**Chapter 1**

**Introduction**

Mental Health can be defined as the way in which people react to the situations they face in life. These situations could be social, emotional or psychological. It adversely affects one’s productivity in professional life, can ruin relationships in personal life, and may end up being a problem to society rather than contributing towards its well-being. As such, it becomes important to understand the various facets of mental health issues, their symptoms and their causes. If we achieve a better understanding of these symptoms, their causes, and how they may be correlated to one another, we can provide some efficient solutions within the limitations of the technology available today or avoid these issues altogether. Diagnosis of mental health illnesses is hard as there is no objective metric to determine one’s mental health status. It is dependent on the subjective knowledge of the psychiatrist. “Prevention is better than cure”.

NMHS has revealed that every sixth Indian needs mental health help [1]. To effectively treat and diagnose mental health disorders, diagnosis of the disorder has to be done at an early age, for this reason the project focuses on the age group 5-24 years. As this age group is highly prone to mental illness and do not have the requisite life experience to handle the stress and anxiety. This generation plays a crucial role in the development of one’s personality, which in turn affects the productivity of the nation. This younger generation is responsible for carrying the nation to glory in the coming years. For the sake of our study we contacted a famous psychiatrist in North Karnataka, Dr. Anand Pandurangi, Shanti Dhama. Our project mainly focuses on the data obtained in North Karnataka with Collaboration of the doctor.

Karnataka has experienced a change in culture in recent years as rapid industrialization and urbanization has attracted many investors and entrepreneurs. All these factors have also caused a change in the mental health conditions experienced by the inhabitants of Karnataka [2]. Most commonly observed mental health conditions among the inhabitants are Schizophrenia which is typically accompanied by aggressive nature, suicidal thoughts, emotional instability and irritation. Another commonly observed condition was Depression, which is typically accompanied by loneliness, hopelessness, anxiety and restlessness. Such conditions are easily treatable today with the cooperation of the patients and provided the patient is diagnosed at an early age. If these disorders are not treated, they may get worse over time and lead to other disorders like substance abuse, or suicidal nature. WHO also revealed in a survey conducted in 2017 [4] that, mental health workforce in India (per 100,000 population) include psychiatrists (0.3), nurses (0.12), psychologists (0.07) and social workers (0.07). For a country with an ever-growing population of 1.3 billion, this seems to be too low. Importance must also be given to raising awareness for mental healthcare at a national level. National mental health policies must be in place which not only promote treatment of mental health disorders but provide overarching directions on broader issues for ensuring mental health promotion.

* 1. **Literature Survey**

This section explains the literature survey of the papers that are used for our Framework for automatic diagnosis psychological disorder. There have been many studies conducted by various organizations around the world to improve the mental health care. Such as prevalence of long-stay [3]. The data for this research was readily available from a pilot trial conducted by administrative systems. Their study concluded that a large number of patients who were admitted were either kept for longer than necessary or were placed in a higher security section than what would have been warranted. The technique used in predicting these was a two-way logistic regression with the security level as the level two analytical unit. The limitation of this study was that there was no diagnostic data for this study.

Ms.Sumathi M.R. &Dr. B. Poorna in [5] performed a prediction of Mental Health disorders among children using 8 machine learning techniques, This study focused mainly on the 5 most popular mental health illnesses amongst children - Attention Problem, Academic Problem, Anxiety Problem, Attention Deficit Hyperactivity Disorder (ADHD) and Pervasive Developmental Disorder (PDD), The factors, symptoms and various test results that are observed by the professionals are given as input to the techniques and the psychological problem diagnosed is retrieved as their output. Successful and accurate predictions were generated by 3 machine learning techniques viz. Multilayer Perceptron, Multiclass Classifier, and LAD Tree. The research was performed on a small dataset (n=60).

Various researchers in the USA and China studied data about misdiagnosis among patients and its effects in [6] and attempted to provide an intelligent solution to the same by introducing a semi-autonomous way that aids in the preliminary diagnosis of the patients. They achieved the same by integrating the technologies of the genetic algorithm, classification data mining, and machine learning. The genetic algorithm was used here to extract the information regarding the mental illness from the symptoms entered. This study focuses on the mental health issues illustrated in ‘Diagnostic Statistical Manual IV, Fourth Edition Text Revision (DSM-IV-TR). Keywords from this textbook about each disease have been loaded into a database. A genetic algorithm has been developed to extract keywords from a patient’s symptoms (using BPEL), these keywords are matched with text in the database to extract a classification. The algorithm is tested on a set of sample data using the same random sample set of 15 symptoms with 600 words each. The algorithms produced a better fitness on average compared to the K-Means Algorithm (Avg. Fitness of 1.569 compared to Avg. Fitness of 1.678 of K-Means). The limitation of this research is that the study has concentrated on a sample of 15 symptoms.

In [7] multi-criteria decision support method called MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique) is used. This study emphasizes on the difficulty in diagnosis of the mental disorder and its severity, and tries to improve the quality of the diagnosis, focusing on the Obsessive Compulsion Disorder, MACBETH was used to generate set of rules. The limitation of this study is that it is limited to OCD and a new expert system will have to be developed in order to diagnose other diseases.

In [8], to predict the mental health disorder the paper describes the use of machine learning technique called Bayesian network and Fuzzy Clustering. The dataset for this research is collected from a Psychologist, consisting 77 of children profile after pre-processing. Applying feature selection algorithm to select the relevant attributes a Bayesian Network was constructed using GeNie tool and the parameters were learned from the dataset. The model was validated with 2-fold cross-validation. This network is called the Expert Knowledge-based Bayesian Network (EKBN). Using GeNie tool, a BN has generated automatically from data (DBN). The predictive accuracies of EKBN and DBN are 48% and 58% respectively.

* 1. **Motivation**

The following factors motivated us to undertake this project:

* Impact of psychological illness:

As per WHO an estimated 2.6 million young people aged 10 to 24 years die each year due to psychological illness and much greater number of young people suffer from behavioural illnesses which hinder their ability to develop to their full potential. At least 20 percent of young people are likely to experience some form of psychological illness- such as Depression. Hence there is a need for more effective medical treatment which can prevent these complications.

* In a study conducted in India [5], according to it, India has highest suicide rate among people aged between 15 years and 29 years, about 30-40 Indians in every 100000 in this

age group kill themselves. The below chart gives a graphical representation of the suicidal rates.

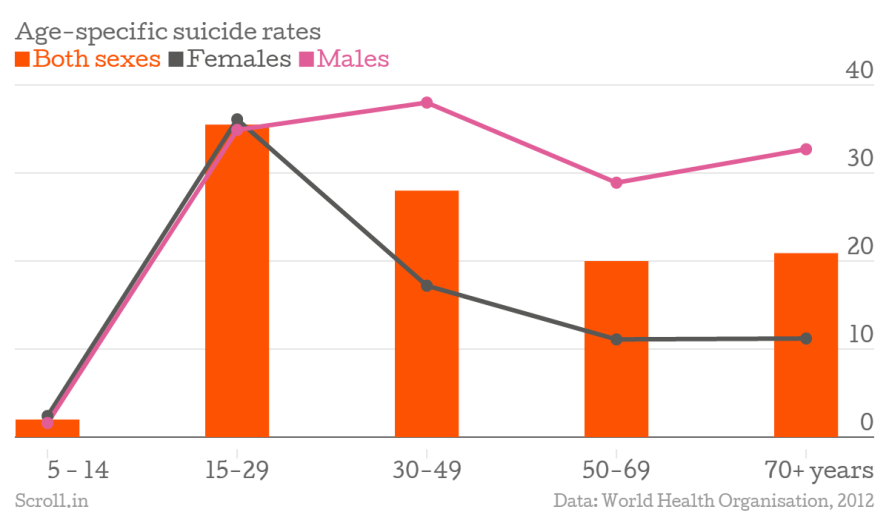


Figure 1: Age to suicide rate.

Figure 1 shows that the suicide rates are the highest in the age group of 15-29 among both genders.

**1.3 Problem Statement**

To develop a framework to automatically generate a patient's case card database, predict his/her psychological disorder and analyse the symptoms and causes.

**1.4 Objectives**

There are 3 main objectives set for this project:

•To automatically generate case card database based on patient’s symptoms specific to a psychologist.

•To Predict the patient’s psychological disorder.

•To analyse the patient’s psychological disorder based on influencing factors (Age, region, and symptoms).

**Chapter 2**

**Requirement Analysis**

This is a brief overview of requirement analysis of the project

**2.1 Functional Requirements**

The following are the functional requirement of the project:

* The user (receptionist) shall be able to enter the basic patient details in a Web Application.
* The doctor shall be able to view the basic patient details and make an entry of diagnosis and prescription in the tablet.
* The system shall generate the case card for each patient and store it in a database.
* The system shall extract keywords (symptoms) from the case card.
* The system shall show the patient disorder information and visit history to the doctor.
* The system shall build a Classification Model.
* The system shall predict the disorder.
* The system shall send the symptoms to model running on the cloud as the input and get the predicted disorder as output (which help the doctor in getting the disorders in patient).

**2.2 Non-Functional Requirements**

The following are the non-functional requirement of the project:

* The system shall able to accurately predict the mental illness 95% of the times.
* The system shall be quick and responsive, the system shall be able to predict the disease within 2 seconds.

**2.3 Proposed System Model**

There are 3 main modules in the system. Those are Processing, Model Building and Impact Analysis.

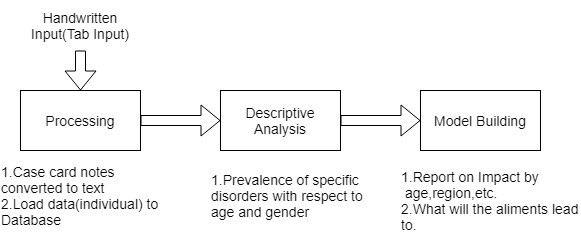


Figure 2: Proposed system model.

The figure 2 shows the process involved in the system. The processing module takes the input of the handwriting. Which is processed and converted into text and loaded into the database. The data is then preprocessed. Based on overall data collected from the above process the system will provide descriptive analytics about the occurrence of the psychological disorder in the patient based on gender, region, and age-groups. The last module, Model Building is done by extracting the features from the data in database. Features include the symptoms the patient is suffering from. And the prediction is made on this pre-processed data.

The input to the system is the data extracted from the case cards that the doctor produces by interviewing the patient, this data is fed into the classifier model which in-turn gives the possible disorders with respect to the input.

**2.4 Data Description**

The data is collected from Psychology Clinic: Shanti Dhama, Dharwad. The data collected from the clinic contained the diagnoses and comments written by the psychiatrist. A set of symptoms were identified from these comments. A dataset was generated with the aforementioned symptoms as attributes, along with preliminary details like age, gender and occupation. The number of attributes in the data was 42 and 350 patient details were collected.

Each tuple in the dataset represents a patient and the symptoms observed by the Psychiatrist are represented in binary form, 0 for absent and 1 for present.

The attributes in the dataset are shown in Table 1.

Table1: Data Description.

|  |  |
| --- | --- |
| Attributes | Description |
| Case no  Age  Gender  Place  Occupation  Issues  Depression  Dull  Self-talking  Aggression  Irritated  Suicide attempts  Attention seeker  slow learner  physically weak  anxiety  panic attack  impulsive  emotionally disturbed  lazy  lost interest to studies  Pressure  Laugh to much  Sleeplessness  loose balance  negative thinking  repeat same action  shy  restlessness  fear  forgetness  abusive language  headache  epilepsy  cry  tension  worry  wandering tendency  physical pain | ID of the patient  Numeric: ideally from 5 to 24  Binary: 1 for male; 0 for female  String: patient’s native region  Nominal: work of patient or standard if patient is a student  Nominal: problems faced by the patient due to his mental illness  Binary: facing depression or no depression (1 or 0)  Binary: acting very dull or not dull (1 or 0)  Binary: patient self-talks or doesn’t self-talk (1 or 0)  Binary: patient has wild behaviour or normal (1 or 0)  Binary: patient gets angry often (1 or 0)  Numeric: no of times the patient tried to commit suicide  Binary: patient try hard to get people’s attention (1 or 0)  Binary: patient is a slower learner or no (1 or 0)  Binary: patient’s physical condition weak or not weak (1 or 0)  Binary: patients gets anxious (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: patient unable to stand properly or not (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: patient talks in rude, insulting language or not (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0)  Binary: yes or no (1 or 0) |

Table 1 show the attribute of the dataset.

**2.4.1 Data Pre-processing**

The initial data collected from the local clinic had many categorical data. The initial step was to convert the handwritten raw text into the form a data frame, or in other words, a systematic arrangement in a table. This was done by identifying a few words which would later be mapped into the data frame. For the diagnosis of each patient, as a pre-processing step, we mapped the categorical data to a binary data with “yes or no (1 or 0)” as shown in table 1 by extracting the keywords from the diagnosis written by the doctor. Some of the attributes of the dataset were blank as there was no information for those in the source data, so the blank tuples were replaced with Null.

For a relatively smaller size of data (N=350), the highly correlated features can produce harmful bias in the data. This can subsequently affect the performance of the predictive model. Pearson’s correlation heat map revealed some highly correlated features like ‘suspiciousness’ and ‘delusion’ with a Pearson’s correlation of 0.75.

A statistical technique called multicollinearity analysis was adopted to tackle this issue. One feature is predicted from the other features with a substantial degree of accuracy in the technique. A quantity called the Variance Inflation Factor (VIF) for each feature in the predictive model was calculated. It is the ratio of variance in a model with multiple features, divided by the variance of a model with one feature alone.

Suppose there is a dependent variable ‘Y’, with independent variables, X1and X2.

So, the regression equation is as shown in Equation (2.1):

………………………........ (2.1)

Where is the change in Y for a 1-unit change in, while is held constant

Where is the change in Y for a 1-unit change in, while is held constant.

It is required to find VIF, the formula to find the same is shown in Equation (2.2).

VIF= ……………………………………….......... (2.2)

Where Ri is the coefficient of determination obtained by regressing Xi over all the other independent variables.

For instance, R1 is obtained by regressing X1 over all other independent variables, as shown in Equation (2.3).

X1= α0 + α1 X0 + α2 X2 + ε ……………………………… (2.3)

For severely correlated features a very high VIF score of 5 or higher was obtained. All the features which produced VIF of higher than 5 were eliminated in the model.

Table 2: VIF factors of the first few features.

|  |  |  |
| --- | --- | --- |
|  | VIF Factor | Features |
| 1 | 3.59 | Depression |
| 2 | 8.545 | Dull |
| 3 | 4.84 | Self-talking |
| 4 | 5.86 | Irritated |
| 5 | 3.75 | Attention seeking |
| 6 | 4.64 | Slow learner |
| 7 | 8.84 | Suspiciousness |
| 8 | 5.05 | Physical weak |
| 9 | 6.53 | Anxiety |
| 10 | 4.26 | Impulsive |
| 11 | 6.34 | Wandering tendency |
| 12 | 7.95 | Delusional |
| 13 | 9.784 | Emotionally disturbed |

Table 2 shows the VIF factors obtained for the first few features. All the features with VIF factor of above 5 were eliminated. For instance, ‘Emotionally disturbed’ has a VIF Factor of 9.784. Since this value is above 5, we have eliminated ‘Emotionally disturbed’ from the feature space. The resulting data had 21 features left.

**Chapter 3**

**Design**

This chapter provides information about the specification of the design. Each of the modules used in the project are explained in brief.

**3.1 System Description**

System description is a high-level overview of the working of the overall system. The system starts with the creation of the case card for each patient with his/her unique ID. The doctor can access these patient IDs from his end via a web application on an Apple iPad Pro. All the data about the patient is stored in a cloud database including the diagnosis of the patient (the data is stored in the database in a predetermined format as mention in the above data description). Using the data, a predictor model makes a prediction of the mental disorder based on the symptoms observed in the patient.

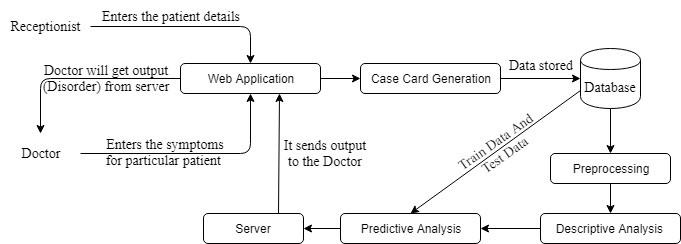


Figure 3: Overview of the Project.

The Figure 3 explains about the way the case card is going to be generated for a patient. The receptionist enters the preliminary details of the patient and this is stored in the database. The doctor then accesses the case card of the patient and enters his diagnosis and prescription. Meanwhile, a predictive model running on the cloud, performs real time feature extraction and suggests the predictive output as a possible diagnosis to the doctor.

**3.2 Hierarchical Design**

Hierarchical Design shows the level-wise flow of the project. There are three levels in the hierarchical model. Initially the Web Application which is the important component of collecting the data and representing the result. It is followed by the second layer, which is the data layer which stores the data required for the application. The next layer, accessed this data and performs predictive analyses. And finally, the output is displayed to the user on the Web Application.

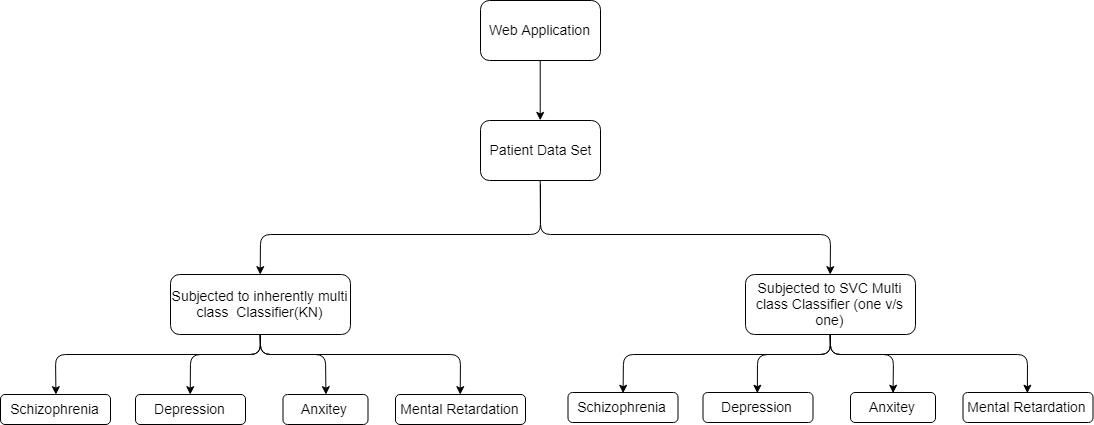


Figure 4: Hierarchical Model.

The proposed system can be thought of as a hierarchical model as shown in the above figure 4. At the topmost layer, we have the web application with which most of the interaction between the system and users happens. After the doctor completes the patient’s case card the dataset gets updated with the new entry. The patients’ data is subjected to two types of classifiers which are:

1. Inherently multiclass classifier.
2. One vs. One multiclass classifier.

These classifiers classify the patient into one of Depression, Schizophrenia, Mental Retardation or Anxiety Disorder.

**3.3 Sequence Diagram**

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.

There are two different procedures in the system for which Sequence Diagram can be defined, they are as follows:

* Receptionist interaction with the server in Figure 5.

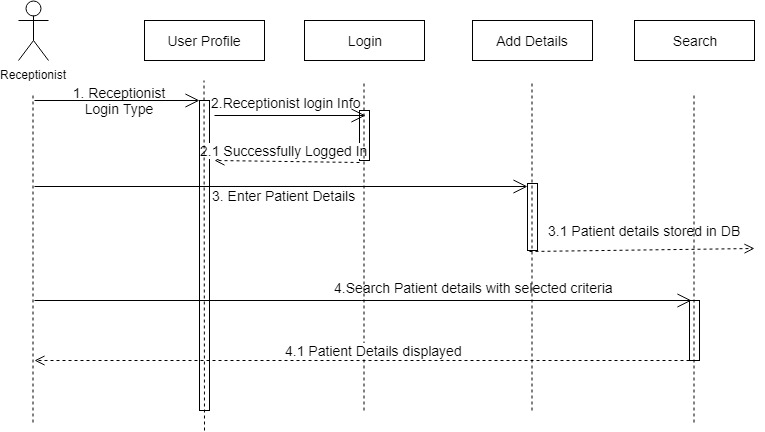


Figure 5: Receptionist Sequence Diagram.

Figure 5 shows, the interaction of the Receptionist with the Web Application. The Receptionist first logs in to the application and for each new patient enters the preliminary patient details and these details are stored in the database. The Receptionist can access the patient details from the database and assign appointment fixtures to the patients.

* Doctor’s interaction with the server, that is to diagnose append the patient results to DB in Figure 6.

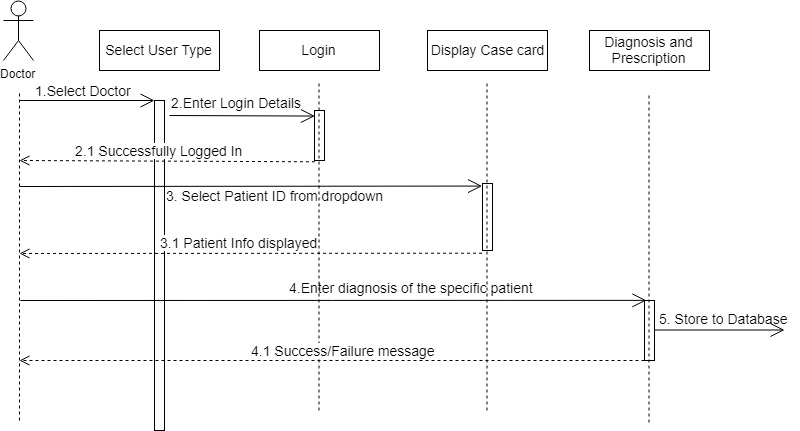


Figure 6: Doctor’s Sequence Diagram.

The Figure 6, shows the Sequence of interactions of the Doctor with the Web Application. The Doctor logs in to the system. Once he is logged into the system, he is displayed a list of patients who have appointments on the particular day. The Doctor then selects the patient and then accesses their case cards and performs diagnosis and prescribes medicine. And these details are then stored in the database.

**3.4 Entity Relationship diagram**

The entity relationship shows the relation between the individual entity and their attribute. It also shows relationship with other entities and the cardinality of the relationship.

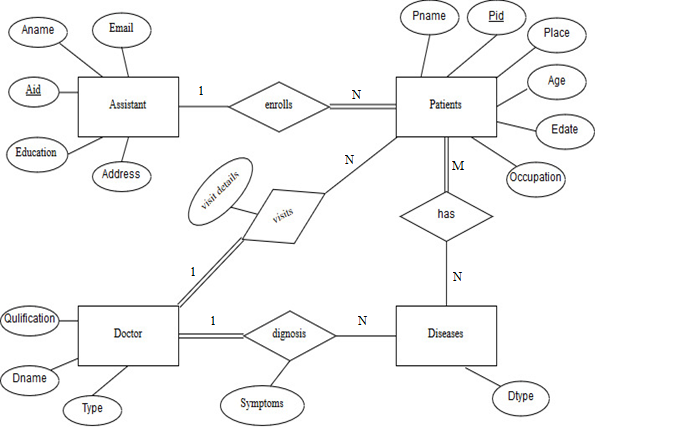


Figure 7: Entity Relationship Diagram.

Figure 7 shows the various entity, relationship between the entity and the attributes of the different entity. The main entities in the database are, ‘Assistant’, ‘Patient’, ‘Doctor’, ‘Disease’. The assistant enrols the patient. Patient visits the Doctor. Doctor diagnoses the disease which the patient has.

**3.5 System Functionalities**

The functionalities of the system are split into the following components:

**3.5.1 Web Application**

The Web Application is the key component of the project. It displays the results and also it provides the interface for collecting data. Its major functionalities are as follows:

* Receptionist creates a case card of the patients with preliminary details of the patients and it gets stored in a MongoDB database.
* The doctor accesses the details of the patients from the same web application to append the diagnose of the patients.
* Fetch the results obtained from the trained model from the server and display the result to the doctor.

**3.5.2 Selecting Relevant Attributes**

The data collected from the Web Application undergoes the pre-processing algorithms to determine only the ones which are relevant to the model. Its functionalities are as follows:

* Input: CSV file containing tuples of all the attributes refer table 1.
* Processing: Computed VIF (variance inflation factors) using multicollinearity.
* Output: attributes having VIF less than 5.

**3.5.3 Building K-Neighbours classifier**

It is one of the models built to classify the symptoms into a disorder, the functionalities of this model are:

* Input: CSV file as input which consist of patient details and a column to predict.
* Processing: features such as mean and standard.
* Output: Predicting the mental disorder of the patients when the symptoms are provided as input.

**3.5.4 Prediction by multiclass SVM classifier**

It is one of the models built to classify the symptoms into a disorder, the functionalities of this model are:

* Input: CSV file as input which contains the patient details and a column to predict.
* Processing: tuning the parameters like ‘kernel’.
* Output: Predicting the mental disorder of the patients when the symptoms are provided as input.

**Chapter 4**

**Implementation and Results**

**4.1 Web application and Database**

To create a web application interface for this project, we have made use of MEAN (MongoDB, Express.js, AngularJS, Node.js) which is an opensource JavaScript software stack for building dynamic websites and applications.

Mean stack includes creating backend server to handle the request sent from the front end angular2.By building APIs which handles the request sent by the services from the front end angular2 to the backend NodeJS managed by the express for different routing path.

MongoDB is used to store the data as it NoSQL database, where it is stored in the form of JSON object and performance won’t degrade for the large data.

After building the application it was deployed in the cloud service provider Heroku, the predictor model in the PythonAnywhere since most of the web APIs and libraries are supported as Platform as a service (PAS).

Using the web application, we are able to meet the requirements of our project

* To be able to generate a case card for a patient.

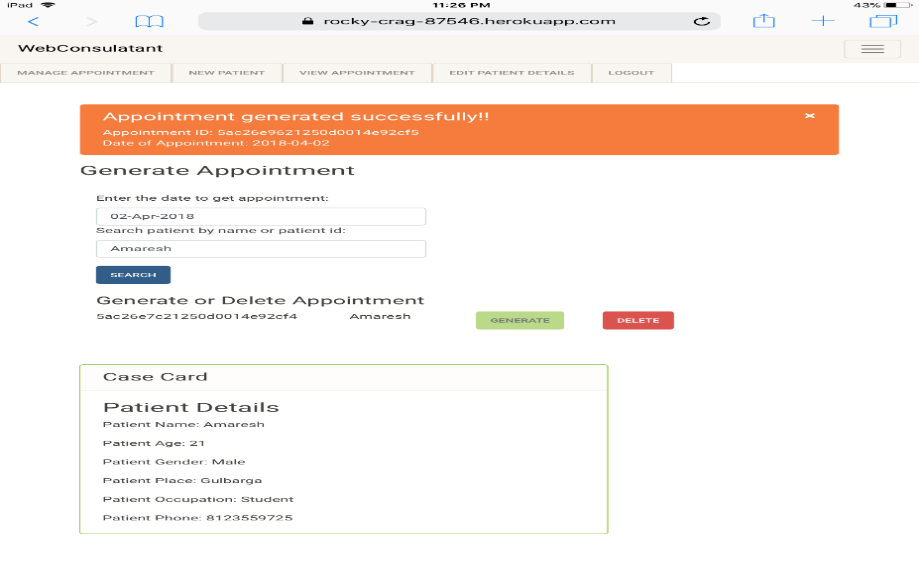
****

Figure 8: Generation of Case Card.

Figure 8 shows the interface of the web application using an iPad, the receptionist will be able to generate a case card for every new patient, and the receptionist will be able to assign every new patient a unique IDs and an appointment time as well as collecting the primary data of the customers.

* Doctor appends the case card with his diagnose of the patient, he can search for the patient having the appointment by the key ID.

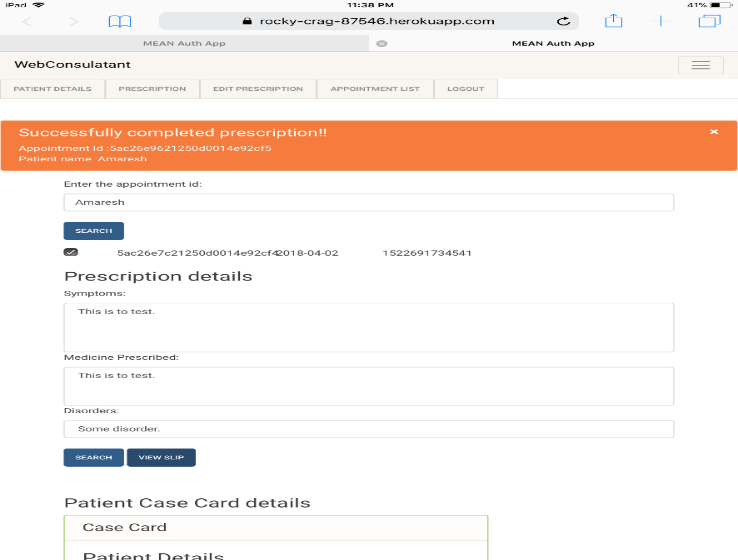


Figure 9: Doctor’s page.

In Figure 9 the doctor will be able to search for the patient who has an appointment ready, he is also able to append the case card with his diagnose about the patient’s symptoms and disorders.

* After the diagnosis of the patient by the doctor, the final form of the case card is generated which contain all the details of the patient like ID, Name, Age, Sex, Appointment Details, Symptoms, Disorder, and Prescription? This can be viewed by both the doctor and the receptionist Figure 10.

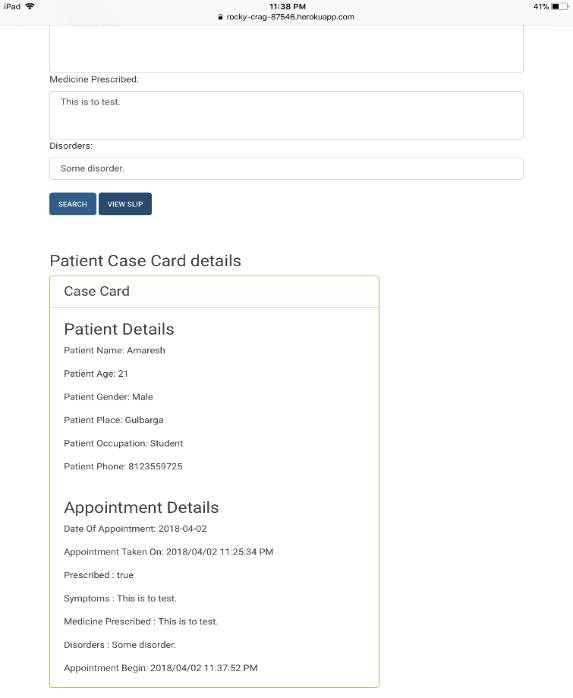
****

Figure 10: Display of Case Card.

**4.2 Descriptive Analysis**

Descriptive analysis used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with graphics analysis.

While we were studying the data that was available for our project, we made some analysis and found some trends in our data. The dataset which we had contained the patient details belonging to many kinds of different mental disorders. Within our data we found out that some of the disorders are more common like Depression, Schizophrenia, Mental Retardation, Seizure were found to have the most numbers of cases. Figure (11)

Furthermore, interesting trends about these top Disorders are discovered, which are backed by the already performed researchers in this domain.

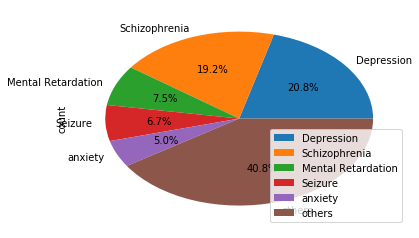
****

Figure 11: Pie chart showing the Disorder spread.

Depression as the most commonly observed psychological Disorder. [9] Has also indicated that Depression is the leading cause of disability worldwide.

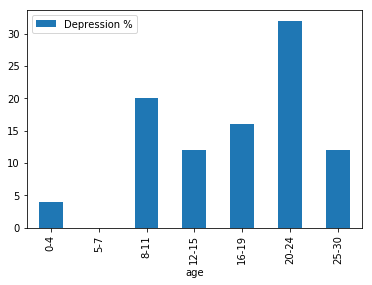


Figure 12: Depression vs. Age.

Figure 12 indicates that Depression is mostly observed in the early 20s (Young adults) [9] [10].

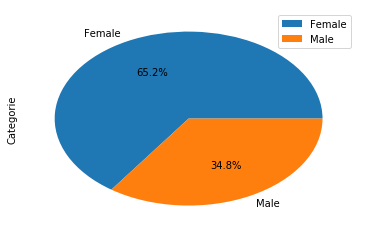
****

Figure 13: Pie chart of Male to Female case of Depression.

The Figure 13 shows that depression is more prevalent in females than in males this may be because of many factors specific to a biological functionality of the male and female [11].

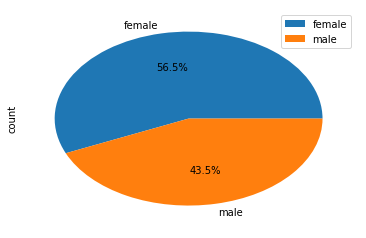


Figure 14: Pie chart of Male to Female case of Schizophrenia.

The Figure 14 shows that schizophrenia which is the second most common disorder in the dataset had a similar visualization to the Depression.

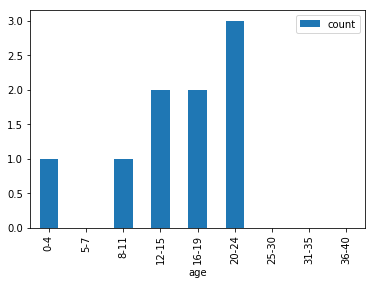


Figure 15: Bar graph of Age Distribution in Mental Retardation.

Mental Retardation begin in childhood or adolescence before the age of 18. And most commonly seen to persist throughout adult life in Figure 15.

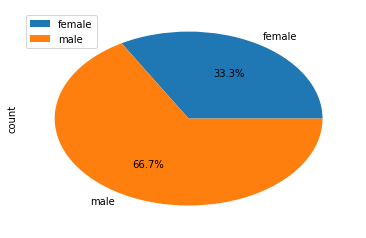


Figure 16: Mental Retardation among Male and Female.

The Figure 16 shows that mental retardation is found to be observed more in male than in females.

**4.3 Association Rule mining**

Association is a data mining function that is used to determine the co-occurrence of two or more items in a dataset. These are expressed by Association rules and the co-occurrence of these items is objectively determined using certain metrics like Support and Confidence.

Association Rules are used in the project to determine the co-occurrence between various symptoms for a specific disorder or the co-occurrence of various symptoms that lead to a specific disorder. Kulczynskci interestingness measure is used to determine the usefulness of the rules.

If Kulczynskci is near 0 or 1. This, can conclude that the association rule is interesting, it can be said that it is negatively or positively associated respectively. If Kulczynskci score is close to 0.5, the respective rule is considered to be skewed and uninteresting with respect to the study.

Association can exist between two objects or more than two objects. A two-object association rule is represented as X->Y, which is read as X implies Y or X leads to Y. Similarly, a three-object association rule can be expressed as X->Y, Z or X, Y -> Z.

Rules mined with minimum support of 20% and minimum confidence of 60%.

Table 3: Two symptoms rules for Schizophrenia

|  |  |  |  |
| --- | --- | --- | --- |
| SL NO | Two symptom rules for Schizophrenia | Conf | Kulc score |
| 1 | self-talking 🡪 aggression | 0.87 | 0.78 |
| 2 | impulsive 🡪 aggression | 1.00 | 0.65 |

|  |  |  |  |
| --- | --- | --- | --- |
| SL | Two symptom rules for depression | Conf | Kulc |
| 1 | low appetite 🡪 panic attack | 1.00 | 0.90 |
| 2 | anxiety 🡪panic attack | 1.00 | 0.65 |

Table 4: two symptom rules for Depression

Table 5: Three symptoms rules for Schizophrenia.

|  |  |  |  |
| --- | --- | --- | --- |
| SL NO | Three symptom rules for Schizophrenia | Conf | Kulc score |
| 1 | irrelevant behaviour 🡪 self-talking, suspiciousness | 0.7 | 0.75 |
| 2 | lazy, aggression 🡪 abusive language | 1.00 | 0.75 |

Table 6: Three symptom rules for Depression.

|  |  |  |  |
| --- | --- | --- | --- |
| SL NO | Three symptom association rules for Depression | Conf | Kulc score |
| 1 | low appetite 🡪 panic attack, sleeplessness | 1.00 | 1.00 |
| 2 | aggression, sleeplessness 🡪 fear | 1.00 | 1.00 |

**4.4 Predictive Model**

The data is split using 80:20 split for training and testing respectively. Models are trained to predict the disorder based on the symptoms. The models are evaluated using Precision, which represents the ratio of predictions that turn out to be positive in the group of records that were predicted to be positive by the model, precision is sometimes also referred to as the positive predictive value (PPV). The second metric used is Recall. Recall is a measure of actual true positive rate, which is the ratio of correct instances among all instances that actually belong to the relevant class. The third metric is F1-score also called as f-measure. F-measure is the harmonic mean of precision and recall. An ideal model would produce a F-measure of 1. GridSearchCV is used to find the optimal parameters for the model.

The first model used in our study is the KNeighborClassifier. It is a non-generalized learning model, sometimes also called as an instance-based learning model. It does not try to develop an internal statistical mode like in Linear Regression with a generalized cost function, but simply stores instances of the training data. These instances of the data are assigned to a particular class using one of three algorithms: Brute Force, K-D tree, Ball-tree. The algorithms used in the study is the Brute force algorithm which utilizes Euclidean distance to calculate the distance between the point of interest and to all training set where for n samples in d dimensions. Classification is computed from a simple majority vote of the nearest neighbours of each point: a query point is assigned the data class which has the most representatives within the nearest neighbours of the point.

https://hlab.stanford.edu/brian/euclid2.gif………………………………. (4.1)

Euclidean distance between two points i and j in an n-dimensional Euclidean plane is given by the equation above. This model is an inherently multiclass classifier which suits the type of classification for our study. This model yielded a precision of 0.58, recall of 0.77 and F-1 of 0. 657. The second model considered in the study is the LogisticRegressionCV. This model is an implementation of logistic regression using bilinear, newton-cg, and sag of lbfgs optimizer. Also, this model is cross validated by default which can help tackle the problem of overfit or underfit. Under fitting is a particular problem for smaller datasets such as the one considered in this study. Also, Logistic Regression based models consider the entire feature vector during training. The model produced a recall of 0.90, precision of 0.73 and F-1 Score of 0.80. The third model used for the study is the Gradient Boosting Mode. GB is a forward stage-wise learning model; one of GB’s greatest advantages is that this can use the arbitrary loss functions for each stage in the GB algorithm.

The following equation depicts the typical loss function called squared error loss function:

h(x) =………………………. (4.2)

Here y is expected value and (x) is the predicted value h(x) is the loss which gradient descent will minimise. Here every stage will correct its predecessor hence the final h(x) loss will be reduced. In each stage, a select number of regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. The model produced a precision of 0.72, recall of 0.8 and F-1 Score of 0.757. LogisticRegressionCV seemed to be the best model obtained with the current amount of data available, and the same is used in the PythonAnywhere cloud which sends real-time prediction to the doctor.

Table 7: Model Performances

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Recall | Precision | F-1 Score |
| KNeighborClassifier | 0.762 | 0.586 | 0.657 |
| LogisticRegressionCV | 0.897 | 0.730 | 0.804 |
| GradientBoostClassifier | 0.8 | 0.722 | 0.757 |

Table 6 shows the performance of each model used with respect to their Precision, recall and F-1 Metric.

**Chapter 5**

**Limitations**

The limitations of our project are as follows:

1. A limited amount of data is available.
2. Collection of Data is time-consuming and tedious.
3. The project focusses on 4 disorders only, due to the insufficiency of data available to reliably predict all the disorders observed.
4. Model is built on static data, model needs to be rebuilt when data is increased.
5. Dependency on a reliable internet connection to query prediction from the model.
6. Technology knowledge required to operate the system.
7. Disorders which are not among the four common disorders are still incorrectly classified into one of the four.

**Chapter 6**

**Conclusion**

We worked on a system which required the interaction of multiple disciplines of Computer Science from Data Mining, Cloud Computing to Web Application Development. A better understanding of the importance of Psychological Health and the impact of the disorders was obtained which was lacking prior. Hands on experience on new and modern computing technology like AngularJS, MongoDB, and Express was obtained. Due to platform dependency for different services, namely the predictive model and web application, we had to learn REST APIs to build a communication channel for passing messages between the two clouds.  A statistical technique called Multicollinearity was used to reduce the feature space. On this reduced feature space, three machine learning models were deployed to classify the disorder. In the future, we would like the application to be more interactive and the predictive model to be dynamic. We would also like to work on improving the prediction model to classify more disorders and also their severity.

**Chapter 7**

**References**

[1]Nimhansreport[www.nimhans.ac.in/sites/default/files/u197/NMHS%20Report%20%28Prevalence%20patterns%20and%20outcomes%29%201.pdf](http://www.nimhans.ac.in/sites/default/files/u197/NMHS%20Report%20%28Prevalence%20patterns%20and%20outcomes%29%201.pdf)

[2] The Hindu. (2018). *Seminar on economic development and socio-cultural change today*. [online] Available at: http://www.thehindu.com/todays-paper/tp-national/tp-karnataka/Seminar-on-economic-development-and-socio-cultural-change-today/article16006930.ece [Accessed 19 Feb. 2018].

[3] Bogic, M., Njoku, A. & Priebe, S. BMC Int Health Hum Rights (2015) 15: 29. https://doi.org/10.1186/s12914-015-0064-9

[4] South-East Asia Regional Office. (2018). *Mental health in India*. [online] Available at: http://www.searo.who.int/india/topics/mental\_health/about\_mentalhealth/en/ [Accessed 4 Jun. 2018].

[5] Ms. Sumathi M.R, Dr. B. Poorna, “prediction of mental health among children using machine learning” International Journal of Advanced Computer Science and Applications, 2016.

[6] GhassanAzar, Clay Gloster,” Intelligent Data Mining and Machine Learning for Mental Health Diagnosis using Genetic Algorithm”,2014

[7] Luciano Comin Nunes, Placido Rogério Pinheiro” An expert system applied to the diagnosis of psychological disorders”, IEEE, 2009

[8] <https://scroll.in/article/694364/four-charts-show-why-indias-youth-suicide-rate-is-among-the-worlds-highest,March,2018>, March, 2018

# [9] Pillai A, Andrews T, Patel V. “Violence, psychological distress and the risk of suicidal behavior in young people in India”,Int J Epidemiol. 2009 Apr

# [10] Sumathi M.R and PoornaB“Prediction of Mental Health Problems Among Children using Fuzzy Clustered Bayesian Model”, International Journal of Control Theory and Applications

# [11][https://adaa.org/about-adaa/press-room/facts-statistics#](https://adaa.org/about-adaa/press-room/facts-statistics), March,2018

# [12] <https://www.childtrends.org/indicators/young-adult-depression/>, March 2018

# [13[] https://www.medscape.com/viewarticle/871949](file:///C:\Users\user\AppData\Local\Temp\Rar$DIa15324.9785\%5d%20https:\www.medscape.com\viewarticle\871949), March 2018

# [14] [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4478054](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4478054/), March 2018

# [15] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2563255/>, March 2018