ML project demo

Team: G13

Chaitanya Chakka -2019A7PS0171H
M. Sheethal Reddy - 2019A7PS0159H
V. Shashivardhan Reddy -2019A7PS0003H
D. Pranay Kumar - 2019A7PS0006H
Vansh Madan - 2018B4A70779H

Problem Definition

- Predicting Individuals Mental Health using different Machine Learning Methods
- Testing and comparing accuracies obtained from the implemented models
- **Objectives** of the project:
  - To develop a mental health prediction tool
  - Predicting whether a person is likely to be depressed or not.
  - Implementing the models using
    - Logistic regression
    - Random Forest
    - SVM
    - Vote Ensembling

- **Dataset** : The dataset used in this study is provided by the ZINDI competition https://zindi.africa/competitions/busaramental health-prediction-challenge/data
  This is based on a survey study made by Busara Center in western Kenya.
  The data contains more than 70 features, including information about the participants, such as health, economic activity, financial flows, and household composition.
  The Training set contains 1,143 instances, where after splitting 800 cases were for the training set and 343 for the validation set, while the testing set has 286 cases.

- Metrics used for comparison:
  - F1-Score
  - Accuracy

Purpose and Need for project

- According to the World Health Organization (WHO), 264 million people are affected by depression globally. It is challenging to address such disease in countries with a lack of mental health resources, such as Kenya.
- Depression affects several countries with a lack of knowledge about the disease and lack of resources, such as psychiatrists, psychiatric nurses, mental psychologists.
- The lack of resources about the disease causes the difficulty of diagnosis and producing an efficient treatment, which eventually increases the number of cases.
- Almost half of the Kenyan population is aged 19 or younger. It is more necessary to have more knowledge correlates of both anxiety and depression in such a country

Purpose and Need for project

- It was approved that analyzing individual behavior using machine learning may help improve diagnostics and have an effective treatment.
- Untreated mental illness can lead to increased medical expenses and affect individual performance at school or work.
- It was also noticed it was the cause of increasing the risk of suicide
- It also noted that depressed people might also have several physical conditions without any apparent physical cause . It is considered one of the most common mental disorders and one of the leading causes of disability worldwide.

Literature review

Research papers

- Mental Anxiety and Depression Detection during Pandemic using Machine Learning.

  https://www.ijser.org/researchpaper/Mental-Anxiety-and-Depression-Detection-during-Pandemic-using-Machine-Learning

- Behavioral Modeling for Mental Health using Machine Learning Algorithms

  https://link.springer.com/article/10.1007/s10916-018-0934-5

- Application of Machine Learning methods in Mental Health detection.

  https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9214815&isnumber=8948470

- Mental health monitoring with multimodal sensing and machine learning

  https://www.sciencedirect.com/science/article/pii/S1574119217305692?via%3Dihub

- Predicting Individuals Mental Health Status in Kenya using Machine Learning Methods

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9464608&isnumber=9464530

(We implemented this paper for our project purpose.)

Predicting Individuals Mental Health Status in Kenya using Machine Learning Methods

Feature Extraction

1. Random Forest Classifier was used to select the most important features.
2. We dropped the least important features; the training dataset has 51 features for 1143 records and the testing dataset has 51 features for 286 records

Models

1. Embedded Method - Random Forest Importance

Results and Conclusions:

1. To evaluate the model's performances on the validation dataset, we used F1-score and Accuracy metrics.
2. SVM with RDF kernel, Random Forest Ada Boosting, and VotingEnsemble models scored the same f1-score and accuracy with 0.78 and 85% .
3. Naive Bayes had the worst model performance evaluation scores with f1-score 0.26 and 25% accuracy, respectively.
4. In our experiment, the ages were in the range of 17-91. The average of the ages was 34. Most of the participants who are unfortunately suffering from depression are younger than 40 .

Data preprocessing

Why do we need data preprocessing

- It transforms text into a more digestible , useful and efficient format. Machine learning algorithms can perform better on preprocessed data.
- Answers in the survey contains a lot of information and opinions.
- Raw answers without preprocessing are highly unstructured, may contain redundant information , spelling errors and short forms.
- To overcome these issues, preprocessing of these answers is performed in multiple steps each solving a problem in the dataset.
- Without preprocessing might lead us to wrong classification of the data.
- If preprocessing is done correctly it can help in increasing the accuracy of the ML tasks.

Our implementation for preprocessing

- The data has 75 features including information about economic status, financial flow, etc. To prevent problems arising due to the curse of dimensionality, we conducted feature selection on the data.
- For this we used the well known embedded technique for feature selection - Random forest importance
- Before this, we had to fill missing cells in the data. Since the paper we took for reference didn't mention any specific method filling data, we decided to fill the values with mode of the value.
- After this, we ran the random forest importance and reduced the feature space from 75 to 51 with a threshold value of 0.06.

Distribution of work

Phase-1 work distribution:

- We have each read a research paper related to the topic and came up with the idea of what models to implement at the end.

Phase-2 work distribution :

- Text Preprocessing using Random Forest and SVM model Implementation was done by **Chaitanya Chakka**
- Logistic Regression Implementation was done by **K.S.V.L Druthi**
- Decision Tree Implementation was done by **D. Pranay Kumar**
- Random Forest Implementation was done by **V. Shashivardhan Reddy**
- Gradient Boosting classifier Implementation was done by **M. Sheethal Reddy**
- Vote Ensembling Implementation was done by **Vansh Madan**

Thank you