**[DAB402 - Capstone Project](https://lms.stclaircollege.ca/webapps/blackboard/execute/launcher?type=Course&id=_39477_1&url=" \t "_top)**

Professor - Muhammad Shahid

Assessment3 – Detailed Data Assessment

**Prediction of mental health based on severity**

**Problem Statement:**

Understand the impact of poverty, violence, civil conflict, and migration on mental, neurological and substance use disorders and reduce the duration of untreated illness by developing culturally sensitive early interventions for mental, neurological and substance use disorders across settings.

**Datafication:**

Age of the data is becoming younger day-by-day. A person searching for some of his interests on the web in his free time would get these same recommendations once if he opens his phone while travelling. This is just a small example of datafication. With the development of technology, accessing and using the data has become so easy and at the same time yielding negative affects if not used for a good cause or with improper usage.

Analysing the data that is fictional in nature gives us a way to apply the technical and the domain knowledge that we have learnt, but working on real time data considering a societal issue as an objective provides a way to extend a solution besides the former. We take this advantage to work on a real-time data.

**Mental health dataset:**

With great efforts for the data, we received our final data from Statistics Canada (Canada, n.d.). In brief, it is an EFT (Electronic File Transfer) public usage data file that is authenticated. Data is of Canadian Community Health Survey (CCHS). We were guided with a toolkit (SPS application) documentary that helped us in opening the dataset being the dataset, sps and sass formatted files. Dataset covers 2 years of data i.e., 2015 and 2016. This has all the information for all the health components of a person participated in the survey. This survey data covers all the provinces in Canada. Survey is conducted by interviewing the respondent by a list of questions that are provided in the questionnaire document. A response of 110,095 individuals is obtained from the survey. Raw version of this dataset is called a master file and it has whole information of the respondent. In order to preserve the analytical value of the data, several controls are implemented in creating the PUMF (Public usage metadata file). This has no personal information such as name, telephone number, whether a pregnant etc. The PUMF file has around 1024 columns and more than 1.2 lakh rows. It has columns that are self- reported and derived from the other columns. This dataset covers mostly whole number valued columns ranging from 1-6 as its values. Out of these columns, we are trying to figure out the columns that would be related for our analysis purpose.

The data that we received is given a tag “final selected data” by looking at:

1. Columns it had specifically that are related for our problem statement analysis
2. Size of the data
3. Quality of the data (a quality assurance document is also being received from the resource channel)

**Ethical principles related to our data:**

Our dataset is completely ethical in a way that it follows all the five framing guidelines (consent, clarity, consistency and trust, control and transparency, consequences) necessary (Patil, n.d.) This is completely secure and reliable too.

* This dataset is collected with user’s agreement and this being mentioned in the Questionnaire document. We had this dataset into our hands after being completing all the procedures officially. Hence, it follows the consent with the user.
* A particular note which specifies the main goal for the data collection is being specified to the user with clarity before collecting the data. We had an official mail communication stating our need for the dataset clearly before the data was being handed over to us. This paves a way for the clarity guideline.
* This data is being collected by the government officials and this is a huge program that took place in all the provinces of Canada for the benefit of society. Hence, there is no chance for the break of trust and consistency of the user. We as a team will be working on this dataset only for our analysis with no misuse and hence forth, we will maintain the same level of consistency and trust. This in a straight way covers the third framing guideline i.e., consistency and trust.
* Even though data is available for public usage, it had certain authentication procedures that are to get followed to access the dataset to control and cover up the transparency. We have received this dataset by stating our need honestly and hence we would maintain the control over the dataset by only providing an access only to our team working on it.
* As this data is collected from public in the society, in order to control the ethical principles, personal information is being hidden and hence this dataset is now free to use with no other consequences because this would not harm the users in any way by lightening a way to cover the fifth framing principle, consequences of the data.

**Challenges:**

* Columns related to our project are to be picked from the whole bunch of the variables in the dataset.
* Data cleaning tool that would be capable enough for the huge size of the dataset is to be chosen and cleaning is to be done.
* Appropriate target variable is to be chosen for the prediction.
* Different techniques and algorithms are to be implemented to have the best outcomes.
* Collaboration with mental health associations are to be done in order to have an extended public usage website which gives them the information of the respective organization they can visit to.

**References:**

Electronic File Transfer Service -  Welcome to the Electronic File Transfer Service - Attention. (2020). Retrieved 19 February 2020, from <https://www75.statcan.gc.ca/eft-tef/en/operations> (Canada, n.d.)

DJ Patil, M. (2020). The five Cs. Retrieved 20 February 2020, from <https://www.oreilly.com/radar/the-five-cs/> (Patil, n.d.)

Nesstar WebView. (2020). Retrieved 20 February 2020, from <http://odesi2.scholarsportal.info/webview/>

**Team:**

Kavya Bulusu (0734186)

Vanita Patel (0734890)

Jinali Patel (0734623)