Milestone 2

Progress report

Team

1. Swet Shah
2. Vinayaka S Gadag
3. Chaitanya Shekhar Deshpande
4. Abhinav Kumar
5. Nicholas Faro

Tasks:

- Random Forest Classifier and Research
    - Swet Shah
- SVM and Research and Data pre-processing
    - Chaitanya Shekhar Deshpande
- KNN and Research
    - Vinayaka S Gadag
- Decision Tree and Research
    - Nicholas Faro
- AdaBoost Classifier and Research
    - Abhinav Kumar
- General Documentation and Research will be done by all members.

Current Progress:

- Research, Literature Survey and collecting the data.

- Data Pre-Processing has been completed. We have utilized Ordinal Encoder and OneHotEncoder(using pd.Dummies) wherever applicable to make the data more usable for Machine Learning models.

- Also, a sample SVM model has been applied on the Training dataset to check if the data is ready for further processing.

- Discussion on algorithms and splitting the tasks between the team members. We have decided to use 5 models of Machine Learning: KNN, SVM, Decision Tree, Random Forest and AdaBoost Classifier. Based on several approaches towards all these models, we'll try to generate a model that gives maximum accuracy.

- Weekly retrospective meeting on the progress

Questions or comments:

- How much data is available?
    - Data available is quite less but it is tangible. (2111)
    - We can apply most of the algorithms except Neural networks as size of the data is less. The probability of getting a high accuracy will be very low in case of neural networks, hence we will not be applying it.

- Is it enough to reasonable develop your system?
    - Yes, we can develop a reasonable system. Since it is a multiclass classification problem and we have 2111 data samples, it should be sufficient to use decent Machine Learning models to predict good outputs. As we will be implementing 5 different models: KNN, SVM, Decision Tree, Random Forest and AdaBoost Classifier, we would also come to know the pros and cons of each model and find out which one gives maximum accuracy.

- Research Questions:
    - As we have data from all countries combined into 1 dataset, we will not be able to distinguish separately between countries.
    - Our dataset consists of several features related to Eating Habits such as: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The other attributes which are related with the physical conditions are: calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS). The remaining attributes are: Gender, Age, Height and Weight.
    - So as research questions, we will try to estimate which is the strongest indicator of high obesity in each category (Eating Habits, Physical Conditions and Others) and will try to evaluate which indicator can be brought under control.
    - We could also evaluate several common factors for a particular age as they determine the lifestyle of a person. Like at a particular age group how frequent is it that a person consumes Alcohol or if he/she Smokes. Another example would be how are people with less BMI balancing well with the calorie consumption and frequency of Physical activities and Mode of Transport used, as this can be implemented by people having higher BMI to reduce it.
    - We will be using several plotting methods like Histogram, Box Plots, Quartile Plots and Scatter Plots to observe and conclude each observation for different models.

Our GITHUB Link:

https://github.iu.edu/CSCI-P556-Spring-2021/P556-group18