification.

### *4.3 Support Vector Regressor (SVR)*

Support Vector Regressor uses same principles as SVM [16], but instead of classification it is used for regression problems. IN SVM hyperplane is a line separating two classes but in SVR this line is used to predict the continuous output.
Advantages of SVR are:

- SVR is easy to implement and low on computational cost.
- It can improve the prediction accuracy by measuring the confidence in classification.
- Decision model can be easily updated
- Can use multiple classifiers trained on the different types of data using the probability rules.

### *4.4 Model Training and Evaluation*

#### 4.4.1 Grid-Search for Hyperparameter optimization to reduce overfitting

Hyper-parameters are parameters that are not directly learnt within estimators. In scikit-learn they are passed as arguments to the constructor of the estimator classes. To select the best hyperparameter i.e. ccp_alpha in case of Decision tree and Random Forest and hyperparameter C in case of Support vector regressor I have used Grid-search [17] to get most accurate predictions. Selection of the best hyperparameter was done by using grid search methods: best_score_ and best_params_.

#### 4.4.2 Cross-validation for model evaluation

Cross-validation (CV) [18] is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. For model selection and evaluation Nested cross validation [19] is used. This means that there are two instances where cross validation is happening. First is outer CV where K-fold cross validation is diving the dataset into K training and test sets. Another instance is when GridSearchCV is fit to data, cross-validation is done internally to select hyper parameters.
Since the size of the dataset is not very large (Maximum number of rows in any wave is 8449), I have used 5-fold cross-validation for reporting performance scores of all models as it would allow 80% of the data to be used for training and 20% for testing.

### *4.5 Performance Metrics and Model Comparison*

#### 4.5.1 R-Squared: Coefficient of Determination

The coefficient of determination, denoted $R^2$ or $r^2$ and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable. This correlation, known as the "goodness of fit. R-squared has the useful property that its scale is intuitive: it ranges from zero to one, with zero indicating that the proposed model does not improve prediction over the mean model, and one indicating perfect prediction.

#### 4.5.2 Adjusted R-squared

Adjusted R-squared is the modification of R-squared and it is adjusted for the numbers of predictors in the model. R-squared value increase if we increase the number of independent variables. Adjusted R-squared increases only if a significant variable is added.

#### 4.5.3 Root Mean Squared Error (RMSE)

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data– how close the observed data points are to the model's predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. RMSE is defined as (2):

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\boldsymbol{Predicted_i} - \boldsymbol{Actual_i})^2}{\boldsymbol{N}}} \qquad (2)$$

where N is number of observations.

#### 4.5.4 Mean Absolute Error (MAE)

MAE is the simplest regression error metric which describes the typical magnitude of the residuals without considering direction. It is the most intuitive metric as it is the absolute difference between the actual and predicted value. MAE is defined as (3)

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \left| \boldsymbol{y_j} - \boldsymbol{\hat{y}_j} \right| \quad (3)$$

Where $y_j$ is the predicted value, $\hat{y}_j$ is the actual value and n is the number of observations.

## V. Results

### 5.1 *Model Performance and Comparison*

Models used in this study i.e. Decision Tree, Random Forest and SVR when trained only with Screen-time features, gave poor performance. So, further features related to screen-time, device ownership, purpose of use and socio-economics factor, which I received after implementing VIF, were fed into the three models and the performance improved as compared to when only fed with screen-time features.

Table 2 shows the grid-search best score for all models. Parameter which gives the best score is the hyperparameter for the corresponding model. The three models were then trained using the best hyperparameters score found by grid search, to get most accurate predictions for those features. Fig.3 shows plot for ccp_alpha vs 5-fold cross validation test r-squared for Random Forest (Child cohort Wave1)

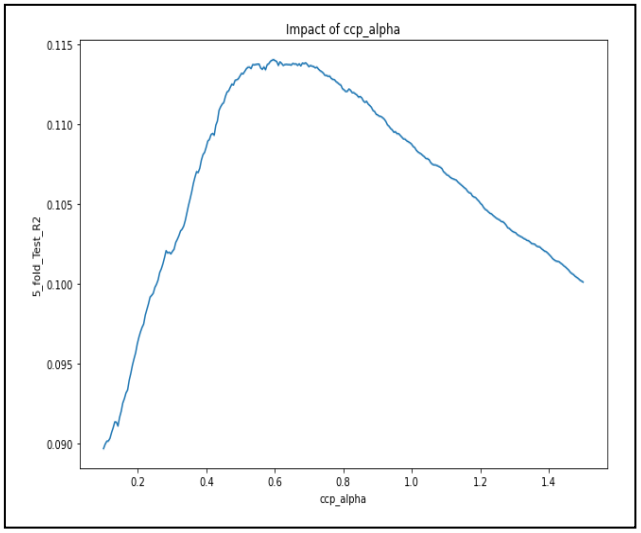


Fig. 3. Plot to show ccp_aplha Vs 5-fold 5-fold cross validation test r-squared for Random Forest

| Child Cohort Wave 1 | Best Score | Best Parameter |
|---|---|---|
| **Decision Tree** | 0.093669 | 0.8424242 |
| **Random Forest** | 0.114049 | 0.5963211 |
| **SVR** | 0.120023 | 1.6929293 |
| | | |
| **Child Cohort Wave 2** | **Best Score** | **Best Parameter** |
| **Decision Tree** | 0.092227 | 0.9909091 |
| **Random Forest** | 0.133912 | 0.8117057 |
| **SVR** | 0.141406 | 1.6545455 |
| | | |
| **Child Cohort Wave 3** | **Best Score** | **Best Parameter** |
| **Decision Tree** | 0.125969 | 0.0024545 |
| **Random Forest** | 0.151151 | 0.0030468 |
| **SVR** | 0.146383 | 0.1050505 |
| | | |
| **Infant Cohort Wave 5** | **Best Score** | **Best Parameters** |
| **Decision Tree** | 0.082139 | 1.040404 |
| **Random Forest** | 0.113111 | 0.5963211 |
| **SVR** | 0.078918 | 2.7272727 |

Table 2 Grid Search Scores for all Models

#### 5.1.1 Performance Metrics

Tables (Table 3,4,5,6) given below describes the performance metrics of prediction for each wave.

| Child Cohort Wave1 | Test | | 5-fold cross validation | |
|---|---|---|---|---|
| | **RMSE** | **MAE** | **R^2** | **Adj R^2** |
| **Decision Tree** | 19.8044 | 16.209 | 0.0936 | 0.0620 |
| **SVR** | 19.2402 | 15.522 | 0.1200 | 0.0893 |
| **Random Forest** | 17.98778 | 14.816 | 0.11405 | 0.0831 |

Table 3 Performance Metrics for Child Cohort Wave 1

| Child Cohort Wave2 | Test | | 5-fold cross validation | |
|---|---|---|---|---|
| | **RMSE** | **MAE** | **R^2** | **Adj R^2** |
| **Decision Tree** | 21.2295 | 17.558 | 0.08961 | 0.06742 |
| **SVR** | 20.73242 | 16.892 | 0.1414 | 0.12047 |
| **Random Forest** | 19.5916 | 16.278 | 0.13392 | 0.1128 |

Table 4 Performance Metrics for Child Cohort Wave 2

| Child Cohort Wave3 | Test | | 5-fold cross validation | |
|---|---|---|---|---|
| | RMSE | MAE | R^2 | Adj R^2 |
| **Decision Tree** | 1.137 | 0.9484 | 0.1044 | 0.0933 |
| **SVR** | 1.0923 | 0.89511 | 0.14638 | 0.1358 |
| **Random Forest** | 1.0633 | 0.887 | 0.15115 | 0.14068 |

Table 5 Performance Metrics for Child Cohort Wave 3

| Infant Cohort Wave5 | Test | | 5-fold cross validation | |
|---|---|---|---|---|
| | RMSE | MAE | R^2 | Adj R^2 |
| **Decision Tree** | 16.9958 | 13.1425 | 0.07923 | 0.0453 |
| **SVR** | 16.5872 | 11.8464 | 0.07891 | 0.04501 |
| **Random Forest** | **15.388** | 12.0769 | 0.1131 | 0.08046 |

Table 6 Performance Metrics for Infant Cohort Wave 5

### 5.1.2 Model Comparison

From above table, it is can be seen that the test RMSE value of Random Forest model is smaller than the test RMSE value for other two models in all waves. Also, the cross-validation r-squared value of Random Forest is greater than other models for two of the waves and smaller than SVR for other two waves, though the difference is not significant. Therefore, Random Forest was chosen for final prediction of the academic scores.

Table 7 shows the actual and predicted academic scores obtained by Random Forest Model for one of the child cohort waves and for infant cohort wave:

| Child Cohort | | Infant Cohort | |
|---|---|---|---|
| Actual Scores | Predicted Scores | Actual Scores | Predicted Scores |
| 50 | 67.88939 | 100 | 82.9514 |
| 90 | 73.45734 | 97.5 | 84.44936 |
| 70 | 65.77764 | 77.78 | 77.54668 |
| 20 | 46.08523 | 90 | 82.20287 |
| 60 | 53.05738 | 70 | 69.54606 |
| 45 | 50.31363 | 88.89 | 81.19632 |
| 65 | 59.80522 | 97.22 | 78.65834 |
| 45 | 46.14564 | 67.5 | 81.05694 |
| 60 | 52.58231 | 70 | 69.21216 |
| 35 | 46.14491 | 45 | 66.87512 |
| 45 | 43.57585 | 77.5 | 73.00951 |
| 55 | 52.74911 | 97.5 | 84.04824 |

Table 7 Comparison of Actual Vs Predicted Academic Scores

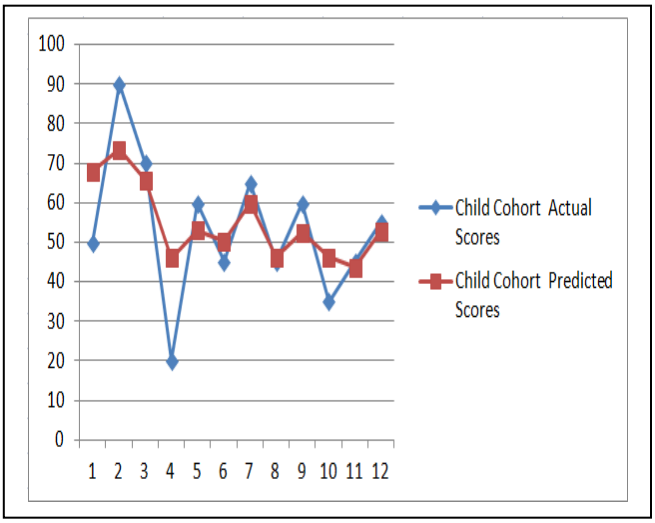


Fig 4. Actual Scores Vs Predicted Scores Child Cohort

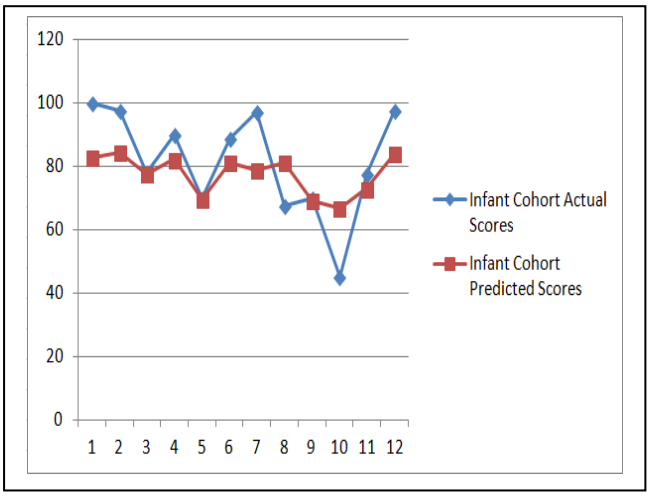


Fig 5. Actual Scores Vs Predicted Scores Child Cohort

### 5.1.3 Feature Importance Using Random Forest

Table 8 shows the top 5 features for each wave as per feature importance derived from Random Forest Model. It can be seen that the features which were most prominent in the prediction for academic performance included Socio-economic features: parental education, child gender and annual income and screen-time features like Child having TV in Bedroom, time spent on mobile and internet. It shows that the parental education and annual income are most crucial for a child's academic performance, while screen-time and use of screen-time affect academic performance to quite an extent.

| | Feature | Feature Importance |
|---|---|---|
| **Child Cohort Wave 1** | **Parental Education** | 0.30797 |
| | **Annual Income** | 0.142346 |
| | **TV in Child Bedroom** | 0.128766 |
| | **Child Gender** | 0.038073 |
| | **Time using Mobile Phone** | 0.030334 |
| **Child Cohort Wave 2** | **Feature** | **Feature Importance** |
| | **Parental Education** | 0.316426 |
| | **Time using Computer** | 0.106066 |
| | **Annual Income** | 0.104798 |
| | **Child Gender** | 0.07468 |
| | **TV in Child Bedroom** | 0.06187 |
| **Child Cohort Wave 3** | **Feature** | **Feature Importance** |
| | **Child Gender** | 0.290278 |
| | **Parental Education** | 0.214182 |
| | **Time spent online (Weekday)** | 0.12214 |
| | **Time spent on TV (Weekday)** | 0.098937 |
| | **Annual Income** | 0.08349 |
| **Infant Cohort Wave 5** | **Feature** | **Feature Importance** |
| | **Parental Education** | 0.359571 |
| | **Annual Income** | 0.21652 |
| | **Own Mobile Phone** | 0.040358 |
| | **Purpose of screen time** | 0.036736 |
| | **Time spent playing on Internet** | 0.035951 |

Table 8 Feature Importance

### 5.2 *Analysis for Infant Waves*

Fig.6. and Fig. 7. are the correlation Matrix for the analysis on Infant Wave 2 and wave 3, respectively. It was observed that the picture recognition test is negatively correlated to the average screen-time, screen-time on TV, and the time infant spend playing mobile or computer games by themselves while supervised computer gaming and study on computer are positively correlated.

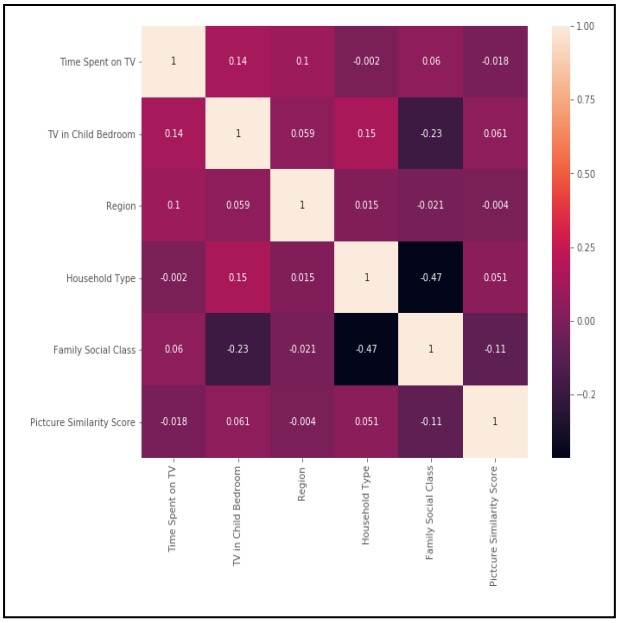


Fig. 6. Correlation Matrix for Infant Wave 2

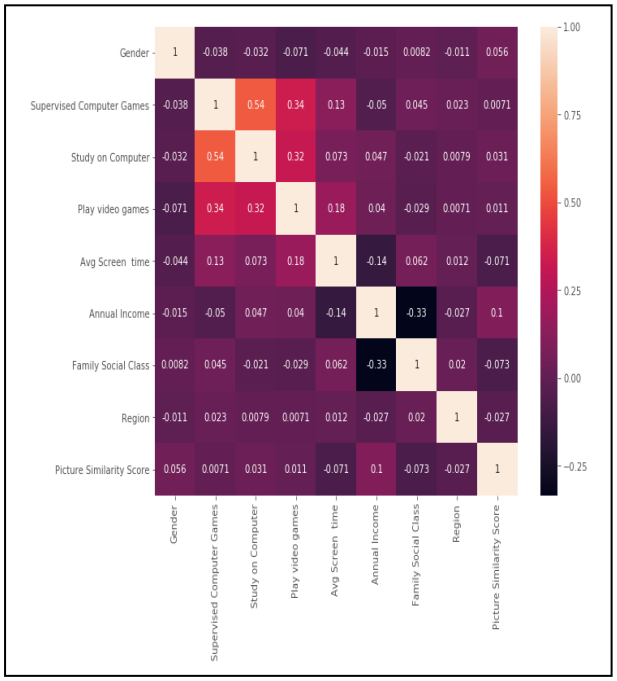


Fig. 7. Correlation Matrix for Infant Wave 3

## VI. CONCLUSION AND FUTURE WORK

In this study, I developed a model using Random Forest to predict the academic performance of children in Ireland based on their screen-time. Based on the machine learning experiments on the data, it can be concluded that Random Forest when used with the hyperparameters gives best prediction amongst models like Decision Trees and Support Vector Regression. Furthermore, it is observed that there is considerable predictability power in screen-time to predict the

academic performance. The study on infant cohort data shows that there is negative correlation between Screen time and total ability score for picture similarities, which means screen time affects Total ability score for picture similarities adversely, which is in accordance with Chandra, M. et al (2016) and Sharif, I. et al (2006). It can also be concluded that socio-economic factors, device ownership and purpose for which screen-time is dedicated to, also must be considered. So, screen-time alone cannot be used for prediction of academic scores as discussed in studies by Casey A. et al. (2012). There are some limitations to this study as features like children's reading habit, all round academic performance as per teachers, and further analysis of socio-economic factors could have been considered for better understanding of effect of screen-time on academic performance. Another limitation of this study is that the prediction of academic performance could not be done across the waves.

The model developed in this study is a simpler and first one on Growing up in Ireland data set which can be enhanced by further feature engineering on the variables related to the study and introduction of more training data from future waves of the survey. Future work for this study can be development of more sophisticated prediction model using Neural Networks. Further research on this data can be prediction of the academic performance of Irish children from their early life screen-time as done by Pagani, L. et al (2010).

## REFERENCES


[1] Moderate use of screen time can be good for your health, new study finds : https://www.oii.ox.ac.uk/news/releases/moderate-use-of-screen-time-can-be-good-for-your-health-new-study-finds/

[2] Is Screen Time Altering the Brains of Children? https://www.healthline.com/health-news/how-does-screen-time-affect-kids-brains

[3] Growing Up in Ireland : National Longitudinal Study of Children https://www.growingup.ie/

[4] Seraphim Dempsey, Seán Lyons & Selina McCoy (2019) Later is better: mobile phone ownership and child academic development, evidence from a longitudinal study, *Economics of Innovation and New Technology*, 28:8, 798 815, DOI: 10.1080/10438599.2018.1559786

[5] Casey, A., Layte, R., Lyons, S. and Silles, M., 2012. Home computer use and academic performance of nine-year-olds. *Oxford Review of Education*, 38(5), pp.617- 634.

[6] Peiró-Velert C, Valencia-Peris A, González LM, García-Massó X, Serra-Añó P, Devís-Devís J. Screen media usage, sleep time and academic performance in adolescents: clustering a self-organizing maps analysis. *PLoS One*. 2014;9(6):e99478. doi:10.1371/journal.pone.0099478

[7] Sharif, I. and Sargent, J.D., 2006. Association between television, movie, and video game exposure and school performance. *Pediatrics*, *118*(4), pp.e1061-e1070.

[8] Chandra, M., Jalaludin, B., Woolfenden, S., Descallar, J., Nicholls, L., Dissanayake, C., Williams, K., Murphy, E., Walter, A., Eastwood, J. and Eapen, V., 2016. Screen time of infants in Sydney, Australia: a birth cohort study. *BMJ open*, *6*(10), p.e012342.

[9] Pagani, L. S., Fitzpatrick, C., Barnett, T. A., & Dubow, E. (2010). Prospective Associations Between Early Childhood Television Exposure and Academic, Psychosocial, and Physical Well-being by Middle Childhood. Archives of Pediatrics & Adolescent Medicine, 164(5). doi:10.1001/archpediatrics.2010.50

[10] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[11] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[12] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[13] Variation Inflation Factor: https://online.stat.psu.edu/stat462/node/180/

[14] Pedregosa et al. Decision Trees, JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/stable/modules/tree.html

[15] Breiman Leo, Cutler Adele. Random Forests https://medium.com/datadriveninvestor/decision-tree-and-random-forest-e174686dd9eb

[16] https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

[17] https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e

[18] https://machinelearningmastery.com/k-fold-cross-validation/

[19] https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/