# DS PROJECT DOCUMENTATION

**Project Name :** To determine the wellbeing for young children during Covid lockdown.

**Project Batch ID :** DS\_Project\_Team\_66

**Participant Name :** Prahasith Sai Dachapally

**Problem Statement:**

Governments across different countries anticipate more lockdowns. It is a very concerning issue for the wellbeing of young children in lockdown situations as they are removed from their friends and play environments.

The task you have been assigned is to examine the factors that determine the wellbeing for young children. To ensure that a biased interpretation does not get involved you are expected to work on:

(a) A standard statistical analysis using regression, and

(b) Factors that might be important to children, but which adults are not aware of, are to be included in the model.

**Survey:**

A questionnaire filled by children with parent consent during their usual school day. Confidentiality of the data is utmost important hence, data is masked. Only the answers for the questionnaire are provided. Parents were also invited to complete a questionnaire about their child.

You have been given:

1. The raw data of a survey with primary school children when they were in the first lock down.

2. The data dictionary.

**Tasks:**

* Data Pre-processing - You will need to prepare the data set so you can analyze it (e.g. Convert string variables to numerical variables).
* Calculate child wellbeing. Q24 is the Me and My Feeling measure

• Look at what determines wellbeing with both

a) A continuous measure, and

b) A binary measure (using logistic regression)) using a hypothesis driven approach.

• Look at what determines wellbeing using a data-driven / machine learning approach (e.g. Decision trees)

• Present and interpret your findings.

**Assessment Criteria:**

* If you/your team use more than one language (e.g., BI, Python and SQL) or multiple files, please create a zip file containing all the scripts and submit the zip file instead.

• Assessment will be based on the structural approach taken by the team for problem solution including the dashboard (reporting), documentation, and codes developed (optimized).

• Recommended reporting structure to follow: Introduction, Methods, Results, Discussion

**Sol:**

**a) Business Objective:** To examine the factors that determines the wellbeing for young children during covid lockdown.

**Business constraints:** Choosing the top most relevant data.

**b) Data dictionary: (Secondary Data Source)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S No** | **Name of feature** | **Data Type** | | **Relevance** |
| 1 | ID | Discrete | Nominal | Irrelevant |
| 2 | Are you still going to school? | Discrete | Nominal | Relevant |
| 3 | Do you have any other children living in your house with you? | Discrete | Binary | Relevant |
| 4 | How many people live in your home with you (including adults)? | Discrete | Count | Relevant |
| 5 | What year are you in now? | Discrete | Ordinal | Relevant |
| 6 | Gender | Discrete | Binary | Relevant |
| 7.1 | Year | Discrete | Nominal | Irrelevant |
| 7.2 | Month | Discrete | Nominal | Irrelevant |
| 7.3 | Day | Discrete | Ordinal | Irrelevant |
| 8.1 | 1. What did you eat for breakfast YESTERDAY? | Discrete | Nominal | Relevant |
| 8.2 | 2. Did you eat any fruit and vegetables YESTERDAY? | Discrete | Ordinal | Relevant |
| 8.3 | 3. How many times did you brush your teeth YESTERDAY? | Discrete | Count | Relevant |
| 8.4 | 4. What time did you fall asleep YESTERDAY (to the nearest half hour)? | Continuous | Ratio | Relevant |
| 8.5 | 5. What time did you wake up TODAY (to the nearest half hour)? | Continuous | Ratio | Relevant |
| 8.6 | 7. In the last 7 days, how many days did you watch TV/play online games/use the internet etc. for 2 or more hours a day (in total)? object | Discrete | Count | Relevant |
| 8.7 | 8. In the last 7 days, how many days did you feel tired? | Discrete | Count | Relevant |
| 8.8 | 9. In the last 7 days, how many days did you feel like you could concentrate/pay attention well on your school work? | Discrete | Count | Relevant |
| 8.9 | 10. In the last 7 days, how many days did you drink at least one fizzy drink (e.g. coke, sprite, thumsup)? | Discrete | Count | Relevant |
| 8.10 | 11. In the last 7 days, how many days did you eat at least one sugary snack (e.g. chocolate bar, sweets)? | Discrete | Count | Relevant |
| 8.11 | 12. In the last 7 days, how many days did you eat take away foods (e.g. Chinese takeaway)? | Discrete | Count | Relevant |
| 10 | 13. On a scale of 0 to 10 (0 being not very safe and 10 being very safe), how safe do you feel playing in your area? | Discrete | Ordinal | Relevant |
| 11.1 | 14. From your house, can you easily walk to a park (for example a field or grassy area)? | Discrete | Binary | Relevant |
| 11.2 | 15. From your house, can you easily walk to somewhere you can play? | Discrete | Binary | Relevant |
| 11.3 | 16. Do you have a garden? | Discrete | Binary | Relevant |
| 12 | 17. How often do you go out to play outside? | Discrete | Ordinal | Relevant |
| 13 | 18. Do you have enough time for play? | Discrete | Ordinal | Relevant |
| 14 | 19. What type of places do you play in? | Discrete | Ordinal | Relevant |
| 15 | 20. Can you play in all the places you would like to? | Discrete | Ordinal | Relevant |
| 16 | 21. Do you have somewhere at home where you have space to relax? | Discrete | Binary | Relevant |
| 17.1 | 22. Tell us if you agree or disagree with the following: [I am doing well with my school work] | Discrete | Binary | Relevant |
| 17.2 | 22. Tell us if you agree or disagree with the following: [I feel part of my school community] | Discrete | Binary | Relevant |
| 17.3 | 22. Tell us if you agree or disagree with the following: [I have lots of choice over things that are important to me] | Discrete | Binary | Relevant |
| 17.4 | 22. Tell us if you agree or disagree with the following: [There are lots of things I'm good at] | Discrete | Binary | Relevant |
| 18.1 | Your Health | Discrete | Ordinal | Relevant |
| 18.2 | Your School | Discrete | Ordinal | Relevant |
| 18.3 | Your Family | Discrete | Ordinal | Relevant |
| 18.4 | Your Friends | Discrete | Ordinal | Relevant |
| 18.5 | Your Appearance (how you look) | Discrete | Ordinal | Relevant |
| 18.6 | Your Life | Discrete | Ordinal | Relevant |
| 19.1 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I feel lonely] | Discrete | Ordinal | Relevant |
| 19.2 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I cry a lot] | Discrete | Ordinal | Relevant |
| 19.3 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I am unhappy] | Discrete | Ordinal | Relevant |
| 19.4 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I feel nobody likes me] | Discrete | Ordinal | Relevant |
| 19.5 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I worry a lot] | Discrete | Ordinal | Relevant |
| 19.6 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I have problems sleeping] | Discrete | Ordinal | Relevant |
| 19.7 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I wake up in the night] | Discrete | Ordinal | Relevant |
| 19.8 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I am shy] | Discrete | Ordinal | Relevant |
| 19.9 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I feel scared] | Discrete | Ordinal | Relevant |
| 19.10 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I worry when I am at school] | Discrete | Ordinal | Relevant |
| 19.11 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I get very angry] | Discrete | Ordinal | Relevant |
| 19.12 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I lose my temper] | Discrete | Ordinal | Relevant |
| 19.13 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I hit out when I am angry] | Discrete | Ordinal | Relevant |
| 19.14 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I do things to hurt people] | Discrete | Ordinal | Relevant |
| 19.15 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I am calm] | Discrete | Ordinal | Relevant |
| 19.16 | 24. Remember, there are no right or wrong answers, just pick which is right for you. [I break things on purpose] | Discrete | Ordinal | Relevant |
| 20 | 25. Are you able to keep in touch with your family that you don't live with? (grand parents, Uncle, Aunt, Cousins, etc) | Discrete | Binary | Relevant |
| 21 | 26. Are you able to keep in touch with your friends? | Discrete | Binary | Relevant |
| 22 | 27. If yes, how are you keeping in touch (tick all you use)? | Discrete | Ordinal | Relevant |

**c) Data Preprocessing:**

* Drop the attributes which are irrelevant from the dataset.
* Customize the column names
* Perform Text Mining to get relevant information for corresponding columns
* Map the labels as per the their scores in the given columns of the dataset
* Perform label encoding for the categorical features where ever it is applicable
* Perform all the required operations for the given columns as mentioned in the Sheet2 of the given dataset
* Check for NaN values, if any present perform imputation.

**d) Exploratory Data Analysis (EDA):**

* Perform pair plot for the finalized dataset & identify for co linearity between predictors.
* Used Bar plot for plotting accuracy scores & f1-scores of different models.
* Build correlation coefficient matrix for dataset for checking of co linearity problem.

**e) Feature Engineering:**

* Using "K-Best & Chi2" Algorithm - Gives most significant features with respect to target variable
* Identify the features which are most relevant for model building based on their scores & drop the remaining features from the dataset
* Check the correlation coefficient,(|r|) between input features, if any present (|r| > 0.85) drop those features which has least impact on target variable

**e) Model Building:**

* Classification Models used are:

(1) Logistic Regression (Multinomial) with Stratified K Fold

(2) Decision Tree with (XG Boosting + Randomized Search CV)

(3) Random Forest Classifier with Stratified K Fold

* Import the required libraries for model building
* Split the entire dataset into train & test (with test size = 0.3)
* Build the different models using train data & improve the model performance using hyper parameter optimization/XG Boosting/Randomized Search CV/stratified k fold techniques
* Run the model using test data & compare the prediction values with actual values of test data
* Check for accuracy scores & f1-scores for goodness of the different models.
* Select the best model for the above business problem
* attachment of python code for model building is available with documentation

**f) Results:**

* **For Emotional dataset:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S No.** | **Model** | **Accuracy** | | **f1-score** | |
| **Emmotinal** | **Behavioral** | **Emmotinal** | **Behavioral** |
| 1 | Logistic Regression (Multinomial) with Stratified K Fold | 0.91 | 0.61 | 0.86 | 0.59 |
| 2 | Decision Tree with (XG Boosting + Randomized Search CV) | 0.89 | 0.64 | 0.85 | 0.62 |
| 3 | Random Forest Classifier with Stratified K Fold | 0.91 | 0.65 | 0.86 | 0.61 |

**g) Factors affecting the wellbeing of child:**

* **For Emotional feelings:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S No.** | **Features** | | **Co linearity between input Features** | **Influence of input feature on target variable** |
| 1 | Continuous measures | 1) Age  2)Sleeping | No Correlation | Least influence |
| 2 | Binary measures | 1) Gender  2) Any children with you  3) Can you walk to park  4) Keep in touch with your friends  5) Can you walk where you play  6) keep in touch with family  7) Do you have garden | Good correlation exists between 'Can you walk to park' & 'Can you walk where you play'. | Least influence |
| 3 | Other than above top ranked features influence on target variable using K-Best & Chi2" Algorithm | 1) Do you get tired (82.19)  2) Is it safe playing in your area (32.33)  3) Your health (30.62)  4) Your appearance (29.50)  5) Your school (27.50)  6) Do you finish your home work (24.54)  7) Important things (18.50)  8) Your life (17.82)  9) Activities (17.76)  10) I'm good at (16.54)  11) Your friends (16.51)  12) Do you have drinks (8.84)  13) Are you part of school community (8.51)  14) Did you ate fruits, veggies (6.42)  15) Are you doing well your homework (4.98) | No Correlation | Highest Influence |

**Summary:** Child should focus on the factors which influence them emotionally mentioned in the above table & should take necessary measures to enhance their wellbeing

* **For behavioral feelings:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S No.** | **Features** | | **Co linearity between input Features** | **Influence of input feature on target variable** |
| 1 | Continuous measures | 1) Age  2)Sleeping | No Correlation | Least influence |
| 2 | Binary measures | 1) Gender  2) Any children with you  3) Can you walk to park  4) Keep in touch with your friends  5) Can you walk where you play  6) keep in touch with family  7) Do you have garden | Good correlation exists between 'Can you walk to park' & 'Can you walk where you play'. | Least influence |
| 3 | Other than above top ranked features influence on target variable using K-Best & Chi2" Algorithm | 1) Do you get tired (103.75)  2) Your appearance (66.58)  3) Your school (60.66)  4) Are you part of school community (44.51)  5) Important things (37.28)  6) Is it safe playing in your area (36.83)  7) Your health (34.76)  8) Do you finish your home work (34.19)  9) Are you doing well your homework (30.90)  6) Your life (29.80)  7) (18.50)  8) Your friends (29.19)  9) I'm good at (28.46)  10) Activities (16.50)  11) Your family (8.08)  12) Do you have drinks (8.84)  13) Do you go to play outside (7.94)  14) Can you play in all places (7.59)  15) No of times brushing teeth (6.29)  16) Do you have drinks (4.86)  17) Yesterday's breakfast (4.00) | No Correlation | Highest Influence |

**Summary:** Child should focus on the factors which influence them behavior mentioned in the above table & should take necessary measures to enhance their wellbeing