A Project Report

on

Using Data Science To Calculate Accuracy of Online Depression Detection Questionnaire

Submitted in partial fulfillment of the

requirement for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

Under The Supervision of
Mr. Hradesh Kumar :
Assistant Professor

Submitted By

Somesh Bachani
18SCSE1180018

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
10, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project, entitled **"USING DATA SCIENCE TO CALCULATE ACCURACY OF ONLINE DEPRESSION DETECTION QUESTIONNAIRE"** in partial fulfillment of the requirements for the award of the **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE and ENGINEERING** submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of month, **JULY-2021 to DECEMBER-2021**, under the supervision of **Mr. HRADESH KUMAR, Assistant Professor, Department of Computer Science and Engineering**, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the project has not been submitted by me/us for the award of any other degree of this or any other places.

18SCSE1180018 SOMESH BACHANI

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

(Mr. Hradesh Kumar, Assistant Professor)

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of **SOMESH BACHANI: 18SCSE1180018** has been held on _____ and his work is recommended for the award of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE and ENGINEERING**.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: DECEMBER, 2021

Place: Greater Noida

Abstract

Depression is a common mental disorder. Globally, it is estimated that 5.0% of adults suffer from depression.Depression is a leading cause of disability worldwide and is a major contributor to the overall global burden of disease. Major depressive disorder (MDD), the clinical term for depression, is one of the most common mental health conditions, affecting an estimated 350 million people in all age groups. So there exist online tools for diagnosing depression. One such questionnaire for Depression is the Patient Health Questionnaire (PHQ-9). This Project will use data science to find out how accurate these questionnaires are based on questionnaire results, age, sex, medical illness history and diagnosed for depression by physician(Target) our goal is to find out if the questionnaire should be the first step in the roadmap to seek medical help. This project will be using various machine learning tools and functions to help find accuracy of online Depression detection systems such as KNN, SVM, DT and Logistic regression.The expected final outcome for this project will be a statistical analysis of output of depression detection questionnaire to that of clinical diagnosed data. Future scope of the project includes calculation of accuracy for systems available in future and be a tool for improvisation.

Table of Contents

List of Tables

List of Figures

Acronyms

PHQ-9	Patient Health Questionnaire
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
MDD	Major depressive disorder
DT	Decision Tree
MIL	Mental Illness history

CHAPTER-1

Introduction

Depression is a common mental disorder. Globally, it is estimated that 5.0% of adults suffer from depression.Depression is a leading cause of disability worldwide and is a major contributor to the overall global burden of disease. Major depressive disorder (MDD), the clinical term for depression, is one of the most common mental health conditions, affecting an estimated 350 million people in all age groups. So there exist online tools for diagnosing depression. One such questionnaire for Depression is the Patient Health Questionnaire (PHQ-9). This Project will use data science to find out how accurate these questionnaires are based on questionnaire results, age, sex, medical illness history and diagnosed for depression by physician(Target) our goal is to find out if the questionnaire should be the first step in the roadmap to seek medical help. This project will be using various machine learning tools and functions to help find accuracy of online Depression detection systems such as KNN, SVM, DT and Logistic regression.

Formulation of problem

There are ways to detect depression through online means but in many cases they might give false positives or false negatives thus finding the accuracy of such a system will allow users to find peace of mind. This method of finding accuracy will use various classification methods to find accuracy and will cross validate each of them with one another. Calculation of accuracy of online depression detection questionnaires will determine the potency of online self help methods and will determine the usefulness of these questionnaires as the first step towards improving mental health.

Tools Used

Using module SKlearn various classification models are used they are:-

1. KNN
2. SVM
3. Decision Tree
4. Logistic Regression
5. Cross-validation(evaluating estimator performance)

KNN :- In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951 and later expanded by Thomas Cover.It is used for classification and regression. In both cases, the input consists of the k closest training examples in a data set. The output depends on whether k-NN is used for classification or regression.

SVM :- In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Decision Tree :- A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Logistic Regression :- In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model.

Cross-validation(evaluating estimator performance) :- Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called overfitting. To avoid it, it is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a test set X_test, y_test. Here is a flowchart of typical cross validation workflow in model training. The best parameters can be determined by grid search techniques.

CHAPTER-2

Literature Survey

For the reading and referring we have referred to a paper published in National House of medicine : Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. Using this as a source I have tried to find out about the characteristics used for detecting depression and which of them affects the output most. I have tried to balance out the criteria required for calculating the output. According to the paper, eligible studies were compared to PHQ-9 scores with major depression diagnoses from validated diagnostic interviews. Primary study data and study level data extracted from primary reports were synthesized. For PHQ-9 cut-off scores 5-15, bivariate random effects meta-analysis was used to estimate pooled sensitivity and specificity, separately, among studies that used semi structured diagnostic interviews, which are designed for administration by clinicians; fully structured interviews, which are designed for lay administration; and the Mini International Neuropsychiatric (MINI) diagnostic interviews, a brief fully structured interview. Sensitivity and specificity were examined among participant subgroups and, separately, using meta-regression, considering all subgroup variables in a single model.

CHAPTER-3

Project Design

Dataset

Dataset is comprises of age, sex, questionaire result [result of online PHQ-9(patient health questionaire), mental illness history and target. This was collected via a google form to accept responses from multiple people between ages of 14-70. This dataset comprises data from over 650 people. Below is a sample data entry for a person's records.

Table 1 Dataset features distribution.

Age	Sex	Questionnaire result	Mental Illness history	Target
21	M	2	1	1

Above data tells us about a person who is 21 year old and is Male and scored 2 on questionnaire results and has mental illness history in the family and was diagnosed positively for having depression. Using data like this and with help of machine learning we will be able to find accuracy of online systems by cross validating their score with output.

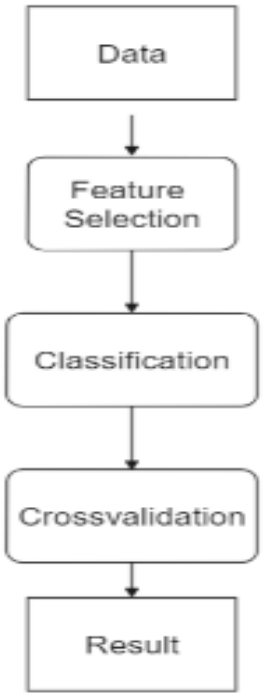
Here Dataset Feature Description Table for the mentioned Dataset.

Table 2 Dataset Feature Description

Feature	Description	Type
Age	Age in years	Numerical
Sex	1: Male 0: Female	Categorical
Questionnaire Result	0: No Depression 1: Mild 2: Moderate 3: Severe	Categorical
Mental Illness history	0: No 1: Yes	Categorical
Target	0: No (Person does not have diagnosed depression) 1: Yes (Person has diagnosed depression)	Categorical

Architecture Design

We have applied four Machine Learning Techniques on the dataset : Logistic regression, Support Vector Machine, Neighbors and Decision Tree. Three performance measures were calculated: Accuracy, AUC and F1 Score using the True Positives, True Negatives, False Positives, and False Negatives parameters, that were extracted from the confusion matrices. Our methodology is described as follows: We take each technique, then we apply it on our custom dataset, then we extract the confusion matrices, and calculate the performance measures. Finally we compare the results. The steps followed by our methodology are shown as follows in the following Figure:-
Fig 1



The architecture diagram includes :-

- Data
- Feature selection
- Classification
- Cross Validation
- Result

Methodology

The methodology consists of detailed steps they are:-
1. Using Borutapy for Feature Selection.
2. Performing Regression using Multiple Models and Comparing them.
3. Performing Cross validation to compare all of them.

Using Borutapy for Feature Selection

Feature selection is one of the most important steps in machine learning. When inputting features into a model, the goal is to feed the model with features that are relevant for predicting a class. Including irrelevant features poses the problem of unnecessary noise in the data, resulting in lower model accuracy. Generally, we use statistical-based feature selection methods such as ANOVA or the Chi-squared test, evaluating the relationship between each predictor variable with the target variable. With boruta, the features are curated down to the 'all relevant' stopping point and not the 'minimal optimal' stopping point. An all relevant variable being not redundant when used for prediction. Boruta is based on two ideas, shadow features and binomial distribution. When using boruta, the features are not evaluated with themselves but with a randomized version of them. For the binomial distribution idea, boruta takes a feature that we aren't aware of being useful or not and either refuses or accepts the feature based on three areas defined by selecting the tails of the distribution, 0.5 percent for an example.

- Area of refusal: area where features are dropped as they are considered noise.
- Area of irresolution: area where boruta is indecisive about the features
- area of acceptance: area where the features are considered predictive

Borutapy module lets us select features based on their weight. Helping us determine the most important features of a dataset based on their impact on target values. It uses fit, transform or , to run the feature selection. Borutapy is a feature selection method that relies on every criterion, while there exist other methods for feature selection they are minimal optimal; therefore it tries to select all features that carry information usable for prediction, rather than finding all the possible compact subset of features on which some of the classifier might have a minimal error.

Boruta methodology

- Begins by providing randomness to the features by creating duplicate features and shuffling the values in each column to remove their correlations with the response. (Shadow Features)
- Trains a random forest classifier on the dataset and calculates the relevance/importance by gathering the Z-scores.
- Finds the maximum Z-score among shadow attributes and assigns a hit to every attribute with a Z-score higher than the maximum Z-score of its shadow feature. (accuracy loss divided by the standard deviation of accuracy loss)
- Take each attribute that has not yet been determined important and perform a two-sided test of quality with the maximum Z-score among shadow attributes.
- Tags the attributes with an importance level lower than MZSA as ‘unimportant’.
- Tags the attributes with an importance level higher than MZSA as ‘important’.
- Removes all shadow attributes
- Repeats this process until the importance is computed for all attributes or the algorithm reaches the set number of iterations

Performing Regression using Multiple Models and Comparing them.

We have applied four Machine Learning Techniques on the dataset : Logistic regression, Support Vector Machine, Neighbors and Decision Tree.

Logistic Regression

Logistic regression is the right algorithm to start with classification algorithms. Even Though, the name 'Regression' comes up, it is not a regression model, but a classification model. It uses a logistic function to frame binary output model. The output of the logistic regression will be a probability ($0 \leq x \leq 1$), and can be used to predict the binary 0 or 1 as the output (if $x<0.5$, output= 0, else output=1).

Basic Theory :

Logistic Regression acts somewhat very similar to linear regression. It also calculates the linear output, followed by a stashing function over the regression output. Sigmoid function is the frequently used logistic function. You can see below clearly, that the z value is the same as that of the linear regression output in Eqn(1).

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots.$$

$$h(\theta) = g(z)$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

The $h(\theta)$ value here corresponds to $P(y=1|x)$, ie, probability of output to be binary 1, given input x. $P(y=0|x)$ will be equal to $1-h(\theta)$.

when the value of z is 0, g(z) will be 0.5. Whenever z is positive, $h(\theta)$ will be greater than 0.5 and output will be binary 1. Likewise, whenever z is negative, the value of y will be 0. As we use a linear equation to find the classifier, the output model also will be a linear one, that means it splits the input dimension into two spaces with all points in one space corresponding to the same label.

The figure below shows the distribution of a sigmoid function.

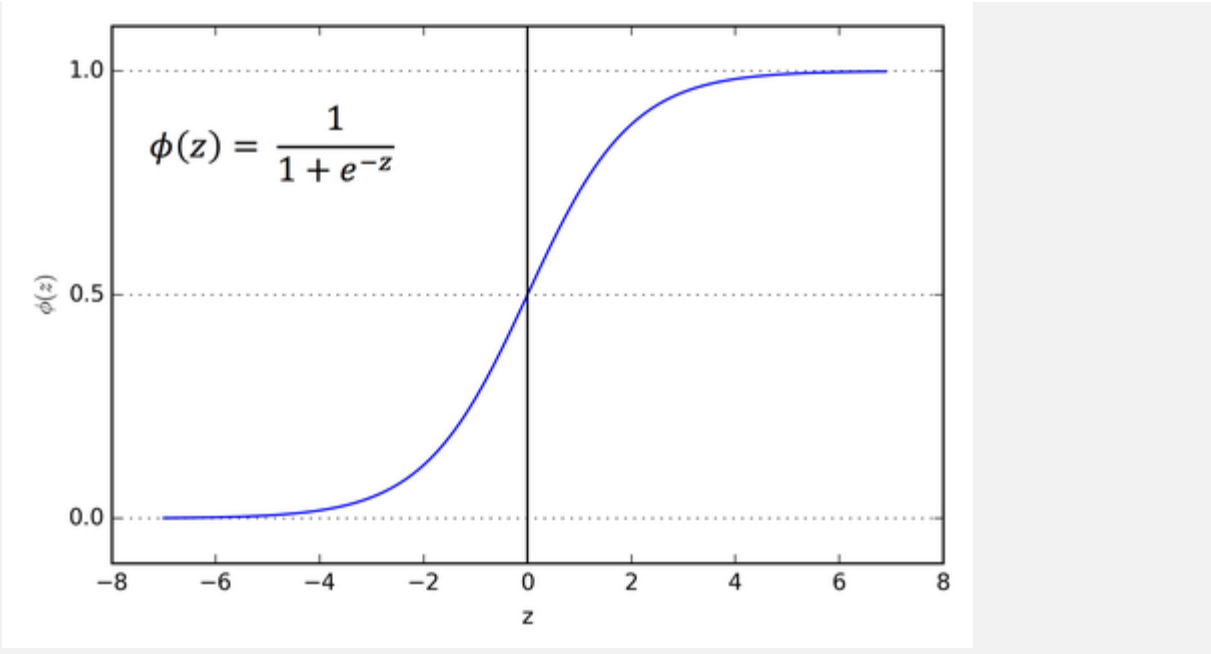


Fig 2 Sigmoid Function

Loss function :

We can't use mean squared error as a loss function(like linear regression), because we use a non-linear sigmoid function at the end. MSE function may introduce local minimums and will affect the gradient descent algorithm.

So we use cross entropy as our loss function here. Two equations will be used, corresponding to y=1 and y=0. The basic logic here is that, whenever my prediction is badly wrong, (eg : y' =1 & y = 0), cost will be -log(0) which is infinity.

$$J(\theta) = \frac{1}{m}\sum cost(y', y)$$

$$cost(y', y) = -log(1 - y') \quad if \ y = 0$$

$$cost(y', y) = -log(y') \quad if \ y = 1$$

In the equation given, m stands for training data size, y' stands for predicted output and y stands for actual output.

Advantages :

- Easy, fast and simple classification method.
- θ parameters explains the direction and intensity of significance of independent variables over the dependent variable.
- Can be used for multiclass classifications also.
- Loss function is always convex.

Disadvantages :

- Cannot be applied on non-linear classification problems.
- Proper selection of features is required.
- Good signal to noise ratio is expected.
- Collinearity and outliers tamper the accuracy of the LR model.

Hyperparameters :

Logistic regression hyperparameters are similar to that of linear regression. Learning rate(α) and Regularization parameter(λ) have to be tuned properly to achieve high accuracy.

Assumptions of LR :

Logistic regression assumptions are similar to that of linear regression models. please refer to the above section.

Comparison with other models :

Logistic regression vs SVM :

- SVM can handle non-linear solutions whereas logistic regression can only handle linear solutions.
- Linear SVM handles outliers better, as it derives maximum margin solution.
- Hinge loss in SVM outperforms log loss in LR.

Logistic Regression vs Decision Tree :

- Decision trees handle collinearity better than LR.
- Decision trees cannot derive the significance of features, but LR can.
- Decision trees are better for categorical values than LR.

Logistic Regression vs KNN :

- KNN is a non-parametric model, where LR is a parametric model.
- KNN is comparatively slower than Logistic Regression.
- KNN supports non-linear solutions where LR supports only linear solutions.
- LR can derive confidence levels (about its prediction), whereas KNN can only output the labels.

K-nearest neighbors

K-nearest neighbors is a non-parametric method used for classification and regression. It is one of the most easy ML techniques used. It is a lazy learning model, with local approximation.

Basic Theory :

The basic logic behind KNN is to explore your neighborhood, assume the test datapoint to be similar to them and derive the output. In KNN, we look for k neighbors and come up with the prediction.

In case of KNN classification, a majority voting is applied over the k nearest data points whereas, in KNN regression, mean of k nearest datapoints is calculated as the output. As a rule of thumb, we select odd numbers as k. KNN is a lazy learning model where the computations happen only at runtime.

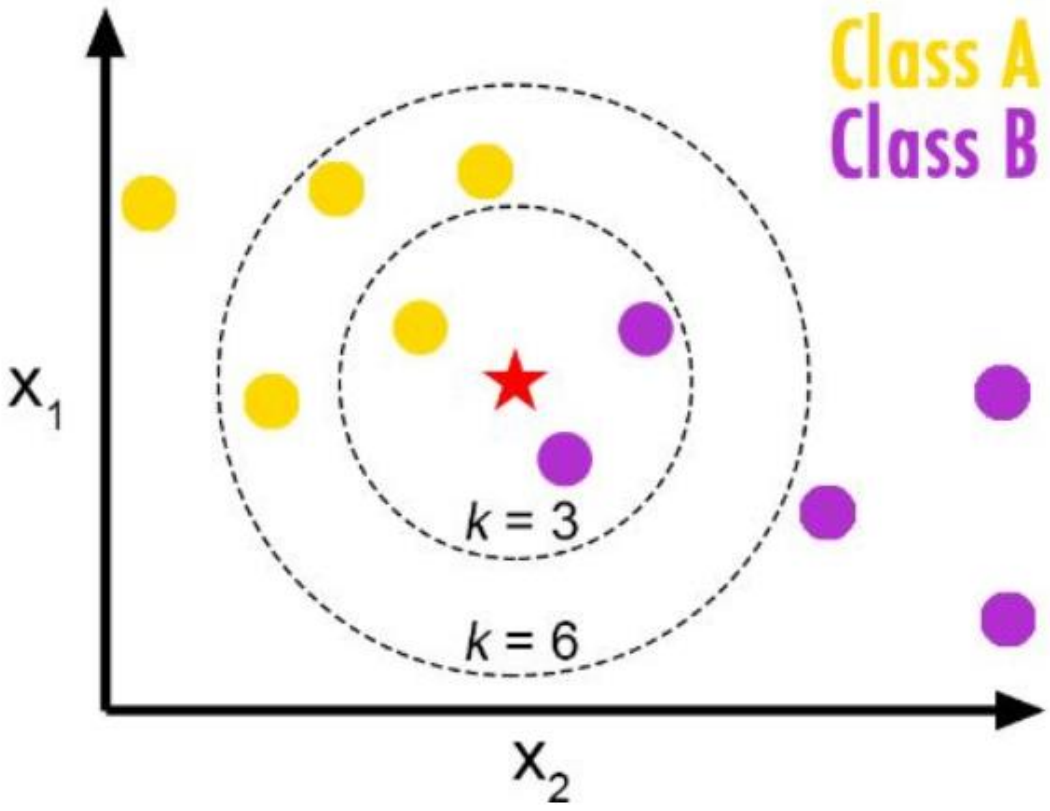


Fig 3 KNN graph

In the above diagram yellow and violet points correspond to Class A and Class B in training data. The red star points to the test data which is to be classified. when k = 3, we predict Class B as the output and when K=6, we predict Class A as the output.

Loss function :

There is no training involved in KNN. During testing, k neighbors with minimum distance, will take part in classification /regression.

Advantages :

- Easy and simple machine learning model.
- Few hyperparameters to tune.

Disadvantages :

- k should be wisely selected.
- Large computation cost during runtime if sample size is large.
- Proper scaling should be provided for fair treatment among features.

Hyperparameters :

KNN mainly involves two hyperparameters, K value & distance function.

- K value : how many neighbors to participate in the KNN algorithm. k should be tuned based on the validation error.
- distance function : Euclidean distance is the most used similarity function. Manhattan distance, Hamming Distance, Minkowski distance are different alternatives.

Assumptions :

- There should be clear understanding about the input domain.
- feasibly moderate sample size (due to space and time constraints).
- collinearity and outliers should be treated prior to training.

Comparison with other models :

A general difference between KNN and other models is the large real time computation needed by KNN compared to others.

KNN vs SVM :

- SVM takes care of outliers better than KNN.
- If training data is much larger than no. of features(m>>n), KNN is better than SVM. SVM outperforms KNN when there are large features and lesser training data.

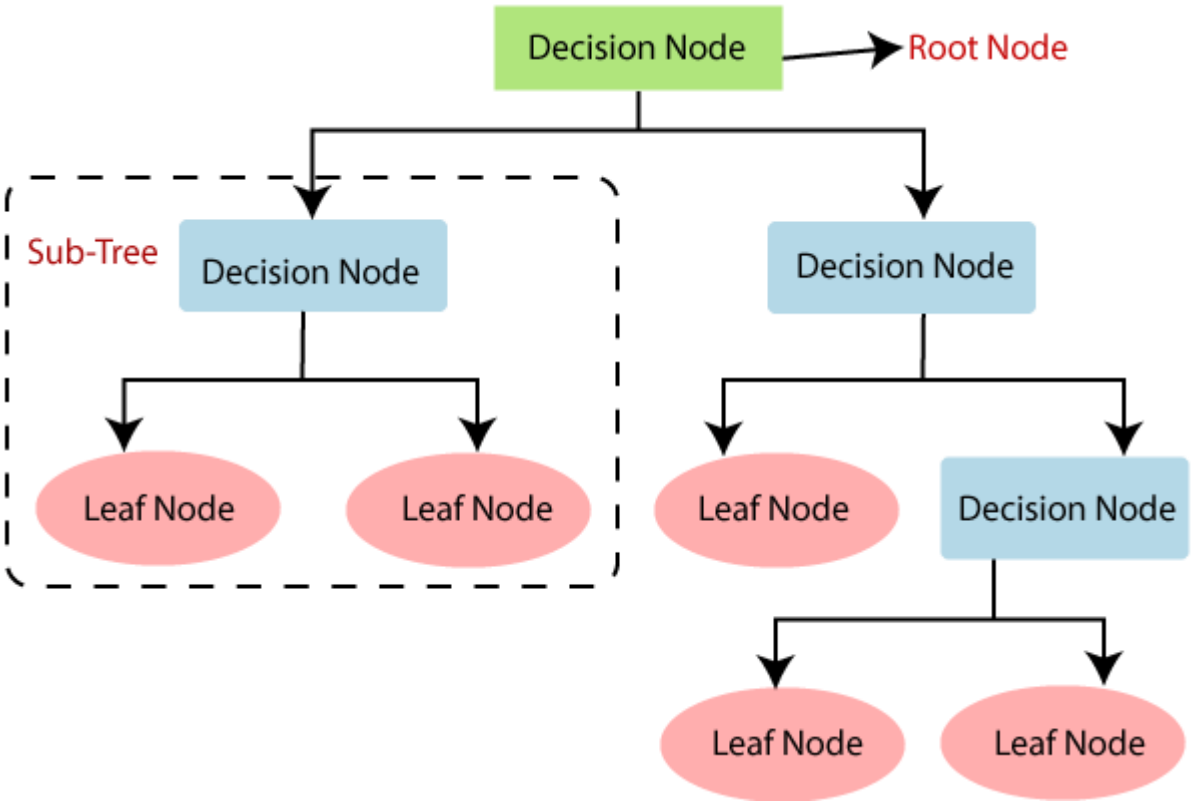
Decision Tree

Decision tree is a tree based algorithm used to solve regression and classification problems. An inverted tree is framed which is branched off from a homogeneous probability distributed root node, to highly heterogeneous leaf nodes, for deriving the output. Regression trees are used for dependent variables with continuous values and classification trees are used for dependent variables with discrete values.

Basic Theory :

Decision tree is derived from the independent variables, with each node having a condition over a feature.The node decides which node to navigate next based on the condition. Once the leaf node is reached, an output is predicted. The right sequence of conditions makes the tree efficient. entropy/Information gain are used as the criteria to select the conditions in nodes. A recursive, greedy based algorithm is used to derive the tree structure.

Fig 4 Decision Tree



In the above diagram, we can see a tree with a set of internal nodes(conditions) and leaf nodes with labels(decline/accept offer).

Algorithm to select conditions :
For CART(classification and regression trees), we use the gini index as the classification metric. It is a metric to calculate how well the data points are mixed together.

$$giniindex = 1 - \sum P_t^2$$

The attribute with maximum gini index is selected as the next condition, at every phase of creating the decision tree. When the set is unequally mixed, the gini score will be maximum.

- For Iterative Dichotomiser 3 algorithm, we use entropy and information gain to select the next attribute. In the below equation, H(s) stands for entropy and IG(s) stands for Information gain. Information gain calculates the entropy difference of parent and child nodes. The attribute with maximum information gain is chosen as the next internal node.

$$H(s) = -\sum P_c \cdot log(P_c)$$
$$IG(s) = H(s) - \sum_t P_t \cdot H(t)$$

Advantages :

- No preprocessing needed on data.
- No assumptions on distribution of data.
- Handles collinearity efficiently.
- Decision trees can provide an understandable explanation over the prediction.

Disadvantages :

- Chances for overfitting the model if we keep on building the tree to achieve high purity. Decision tree pruning can be used to solve this issue.
- Prone to outliers.
- Trees may grow to be very complex while training complicated datasets.
- Loses valuable information while handling continuous variables.

Hyperparameters :
Decision tree includes many hyperparameters and I will list a few among them.
criterion : which cost function for selecting the next tree node. Mostly used ones are gini/entropy.
max depth : it is the maximum allowed depth of the decision tree.
minimum samples split : It is the minimum node required to split an internal node.
minimum samples leaf : minimum samples that are required to be at the leaf node.

Comparison with other Models :

Decision tree vs KNN :

- Both are non-parametric methods.
- Decision tree supports automatic feature interaction, whereas KNN cant.
- Decision tree is faster due to KNN's expensive real time execution.

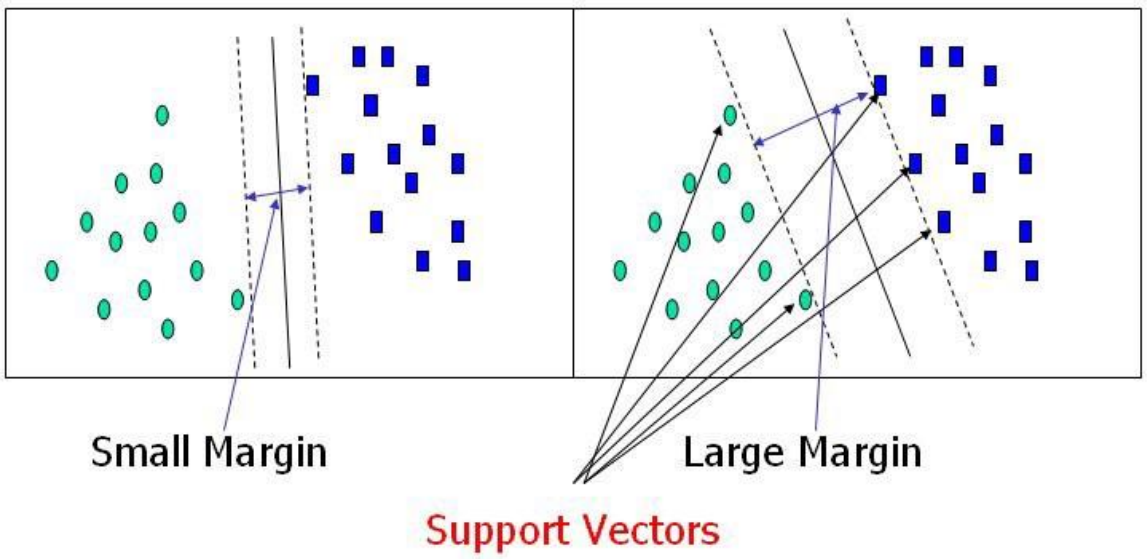
Decision tree vs SVM :

- SVM uses kernel tricks to solve non-linear problems whereas decision trees derive hyper-rectangles in input space to solve the problem.
- Decision trees are better for categorical data and it deals with collinearity better than SVM.

Support Vector Machine

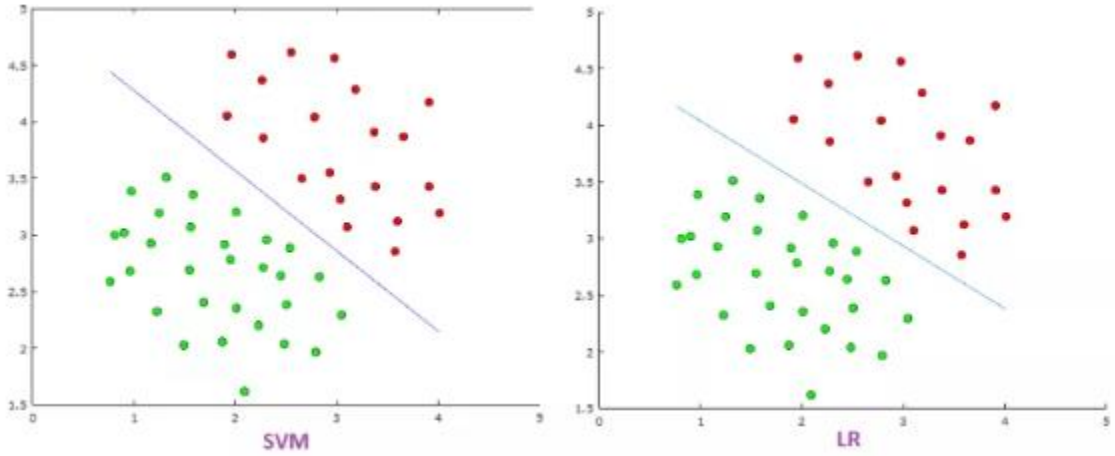
The support vector machine is a model used for both classification and regression problems though it is mostly used to solve classification problems. The algorithm creates a hyperplane or line(decision boundary) which separates data into classes. It uses the kernel trick to find the best line separator (decision boundary that has the same distance from the boundary point of both classes). It is a clear and more powerful way of learning complex nonlinear functions.

Fig 5 SVM margin comparison



Difference between SVM and Logistic Regression

- SVM tries to find the “best” margin (distance between the line and the support vectors) that separates the classes and this reduces the risk of error on the data, while logistic regression does not, instead it can have different decision boundaries with different weights that are near the optimal point.



- SVM works well with unstructured and semi-structured data like text and images while logistic regression works with already identified independent variables.
- SVM is based on geometrical properties of the data while logistic regression is based on statistical approaches.
- The risk of overfitting is less in SVM, while Logistic regression is vulnerable to overfitting.

Three performance measures were calculated: Accuracy, AUC and F1 Score using the True Positives, True Negatives, False Positives, and False Negatives parameters, that were extracted from the confusion matrices. Our methodology is described as follows: We take each technique, then we apply it on our custom dataset, then we extract the confusion matrices, and calculate the performance measures. Finally we compare the results.

Performing Cross validation to compare all the models

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
    3.1. Take the group as a hold out or test data set
    3.2. Take the remaining groups as a training data set
    3.3. Fit a model on the training set and evaluate it on the test set
    3.4. Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

CHAPTER-4

Results and Discussion

Using Borutapy for feature extraction determined that the most important feature was Questionnaire results and the Top Features Included Age, Sex, Questionnaire results and Mental Illness history. After that, I computed the score using various models and compared them. Table 3 Shows the relative scores Accuracy, AUC and F1 score.

Table 3: Model Comparison

	Accuracy	**AUC**	**F1 Score**
Logistic Regression	0.9333	0.957672	0.9333
KNN	0.9333	0.971892	0.926829
Decision Trees	0.9111	0.951058	0.900000
Support Vector Machine	0.9333	0.968695	0.9333

Logistic regression had a better F1 score compared to others with 93.33% Accuracy and 93.33% F1 Score.
Support Vector Machine had a Consistent average score with 93.33% Accuracy, 96.86% AUC and 93.33% F1 Score.
K Nearest Neighbors was the most accurate and had a better AUC score compared to others with 93.33% Accuracy and 97.18% AUC.

The Decision Tree Was the least Accurate with only 91.11% accuracy.

Upon closer inspection we find that most of the models give about the same accuracy with exception of the Decision Tree algorithm. There are minor differences in AUC scores and F1 scores. We compared different models using graphs.

In our study and in terms of performance, we conclude that KNN is the best model for Validation of Depression Detection questionnaire as it was the most accurate and had the largest AUC score, But with parameter tuning, the presented results may change, and the performance of each model could be improvedAfter Comparing Scores and using cross validation we have received an output of accuracy with 92.39%.

Using Data Science we have gained various Insights via graphical representation of the dataset some of the insights include: More men have depression compared to women, ages between 21 and 39 are more prone to suffer from depression, on a scale of 0-3 on questionnaire results if the output is 3 they are more likely to have depression, People are likely to suffer from mental Illness if they have a history of it in the family. Using Machine learning techniques like KNN, SVM, Decision Tree and Logistic Regression we have determined that the online depression detection questionnaire works with an accuracy of 92.39%(based on cross validation scores).

CHAPTER-5

Conclusion

In this paper we tried to focus on the most common mental health disorder which is depression and presented a literature survey to determine the accuracy of depression detection systems (In this case PHQ-9). The regression Techniques discussed here are KNN, SVM, Decision Tree and Logistic Regression. Using These techniques we have been able to determine the accuracy of online depression detection tools. Using this system has allowed us a better insight on depression detection tools. So finally using our methodology we have determined that Online Depression Detection Questionnaire PHQ-9 does Work with an accuracy of 92.39% (based On cross validation score).

References

1. . Depression. https://www.who.int/news-room/fact-sheets/detail/depression Accessed 16 Dec 2021
2. Machine Learning Mastery. https://machinelearningmastery.com/k-fold-cross-validation/ Accessed 17 Dec 2021
3. Page 181, An Introduction to Statistical Learning, 2013.
4.Medium article on SVM VS logistic regression https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16