DOKUZ EYLÜL UNIVERSITY
ENGINEERING FACULTY
DEPARTMENT OF COMPUTER ENGINEERING

CME 4416
INTRODUCTION TO DATA MINING

PREDICTING DEPRESSION WITH MACHINE LEARNING

by
Zeynep KAYA
Nilay YÜCEL

May, 2022
İZMİR

CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1. Problem Definition

The main characteristics of depression, also called major depression or clinical depression, are unpleasant mood, lack of interest and pleasure, hopelessness and pessimism. In general, the risk of depression is between 3-5.8%. The lifetime risk is 3-12% in men and 10-26% in women.

It has been observed that depression can cause physical changes in the brain of patients. The exact cause of depression is not yet known. Although it is known that hereditary factors play a role in its development, studies are continuing about which genes cause it. Among the biological causes, besides genetic predisposition, there are events that affect the mood of the individual such as some drugs, diseases, hormones, birth, menopause.

An individual's social life can also trigger depression. According to studies, negative life events, loss of parents in the early period, insufficient social support, spouse, family, work problems, previous depression and being a woman are among the important triggers of this situation.

A physical examination is the first step in diagnosing depression. The doctor asks questions about the person's health status. In some cases, the cause of depression may be physical health problems. Laboratory tests are needed to determine this. At this stage, the working level of the thyroid glands is checked by applying a test called "complete blood count". In the next stage, an interview is conducted to obtain information about the individual's feelings, thoughts and behaviors through a psychiatric evaluation.

There are many different types of major depression. Since the treatment methods differ according to the type of depression, it is extremely important to determine the causes of depression and to make the correct diagnosis. Usually, drug therapy and psychotherapy show effective results on patients. In some cases, these treatments may not be enough. In this case, the person may need to be treated under surveillance by staying in the hospital for treatment.

1.2. Project Description

The project being worked on aims to develop a machine learning model that will predict whether people are depressed by examining their data. As a result of the researches, it has been observed that many factors are effective on depression. There are some scales developed for diagnosis. These scales contain too many items and questions. The correct interpretation and evaluation of this content is a long and tiring process.

With the machine learning model to be developed, necessary evaluations will be made on the available data. As a result of these evaluations, the ratio of the effect of the answers to the questions asked on the result will be determined and the most accurate estimates will be made easier and faster. In addition, it is aimed to determine the evaluation criteria that do not affect the results and to shorten the evaluation processes of the patients.

The development process of the model started with a detailed investigation of the subject. In order to get to know the data set, information about the data to be processed was collected. Pre-processes to be made were determined by interpreting in more detail with visualizations. Preprocessing steps were carried out and the data set was made ready for machine learning algorithms. Machine learning algorithms to be used for the data set will be determined as a result of literature review and necessary research.

CHAPTER TWO

DATASET INFORMATION

The data set on which the studies will be conducted contains information that will affect the emotional state of the person, especially the personal, social and physiological information of individuals aged 18 and over. In the column determined as the target column, a result is returned regarding whether the person is depressed or not based on the data.

The data set, consisting of 36259 lines and 492 attribute in total, includes general information such as the person's age, gender, education level, place of residence, family, as well as the results of blood tests, diseases and drugs used. Since it is difficult to explain and understand such a large data set, it was decided to reduce the data set with detailed examinations and research. For this reason, studies in the diagnosis of depression were investigated and it was decided to keep data of great importance at the diagnosis stage.

```
[3]: raw_df.shape

[3]: (36259, 492)
```

Figure 1.2.1 Dataset Shape

In this section, information will be given about the data set that has been dimensionally narrowed. It will be explained what is done during the dimentionality reduction phase in preprocessing processes.

```
[5]: df.shape

[5]: (36259, 105)
```

```
 104  Depression                          object
dtypes: float64(52), int64(2), object(51)
memory usage: 29.0+ MB
```

Figure 1.2.2 Dataset Description

Consisting of 36259 rows and 105 columns, the data set contains 52 float64, 2 int64, and 51 object data. The column named 'Depression' from the Object type data is the target column of our dataset. This column contains information about whether the person is depressed.

Float64 consists of blood, weight and height values, object consists of answers to disease questions (yes/no) that will create categorical data and other personal information.

Gender	object	Sodium	float64
Age	int64	Potassium	float64
Race	object	Chloride	float64
Education Level	object	Osmolality	float64
Marital Status	object	Globulin	float64
Pregnant	object	White BCC	float64
Household Income	object	Lymphocyte Percent	float64
Asthma	object	Monocyte Percent	float64
Anemia	object	Neutrophils Percent	float64

Figure 1.2.3 Data Types

CHAPTER THREE

DATA PREPROCESSING

3.1. Dimentionality Reduction

It was determined that 324 of the columns were reserved for the drugs used. Most patients got bored with such a long question-and-answer process that more than half of these columns were left blank. In addition, since detailed information about diseases and blood values was already obtained, it was decided that these columns were not very necessary.

At the same time, it was noted that additional questions were asked in other columns for information that contained the same information or could be extracted from one column. It was decided to remove such redundant columns.

As a result of these operations, 105 attributes remained in the data set.

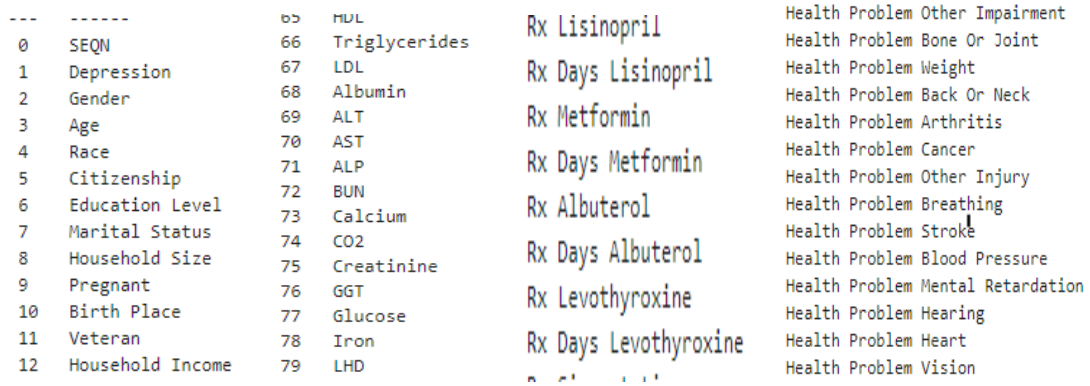

Figure 3.1.1 Dataset Attributes

3.2. Filling the Missing Values

When the fullness of the data set is examined, it is said that there are no null values. However, when examined in detail, it is observed that the values that should remain empty are filled with values such as “Missing”, “0”. These values should be determined and filled in the most accurate way.

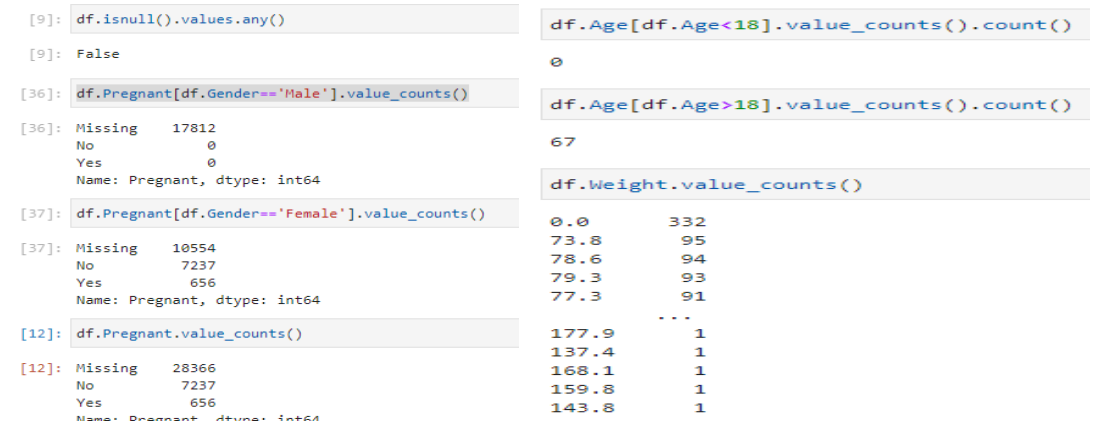

Figure 3.2.1 Missing Values

3.3. Outlier Detection

Outliers data were determined for the numerical values in the data set. This process was done using the interquartile range. After determining the lower limit and upper limit for all numerical attributes, outliers values were determined and cleared. It was found that too many inconsistent entries were made while collecting the data because the dataset contains too many attributes. For this reason, 22.853 of the 36.259 records in total were determined as outliers, as can be seen in *Figure 3.3.1*. Since there is enough data and it is desired to work in the most accurate range, the rows with all outlier values are removed from the data set.

Figure 3.3.1 Outliers

3.4. Normalization

The normalization shown in *Figure 3.4.1* was made for the numerical values after removing the inconsistencies in the data set, filling the missing values and clearing the outliers. 0-1 normalization was used as a method.In this way, the data has been brought into a regular format.

Gender	Age	Race	Education Level	Marital Status	Pregnant	Household Income	Asthma	Anemia	Ever Overweight	...	Health Problem Senility	Marijuana Use	Cocaine Use	Heroine Use	Meth Use	Inject Drugs	Rehab Program	Current Smoker	Household Smokers
Female	0.388060	Black	Some College or AA Degree	Married	No	$75K+	No	No	No	...	No	No	No	No	No	No	No	No	1.0
Male	0.820896	White	High School	Married	No	Below $25K	No	No	Yes	...	No	No	No	No	No	No	No	No	1.0
Male	0.014925	White	Some College or AA Degree	Never Married	No	$75K+	No	No	No	...	No	No	No	No	No	No	No	No	1.0
Male	0.910448	White	High School	Divorced	No	Below $15K	No	No	No	...	No	No	No	No	No	No	No	Never	1.0
Female	0.313433	White	9-11th Grade	Married	No	Below $45K	No	No	No	...	No	Yes	No	No	No	No	No	No	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Female	0.850746	White	College Graduate or Above	Divorced	No	$75K+	No	No	Yes	...	No	No	No	No	No	No	No	No	0.0
Male	0.223881	White	High School	Married	No	Below $65K	No	No	No	...	No	Yes	Yes	No	Yes	No	No	Every Day	0.5
Female	0.850746	White	Some College or AA Degree	Divorced	No	Below $25K	No	No	Yes	...	No	No	No	No	No	No	No	Never	0.0
Male	0.626866	Black	9-11th Grade	Partner	No	Below $25K	No	No	Yes	...	No	No	No	No	No	No	No	No	0.0
Male	0.358209	Mexican	High School	Separated	No	Below $25K	Yes	No	Yes	...	No	No	No	No	No	No	No	Never	0.0

Figure 3.4.1 Normalization

3.5. Transformation

The categorical values in the data set were converted with one-hot encoding before running the models on the data set. *Figure 3.5.1* shows the final state of the data set.

...	Meth Use_Yes	Inject Drugs_No	Inject Drugs_Yes	Rehab Program_No	Rehab Program_Yes	Current Smoker_Every Day	Current Smoker_Never	Current Smoker_No	Current Smoker_Some Days	Depression
...	0	1	0	1	0	0	0	1	0	Not Depressed
...	0	1	0	1	0	0	0	1	0	Not Depressed
...	0	1	0	1	0	0	0	1	0	Not Depressed
...	0	1	0	1	0	0	1	0	0	Not Depressed
...	0	1	0	1	0	0	0	1	0	Not Depressed

Figure 3.5.1 Transformation

3.6. Over Sampling

As a result of the target column scatterplot examinations, it was observed that the 'not depressed' cases were 13 times more than the 'depressed' ones. In this case, it was decided to apply an over sampling or under sampling process so that the weight does not shift to a single target.

As an oversampling technique, SMOTE process was applied and synthetic data was produced so that the target value distribution was equal (*Figure 3.6.1*).

```
Before SMOTE : Counter({'Not Depressed': 12464, 'Depressed': 942})
After SMOTE : Counter({'Not Depressed': 12464, 'Depressed': 12464})
```

Figure 3.6.1 Smote

CHAPTER FOUR

MODELLING

4.1. Algorithms for Modelling

Machine learning models are used to draw meaningful conclusions from the data collected while data mining. After the preprocessing is done, the data is brought into a format that the models can understand and machine learning models can be developed on it. Within the scope of this project, a total of 3 different algorithms have been used, including, Support Vector Machine, Gradient Boosting Machine, and Neural Network algorithms.

4.1.1. Support Vector Classifier

It is often used to solve Classification problems. In its basic logic, the data set is divided into classes with the help of a hyperplane as shown in the *Figure 4.1.1*. In addition to this hyperplane, support vectors defined at a certain distance are also used. Whichever of the dividing regions the value to be estimated enters, the class label of the value becomes the label of that region.

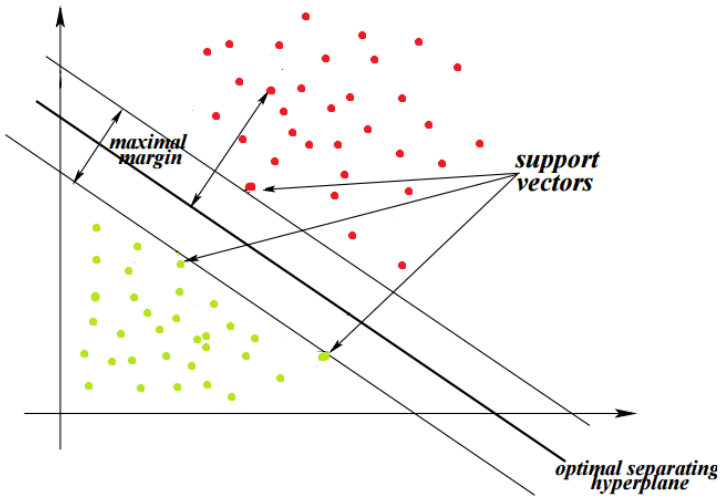

Figure 4.1.1 Support Vector Machine

4.1.2. Neural Network

A neural network works similarly to the human brain's neural network. A "neuron" in a neural network is a mathematical function that collects and classifies information according to a specific architecture. The network bears a strong resemblance to statistical methods such as curve fitting and regression analysis.

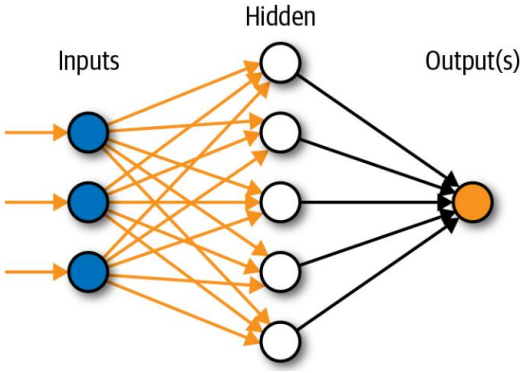

Figure 4.1.2 Neural Network

4.1.3. Gradient Boosting Machine

It is a type of boosting algorithm. As shown in the *Figure 4.1.3*, the general working principle is based on trying to correctly estimate the incorrectly predicted values each time with different iterations.

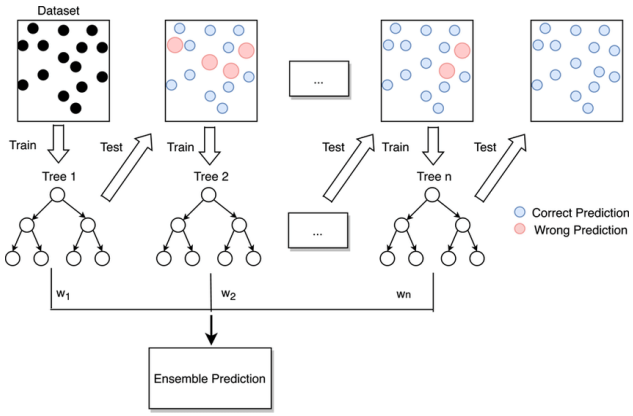

Figure 4.1.3 GBM

4.2. Test/Experiments

Modeling algorithms mentioned in Part 4.1 were run and models were obtained. While modeling, the reliability of the accuracy was also observed by using grid search, cross validation and some measurement matrices.

4.2.1. Grid Search

The algorithms mentioned in the above section 4.1 work with more than one parameter. These parameters can have different values. According to the data set used, the most suitable ones from these parameters should be selected and the model is trained. In this way, the best accuracy can be achieved.

However, it is impossible to predict from the beginning which values of the parameters will give the best results for which data set. For this reason, grid search is used. Grid search looks at which parameters give better results for that data set and shows us the best parameter values to use. Table 4.1-1 shows the best parameter values obtained as a result of grid search.

Table 4.21-1 Best Parameters

Algorithms	**Best Parameters**
SVC	C: 10 degree: 3 gamma: 0.01 kernel: poly
Neural Network	hidden_layer_sizes: (100, 100, 100) activation:logistic alpha: 0.1 solver: adam
GBM	learning_rate: 0.01 max_depth: 50 n_estimators: 500 subsample: 0.5

4.2.2. Cross Validation

Cross validation is a method that measures whether the high performance of the model is random or not. The data that is split into train and test while creating a model may not be randomly split. When this happens, the model has an overfitting problem.

To prevent this situation, the cross validation method is used. The n-fold cross validation divides the data set into n pieces and uses each of these pieces as a test set, respectively. The remaining n-1 pieces serve as the train set. In this way, it extracts n different accuracy values and presents their averages. This average can be considered as the most realistic accuracy value.

4.2.3. Comparision

As a result of the researches and studies, classifiers were made with 3 machine learning algorithms. As seen in *Table 4.3-2*, accuracy, measurement values and cross validation score were recorded before and after Grid Search.

When the values in the *Table 4.3-2* are examined, it is observed that all 3 algorithms give very consistent and high results. Based on this, it can be said that there is no overfitting problem in the models.
At the same time, if the ranking is made, it is seen that the best and most consistent result is reached in the **GBM** model.

Table 4.23-2 Result Comparision

Algorithms	Before Grid Search				After Grid Search				
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	CV Score
GBM	**96%**	**96%**	**96%**	**96%**	**97%**	**97%**	**97%**	**97%**	**97%**
SVC	96%	97%	97%	97%	96%	97%	97%	97%	95%
Neural Network	94%	95%	95%	95%	96%	97%	97%	97%	93%