July 2019

Statistical Learning of Biomedical Non-Stationary Signals and Quality of Life Modeling

Mahdi Goudarzi
University of South Florida, goudarzim@mail.usf.edu

Statistical Learning of Biomedical Non-Stationary Signals and Quality of Life Modeling

by

Mahdi Goudarzi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Mathematics and Statistics
College of Art and Sciences
University of South Florida

Major Professor: Chris P. Toskos, Ph.D.
Kandethody M. Ramachandran, Ph.D.
Lu Lu, Ph.D.
Yicheng Tu, Ph.D.

Date of Approval:
July 2, 2019

Keywords: Biomedical Signal Processing, Data Mining, Tree-based Methods,
Supervised-Unsupervised Learning, Quality of Life, Non-Homogeneous Poisson Process

DEDICATION

To my parents, my brothers, my sisters,

AND

My beloved wife, Elaheh

ACKNOWLEDGMENTS

First of all, I would like to thank my very supportive advisor, Professor Chris P. Tsokos, who always brights up the vague path with full of obstacles in my professional career. I have evermore felt his unwavering support and guidance during my entire PhD studies both in academical and personal manner. As a scientific father, his positive and optimistic personality has always been my inspiration. I always appreciate his trust in my abilities to discover new problems that solutions to them could impact wide range of applications. I also want to thank the very respectful committee members, Dr. Kandethody Ramachandran, Dr. Lu Lu, and Dr. Yicheng Tu for providing their valuable inputs regarding this research work. I really appreciate Dr. Yicheng Tu to chair my defense as well as serving in my committee. Also, I would like to express my gratitude towards my family, my parents, my brothers and my sisters for the encouragement which helped me in completion of this study, My beloved and supportive wife, Elaheh, who is always by my side when times I needed her more. Finally, I want to thank my friends specially Dr.Abolfazl Saghafi who make the department a place with full of intellectual and academic discussions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Statistical learning is a set of tools for modeling and understanding complex datasets. It is a recently developed area in statistics and blends with parallel developments in computer science and, in particular, machine learning.

The classification of biomedical non-stationary signals such as Electroencephalogram (EEG) is always a challenging problem due to their complexity. The low spatial resolution on the scalp, curse of dimensionality, poor signal-to-noise ratio are disadvantages of working with biomedical signals. EEG signals are unstructured data which needs preprocessing steps to extract informative features which are measurable and predictive. In the first two chapters of this dissertation, EEG signals that are recorded in 14 different locations on the scalp are utilized to detect random eye state change. We investigate this EEG data from two perspectives i.e., classification of raw data with and without feature extraction. In one of the methods, we bypass the feature extraction phase. SPI index, which is a transformation adapted from meteorology sciences, is implemented to transform data into a more appropriate space. Then, a Bayesian analysis of non-homogeneous Poisson process (NHPP) in a presence or absence of a change-point (open to close or vice versa) is developed using MCMC. We apply the power-law function as intensity function of NHPP models. The final classifier is

a model selection process between two NHPP models. In each time frame the best model, which fits to the data better, is selected. The accuracy of 74% is the best performance of the-state-of-art model.

In the second method, some features are extracted from EEG data based on fast Fourier transformation. We take into consideration all of the aforementioned difficulties and developed a three-layer classifier which is capable of solving the complexity of EEG signals (high dimensionality, noise, and poor spatial information) one by one in each step. Reduction of the number of signals from 14 to 5, with an accuracy 96% on one-second on reframed data in less than 3 seconds as well as extracting useful information from all channels (even those that seem uninformative in the first look) are main contributions of this method.

In addition to EEG data, the health-related problems are also explored in this dissertation in terms of their impact on the quality of life. The data consists of socio-demographic information as well as psychological background of 1080 individuals from different regions of Italy. This data is analyzed using supervised and unsupervised learning. The supervised learning method is a combination of classical non-parametric and machine learning methods to predict the general quality of life with an accuracy of 83%. The developed model is very informative and useful for either individual to monitor and improve their quality of life or for the administrative group to distribute their sources wisely and directly to the right target group.

In unsupervised learning, the group of people is clustered to three different categories according to their similarity in socio-demographic, health, and psychological information.

The implemented model is based on the K-medoids clustering. Such clusters can be used to have better understanding of the population for further analysis.

CHAPTER 1 : INTRODUCTION

The subject study consists of four chapters and given below are brief introductions.

1.1 Change-Point Detection of Biomedical Signals Using non-Homogeneous Poisson Process

This chapter is the continuation of the previous chapter on EEG signal processing and classification. One of the most important and challenging steps of supervised learning, in either classification or regression, is feature engineering. The main two hidden difficulties in this process are to extract informative as well as easy to access and measurable features. This step can be difficult specially for analyzing unstructured data such as noisy and high dimensional biomedical signals. A variety of methods have been used to solve this problem including (Sabancı and Koklu [1], Saghafi et al. [2], Rösler and Suendermann [3], Arvaneh et al. [4]) which are heavily dependant on feature extraction. However, in this chapter, we proposed the-state-of-the-art model to detect abnormality and classify raw EEG signal without implementing any feature engineering. In our model, we bypass the feature extraction by transforming the data into another space. The model has three main steps. In the first

step, the data points in the raw data are separated using Standardized Precipitation Index (SPI) transformation (McKee et al. [5], Wambua et al. [6]). The SPI has application in the detection of drought period in meteorology science. A peak point in EEG signal is similar to the drought time in climate data. In the next step, the signal is divided into equal frames. Then, two sub-models are developed. The first sub-model considers a hypothetical random change point and two non homogeneous Poisson process models (NHPP) with two parameters before and after the change-point. The second sub-model is a NHPP without any change-point. In the last step, two models are fitted to signals and the model with lower Deviance Information Criterion (DIC) is selected and as a result any change-point is detected. All the parameters are estimated using Bayesian method and GIBS sampling. The accuracy of 74% and no feature extraction are major results of the proposed model. The anomaly detection is real-time with maximum 2 seconds delay.

1.2 Ensemble Learning of Biomedical Signals Using Fast Fourier Transformation

The electrical activity of the brain is monitored by means of electroencephalogram (EEG) signals. It has substantial application in diagnosis of the abnormalities related to the brain such as epilepsy, sleep disorders, depth of anesthesia, coma, encephalopathy, and brain death. The amount of information captured by EEG signals makes the time of analysis a challenging problem for scientists. The main target is to detect abnormality by investigation of EEG signals via classification method. The classification of these types of signals has

been studied from different perspectives (Townsend et al. [7], Ghosh-Dastidar and Adeli [8], Wang et al. [9]). Instance-based and frame-based are the two major approaches to detect abnormalities, but the time of analysis has been investigated with less attention than achieving the best accuracy of classifiers. On the other hand, the curse of dimensionality and noisy data are the two main challenges in the analysis of EEG data. The signal-to-noise of ratio EEG data is very poor which makes it difficult to extract useful information to address the subject of interest. Another challenge with EEG signals is the non-exact spatial problem which refers to difficulty in locating the exact spot on the cerebrum that generates the signal. One of the fields that has attracted attention recently is eye status detection and drowsiness monitoring which has significant application in designing an effective warning system in sensitive fields such as safe driving, among others. Most studies in this field deal with one or two of the aforementioned difficulties of EEG signals, for example the accuracy of the developed model (Subasi and Ercelebi [10], Subasi and Gursoy [11]), dimension reduction and accuracy (Arvaneh et al. [4]), handling noise (Xu et al. [12]), and dealing with poor spatial resolution (Edlinger et al. [13], Burle et al. [14]). However, in this chapter, different methods are incorporated to tackle the all issues of working with EEG signals. In our novel model, different methods have been utilized either in preprocessing of data or in the analysis phase to optimize the result as well as reduce the time of analysis.

Multi-status EEG signals contain some change-points (mostly random points) in which the status of the signal is different before and after. Normal-Non-normal (Guo et al. [15]) and open-close eyes (Saghafi et al. [2]) are binary examples, and Normal-ictal-spike

(Vincent et al. [16]) is an example of three-status signal.

The EEG eye state corpus from UCI Machine Learning Repository created by (Frank [17]) is utilized in this study. It contains 14 sensors distributed symmetrically on the scalp. One column in this data is considered as ground truth labeled by an expert. This label is either 0 (open eyes) or 1 (closed eyes). Several models have been developed on this data to acquire maximum accuracy such as (Sabancı and Koklu [1], Saghafi et al. [2], Rösler and Suendermann [3]).

The developed model in this dissertation consists of two layers namely training a pool of base classifiers and then ensemble of them to improve the accuracy. The first step focuses on solving the curse of dimensionality by means dividing each raw signal to equal-length frames. Then, each frame is filtered using a frequency band that can give the most informative features based of spectral density. The result of first layer is the best classifier and best frequency band for each signal. In the next step the issues of signal-to-noise ratio and non-exact spatial are resolved through developing a set of tree-based classifiers. Finally, the result of five most accurate signals are combined to increase the accuracy of a single signal. The accuracy of 96% on one-second frames obtained in less than 3 seconds is the best result of the-state-of-the-art method.

1.3 Quality of Life: Statistical Analysis and Modeling of Psychological General Well-being Index via Supervised Learning

The quality of life (QoL) is defined by World Health Organization (WHO) as an individual's perception of their position in life in the context of culture and value systems in which they live and in relation to their goals, expectations, standards, and concerns (Group et al. [18]). The prediction of QoL is necessary to monitor a society well-being while helping individuals in achieving their goals with the maximum satisfaction. A reliable and well-trained model can assist people to correct their path by calibrating their risk factors based on their goals. The QoL can be divided into two main branches, health related quality of life (HRQoL) and general quality of life (QoL). In this chapter, our main focus is on the later type .

For decades, several studies focused on different aspect of QoL. The researchers attempted to answer a spectrum of questions from the definition to scaling and measuring the Qol (Diener [19], Harrington and Loffredo [20], Aaronson [21], Casellas et al. [22], Grossi et al. [23], Veit and Ware [24], Diener et al. [25], Lundgren-Nilsson et al. [26], Compare et al. [27]).

The HRQoL which is not our main concentration, has also received a huge amount of interest. Different studies tried to extract and rank the risk factors contributing to a specific diseases (Yazdi-Ravandi et al. [28], [29], [30], Karlsen et al. [31], D'alisa et al. [32], Logsdon et al. [33], Fenn et al. [34]). However, in this chapter, we developed a predictive model based on machine learning methods which can be implemented for monitoring everyday QoL. The

data utilized in this study is based on Psychological General Well Being Index (PGWBI) survey. The PGWBI questionnaire has been used for several decades as a measure for evaluating the state of mental health by means of a summary score. The data for this study has been collected from 1080 individuals from three different regions in Italy. Different methods such as regression and ANOVA have been implemented to analyze such surveys (Bianchi et al. [35], [36]). But in this chapter, we used several non-parametric methods to bypass assumptions of parametric methods and also to handle the large number of categorical variables in the data. To obtain a better understanding of the subject data and to answer some relevant questions in this regard, we performed non-parametric analysis, the Kruskal-Wallis test and the modern ML method of Random Forest. More specifically, this study tries to investigate how an individuals' socio-demographic information can be interconnected to the QoL measured by PGWBI score.

From the developed model, the quality of life index can be predicted with high accuracy from demographic information and health background alone. The accuracy of 0.83 and ranking of the risk factors contributing to the quality of life index from two different perspectives are the strong results of this study. According to developed model, age, income, education, occupation, region of living, and health background including sciatica, arthritis, and hypertension have the highest contribution to PGWBI score.

1.4 Quality of Life: Unsupervised Machine Learning of Social Data Analysis and Statistical Modeling

Sociological theory and phenomena are often hypothesis-driven in which explaining the cause of the problem is the core of the analysis (Rudin [37]). Sociological events are not very clear to be explained if the population is non-homogeneous. Therefore, it is more advantageous to divide the whole population to sub-populations with more similarities and less variability. Clustering analysis is an appropriate tool to explore such data and find similarities.

Clustering of social data is a useful tool for administrative purpose especially government and insurance companies. The appropriate clustering can facilitate a government's task to allocate limited source of funds to the proper group of people which have similar characteristics in a society. Moreover, insurance companies can create clusters of individuals with similar risk factors for better cost predictions.

Data mining and clustering methods have been widely applied to find hidden patterns in mixed social data. Researchers use clustering method to find a similarity among small producers in six cities in the northeast of Brazil (Maione et al. [38]). The social network is another interesting subject for researchers. The clustering of people in a social network using K-means clustering is a good judgmental tool to find users with similar behavior (Singh et al. [39]). In a human behavior study, the authors study the human social behavior to find similar patterns by means of clustering methods (Ferrara et al. [40]). However, in the most of social experimental design a representative data is collected in form of hybrid data, i.e., continuous

and categorical variables, and this restrains the usage of K-means clustering which is one of the most applicable method in data mining and clustering. K-medoids, the combination of Gower distance (Gower [41]) and K-means, can be used to handle the clustering of hybrid data. This method has a growing popularity among researchers in different areas of interest, (Velmurugan and Santhanam [42], Arora et al. [43]). In the present chapter, we investigate social data clustering using K-medoids clustering with Gower distance to find the similarity among individuals from different regions of Italy which has been used in the previous chapter. In the present chapter, we conclude a separation of individuals into three groups with similar characteristics that it can be a good source of information for government to make critical decisions about different groups of people. Also, our results are very important for individuals to recognize their similarities based on the risk factors on which the clusters have been created. Also, with high accuracy (95%), the QoL index is predicted after creating the similar groups.

CHAPTER 2 : CHANGE-POINT DETECTION OF BIOMEDICAL SIGNALS USING NON-HOMOGENEOUS POISSON PROCESS

2.1 Introduction

Biomedical signals such as Electroencephalogram (EEG) records electrical activity of the brain. The human brains is estimated to have 86 billion neurons in average(Azevedo et al. [44]) which makes it very complex to analyze. EEG signals have been investigated for decades from different perspectives (Pijn et al. [45], Dauwels et al. [46], Dement and Kleitman [47]).

Different models and methods have been implemented on EEG data to extract information to address a certain problem. Working with raw EEG data are experienced with several difficulties such as time of analysis due to its high dimensional data, etc. Therefore, transforming the raw data to another space is necessary to be able to solve the problem in a shorter time. The main challenge in mining the information from EEG data is to extract a set of relevant features to address the subject of interest (Ting et al. [48], Jenke et al. [49]).

Various approaches have been proposed in the literature to achieve higher classification accuracy, such as embedded hidden Markov models (Qin et al. [50]), time series

classification (Wang et al. [51]), pattern recognition (Estévez et al. [52]), and machine learning methods (Rösler and Suendermann [3]; Sabancı and Koklu [1]; Saghafi et al. [2]), among others.

In this chapter, the goal is to detect change point in the EEG status which is a classification problem. The novel analytical proposed model in this chapter can capture the change-point in EEG data without performing any feature engineering on the raw dataset. There are several studies on dataset used in this chapter that majority of them implemented a preprocessing and feature engineering methods to extract features(Rösler and Suendermann [3], Wang et al. [51], Sabancı and Koklu [1], Saghafi et al. [2]).

In our analytical model, after transforming the raw data using a transformation similar to SPI-index, the novel proposed classifier will develop two models on the signals after dividing the signals to sub signals of equal distances. One of the models considers a random change-point in the signal and the second model considers no change point in the data. The two statistical models are Non-homogeneous Poisson process models where their parameters are estimated using Bayesin method and Gibs sampling. The accuracy of 74% without any feature extraction are our major results of the study and the average time delay in change-point detection is only 1.5 seconds.

2.2 Data Pre-processing using SPI index

Anomaly in non-stationary signals can be detected as spike or inter-spike events over time. Therefore, a family of probability distributions such Gamma, Beta, and Normal

distributions can explain the probabilistic behaviour of this type of data. SPI, Standardized Precipitation Index, is used to quantify rainfall for a long-term climate data. The SPI values represents the level of drought in the under study region. There are two main reasons that the biomedical signals data are correlated to climate data; First, the nature of spikes in the biomedical signals are the same as drought in rainfall data which is an abnormal event in the climate data. And, secondly the spike or inter-spike happens gradually over a short time interval and not completely abrupt. SPI was first developed by McKee et al. [5]. Also, the details of the SPI calculation process described by Wambua et al. [6] are necessary, since understanding the threshold of drought season and its modification to biomedical signals is a key part of the classification and failure time definition in our study. The selection of probability distribution is the first step. In this study, the selected distribution is the Gamma probability distribution expressed by its probability density function as:

$$g(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha} e^{\frac{-x}{\beta}}, \qquad for \qquad x > 0, \tag{2.1}$$

where α, β are shape and scale parameters, and x is the raw data. $\Gamma(\alpha)$ is the integral constant calculated by means of:

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy, \tag{2.2}$$

where y is the output function defined in equation 2.1.

In the next step, the cumulative probability distribution of equation 2.1 is given by:

$$G(y) = \int_0^y x^{\alpha-1} e^{\frac{-x}{\beta}} dx. \tag{2.3}$$

Equation 2.1 is defined only for positive values. If there exist any zero value in the data, it can be solved by re-scaling the cumulative probability distribution as follow:

$$H(x) = q + (1+q)G(x;\alpha,\beta), \tag{2.4}$$

which $H(x)$ is the Cumulative probability and q is the probability of a zero value in the data. The cumulative probability was then transformed into a standard normal distribution using an approximate transformation as:

$$SPI = -(k - \frac{c_0 + c_1 k + c_2 k^2}{1 + d_1 k + d_2 k^2 + d_3 k^3}), \quad for \qquad 0 < H(x) \leq 0.5 \tag{2.5}$$

and

$$SPI = +(k - \frac{c_0 + c_1 k + c_2 k^2}{1 + d_1 k + d_2 k^2 + d_3 k^3}), \quad for \qquad 0.5 < H(x) < 1 \tag{2.6}$$

where k is attained by:

$$k = \sqrt{\ln(\frac{1}{H(x)^2})}, \quad for \qquad 0 < H(x) \leq 0.5 \tag{2.7}$$

and

$$k = \sqrt{\ln(\frac{1}{1 - H(x)^2})}, \quad for \quad 0.5 < H(x) < 1 \tag{2.8}$$

where; $c_0 = 2.55517, c_1 = 0.802853, c_2 = 0.010328, d_1 = 1.432788, d_2 = 0.189269,$ *and* $d_3 = 0.001308$. In the study of rainfall data, the SPI values were calculated using a monthly or 3,6,12,24,48 months time step and the threshold criterion as presented in Table 2.1, below.

Table 2.1: Drought classification based on SPI

State	**Criterion**	**Drought classification**
1	2.00 or more	Extremely wet
2	1.50 to 1.99	Very wet
3	1.00 to 1.49	Moderate wet
4	0.99 to -0.99	Near normal
5	-1.00 to -1.49	Moderate drought
6	-1.50 to -1.99	Severe drought
7	-2.00 or less	Extreme drought

In the present study the Table 2.1 will be modified based on the biomedical signals critical values and inter-spike and it will be discussed in section 3.5.

2.3 Model Description

In the previous section, we state that the nature of failure time in the climate data which can be generalized to biomedical signals. In this section the model to capture the failure (spike to inter-spike) is presented. This model can be applied on any data which has binary label. The core idea behind the model is non-homogeneous Poisson process(NHPP) with power law function its intensity function.

2.3.1 NHPP-Power Law Process

The probability of attaining n failures of a system in time interval $(0, t)$ can be expressed as

$$P(x = n|t) = \frac{e^{-\int_0^t \lambda(x)dx}\{\int_0^t \lambda(x)dx\}^n}{n!}, \quad for \quad t > 0. \tag{2.9}$$

There are a variety of choices for the intensity function, $\lambda(t)$, such as power law(PLP), the Musa-Okumoto(MOP), Musa and Okumoto [53], The Goel-Okumoto(GOP), Goel and Okumoto [54], among others. For simplicity, power law will be used in this study as the intensity function. PLP, Power Law Process, intensity function is defined as below:

$$\lambda^{(PLP)}(t|\alpha, \beta) = \frac{\alpha}{\beta}(\frac{t}{\beta})^{\alpha-1}, \qquad for \;\; t, \alpha, \beta > 0, \tag{2.10}$$

and the mean value function which is $\lambda(t|\alpha, \beta) = \frac{d}{dt}m(t|\alpha, \beta)$, is defined by:

$$m^{(PLP)}(t|\alpha, \beta) = (\frac{t}{\beta})^{\alpha}, \qquad for \;\; t, \alpha, \beta > 0. \tag{2.11}$$

It is observable from the definition of $\lambda(t)$ that the status of the system is a function of α which can be constant, decreasing, or increasing if $\alpha = 1, \alpha < 1 \; or \; \alpha > 1$, respectively.

Another model which is generalized from PLP is NHPP with one change-point in the time interval of study. This change-point is the time that there is a significant change in the nature of failures. Therefore, two NHPP is combined, one before change-point and one after. In this new model, the change-point is considered as a random variable which should be

estimated in the parameter estimation phase. All in all, if we consider power law as intensity function for each of NHPP in this model, thus five parameter should be estimated.

The discussion can be summarized mathematically as follow:

Suppose there exists one change-point over the time range $(0, T)$, it means there is a single change-point η making a shift from a NHPP to another. The intensity function of the overall process is defined by,

$$\lambda(t;\theta) = \begin{cases} \lambda_1(t), & 0 \leq t \leq \eta \\ \lambda_2(t), & t > \eta, \end{cases} \tag{2.12}$$

where $\lambda_j(t) = \lambda(t;\theta_j),\ j = 1, 2$ is the intensity functions before and after the change-point respectively and $\theta = (\alpha_1, \beta_1, \eta, \alpha_2, \beta_2)$ is the vector of parameters. By substitution of equation 2.10 in the overall PLP equation 2.12, the intensity function is given by:

$$\lambda(t;\theta) = \begin{cases} \frac{\alpha_1}{\beta_1}(\frac{t}{\beta_1})^{\alpha_1 - 1}, & 0 \leq t \leq \eta \\ \frac{\alpha_2}{\beta_2}(\frac{t}{\beta_2})^{\alpha_2 - 1}, & t > \eta, \end{cases} \tag{2.13}$$

with the corresponding mean value function extracted from equation 2.11 is:

$$m(t;\theta) = \begin{cases} (\frac{t}{\beta_1})^{\alpha_1}, & 0 \leq t \leq \eta \\ (\frac{\eta}{\beta_1})^{\alpha_1} + (\frac{t}{\beta_2})^{\alpha_2} - (\frac{\eta}{\beta_2})^{\alpha_2}, & t > \eta. \end{cases} \tag{2.14}$$

Because of the number of parameters for estimation and the amount of uncertainty, the maximum likelihood estimation is not applicable and therefore Bayesian estimation is used

to acquire estimation of the parameters. In Bayesian parameter estimation, the prior information about the unknown parameters is combined with the current data to have better estimation of the parameters. Bayesian inference derives from three probabilities, prior, likelihood, and posterior which linked by:

$$p(\theta|D_T) \propto p(\theta)L(\theta|D_T), \tag{2.15}$$

where $p(\theta)$ denotes the joint prior distribution and $L(\theta|D_T)$ is the likelihood function and $D_T = \{n; t_1, ..., t_n; T\}$ denotes the set of n failure times of the NHPP in $(0, T)$ and t_is are in increasing order. In an iterative process, the posterior probability distribution is substituted with the prior to have better approximate estimation of unknown parameters.

The likelihood function for θ assuming the truncated conditional probability distribution function is given by (Cox and Lewis [55], Tsokos [56]) can be expressed as follow:

$$L(\theta; D_T) = \prod_{i=1}^{n} \lambda(t_i) e^{-m(T)} \tag{2.16}$$

The equation 2.16 is the likelihood function for the NHPP without any change and by combining the random change-point with two likelihood functions, the likelihood function for the model with the presence of the change-point can be given by:

$$L(\theta; D_T) = \prod_{i=1}^{N(\eta)} \lambda_1(t_i) \times e^{-m_1(\eta)} \times \prod_{i=N(\eta)+1}^{N(T)} \lambda_2(t_i) \times e^{[-m_2(T)+m_2(\eta)]} \tag{2.17}$$

where $N(\eta)$ is the number of failures before the change-point, η.

2.3.2 Proposed Model

The proposed model consists of three main steps, failure definition, NHPP model fitting and model selection. The diagram 2.1, below, illustrates all of three steps. In the first step, the raw data is converted to SPI described in section 2.2, to be appropriate for further analysis. In the next step, the new signal is divided for a set of equal-length frames. For model fitting, because of the number of parameters in the models, two in no-change and five in one-change model, and non-existence closed form for the probability distributions, the Bayesian estimation of parameters described in equations 2.15,2.16,2.17 is applied using Markov Chain Monte Carlo (MCMC) methods. In the middle of dashed loop, there are two models, no-change and one-change models. In the case of no-change model, there is not any prior information about the two parameters of intensity function defined in equations 2.10, 2.16. A uniform distribution in the interval U[0,100] is considered for the parameters α and β to have approximately non-informative priors. As mentioned, since there is no closed form for the joint posterior probability distribution of θ, the simulated samples are obtained from this distribution using standard MCMC methods.

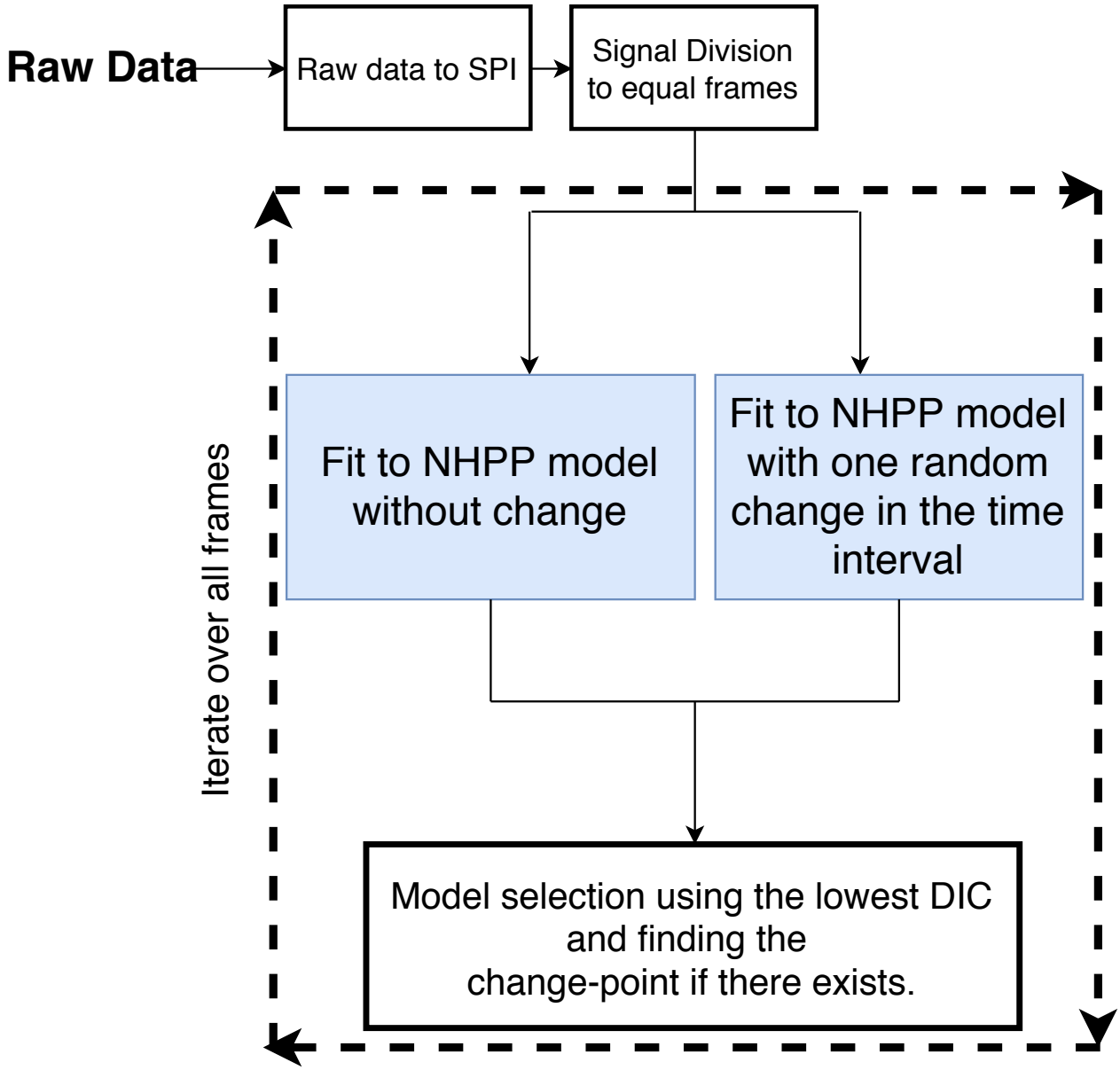

Figure 2.1: NHPP models in detection of change-point

That is , the sample is extracted from the full conditional posterior distribution $p(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n, D_T)$ for $i = 1, \ldots, n$ (Gelfand and Smith [57]).

In the second model with one change-point where properties are described by equations 2.13, 2.14, and 2.17, the joint posterior probability distribution of five parameters is

sampled by means of MCMC. The prior distribution for the intensity function parameters, $\alpha_1, \alpha_2, \beta_1, \beta_2$ are non-informative priors, that is ,$U[0, 100]$ and the same for change-point but on different interval $U(0, T)$, which T is the last failure time of the data.

In the last step in section 2.1, the best model is selected based on a Bayesian adequacy measures such as the Deviance Information Criterion (DIC) (Spiegelhalter et al. [58]) which is an approximation estimator of Bayes factor. The Smaller DIC is led to a better model. Finally , the dashed lines around the two models denote that the process is repeated for all the frames until the end of the signal.

2.4 Analysis of Eye EEG Signal

The EEG eye state corpus from the UCI Machine produced by Frank [17] is employed in this study. The dataset was created using Emotive EPOC shown in figure 2.2.

Figure 2.2: Emotive EPOC headset

The corpus contains 14980 instances(1/128 second) of 15 attributes which 14 of them

denote signals captured from locations on different part of scalp represented in figure 2.3.

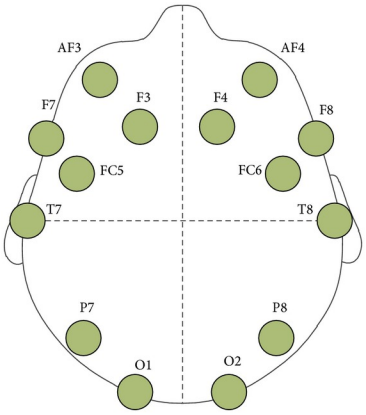

Figure 2.3: Scalp location covered by Emotiv EPOC

The last attribute denotes the eye state (open/closed). The duration of the experiment was 117 seconds recorded at 128 Hz sampling frequency during the individual involved in test opened/closed their eyes at will. The only data preprocessing accomplished on this data is removing 4 instances marked as outlier. These instances were 899, 10387, 11510,and 13180. The models in Figure 2.1 has been developed on raw dataset after one transformation and without any filtering which is common in signal processing.

The core idea of this study is framed on NHPP. In the first step of the analysis, it is necessary to define the failure time according to the proposed data. Every instance in transformed data which is greater than 1 or less than -1 is considered as a failure time. In the next step, the signal is divided into 13 equal-length sub-signals of 3 seconds which they are fed into the proposed model sequentially. The choice of 3 seconds is due to the long-term data which is necessary for SPI analysis ,that is, for each sub-signal, there exists 384 data instance which is enough comparing with climate data analysis transformed by SPI. The first

2000 instances (approximately 15 seconds) of all signals are shown by Figure 2.4. The red vertical lines are division of the signals in 3 seconds and the black vertical line is the random change in the status of the participant's eyes. Therefore in the first 3 seconds interval, there is one change almost in the middle of the interval , in the second one no-change, in the third one, there is one change at the beginning and so on, so forth. The y-label in the all of the subplots of Figure 2.4, the y-axis values are SPI. From Figure 2.4, it is observable that some of the signals such as AF3, AF4, F7, FC6, F4 are more informative to capture the spikes or eye state change which they are defined as failure, than other signals which are located in the middle of the scalp.

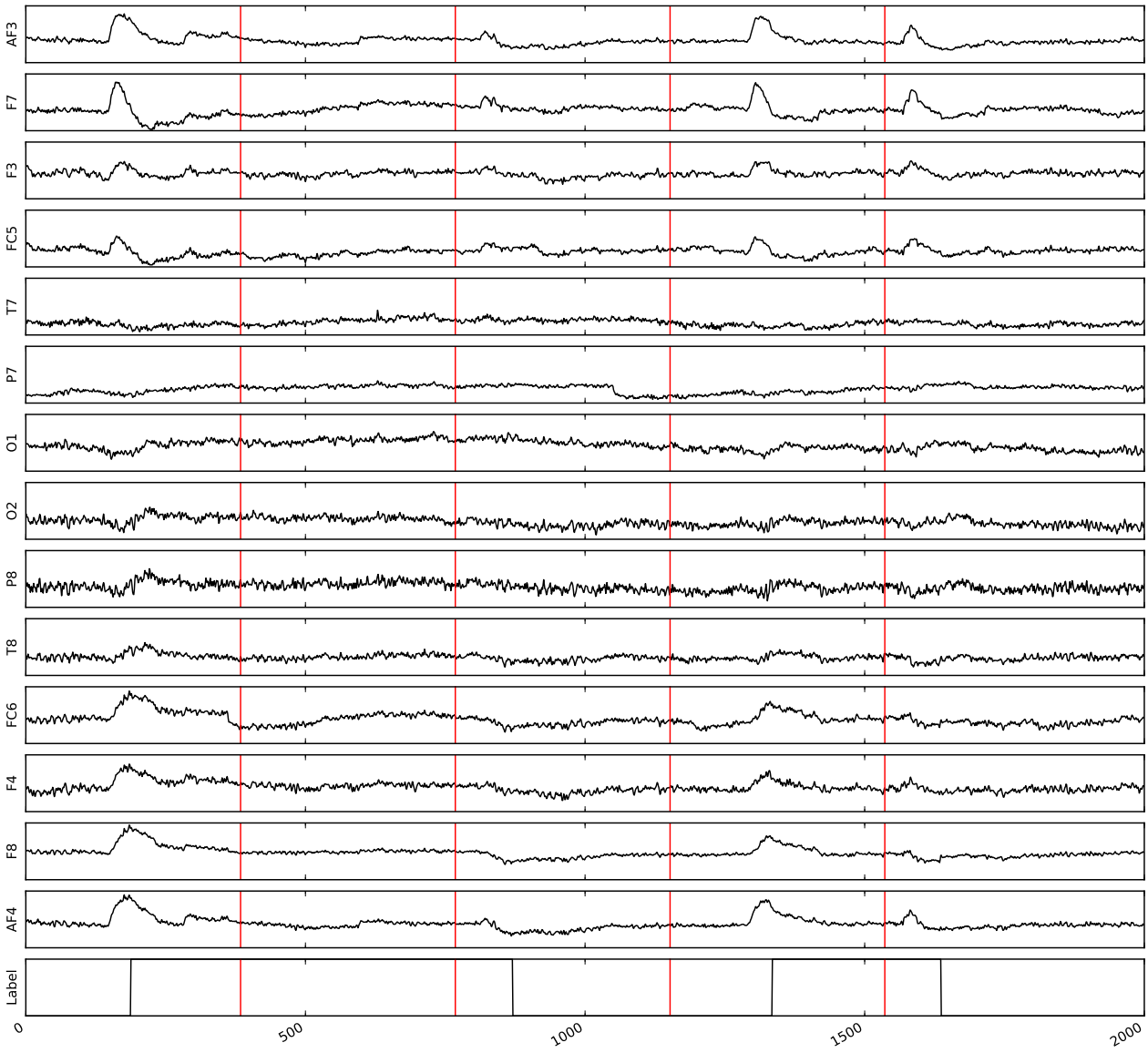

Figure 2.4: The illustration of 15 attributes and 3 seconds signals. SPI>1 and SPI<-1 are considered as failure time. The black vertical lines in Label graph are changing time in eye status.

2.5 Result

All of parameters of interests in this study are τ, α and β which α and β are vectors of parameters if the model with one changes is considered. Another parameter of interest

is the threshold for defining the failure time in signals. The Table 2.1 shows the thresholds for climate and drought data and therefore for implementation of our model is necessary to define a threshold for EEG signal. The cut-off point for biomedical signals is attained by means of 5-fold cross-validation on signals. The best cut-off point which give us the best result is **1** which means every point below -1 and above 1 is considered as failure point.

As a sample, we shows the performance of the proposed model on a three-second frame from **F7** channel. Figure 2.5, 2.6 are the distributions of different parameters of the model. In the first row of Figure 2.5, the first graph is the distribution of the deviance information criterion which is an approximation for the Bayes factor and is used for model selection. The second and third graphs of the distributions of the two unknown parameters. Figure 2.6 consists of three rows and two columns, deviance and parameters' distributions. In this frame, there is a change-point at $T = 188(1.47seconds)$. In terms of model selection in the proposed classifier, DIC of the model with one-change is 345 and for model without change point is 548 and therefore the earlier model is the best for this frame. This model predict a change-point at $T = 155(1.21seconds)$ which is very close to actual change point.

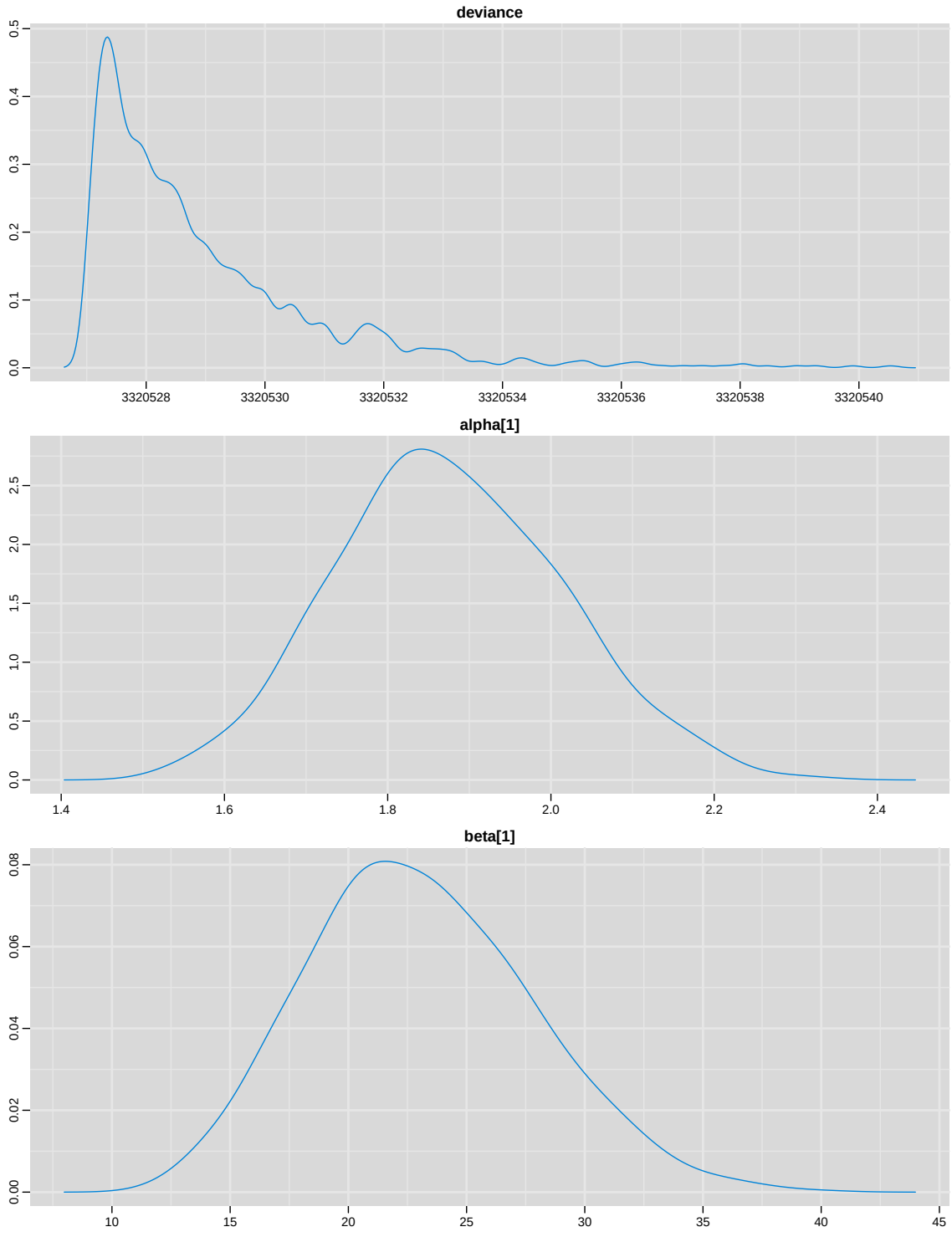

Figure 2.5: The analysis of one 3-frame signal with no change

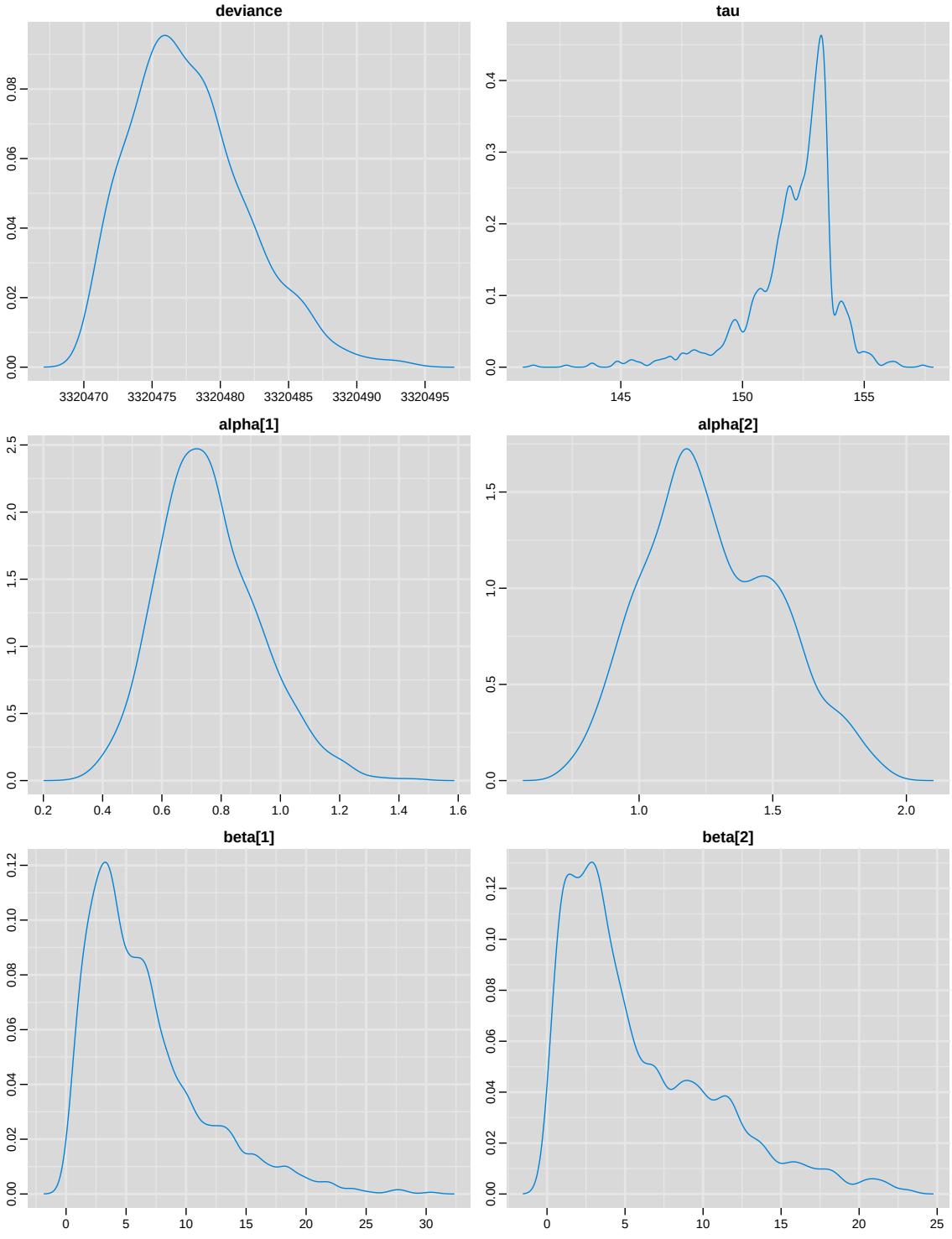

Figure 2.6: The analysis of one 3-frame signal with one change

Table 2.2: The result of classifier on AF3 channel

Time frame (3s)	**Predicted label**	**Predicted CP**	**True label**	**True CP**
"1"	"Change-point"	"215"	"Change-point"	"188"
"2"	"No-change"	"-"	"No-change"	"-"
"3"	"Change-point"	"818"	"Change-point"	"872"
"4"	"Change-point"	"1349"	"Change-point"	"1336"
"5"	"Change-point"	"1595"	"Change-point"	"1638"
"6"	"Change-point"	"2179"	"Change-point"	"2176"
"7"	"Change-point"	"2588"	"Change-point"	"2633"
"8"	"Change-point"	"2884"	"Change-point"	"2900"
"9"	"Change-point"	"3301"	"Change-point"	"3342"
"10"	"Change-point"	"3703"	"No-change"	"-"
"11"	"No-change"	"-"	"No-change"	"-"
"12"	"Change-point"	"4343"	"Change-point"	"4352"
"13"	"Change-point"	"4784"	"No-change"	"-"
"14"	"Change-point"	"5194"	"Change-point"	"5244"
"15"	"No-change"	"-"	"No-change"	"-"
"16"	"Change-point"	"5958"	"Change-point"	"5928"
"17"	"Change-point"	"6222"	"No-change"	"-"
"18"	"Change-point"	"6714"	"Change-point"	"6653"
"19"	"Change-point"	"6964"	"No-change"	"-"
"20"	"Change-point"	"7357"	"No-change"	"-"
"21"	"Change-point"	"7739"	"No-change"	"-"
"22"	"Change-point"	"8419"	"No-change"	"-"
"23"	"Change-point"	"8523"	"No-change"	"-"
"24"	"Change-point"	"9131"	"Change-point"	"9054"
"25"	"Change-point"	"9472"	"No-change"	"-"
"26"	"Change-point"	"9743"	"No-change"	"-"
"27"	"Change-point"	"10130"	"No-change"	"-"
"28"	"Change-point"	"10660"	"No-change"	"-"
"29"	"Change-point"	"10880"	"Change-point"	"11104"
"30"	"Change-point"	"11217"	"No-change"	"-"
"31"	"Change-point"	"11683"	"No-change"	"-"
"32"	"Change-point"	"11927"	"Change-point"	"12074"
"33"	"Change-point"	"12504"	"No-change"	"-"
"34"	"Change-point"	"12736"	"Change-point"	"12726"
"35"	"Change-point"	"13310"	"No-change"	"-"
"36"	"No-change"	"-"	"No-change"	"-"
"37"	"Change-point"	"14166"	"No-change"	"-"
"38"	"Change-point"	"14325"	"Change-point"	"14214"
"39"	"Change-point"	"14687"	"Change-point"	"14956"

Table 2.2 is the result of running Figure 2.1 algorithm on one of the signals, AF3. The first column is the index of the time frame which is three-second frames. The next column is the predicted label of that frame. If the model with the presence of a change point is selected, then the **Changed-point** means there is a changing point in this frame either from open-close or close-open. In the next column the predicted time of change-point (τ) is shown which is the predicted time of change in raw data which it can be converted to seconds by dividing by 128. The fourth column is the true label of frame in the data and the **True CP** is the actual change-point time in the raw data. By matching the predicted label and actual label , the accuracy on this signal channel can be calculated. The accuracy on this channel can be calculated by division of the number of correctly predicted frame by the total number of frames, 39. The accuracy on AF3 56.41% which is close to a weak classifier. We should add other results from the different signals to increase the accuracy. Also, from the table 2.2, we can observe that the classifier is bias to classify the majority of frames as change-point frame. Thus, by adding more signals, the baisness will be reduced.

The next step is the selection of a set of channels to increase the accuracy and decrease the bias of the classifier. Since the channels are highly correlated, therefore only some of them should be included in the final set of channels. The selection of channels is based on their probabilistic behaviour in the first step of modeling procedure and the channels with the similar distribution are clustered in the same group. Figure 2.7 shows the estimation of Gamma distribution parameters in the first step. Each color in this table contains the similar channels with respect to parameters of Gamma distribution. AF3 and AF4 have

close shape and scale parameters. FC6, O1, FC5, and F3 have the same distribution and so on, so forth.

We can deduce from Figure2.7 that five signals, one from each color, are utilized for the final classifier. The final selection is done by choosing five channels randomly which gives the maximum accuracy. The majority vote among 5 channels gives the accuracy of that specific selection and after running this process 1000 times, the final selection and accuracy are gained by the maximum accuracy of 1000 selections and its corresponding selection. The final set of channels is **AF4, FC6, T8, F4, P8** with max accuracy of **74%**.

Signal	alpha	beta
AF4	12851.0	0.33939
AF3	12969.0	0.33166
F7	17759.0	0.22578
F8	19012.0	0.24227
FC6	29477.0	0.14256
O1	37938.0	0.10735
FC5	38795.0	0.10626
F3	39944.0	0.10675
F4	45349.0	0.09436
T8	45824.0	0.09234
P8	55454.0	0.07576
P7	62250.0	0.07422
O2	63333.0	0.07288
T7	65166.0	0.06662

Figure 2.7: Gamma Parameter Estimation

2.6 Contribution

We propose an analytical classifier to detect changes in the brain signals. The main difference between the proposed analytical model and machine learning classifiers is to skip feature engineering step which for EEG data is the main difficulty and generally in any data mining problems. The EEG data are noisy, high dimensional, and has poor signal-to-noise ratio which makes it difficult to extract features time-wise.

In the proposed statistical model, in the first step, we transformed the data using a transformation similar to SPI-index adapted from meteorology and climate data analysis. This transformation separates the raw data to be more readable and proceed to modeling. In the second steps, the transformed data is fed into two sub models, a model that assumes there is no change point in the data and the second statistical model considers a change-point in the frame. The first model is a non Homogeneous Poisson Process with power law function as the intensity function and has two unknown parameters for estimation. The second model has 5 parameters that consists of two parameters for model before change-point, two for after and one for change-point itself. All of the unknown parameters are estimated using Bayesian and GIBS sampling. The estimation is accomplished by sampling from joint posterior probability distribution. The accuracy is 74% using cluster of five channels *AF4, FC6, T8, F4, P8* with 1.5 seconds delay on the average detection and the main result of this classifier. The above findings can be summarized as:

- Several Non-Homogeneous Poisson process, SPI index transformation from climate data analysis and technique of signal processing are combined to create a classifier to

detect changes in EEG data.

- The proposed analytical classifier is implemented on raw data and does not need any feature engineering and extraction which is the most difficult step in data mining, specifically for EEG data.

- The accuracy of 74% using a cluster of only five signals are the most important result of the proposed analytical classifier

CHAPTER 3 : ENSEMBLE LEARNING OF BIOMEDICAL SIGNALS USING FAST FOURIER TRANSFORMATION

3.1 Introduction

EEG (Electroencephalography) is a monitoring method to record the electrical activity of the brain. It has extensive application to diagnose abnormalities related to the brain behaviour such as epilepsy, sleep disorders, depth of anesthesia, coma, encephalopathies, brain death and heart abnormalities. Despite some major disadvantages of EEG signals such as high dimensionality, poor signal-to-noise ratio, and non exact spatial spots, EEG recordings still play an important role in diagnosis of neurological illnesses by representing the neuronal membrane potential with complicated and aperiodic time series.

The classification of these types of signals has been investigated from several perspectives (Townsend et al. [7], Ghosh-Dastidar and Adeli [8], Wang et al. [9]). Instance-based and frame-based are two main approaches to detect abnormalities in the signals, however the time of analysis in the subject area has received less attention than accuracy of the classification method. Whereas, the time required for the analysis of the response is one of the important factors that needs to be considered. The noise level and dimensionality of the data

are two big obstacles during the analysis with respect to time. One of the most demanding area of study is the eye status detection which has direct application in an effective designing of the warning alarm system in sensitive fields such as autonomous vehicle and reliability of driving behavior investigation, among others.

In order to extract relevant information from these signals, a variety of methods in either preprocessing phase or in the analysis have been implemented. It will be very difficult to handle all the the aforementioned problems which are the curse of dimensionality, high volume noise, non-exact spatial and time, in one algorithm and with one unique method. Most studies on EEG deals with one or two the aforementioned difficulties based on the goal specified at the beginning of research. Several studies investigated the accuracy of models developed on EEG signals (Subasi and Ercelebi [10], Subasi and Gursoy [11]), optimizing the number of channels and accuracy at the same time (Arvaneh et al. [4]), handeling poor signal-to-noise ratio (Xu et al. [12]) and dealing with poor spatial resolution (Edlinger et al. [13], Burle et al. [14]).

Multi-status EEG signals contain some change-points (which mostly are random points) in such a manner that the status of the signal before and after each point is different. Normal-Non-normal (Guo et al. [15]), open-close eyes (Saghafi et al. [2]) are binary examples and [Normal-ictal-spike](Vincent et al. [16]) is an example of 3 status signal.

The EEG eye state corpus from UCI Machine Learning Repository created by (Frank [17]) is utilized in this study. Several models have been developed on this data to acquire maximum accuracy such as neural network (Sabancı and Koklu [1]), logistic regression with

MEMD as features (Saghafi et al. [2]), k-star classifier on raw data(Rösler and Suendermann [3]) which majority of them investigated the accuracy of the proposed model. However, the developed model in this paper is targeted to cover all the disadvantages of EEG signals analysis as well as giving the maximum accuracy. Our model consists of three layers which in each layer, we implemented a state of art method to achieve to the final classifier. In the next sections, we shall give a brief review and highlights of the methods that we use during the process of developing our model.

3.2 Classifiers & Signal Processing Methods:

In developing the model in the subject area, we use a combination of classification and signal processing techniques. Classification Algorithms used in the subject area include K-Nearest Neighbors (KNN) Algorithm, Gradient Boosting, Random Forest and AdaBoost. In signal processing techniques we shall use Butter Worth Filtering, Hann Window and Fast Fourier Transform (FFT). Given below is a brief description of each of the methods used in achieving our objective in the subject study.

3.2.1 K-Nearest Neighbors(KNN) Algorithm

k-nearest neighbor algorithm (KNN) is a non-parametric classification algorithm in which input data are separated into several classes and the test sample is classified based on k closest training examples in feature space. When predicting a new data point, the KNN algorithm examines k-neighbors in the training set that have the maximum similarity to

determine the proper class. While there are several similarity measures including Euclidean, Manhattan and cosine similarity, we will use the Manhattan distance for our high dimensional data (Aggarwal et al. [59]) showed that the Manhattan distance metric provides the best discrimination in high-dimensional data spaces.

The Manhattan Distance is defined (for k=1) by:

$$L_k(x) = \sum_{i=1}^{d} (||X_i - Y_i||^k)^{\frac{1}{k}}, \tag{3.1}$$

where $x, y \in R^d, k \in Z$ and d is the dimensionality. ***Gradient Boosting Algorithm***

Gradient Boosting is an algorithm which builds predictive models in the form of an ensemble of weak learners. It can be interpreted as a numerical optimization problem where the loss of the model is minimized. The model is built in a stage-wise fashion by adding one weak learner at a time while leaving the existing weak learners unchanged.

Basically, there are three elements involved in the Gradient Boosting Algorithm: A loss function to be optimized, a weak predictive model to make predictions and an additive model to add weak models to minimize the loss function.

3.2.2 Random Forest

Random forest algorithm is a collection of decision trees where a random combination of features is selected at very node for splitting, hence the name **Random Forest**. Lets denote the training data set by $D = (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ and the feature vector

space by $X_p = (x_{p1}, x_{p2}, ..., x_{pn})$ such that $X \in D \in R^n$. During the training process K number of Bootstrap sample data sets are generated, with replacement, for K number of trees. In order to depict the growth of each tree, Independently and Identically Distributed(IID) random set of vectors $\{\phi_1, \phi_2, ..., \phi_K\}$ is also generated. Let each tree predictor be $h(X, \phi)$. Then, the collection of such predictors $h_1, h_2, ..., h_k$ is the random forest(Patri and Patnaik [60])

Random Forest Algorithm can be summarized as follows:

- Let the number of instances be denoted by N and the number of features by n
- Denote the number of features at a node of the decision tree by m ; where m <n
- Repeat the following steps for each decision tree:
  - Set a subset of the training data with replacement to represent N instances and the rest of the data to measure the error of the tree
  - Repeat the following step for each node of the tree: To determine the decision at the node and calculate the best split accordingly, select m features randomly. Tree pruning is not allowed.
- End

3.2.3 AdaBoost

In the Adaptive Boosting (AdaBoost for short), observations are first weighted in such a way that difficult to classify instances having more weight and well-handled instances

having less weight. The model is sequentially built adding new weak learners with the focus of training the instances which are difficult to classify.

3.2.4 Signal Processing: Butterworth Filter

One of the main disadvantages of working with biomedical signals, specifically EEG signals is the poor signal to-noise ratio. To assist with this problem, the researcher should remove some frequency bands from the signals. For example, in Eye signal, for detection of drowsiness only frequency band less than 15Hz is necessary to be investigated. Therefore, Butterworth bandpass filter is implemented on signals to remove all unwanted information.The Butterworth Filter method, which was first introduced by physicist Stephen Butterworth, is a signal processing filter useful in making a frequency response as flat as mathematically possible in the passband. The range of frequencies that can pass through a filter is known as passband.

The generalized equation for the frequency response of the nth order Butterworth filter is given by,

$$H_a(j\Omega) = \frac{1}{\sqrt{1 + (\frac{\Omega}{\Omega_c})^{2M}}} \tag{3.2}$$

where M represents the filter order and Ω_c is the cut-off frequency. We shall use this filter in the EEG signals that we are working with.

3.2.5 Signal Processing: Hann Window

A mathematical function that is zero-valued after some chosen interval is known as a window function. In digital signal processing, the Hann function is used as a window function in order to select a series of samples to perform a Fourier Transform. The Hann function is given by,

$$\omega(n) = \frac{1}{2}(1 - \cos(\frac{2\pi n}{N-1})) \tag{3.3}$$

3.2.6 Signal Processing:Fast Fourier Transformation(FFT)

The Fast Fourier Transform (FFT) is an efficient algorithm used to compute the Discrete version of the Fourier Transform.FFT employs a mathematical approximation of a given signal as an infinite combination of sin and cos waves to extract an approximation of the Power Spectral Density (PSD). PSD illustrates the strength of the variations(energy) of signal as a function of frequency[cite]. The FFT and PSD can be applied for feature extraction purpose. If a long signal is divided into equal length signals, then the PSD of a shorter signal can be extracted through the Short Fast Fourier Transform (SFFT). Then the short signals can be compared according to their difference in PSD

3.3 EEG Data

The EEG eye data corpus is utilized in this study. The data has been collected in UCI (University of California Irvine) Machine learning Repository by (Frank [17]). There are

14,980 instances of 15 attributes which 14 of them represent the recorded brain activity from different location on the scalp and one of the last attribute that determines the behaviour of the eye state(closed/open). The experiment has been completed in 117 second at 128Hz sampling frequency where each participant is asked to close their eyes randomly. Four instances which have unusual values and are outside of the main trend of other observations have been removed as outliers. These instances were at 899, 10,387, 11,510, and 13,180 time of recording. Further data explanation and filtering will be presented in Experimental Result section.

3.4 Methodology

As we previously mentioned in the introductory chapter, some of the main challenges in analyzing EEG signals are curse of dimensionality and poor spatial and signal-to-noise ratio. The steps taken to address each of these challenges and their justification are briefly described below.

In our proposed methodology, we treat each signal captured from the electrical activity of brain independently. In sleep disorder problems, frequencies ranging from 0-15 segmented in the three frequency bands: Delta, Theta, and Alpha, have some signs of drowsiness and closed eyes. This implies that to explain the maximum amount of information, we need to consider a sub-interval in the frequency range 0-15. In this approach, the sub-interval with maximum information is detected to increase amount of signal-to-noise ratio. The detail of the detection of this interval will be presented in next section.

As it is common in most of the studies related to biomedical signals, the amount of recorded information is high which makes it complicated to analyze raw data. To address the curse of dimensionality, we first implemented feature extraction prior to the analysis using Fast Fourier Transformation (FFT). This allows us to transfer raw data into feature space and frames. Spatial problem is another issue that arises when handling EEG signals. Specifically, there are some instances where the experts working with EEG signals are not certain as to which part of the brain generate the signals. We employ a cluster-based procedure to reduce the number of sensors placed on the scalp. This enables us to identify the most informative subsets of the recorded signals. Finally, we claim that it is not sufficient to consider only one classifier for all signals. Thus, we select the best classifier for each specific signal from a pool of classifiers which generates the highest degree of accuracy. The first layer of our 3-layer classification models is summarized in the Figure 3.1 which illustrates the procedure to identify the best classifier and the corresponding frequency band.

In Figure 3.1, After feeding the raw EEG to a high-pass a 0.5Hz high-pass filter to remove DC effect, the filtered signal is divided to sub-signals with equal length. In the next step, the frames are labeled. In the dashed loop, two integers are selected randomly and the signal is filtered with a mid-pass filter, then the signal is classified by means of extracted features. The loop is iterated 1000 times for all 14 channels to find the best classifier and the best lower and upper bound cut-off points for filtering.

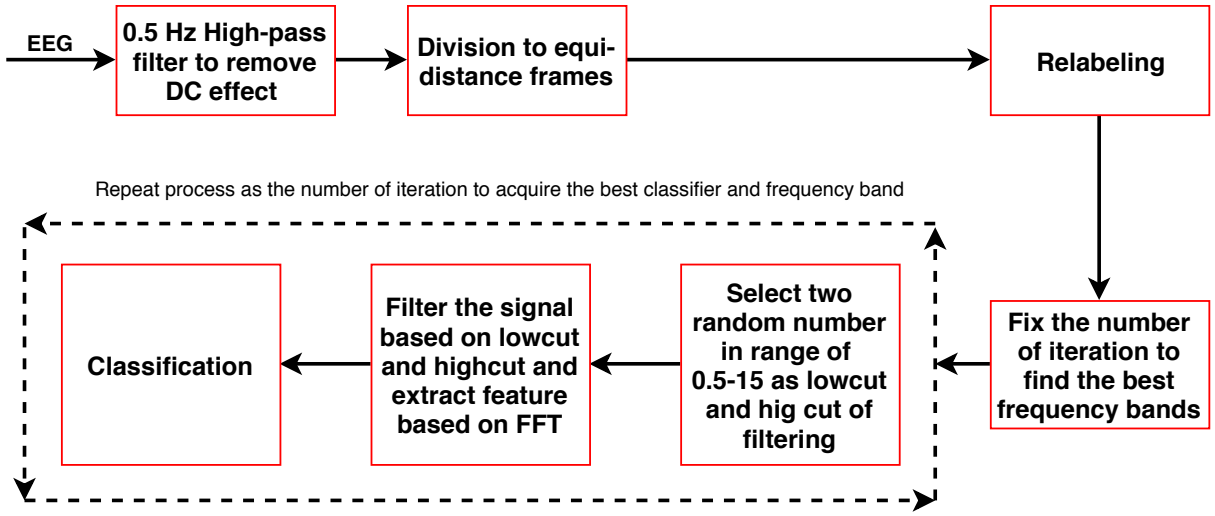

Figure 3.1: Description of the first layer of classifier

In the experimental section, we try to implement the figure 3.1 on the EEG data described in section 4.2. The classification process can be summarized in these steps, **Filtering, Classifiers, Feature extraction, Layer 1, Layer 2 and Layer3**.

3.4.1 Filtering

Comparison of EEG bands

Band	Frequency (Hz)	Location	Normally	Pathologically
Delta	< 4	frontally in adults, posteriorly in children; high-amplitude waves	• adult slow-wave sleep • in babies • Has been found during some continuous-attention tasks[49]	• subcortical lesions • diffuse lesions • metabolic encephalopathy hydrocephalus • deep midline lesions
Theta	4–7	Found in locations not related to task at hand	• higher in young children • drowsiness in adults and teens • Idling • Associated with inhibition of elicited responses (has been found to spike in situations where a person is actively trying to repress a response or action).[49]	• focal subcortical lesions • metabolic encephalopathy • deep midline disorders • some instances of hydrocephalus
Alpha	8–15	posterior regions of head, both sides, higher in amplitude on dominant side. Central sites (c3-c4) at rest	• relaxed/reflecting • closing the eyes • Also associated with inhibition control, seemingly with the purpose of timing inhibitory activity in different locations across the brain.	• coma
Beta	16–31	both sides, symmetrical distribution, most evident frontally; low-amplitude waves	• range span: active calm → intense → stressed → mild obsessive • active thinking, focus, high alert, anxious	• benzodiazepines • Dup15q syndrome [50]
Gamma	> 32	Somatosensory cortex	• Displays during cross-modal sensory processing (perception that combines two different senses, such as sound and sight)[51][52] • Also is shown during short-term memory matching of recognized objects, sounds, or tactile sensations	• A decrease in gamma-band activity may be associated with cognitive decline, especially when related to the theta band; however, this has not been proven for use as a clinical diagnostic measurement
Mu	8–12	Sensorimotor cortex	• Shows rest-state motor neurons.[53]	• Mu suppression could indicate that motor mirror neurons are working. Deficits in Mu suppression, and thus in mirror neurons, might play a role in autism.[54]

Figure 3.2: Different drowsiness frequencies in EEG signals

According to figure 3.2, the frequency band ranges from 0 to 15 can normally be a sign of drowsiness, sleeping or closing the eyes. In our method, the signals are filtered with Butter-worth filter which is a bandpass filter described in section 3.2.4 to remove the noise from the signals and prepare them for our purpose. However, the 14 signals are filtered different from each other. In an iterative process, for each signal a subset of [0,15] will be extracted such that the signal has the maximum information within that band or interval. In other studies, all the signals filtered with one fixed frequency band. In our analysis, for example **AF3** will be filtered in **[9.0, 12.5]**, **AF4** in **[4.5, 7.5]** and etc. The result of filtering band determination is summarized in Table 3.2.

Table 3.1: The result of the first layer of 1-second frames, p117

Channel	**Frequency Band**	**Train Accuracy**	**Classifier**	**Test Accuracy**
FC5	[4, 12.5]	0.712	KNN	0.727
F4	[2.5, 12.5]	0.704	KNN	0.727
AF3	[9.5, 13.5]	0.7	KNN	0.727
T7	[4.5, 12.5]	0.678	AdaBoost	0.727
F7	[2, 4.5]	0.691	KNN	0.681
AF4	[4.5, 8]	0.676	KNN	0.681
O1	[3, 7.5]	0.723	KNN	0.59
FC6	[7.5, 10]	0.651	KNN	0.59
O2	[1.5, 6.5]	0.724	KNN	0.545
P8	[8.0, 8.5]	0.69	AdaBoost	0.545
P7	[7.5, 10]	0.683	KNN	0.545
F3	[2, 8.5]	0.752	KNN	0.5
F8	[8, 9.5]	0.712	GBC	0.5
T8	[1, 2.5]	0.662	KNN	0.454

3.4.2 Classifier

As mentioned , a set of classifiers will be trained in the proposed model. In section **??** these classifiers described with more details and also a short description of classifiers in the pre-processing phase, the parameters of each specific classifier will be tuned by grid search over different. Random Forest, AdaBoost, gradient boosting, and KNN with different parameters are used to predict the true labels in our dataset. The number of neighbors in KNN is set to 3 because it has better performance and the rest of the parameters used without any change in the default values. For gradient boosting classifier, the number of estimators adjust on 1000 and in Random Forest case the number of estimators will are set on 50 and for prevention of over-fitting the maximum depth is adjusted on five and finally for AdaBoost all of the parameters will leave on the default parameters.

3.4.3 Feature Extraction

A set of features are extracted to feed into the classifiers. For each signal in training the learner, mean of signal, range, power of signal are the first three features which they are extracted by calculation of the basic statistics of each signal. By transforming the signal from time domain to frequency domain, another set of features are extracted. For each signal, we proceed with the spectral analysis using discrete Fast Fourier Transform. Since the maximum frequency of signal is 15, the first fifteen frequency bands will be used as filter. The real and imaginary components of each band of the Fourier Transform were mixed into a single magnitude. Therefore, 18 features (frequency bands, mean, range, and power) are

obtained.

3.4.4 Pre-processing and Classification

We begin by dividing the major 117-second signals to three dataset of short signals of 1-second, half-second, and quarter-seconds that we have 117, 234, and 468, respectively. The division with 468 sub-signals are more closer to the real-time classification, but it has less information in each frame which makes the classifier in danger of being less accurate. Therefore, there is a trade-off between having sufficient information and being close to the real-time classification(more accurate classifier). Each of the set of short signals will be used in the first and second layer independently and in the third layer the results of two layers are combined to each other by majority voting. After converting raw data to set of short frame, the eighteen features described in section 3.4.3, are extracted for each frame. The only problem is the label of each frame. Since closed eyes are more important than open, we label the frame "open" if all raw data in that frame are open otherwise it will be labeled "close". In the first layer, in an iterative process on each dataset extracted from 14 signals, the best frequency band and the best classifier for each signal are acquired based on the maximum accuracy on the test data. In the second layer, the frequency band and the classifier from the first layer will be used to retrain the classifier again and a pool of base classifier is obtained. Then, again in an iterative process a set of the best five signals are selected which they give the maximum accuracy by majority vote. Figure 3.3 illustrates the general process of the proposed model.

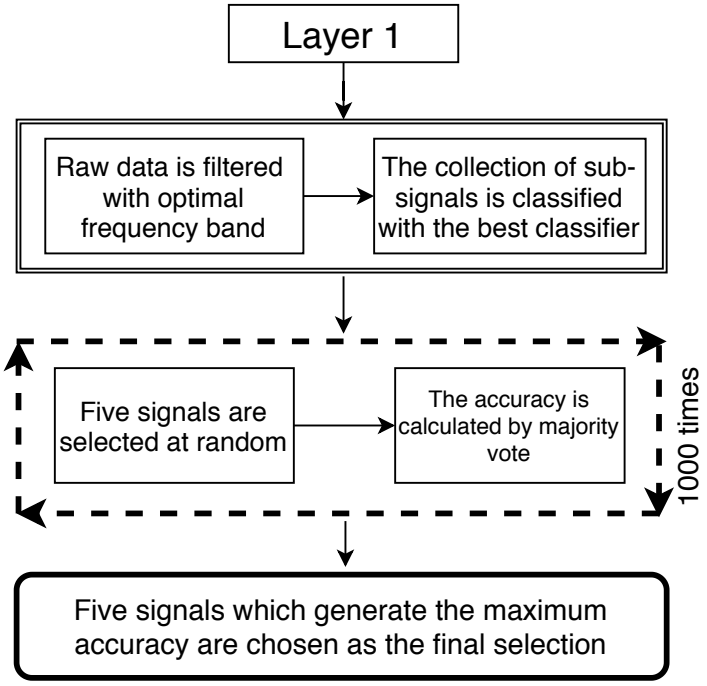

Figure 3.3: The second layer classifier procedure and the final classifier

3.5 Experimental Results

As we have described it above, there are two layers with three different results. In the last step of this procedure two problem will be solved, first the random nature of changing and classification of different frames from the reduced data. P117, P234, and P468 are three data set which are 117, 234, and 468 data frames, respectively. The pool of the base classifiers trained for p117 in the first layer is illustrated in Table 3.2, below:

Table 3.2: The result of the first layer of 1-second frames, p117

Channel	**Frequency Band**	**Train Accuracy**	**Classifier**	**Test Accuracy**
FC5	[4, 12.5]	0.712	KNN	0.727
F4	[2.5, 12.5]	0.704	KNN	0.727
AF3	[9.5, 13.5]	0.7	KNN	0.727
T7	[4.5, 12.5]	0.678	AdaBoost	0.727
F7	[2, 4.5]	0.691	KNN	0.681
AF4	[4.5, 8]	0.676	KNN	0.681
O1	[3, 7.5]	0.723	KNN	0.59
FC6	[7.5, 10]	0.651	KNN	0.59
O2	[1.5, 6.5]	0.724	KNN	0.545
P8	[8.0, 8.5]	0.69	AdaBoost	0.545
P7	[7.5, 10]	0.683	KNN	0.545
F3	[2, 8.5]	0.752	KNN	0.5
F8	[8, 9.5]	0.712	GBC	0.5
T8	[1, 2.5]	0.662	KNN	0.454

In Table 3.2 the first column is the 14 channels, the second column is the best frequency band that the filtered signal under this band has the greatest accuracy given in the third column and the signal is more informative in this interval. For training the classifiers and choose the best one in the fourth column, 5-fold cross validation has been utilized on 80 percent of data. The last column is the accuracy on 20% of the data as test data. It is observable from Table 3.2 that even with one channel such as FC5, F4, AF3, or T7 with accuracy of almost 73% the change of eye status can be detected with the delay of maximum 1 second.

In the second layer, five channels will be selected randomly and the accuracy is calculated by majority vote among them. After repeating of this process 1000 times the best combination of channels are **T7, O1, AF3, FC5, F7** which they generate the accuracy of **96%**. This

accuracy is on the test data which is 20% of one-second frames.

The Table 3.3 contains the result of the same procedure on the p234 and p468.

Table 3.3: The result of the first layer of half-second frames, p234

Channel	**Frequency Band**	**Train Accuracy**	**Classifier**	**Test Accuracy**
AF3	[3, 10.5]	0.669	AdaBoost	0.695
F7	[7, 11.5]	0.612	AdaBoost	0.695
P8	[0.5, 4]	0.628	GBC	0.608
O2	[3.5, 7.5]	0.601	KNN	0.586
F3	[5, 11]	0.649	GBC	0.565
F8	[0.5, 8]	0.64	GBC	0.565
F4	[8, 13]	0.637	GBC	0.565
O1	[9.5, 10.5]	0.616	GBC	0.521
T8	[9, 10]	0.578	KNN	0.521
FC5	[0.5, 7.5]	0.654	KNN	0.500
FC6	[5, 10.5]	0.616	KNN	0.500
AF4	[9, 13.5]	0.670	RF	0.456
T7	[12.5, 13.5]	0.628	AdaBoost	0.456
P7	[4, 4.5]	0.644	KNN	0.434

The result of the first layer classifier for the half-second frames are shown in Table 3.3 and the description is the same as Table 3.2, and it is observed that the obtained accuracy on the last column even for two or three channels are reasonably acceptable, that is, for AF3, F7 and P8. Another difference between this p117 and p234 are the type of classifier used GBC and AdaBoost are more frequent than KNN that had better performance in p117. The same procedure is repeated for the second layer of p234 and accuracy of **83%** is acquired from **F8, O2, O1, F7, AF3**.

Table 3.4 below, is a pool of base classifiers which illustrates the accuracy for the p468.

Table 3.4: The result of the first layer of half-second frames, p468

Channel	Frequency Band	Train Accuracy	Classifier	Test Accuracy
FC5	[1, 11.5]	0.606	GBC	0.695
AF4	[0.5, 12]	0.611	RF	0.630
T7	[12.5, 13]	0.592	KNN	0.608
O1	[8, 12]	0.616	GBC	0.597
T8	[8, 11.5]	0.601	RF	0.586
AF3	[9.5, 11]	0.595	RF	0.565
F4	[13, 13.5]	0.585	KNN	0.565
F8	[7, 11.5]	0.600	GBC	0.554
O2	[9.5, 12]	0.582	AdaBoost	0.554
FC6	[4.5, 11.5]	0.608	KNN	0.543
F7	[4.5,7]	0.585	AdaBoost	0.543
F3	[12, 12.5]	0.628	RF	0.532
P7	[0.5, 4]	0.603	GBC	0.5
P8	[0.5, 11]	0.596	AdaBoost	0.5

and for the second layer of p468, we can obtain the accuracy of **75%** by merging the result of five channels **FC5, AF3, T7, AF4, O1**. The figure3.4 illustrates the five significant channels for each dataset.

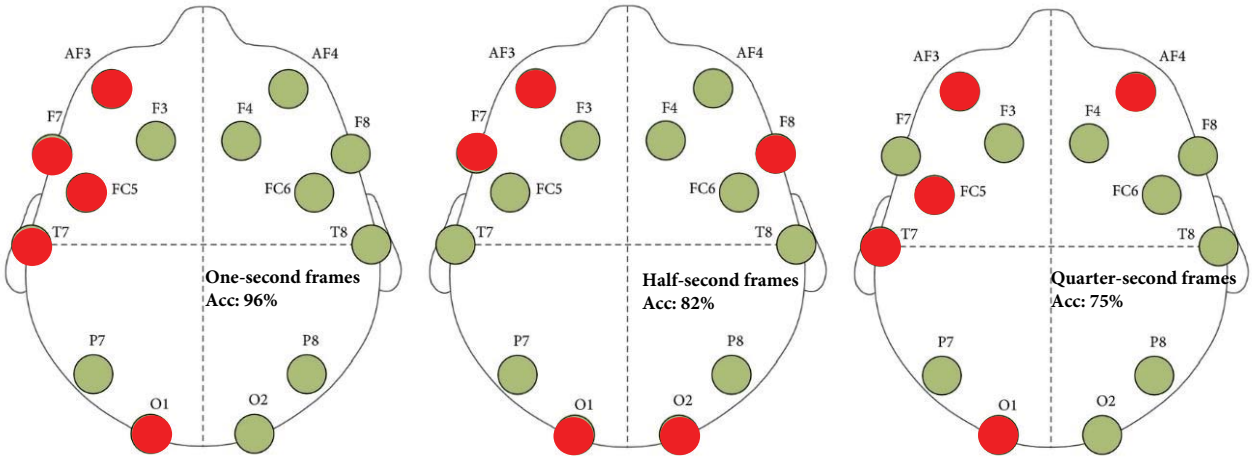

Figure 3.4: The optimized signals which create the maximum accuracy on test data. From left: p117, p234, p468 have accuracy of 96%, 82%, 75%, respectively.

3.6 Contribution

A method was developed on the EEG eye data to handle the disadvantages of analyzing with EEG data. The results presented in the section 3.5, can be accomplished on any dataset like EEG or ECG data. In our results, three different outputs are presented. Accuracy of the proposed classifier on one-second frames is 96%, half-second 82% and quarter-second 75% . The ideal result is on the raw data at the end of the process, but if the delay of maximum one second in detection of the abnormality in the signal is not a big issue, then accuracy of 96% from the first data is enough and more plausible than raw data. The longer frame in the classification process, the greatest accuracy is acquired. The accuracy of 96% is higher than all models and feature extraction methods on this data set (Wang et al. [9],Sabancı and Koklu [1], Saghafi et al. [2]) except the model trained by (Rösler and Suendermann [3]) which took 20 minutes on all 14 signals. The proposed model can detect the abnormality in EEG data with delay of maxim um one seconds. This chapter can be summarized as:

- The curse of dimensionality, poor signal-to-noise ratio and low spatial resolution of EEG signals are handled in the proposed model.
- Correlation among signals are handled using different classifiers such as Random Forest, Adaboost, Gradient Boosting, and K-nearest neighbors.
- Informative features are extracted using Fast Fourier transform.

- The accuracy of 96% on one-second frames, 82% on half-seconds, and 75% on quarter-second frames makes the result of this study very close to real-time classification.

CHAPTER 4 : QUALITY OF LIFE: STATISTICAL ANALYSIS AND MODELING OF PSYCHOLOGICAL GENERAL WELL-BEING INDEX VIA SUPERVISED LEARNING

4.1 Introduction

World Health Organization(WHO) defines the quality of life as an individual's perception of their position in life in the context of culture and value systems in which they live and in relation to their goals, expectations, standards, and concerns Group et al. [18]. Developing a general predictive model for QoL is necessary to monitor a community well-being while aiding individuals in achieving their goals and missions with maximum satisfaction. A well-defined and robust model can assist individuals to correct their path to have a happier life. Health-related quality of life (HRQoL) and general quality of life are the two main branches of QoL studies.

The quality of life has been an important aim of several studies for some decades (Diener [19], Harrington and Loffredo [20]). Initially authors in (Aaronson [21]) introduced some principal factors in how the quality of life scale should be developed. The main factors introduced in that research are: generic vs disease-specific focus, level of data aggregation,

interview vs questionnaire, and response scale. The well being-index is a tool which can be used as a quantity to measure the quality of life (Casellas et al. [22], Grossi et al. [23], Veit and Ware [24]). In (Diener et al. [25]), the authors created a new well-being measurement to assess positive and negative feeling of participants. One of the most significant applications of this measurement is to monitor patients' quality of life before and after treatment (Lundgren-Nilsson et al. [26], Compare et al. [27]). For example, the degree of recovery is a quantity that experts and doctors try to maximize in a shorter time (Grebner et al. [61]).

One of the main questions in this area is what factors have the main contribution in the prediction of quality of life. The generic and disease-based are the two domains that experts concentrate for prediction and maximization of the quality of life. In Yazdi-Ravandi et al. [28], the researchers consider self-efficacy, pain intensity, and pain duration as risk factors in the quality of life prediction of patients with pain disorders. In ([29]), physical, psychological and social components were recognized as main factors in predicting the quality of life for older people.

In most of the studies, the participants have physical health defects. For example, ([30]) shows that the presence of depression, disability, postural instability and cognitive impairment contribute the most in the quality of life for individuals with Parkinson disease. In addition, the influence of clinical and demographic variables on the quality of life of participants with Parkinson disease are also investigated in (Karlsen et al. [31]). In addition, patients with sclerosis and Alzheimer's diseases were also studied. (D'alisa et al. [32]) discussed the risk factors contributed in patient's quality of life with sclerosis and (Logsdon

et al. [33]) analyzed the quality of life of patients with Alzheimer. In another research,([34]) depicted the impact of financial inability for the patients with cancer. The authors took into consideration the impact of the financial burden of cancer on the survivors' quality of life. Regression analysis which needs some predefined assumptions, is one choice to analyze information extracted from this type of questionnaires or interviews (Bianchi et al. [35], ,D'alisa et al. [32]). ANOVA is another popular approach for analyzing this type of data as implemented in (Carotenuto et al. [36]) to compare the mean of quality of life index before and after one month of living and working on the sea in five groups of workers.

In the present study, PGWBI is considered a measurement of the quality of life. This index score is analyzed via two non-parametric methods, Kruskal-Wallis test, and decision tree-based method. By implementing these methods, three main goals are achieved: identifying which variables contribute to the quality of life index, second, ranking the attributable variables as a function of their contribution and finding the most contributing variables. Finally, developing a statistical model to predict the quality of life index without human interference would be highly desirable. In most of the studies related to the quality of life, there are three forms of information collected from individuals, that is, experience, demographic information, health background and PGWBI questionnaire, but on the contrary, environmental and demographic conditions can have a direct influence on quality of life as depicted by (Lawton [62]). In this study, a statistical model is developed to efficiently predict the quality of life index for either administrative or governmental agencies or for individuals interests in understanding and possibly increasing their quality of life.

Our study is arranged as follows. In section 4.2, the collected raw dataset introduced and prepared for the analysis and modeling. Section 4.3 introduces the methods implemented in this study. In the first part, Kruskal-Wallis test and then Random Forest(RF) are reviewed. In section 4.6, the results obtained from the two aforementioned sections are correlated and in Section **??** the models are implemented on data. Figure 4.1,below, illustrates the process that use depicts the main trend of this study.

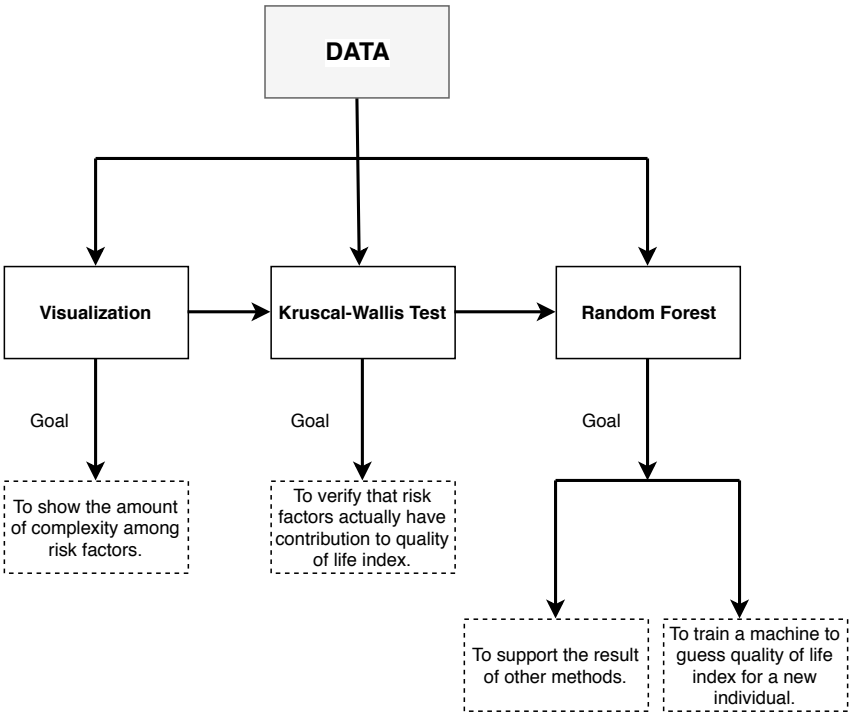

Figure 4.1: General trend of quality of life index

4.2 Description of Dataset

The original data was collected from different regions of Italy by Doxa, the Italian branch of the Gallup International association. The size of the dataset used in this study

comes from 1080 individuals which describes demographic, health, and psychological background. The data also describes the health background and disabilities of each individual and each individual is requested to fill out PGWBI questionnaire as well as demographic information. PGWBI, Psychological General Well-Being Index, is a 22-question questionnaire developed by Harold J Dupuy in 1971 to measure the quality of well-being as a self-representation of internal emotion[PGWI 1971](Ryff [63], Griffin [64]). Information obtained from one individual is independent from another. PGWBI divides each individual's internal response into six categories: Anxiety, Depression, General Health, Self-control, Vitality, and positive well-being. Each question in PGWBI pertains to one of the aforementioned categories and each category contains between three to five questions, with each question having six ranks. Each rank describes how an individuals feels towards the particular question with zero being the least and five the most favourable. In the end, all of the categories' scores are aggregated to generate one number between zero and 110 as the PGWB-Index. All of the demographic information is in discrete format and PGWBI extracted as continuous variable.

The demographic information is given by 10 categorical variables describing the general information. The ten variables are: region, municipal, amplitude, age, gender, education, marital status,occupation, and income. Each region has three levels. Municipal has two levels being capital or non-capital. Amplitude describes the size of the region from "small" to "city". Age is converted to a categorical variable with seven levels of equal age intervals. Education has five levels ranging from "none" to "university". Marital status has four levels ranging from "single" to "widowed". Occupation contains seven levels con-

taining "manager", "employee", "self-employed", "farmer", "retired", "unemployed", and "student". Finally, income has five levels form "low" to "super".

The physical health background is collected by asking some Yes/No questions about specific illnesses and disabilities. This background contains hypertension, heart attack, heart failure, diabetes, angina, cancer, allergy, arthritis, sciatica, blindness, lungs problem, dermatitis, deafness, weakness in arms, depression and mental disorder.

The target or dependent variable is obtained from PGWBI score. The quality of life index or score is considered as a dependent variable where its value can be predicted based on risk factors that have already been extracted.

4.2.1 Descriptive Statistics

In order to verify that there is a relationship between the risk factors and quality of life index, basic statistics and visualization are highlighted.

According to Figure 4.2, all of the variables are considered categorical except for the index being assumed continuous. The last three rows are the numbers generates from raw dataset, the missing data and the size of dataset after cleaning, respectively. The missing data is removed from the raw data set and the final version contains 1080 independent individuals.

Variable	Class	Description
Region	3	North ,Center- South
Municipal	2	Capital, Non-Capital
Amplitude	5	Small, Medium, Big, Very big, City
Class age	7	Seven equal size intervals from 15-93
Gender	2	Male, Female
Education	5	University, High School, Middle School, Elementary School, None
Marital Status	4	Married, Single, Widow, Divorced
Occupation	7	Manager, Employee, Home jobs, Farmer, retired, unemployed, Student
Income	5	Low, Medium, Medium-High, High, Super
Diseases class	5	0: no disease, 1: one disease , 2: two disease 3: three or four diseases 4: more than four diseases
Index	Continuous	Ranges from 0 to 110
Number of independent individuals		1129
Missing data		49
Size of data		1080

Figure 4.2: Basic Summary of Data.

4.2.2 Kruskal-Wallis Test

This section clarifies the selection of risk factors in this study by means of utilizing the Kruskal-Wallis test. The non-parametric version of one-way ANOVA is the Kruskal-Wallis test. This method is used to test the different contributions among the levels of the nominal risk factor with respect to the continuous target variable, the quality of life index. The null hypothesis of the Kruskal-Wallis test is that the mean ranks of the groups are the same. The expected mean rank depends only on the total number of observations (for n observations, the expected mean rank in each group is (n+1)/2). Since all of the risk factors are nominal and the target value is continuous, conducting a KW-test on each nominal risk factor is the best method to find its contribution. Table 4.1, displays the results of KW-test. From the results in Table 4.1, below, and with a significant level,α=0.05, the null hypothesis of KW-test will be rejected. Therefore, statistically speaking, there is a difference among the levels of contribution of the risk factors. The result of this test supports the visualization which has been done in the previous section. All of the variables used in this test are from the prepared data set except for the variable, **Diseases**. The health data are binary variable which they denote the existence of an illness. **Diseases** is the collection of health background features formed by adding the number of health problems from each individuals.

Table 4.1: Kruscal-Wallis Test Result.

Kruscal-Wallis Test			
Variable	**P-Value**	**Variable**	**P-Value**
Region	0.0002304	Cancer	0.02726
Municipal	0.004334	Allergy	0.01109
Amplitude	0.05754	Arthritis	$<$ 2.2e-16
Age Class	0.0006496	Sciatica	$<$ 2.2e-16
Gender	5.73e-09	Blindness	3.41e-05
Education	1.122e-05	Lungs	0.01093
Marital	0.0001179	Dermatitis	0.03757
Occupation	0.0003798	Deafness	0.0003482
Income	1.992e-08	Weak arms	6.481e-07
Hypertension	4.866e-09	Depression	4.618e-09
Heart.attack	0.0003932	Mental.disorder	0.002306
Heart.failure	3.921e-07	Diseases	5.264e-09
Angina	1.034e-05		

4.2.3 Visualization

In any parametric and non-parametric analysis, understanding the nature of the dataset is one of the most important parts of the analysis. For most of the parametric analysis, independent risk factors are assumptions that should be checked before performing

any analysis. This study illustrates the two-way and three-way interactions between different risk factors to understand the complexity of the dataset. By having 10 risk factors, there are a variety of choices to select interaction plots. A certain number of two-way and three way interactions are selected and explained in detail.

Figure 4.3, shows the two-way interaction among the 9 out of 10 risk factors. Kruskal-Wallis test was the first approval for actual contribution of risk factors in the quality of life index. The different interaction graphs plotted on Figure 4.3, are the second confirmations on the main effects of risk factors in this study.The complexity is obvious from the interaction plot. This graph was only one sample of the numerous interaction plots which can be generated. This plot visually proves that the nature of our dataset is very complicated and classical methods are not suitable for this problem. On the other hand, in Table 4.2, the analysis of uncertainty in the form of confidence interval of the mean effect of interactions is outlined. All of the hypotheses tests on existence of meaningful interaction effects are accepted at the level of $\alpha = 0.05$. The vertical line on levels of factors are confidence intervals calculated and shown in Table 4.2.

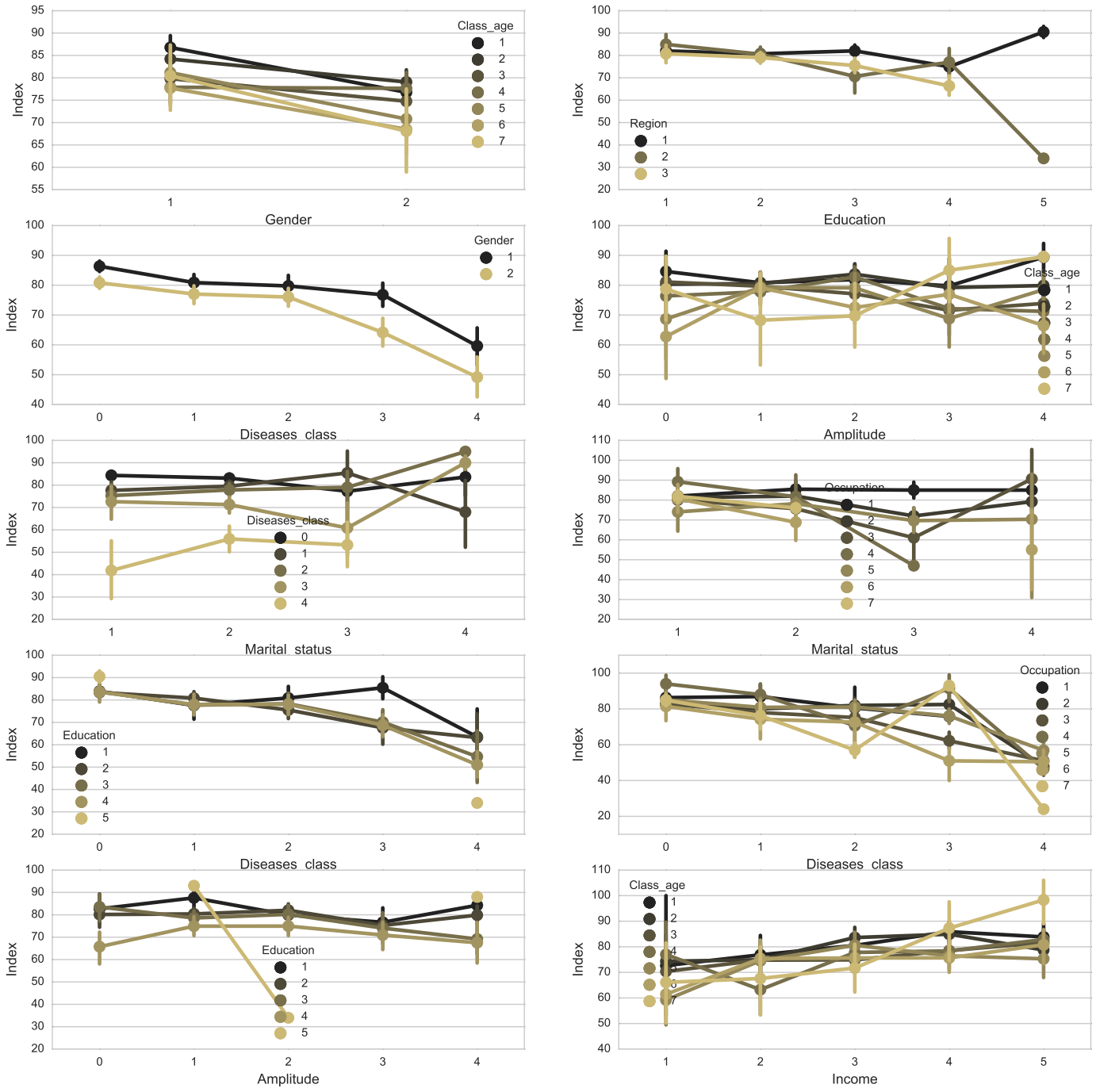

Figure 4.3: Two-way interaction between risk factors

Table 4.2: Uncertainty analysis of interactions .

Interaction	**Confidence Interval**	**P-value**
Gender-Age	[2.3258 , 17.2376]	0.0539
Region-Education	[-127.1427,-0.577]	0.00344
Diseases-Gender	[-13.1839, -1.0035]	0.0515
Amplitude-Age	[6.7080,15.58]	0.01697
Marital.status-Diseases	[3.2584,123.777]	0.0276
Marital.status-Occupation	[-105.4356,-7.5643]	0.038
Diseases-Education	[-30.4936,-.3197]	0.051
Diseases-Occupation	[0.4146 , 64.7696]	0.0577
Amplitude-Education	[-29.9833,-1.8799]	0.00881

4.3 Methods

Supervised learning is the machine learning method that utilizes a known dataset to make predictions. The supervised learning algorithms aims to develop a model that can make predictions from the response values for a new dataset. The supervised learning can be divided to into two main branches, regression for continuous response variable and classification for discrete variables. Classification is implemented in the present study.

The main goal is to train a machine to decide automatically on the response level of a new variable without human interference. The explanation of a family of tree-based methods are presented. The visualization and Kruskal-Wallis test sections proved that parametric analysis cannot be applied to this type of data because of the amount of complexity and interaction among risk factors. In this section, the two types of non-parametric analysis, Decision Tree(DT) and Random Forest(RF) will be discussed with these two methods. A predictive model is developed to predict the quality of life index. First, we will review the tree-based methods.

4.3.1 Trees Based Analysis

Supervised or unsupervised learning are two families of methods which their significant advantage over the classical methods is the ability to handle complex data. In supervised learning, there are two main components, feature space and the target value. The feature space is a collection of risk factors that may or may not have a direct contribution to the target value which is the dependent variable in a classical approach. If the target value is a categorical variable then the supervised learning is a classification problem, otherwise, it is called a supervised regression. On the other hand, in unsupervised learning, there is no target value and we try to find similarity between points . There are several methods which can be applied in both supervised and unsupervised learning. Figure 4.4, below, illustrates the supervised and unsupervised learning in details.

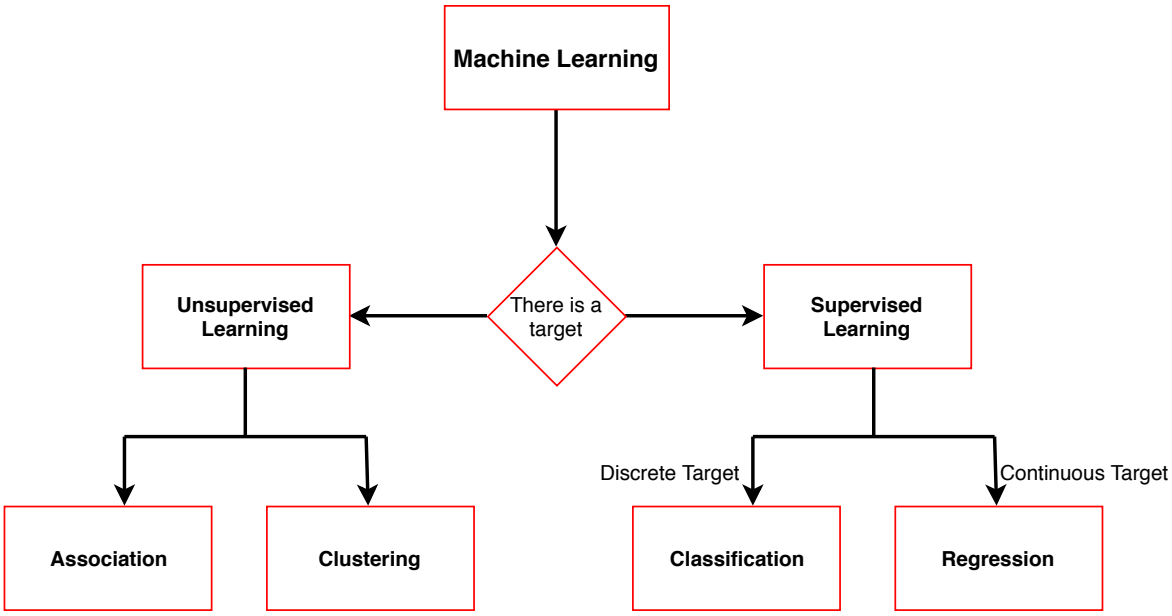

Figure 4.4: The Machine Learning Categorization with respect to the presence of a target value

The present study involves supervised learning. There are several methods which can be applied in this area of machine learning such as linear regression, logistic regression, decision tree, SVM(support vector machine), Naive Bayes, KNN, K-Means, Random Forest, Dimensionality Reduction Algorithms, Gradient Boosting algorithms, GBM, XGBoost, LightGBM, CatBoost, among other. These methods have different applications in the field of study from sciences to engineering (Rabiei et al. [65],[66, 67, 68, 69], Saghafi et al. [2], Jafarian et al. [70]). From the varieties of methods for supervised learning, the sequence of complementary methods, Decision Tree, Bagging and Random Forest will be briefly explained. Figure 4.5, below, shows the general process of supervised learning. The next two sections explain Decision Tree (DT) and Random Forest (RF) in some details.

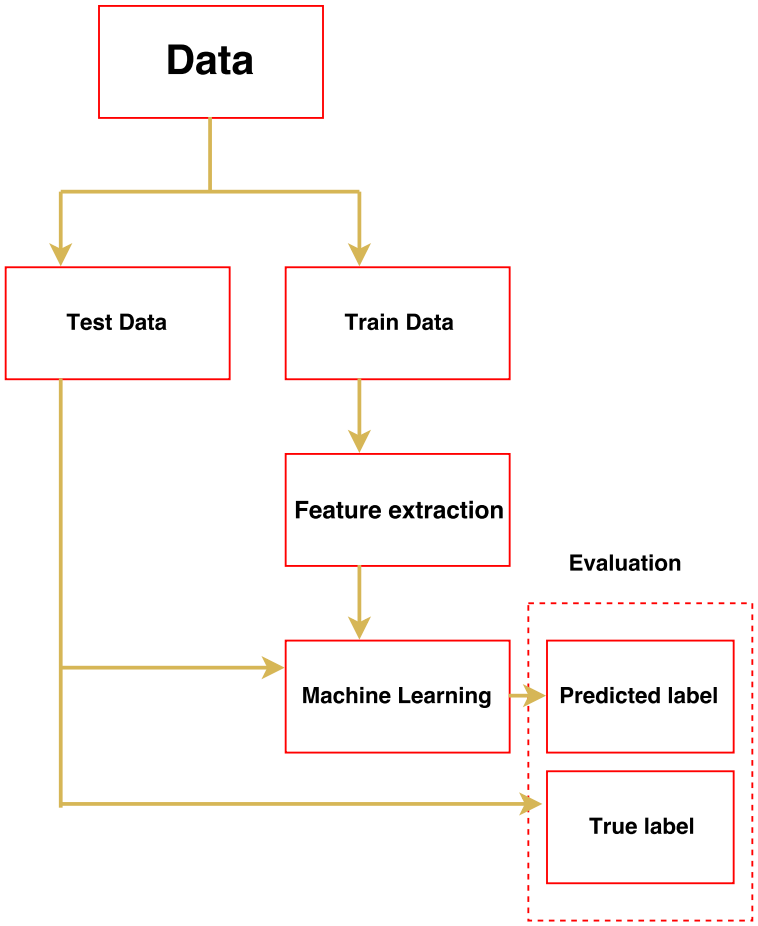

Figure 4.5: The General Process of Supervised Learning

4.4 Decision Tree

The Tree-Based methods partition the feature space into a set of rectangles and then fit a simple model (like a constant) in each one (Hastie et al. [71]). For a growing Decision Tree, DT, the training data is considered as the root node. Then by using a splitting criteria, the risk factors space is divided into two or more sub-spaces. The splitting criterion can be either a mis-classification error , a Ginni index, a cross entropy or deviance (Hastie et al.

[71]). This process will be preceded until there is no more data to be split on. At this level, the leaf nodes are generated and the decision is made.

Let $\mathbf{X} \in^{n*p}$ be a set of observations of risk factors or features and $\mathbf{y}$ be the response variable or , that is ,

$$\mathbf{X(input)} = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ . & & & \\ . & & & \\ . & & & \\ x_{N1} & x_{12} & ... & x_{NP} \end{bmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_N \end{pmatrix} \tag{4.1}$$

each observation is in the form of $(x_{i1}, ..., x_{ip}, y_i) \in^{p+1}, 1 \leq i \leq n$. The main goal is to find a function to predict the response value from a set of features automatically. Suppose that there are partitions of $R_1, ..., R_m$ of ^{p}. We define

$$\hat{P}_{ik} = \frac{1}{N_i} \Sigma_{x_i \in_i} \mathbf{1}_{y_l \in R_i} \tag{4.2}$$

the process of counting the proportion of class K observations in node i. The observations are classified by majority vote in node i, that is,

$$K(i) = argmax_k \hat{p}_{ik} \tag{4.3}$$

Mis-classification error is one of the measures used to determine how good a given partition is (how to split) which is calculated by:

$$\frac{1}{N_i}\Sigma_{x_l \in i}\mathbf{1}_{y_l \neq k(i)} = 1 - \hat{P}_{i,k(i)} \tag{4.4}$$

Generally the process is stopped for a given region, R_i, when there are less than five observations in that region.

One of the most important part of growing a tree is the size of it. A very large tree might over-fit (almost zero error but poor prediction) the data, while a small tree might not capture the important structure of the data. Thus, the size of a tree is a parameter which should be tuned before, during the process or by iterating over different sizes. Figure 4.6, below, describes the generation of a decision tree.

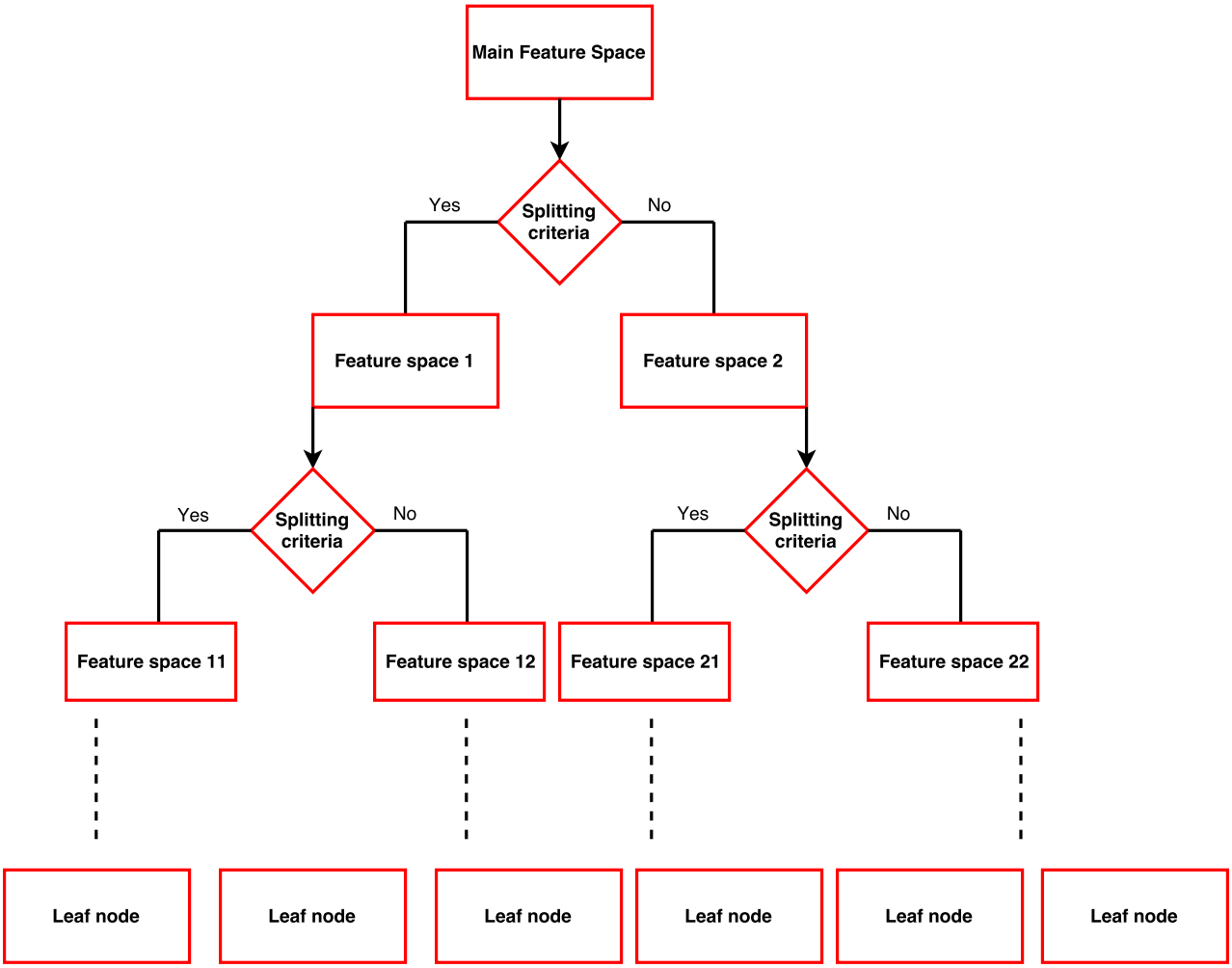

Figure 4.6: The Feature Space (sample space) division in growing of a single tree

The DT's offer several advantages and disadvantages. They are simple to understand, interpret, and they can deal with any type of variables as well. Furthermore, they can be combined with other decision methods to improve the result of modeling. On the other hand, it has high variance and instability. Any small change in the data can lead to a large change in the structure of the optimal Decision Tree. By combining several DT's, two main problems of a DT can be solved, high variance and inaccuracy. There exist several methods

such as AdaBoost(Adaptive Boosting), Bagging(Bootstrap aggregation) and Random Forest which can combine several DT's to boost the result of a single decision tree. In the next section, Random Forest (RF) is explained in some details.

4.5 Random Forest and Bagging

Random Forest was proposed by Hastie et al. [71]) is an extended version of Bagging Method. Thus, before we discuss Random Forest, we shall give a a brief description of Bagging. Bagging is usually used to decrease the variance of a single DT. A group of Decision Trees are trained on a given set of training data by bootstrapping the train data and at the end the results of all the trees are aggregated by a majority vote. The main point and difference of Bagging and Random forest is in the training process. In Bagging, all features (risk factors) are utilized to train each tree. Figure 4.7, below illustrates the general process of Bagging Method.

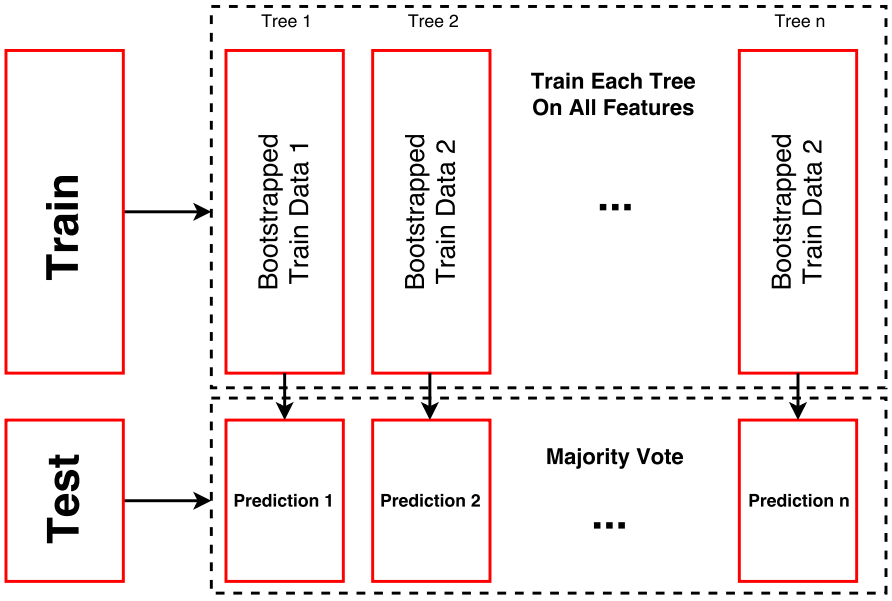

Figure 4.7: Bagging Method Training/Testing algorithm

Let $Z = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ be the train data. Bagging averages this prediction over a collection of bootstrap sample $Z^{*b}, b = 1, ..., B$, defined by

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x) \tag{4.5}$$

In the random forest on the other hand, we are imported another level of randomness that is added to the algorithm by selecting a random subset of features instead of all features. The selection of the features makes the algorithm very robust against over fitting (Hastie et al. [71]) . Also, the best number of random features for different trees in RF developing are log_2^n, $\sqrt{n}$ where n is the number of risk factors. The main advantage of RF is its user friendliness. It means that the only parameters which should be determined prior to run the algorithm are the number of trees and the number of random features.

For building a RF, two steps are necessary: Algorithm:

1. For b=1 to B: (a) Draw a bootstrap sample Z^* of size n from the training data.
2. Output the ensemble of trees T_b b=1,...,B

- Draw n_{trees} Bootstrap samples from the original data.
- For each of the Bootstrap samples, grow a tree. At each node, rather than choosing the best split among all features, randomly sample m_{try} of them.
- By majority vote among trees, the final prediction is determined.

An estimate of the error rate can be obtained, based on the training data, by the following procedure:(Liaw et al. [72])

1. At each bootstrap iteration, predict the data not in the bootstrap sample (what Breiman calls âĂIJout-of-bagâĂİ, or OOB, data) using the tree grown with the bootstrap sample.
2. Aggregate the OOB predictions. On the average, each data point would be out-of-bag around 36 percent of the times, thus we aggregate these predictions. Calculate the error rate, and call it the OOB estimate of error rate.

Visualization of RF is not as good as a single Decision tree, but two important pieces of information can be extracted from a RF, **Variable Importance** and **Proximity Measure**. The random forest algorithm estimates the variable importance by looking at how much prediction error increases when the data for that variable is permuted tree by tree while

other variables are left unchanged. The (i, j) element of the proximity matrix produced by random forest is the fraction of trees in which elements i , j fall in the same terminal node. This property can be used to identify the structure of the dataset. For further readings, (Hastie et al. [71], Liaw et al. [72]) are good sources to clarify the details of this discussion.

4.6 Analysis of the Results

In the analysis part, we implement both DT and RF on the described data. DT is applied in the preprocessing step which is explained in some details. The RF is developed for the main body of analysis and ranking the importance of the variables. Thus, we can summarize the analysis part into preprocessing the data, obtaining the results using RF and conclusion.

4.6.1 Preprocessing of the data

We begin by considering all of the attributable variables as categorical variables with different levels except for the target value, the quality of life index.

Since the goal of this study is a classification version of supervised learning, we convert continuous the index to three levels. This conversion happens by means of DT. We develop DT several times to find the best two points with the minimum error which can be considered as the breaking point. The result of this process leads to this interpretation, '0-50' considered as Low quality of life, '51-92' Medium and '93-110' as High quality of life

index. After discretization of the data, the resulting data is an imbalanced data, so by using "Undersampling-Oversampling" method, we transfer the dataset into a balanced data. Inspecting the dataset at the end, we can see that for the individuals under 18 years old there are some responses to questions in the questionnaire which are not logical. In other words, the oscillation in quality of life index is very high for people less than 18 years old. Therefore we remove this part of the dataset to have a more logical and smoother data. Figure 4.8, below illustrates this non-stationary plot of age versus the index.

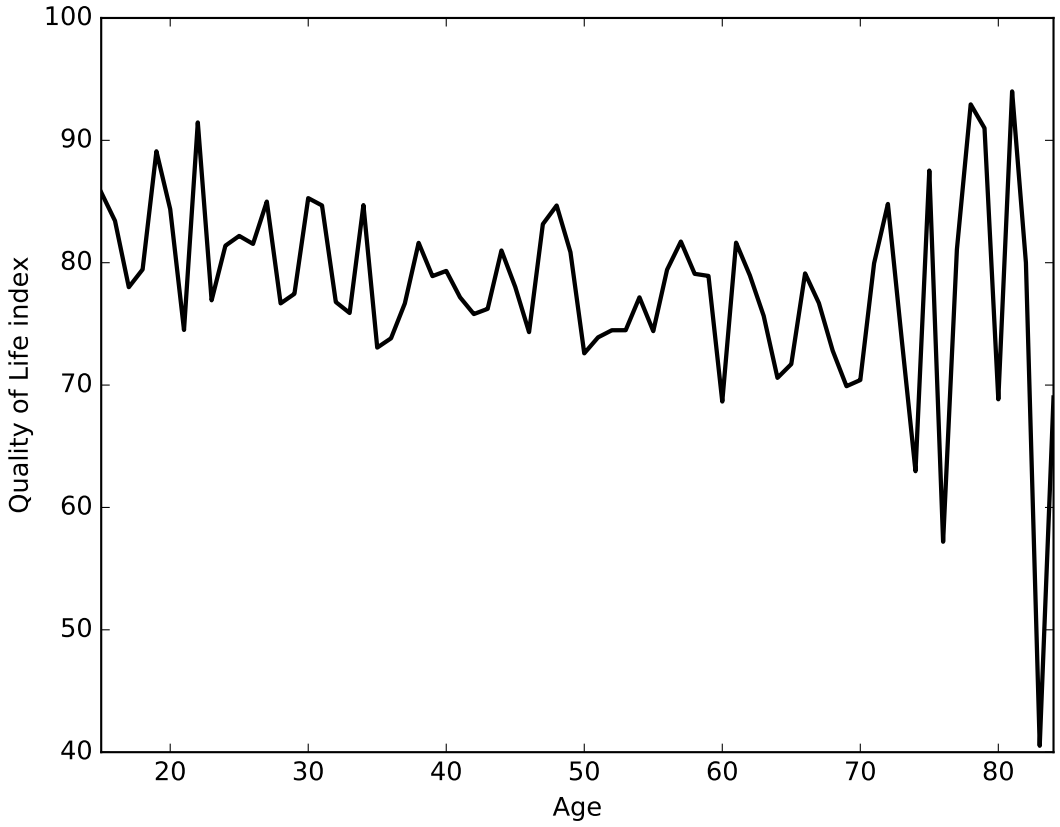

Figure 4.8: The change of quality of life index vs. age

In this figure we can see zero at the beginning means 15 years old. So, the oscillation at the beginning is completely traceable and also at the end, because of lack of data, the behavior of the plot is not very smooth. So we can cut two ends of our dataset to have a smoother dataset. Finally, we have age as a risk factor which ranges from 19 years old to 80 instead of 15-83. Figure 4.9, below summarizes the preprocessing step in one diagram.

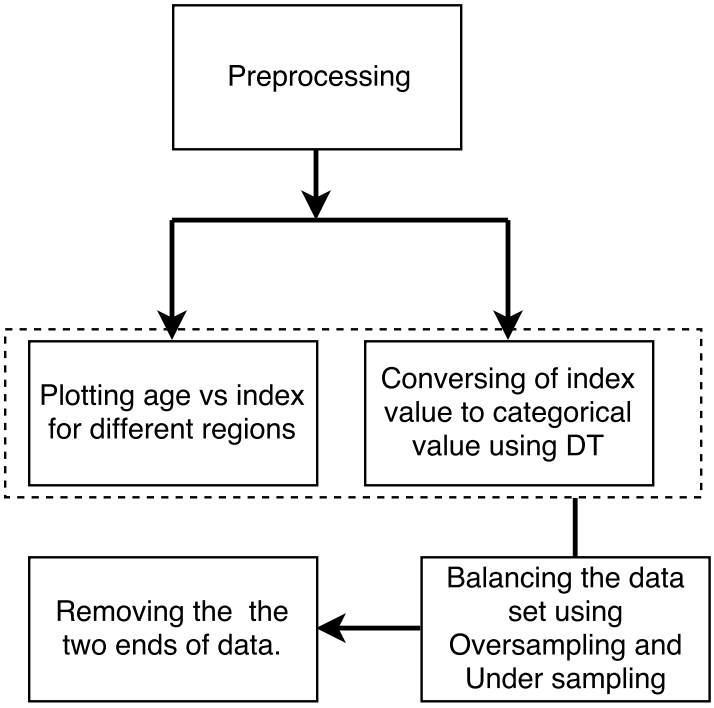

Figure 4.9: Step1: Preprocessing of Quality of Life Index Data

4.6.2 Analysis and Results

The preprocessing of the data set is accomplished using WEKA 3.8.2. Then the extracted dataset from WEKA is imported to Python 2.7 for further analysis. By using SKlearn package in python, a Random forest with different number of iterations will be

trained. The 60 percents of the dataset is separated as train data and the rest is kept as test(20%) and validation data(20%) to validate the final model.

In the next step, we train the RF with a different number of trees as the base classifier and examine them using the test data. The accuracy of the developed RF on test data after 1000 iterations is **83 percent**. Figure 4.10 illustrates the normalized confusion matrix. The confusion matrix shows that our model can capture correctly the High, Medium and Low levels in 61, 94 and 80 percent of the points respectively, but the accuracy is not enough to judge the method.

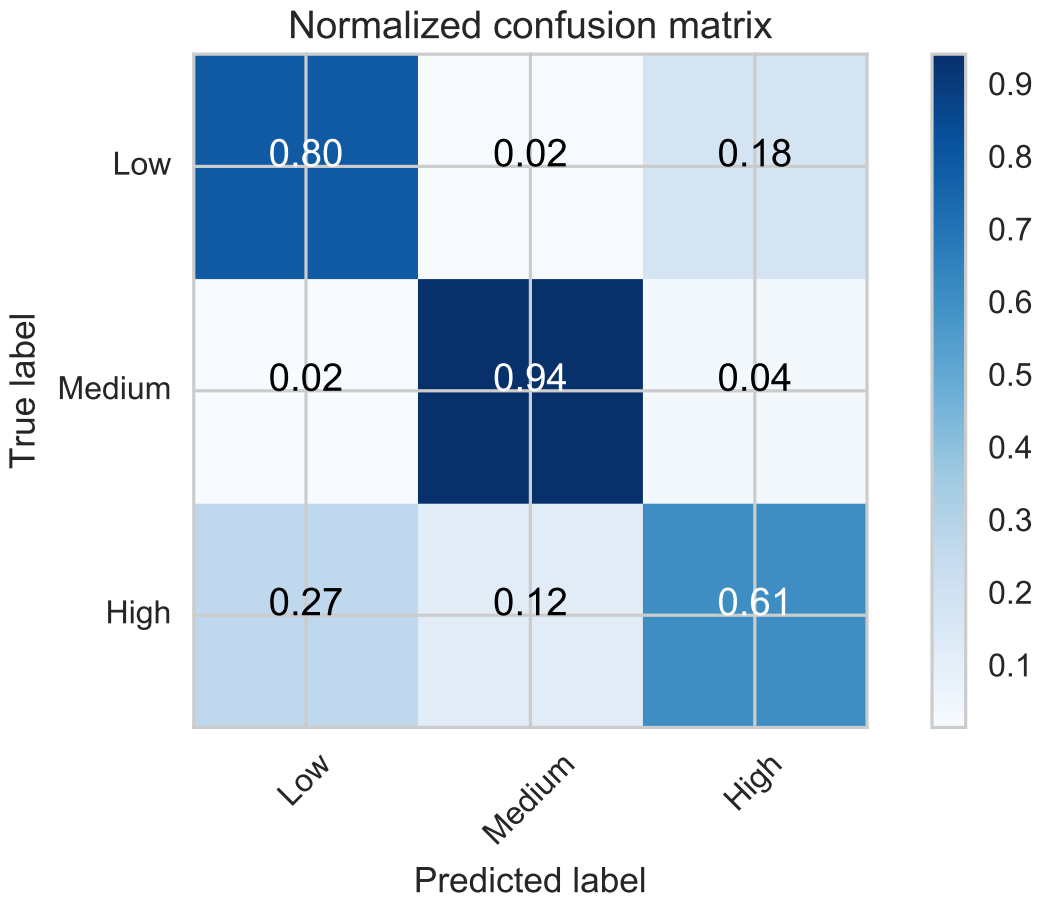

Figure 4.10: Normalized Confusion Matrix

Table 4.3,below shows the detail of classification process for each level of quality of life index. Precision, recall, and F1-scores which the last one is the indicator of a trade-off between Type I and Type II error are shown in this table.

This model can perfectly recall the three levels of the quality of life, Low, Medium, and High in 80, 94 and 61 percent of points respectively and it is precise in 77, 90 and 69 percent of recalled points, respectively. Therefore, on the average (F1-score is the harmonic average of precision and recall) 78, 92 and 65 percent of recalled points are classified precisely.

	Precision	Recall	F1-score
Low	0.77	0.80	0.78
Medium	0.90	0.94	0.92
High	0.69	0.61	0.65
Average	0.80	0.81	0.80

Table 4.3: Table of the result.

The last and one of the most important table and graph resulted from Random Forest is the Feature Importance Rank. Table 4.4 and Figure 4.11 given below, illustrate clearly this importance. From these two plots, we can extract some important information. Each factor should be interpreted independently or by combining other factors. Also, we can investigate this model and its result from an individual point of view or a more broad view such as an administrative unit such as government. We can see that income has the highest contribution to quality of life, but a person who is in his/her late 60s is hard to change income level or education. So instead they should concentrate on health background. From a governmental point of view, since they have more power to change these risk factors, they can decide more efficiently to allocate the limited sources of funds based on the rank of the risk factors.

Table 4.4: Variable importance

Feature	**Contribution**	**Feature**	**Contribution**
Income	0.131	Marital status	0.046
Class-age	0.132	Depression	0.014
Education	0.119	Heart failure	0.012
Occupation	0.101	Weakness in arms	0.011
Region	0.090	Diabetes	0.009
Gender	0.071	Blindness	0.009
Sciatica	0.069	Angina	0.003
Arthritis	0.062	Heart attack	0.003
Municipality	0.057	Mental disorder	0.001
Hypertension	0.050		

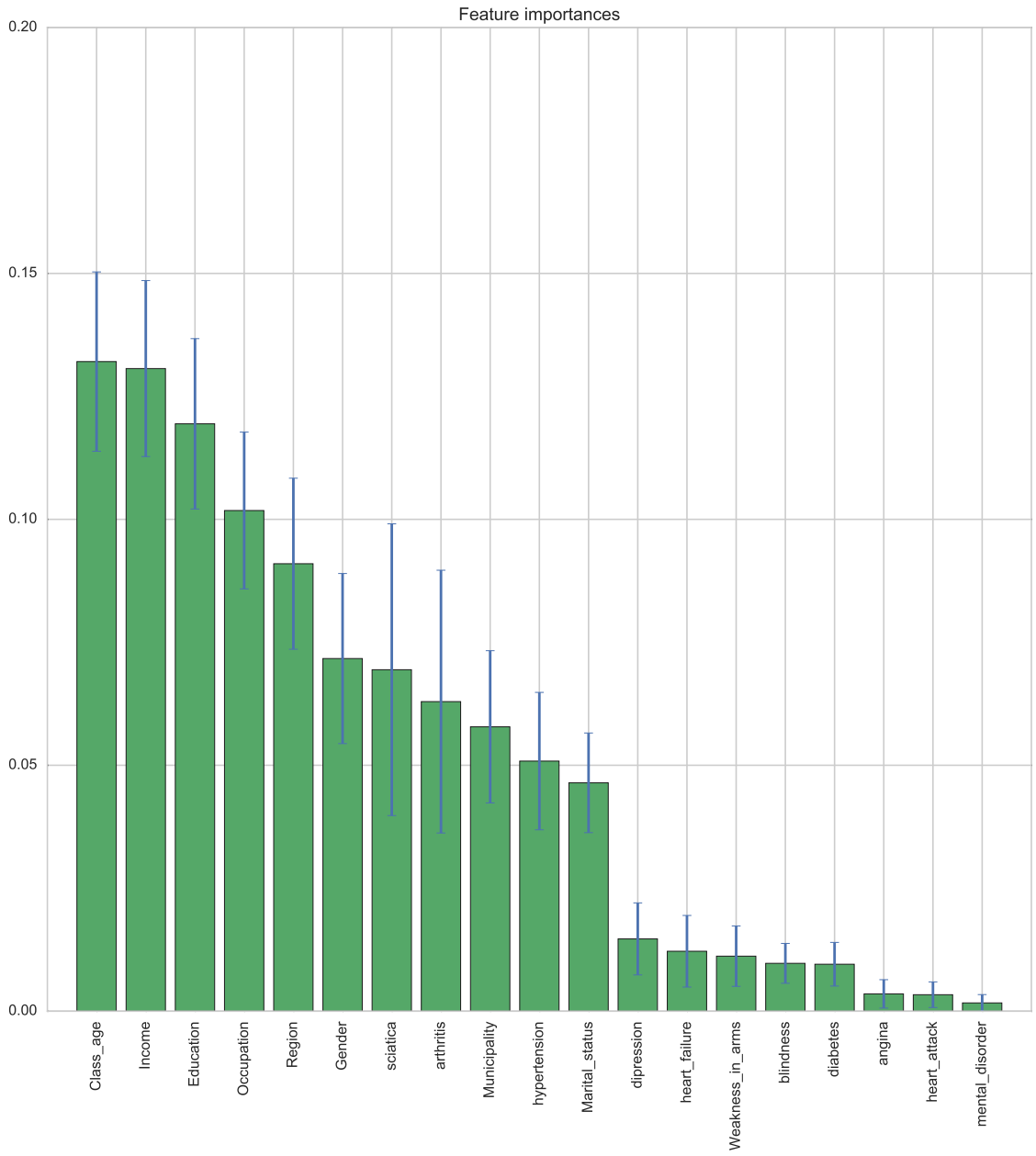

Figure 4.11: Variables Importance Rank

4.7 Model Calibration and Validation

The calibration is a key entity in modeling prediction and validation. In classification, finding the probability of the predicted label will increase confidence on the prediction. The calibration allows us to predict the labels with more confidence on a new input. In other word, a well-calibrated classifier is a probabilistic classifier for which the probability of a label can be directly interpreted as a confidence level. The Sigmoid calibration is used for calibration of the model. The Figure 4.12 shows the calibrated model developed by Random Forest. When performing classification we often want not only to predict the class label, but also obtain a probability of the respective label. This probability gives us some kind of confidence on the prediction. The calibration module allows you to better calibrate the probabilities of a given model, or to add support for probability prediction. For instance, a well calibrated (binary) classifier should classify the samples such that among the samples to which it gave a probability value close to 0.8, approximately 80% actually belong to the positive class. Illustrated figure 4.12 is the standard 2-simplex, where the three corners correspond to the three classes. Arrows point from the probability vectors predicted by an uncalibrated classifier to the probability vectors predicted by the same classifier after Sigmoid calibration on a hold-out validation set. Colors indicate the true class of an instance (red: Low, green: Medium, blue: High). If this classifier is trained on all train data, it is overly confident in its predictions and thus incurs a large log-loss. Calibrating the Random Forest classifier, which was trained on train data, with method="sigmoid" on the validation data set reduces the confidence of the predictions, i.e., moves the probability vectors from the

edges of the simplex towards the center. This calibration results in a lower log-loss.

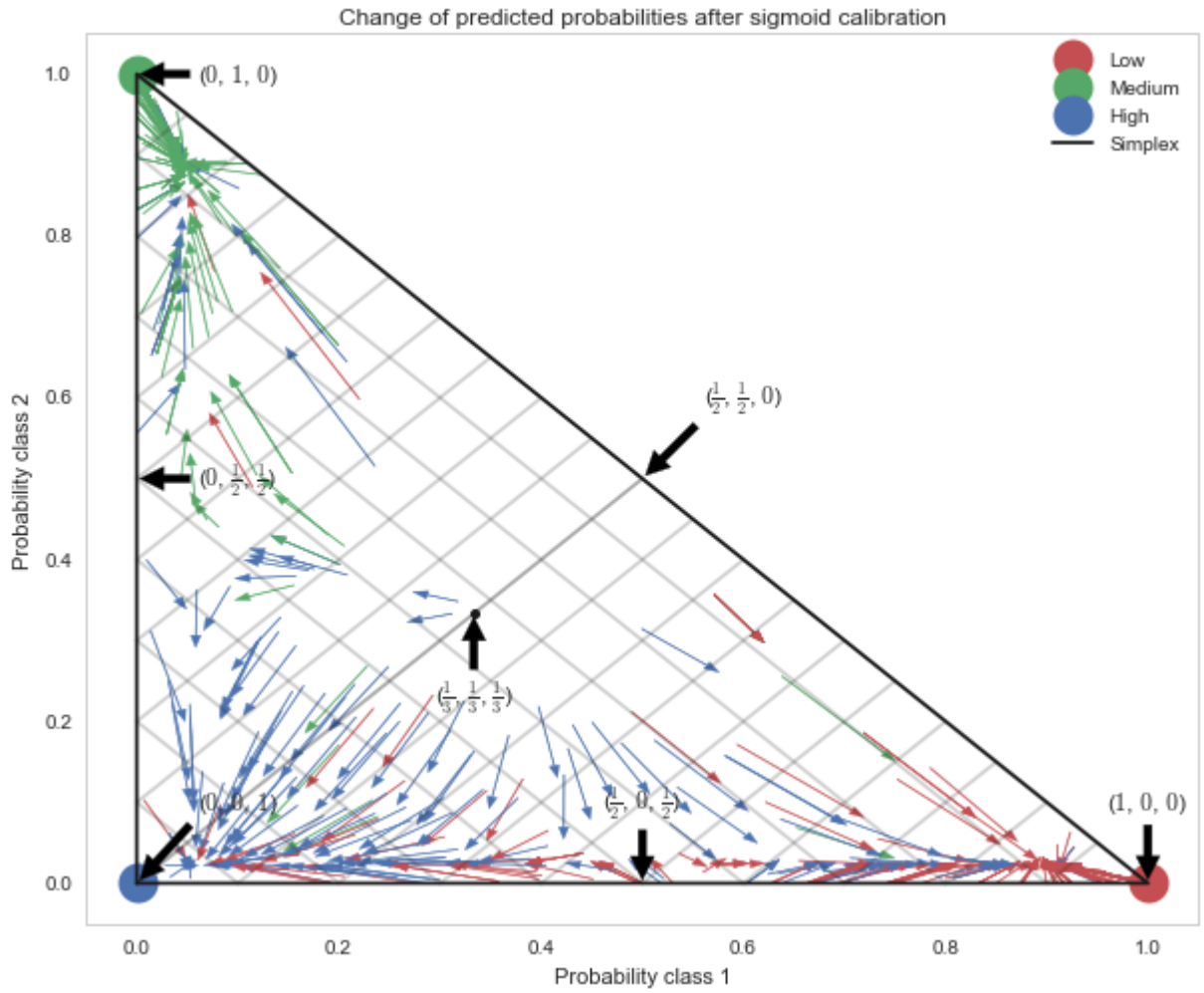

Figure 4.12: Calibration of Classifier Using Sigmoid Function

After the calibration process, the log loss is deducted almost 50% from the 0.807 to 0.487 which is the indication of a more reliable classifier. The reliability plot in Figure 4.12 shows that the probability of the predicted labels have been distributed uniformly after calibration.

A sample from the dataset was considered as unseen data or test data to test the model. The

receiver operating characteristic, ROC curve is plotted to illustrate the performance of the model on the test data. The ROC curve illustrates the trade-off between recall, ability to find all relevant instances in a dataset, and precision, the fraction of relevant instances among the retrieved instances, which identifies the quality and power of the classifier, respectively. Figure 4.13,below illustrates the performance of the classifier on each class of the response variable. This graph confirms the excellent performance of the proposed model. This curve, also gives us more information than accuracy alone, that is, the area under this curve is another trade-off between true positive rate and false positive rate. Furthermore, we can observe that all the curves in ROC curve are near to one which is an indicator of a decent model.

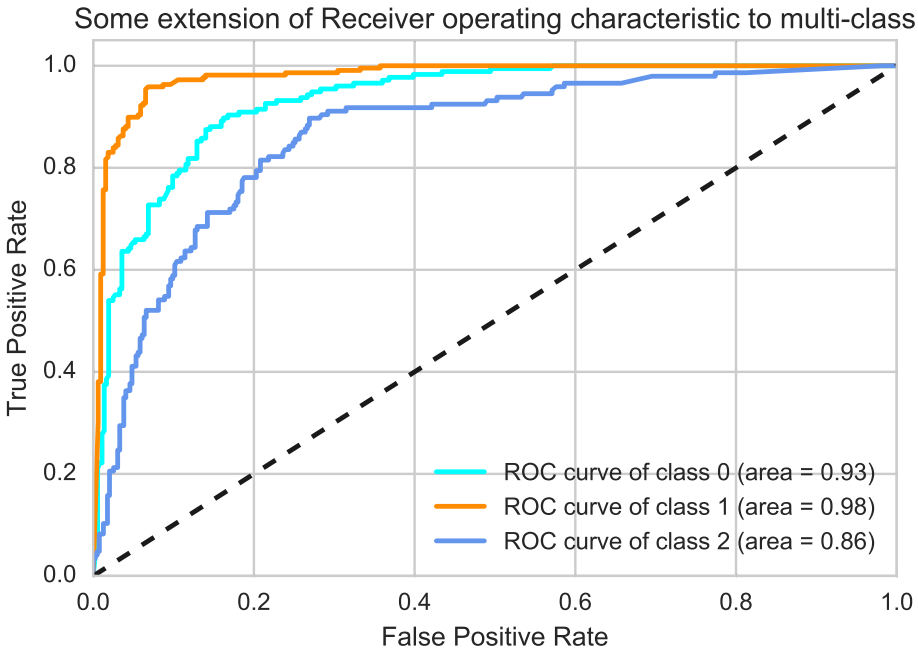

Figure 4.13: ROC curve

4.8 Contribution

In this study, we have demonstrated that several risk factors have a significant contribution to the quality of life. Age, income, education, occupation, region of living, gender, and existence of some of the health problems are the most important risk factors in prediction of the quality of life. Working on the quality of life is an important issue for individuals and governments. For a government, it is important to find the best factors to conduct the source of funds to increase the quality of life. For individuals, we presented a list of risk factors with different ranks that they can work to increase their quality of life. Probably for an individual is not possible to change all of those risk factors but they may choose a subset of them to change based on their abilities.

In this paper, we connected classical methods of statistical analysis with state-of-the-art machine learning models to find the best risk factors which have a contribution to the quality of life index. The developed model using Random Forest can predict the index of a new individual with 83 percent accuracy which can answer all of the questions which we wanted to answer in this study. The model developed in this study was calibrated to reduce 50 percent of the log-loss value and gives more reliable classifier. The same process can be done for a new data set to monitor the status of the quality of life for any specific region for a long time and work on different risk factors found in this study to improve the quality of life. The contribution of this section can be summarized as:

- We developed a calibrated 3-class(Low, Medium, High) predictive model with accuracy of 83% to predict the general quality of life level.

- The recall of 80% and 94% for Low and medium levels which means 80% of low quality of life and 94% of medium level population can be detected by this model.

- The classical statistical models and test, Kruskal-Wallis test conjunction with machine learning methods, Random Forest, are used to model the data.

- The developed model can be used both by administrative organizations such as insurance companies and individuals to monitor the general quality of life over time.

4.9 Acknowledgement

The authors acknowledge the support of Dr. Enzo Grossi in providing us the experimental data set.

CHAPTER 5 : QUALITY OF LIFE:LATEX ERROR: SOMETHING'S WRONG–PERHAPS A MISSING

5.1 Introduction

Sociological theory and phenomena are often hypothesis-driven and explanation of the reason of the problem is the core process of the analysis (Rudin [37]). The reasons behind each sociological event is not straightforward procedure to explain if the population is non-homogeneous. Therefore, it is appropriate to divide the whole population to sub-populations with more similarity and less variability. One of the most appropriate tool for this purpose is unsupervised learning (clustering) in machine learning to group similar individuals.

Machine learning and data mining methods have enormous applications in areas such as medicine (Díaz-Uriarte and De Andres [73], Statnikov et al. [74]), engineering (Rabiei et al. [65], [66],[2],[67],[68], [69]), finance (Jafarian et al. [70]), social science (Gutiérrez et al. [75], Weng et al. [76]). Machine learning can be divided into two main categories, supervised and unsupervised learning. Supervised learning is typically done in the context of classification or regression. In either case, the output which is called the target value and a set of predictors or features $\mathbf{X}$ acquired. The main idea is to define a function which maps $\mathbf{X}$ to $\mathbf{y}$(target

value), $f(X) = y$.

In unsupervised learning, the goal is to learn inherent structure within the data without using explicitly provided labels. Clustering or cluster analysis is the process of grouping a set of individuals or objects in such a way that each group contains the most similar objects.

The combination of demographic, social and health data can be utilized to explain the critical issues of society and predict the future in advance to have a better life. The importance of demography lies in its contribution to helping government and society better prepare to deal for the issues and demands of population growth, aging and more generally improving the quality of life.

Clustering of social data is a useful tool for administrative purpose especially government and insurance companies. The appropriate clustering can facilitate a government's task to allocate limited source of funds to the proper group of people which have similar characteristics in a society. Moreover, insurance companies can create clusters of individuals with similar risk factors for better predictions.

Data mining and clustering methods have been widely applied to find hidden patterns in mixed social data. Researchers use clustering method to find a similarity among small producers in six cities in the northeast of Brazil (Maione et al. [38]). The social network is another interesting subject for researchers. The clustering of people in a social network using K-means clustering and the textual similarity is a good judgmental tool to find users with similar behavior (Singh et al. [39]). In a human behavior study, the authors study the

human social behavior which is a big data to find similar patterns by means of clustering methods (Ferrara et al. [40]). The reduction of high density area of accidents using GIS, Kernel density estimation, and K-means clustering are the main idea of studies in (Anderson [77]). Relating mobility patterns to socio-demographic profiles highlights the importance of finding patterns of movement for finding various administrative strategies (Liebig [78]).

However, in the most of social experimental design a representative data is collected in form of hybrid data, continuous and categorical variables, and this restrain usage of K-means clustering which is one of the most applicable method in data mining and clustering. K-medoids, the combination of Gower distance (Gower [41]) and K-means, can be used to handle the clustering process of hybrid data . This method has a growing popularity among researchers in different areas of interest, (Velmurugan and Santhanam [42], Arora et al. [43]). The authors(Khatami et al. [79]), utilize optimization method and K-medoids on images to propose a new fire detection. A simple and fast version of K-medoids clustering and some experimental results can be found in (Park and Jun [80]).

In the present study, we investigate social data clustering using K-medoids clustering with Gower distance to find the similarity among individuals from different regions of Italy.

5.2 The Statistical Method

In this section, we will give a brief description of the K-means, K-medoids, Gower distance and Johnson family of probabilistic distribution.

The K-means clustering (Park and Jun [80]) is a common method to partition a set

of observation automatically into k groups. Given a set of observations $x_1, x_2, ..., x_n$, where each observation is a d-dimensional real vector, k-means clustering aims to separate the n observations into $k(<< n)$ sets $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ so as to minimize the within-cluster sum of squares (variance). Formally, the objective is to find the minimum variance:

$$arg\ min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} ||\mathbf{x} - \mu_i||^2 = arg\ min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| Var(S_i)$$

, where $\boldsymbol{\mu}_i$ is the mean of S_i.

This is equivalent to minimizing the pairwise squared deviations of points in the same cluster, that is,

$$arg\ min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{\mathbf{x},\mathbf{y} \in S_i} ||\mathbf{x} - \mathbf{y}||^2.$$

The equivalence can be extracted from the identity

$$\sum_{\mathbf{x} \in S_i} ||\mathbf{x} - \mu_\mathbf{i}||^2 = \sum_{\mathbf{x} \neq \mathbf{y} \in S_i} (\mathbf{x} - \mu_i)(\mu_i - \mathbf{y}).$$

Because the total variance is constant, also is equivalent to maximizing the sum of squared deviations between points in different clusters, which follows the law of total variance. In K-means clustering, the number of clusters, K, and the dataset are two inputs into algorithm. The initial estimates for the K centroids are generated by the algorithm. The algorithm iterated between two steps:

- Data assignment

- Centroids Update

In the first step, each point is assigned to the closest centroids using the Euclidean distance, that is,

$$argmin_{c_i \in C} dist(c_i, x)^2. \tag{5.1}$$

In the second step, the centroids are updated. This is accomplished by taking the mean of all the data points assigned to that centroids cluster,

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i. \tag{5.2}$$

The algorithm iterates between these steps back and forth until a stopping criteria is met and it is guaranteed to converge.

K-means clustering is not applicable for hybrid data which include categorical variables because it needs to calculate the minimum Euclidean distance between two data points and the Euclidean distance mostly has been utilized for continuous variables. Even by encoding categorical variables to numeric values, the result of the analysis is difficult to interpret. Thus, we need to use the K-medoids clustering for mixed data, which is a procedure analogous to K-means clustering for finding similar groups when the data includes some categorical, nominal or ordinal, variables.

The main difference is the definition of the distance that is computed between categorical levels as well as the data of continuous columns. A categorical data can be nominal or ordinal, in which either of them is treated differently in the clustering procedure. The

Gower distance (Gower [41]) is the distance which can handle similarity in mixed data, such as the data investigated in this study.

The Gower distance is a combination of two particular distance metrics, Manhattan and Dice distance, that works quite well for any type of variable.

For quantitative(interval) variables, range-normalized Manhattan distance is calculated. Ordinal variables are first ranked, then Manhattan distance is applied with a special adjustment for ties. Finally, nominal variables with k categories are first converted into k binary columns and then the Dice coefficient is used. The Dice coefficient is equivalent to F1 score in supervised learning (Gower [41]).

Each of these specific distances will be scaled to fall between 0 and 1. Then, a linear combination of using some weights defined by the user (most usually an average) is calculated to generate the final distance matrix. The Gower distance is sensitive to non-normality of continuous variables.

We can summarize the Gower distance as follow:

Let $X = (x_1, x_2, ..., x_n), Y = (y_1, y_2, ..., y_n)$ be two observations.

- **Nominal/Binary:** Simple matching coefficient $d(i, j) = \frac{m}{p}$, where

  m is the number of variables that object i in x and j in y mismatch and p is the number of variables.

- **Ordinal:** we use normalized ranks, then like the continuous variables, Manhattan distance can be applied

- **Continuous/Interval-scaled:** For this type of variables, the normalized Manhattan distance is applied,

$$d_1(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_1 = \sum_{i=1}^{n} \frac{|x_i - y_i|}{R}$$

, where R is the range of the variable.

- **Gower Distance** is defined by $d(i, j) = \frac{1}{p} \sum_{i=1}^{p} d_{ij}^{(f)}$.

After introducing the distance matrix, the next step to select the clustering algorithm that is applicable in this case. There are many algorithms that can handle a custom distance matrix generated from the previous section using Gower distance. The method used in this study is PAM, Partitioning Around Medoids, or simply K-Medoids. K-medoids is another version of K-means which uses observations themselves as centers instead of centroids using Euclidean distance. K-medoids can be summarized as follows:

1. Choose K random entities to become medoids.

2. Assign every entity to its closest medoid, using custom distance.

3. For each cluster, identify the observation that would yield the lowest average distance if it were to be re-assigned as the medoid. If so, make this observation the new medoid.

4. If at least one medoid has changed, return to step 2. Otherwise, end the algorithm.

A variety of metrics exist to help choose the number of clusters to be extracted in performing a cluster analysis. Since the dataset in our study does not include any labels, then silhouette width is the most appropriate metrics to evaluate the clustering analysis.

5.2.1 Silhouette Distance

Let $\mathbf{x}$ be an observation and AVG($\mathbf{x}$) be the mean distance between $\mathbf{x}$ and all other data points within the same cluster. Also, let **Inf**($\mathbf{x}$) be the smallest mean distance of $\mathbf{x}$ to all data points in any other cluster, of which $\mathbf{x}$ is not a member. Then silhouette distance is defined as follow:

$$s(\mathbf{x}) = \frac{\mathbf{Inf}(\mathbf{x}) - AVG(\mathbf{x})}{max\{\mathbf{Inf}(\mathbf{x}), AVG(\mathbf{x})\}}.$$

This metric can range from -1 to 1, where higher values are more desirable. The closer s($\mathbf{x}$) is to 1, the better the cluster are resulted. In addition to the Silhouette Distance, the validation of clusters is investigated by finding the probability distribution of the clusters and making comparisons between the different clusters.

Another important entity in the present study in the Johnson family of probability distributions. It is important to justify that all the clusters follow Johnson probability distribution (Johnson [81]) with different parameters. The Johnson family of distributions include three different equations: SU, SB, and log-normal. The subject probability distribution supports any specified measurement of the central tendency such as mean, standard deviation, skewness and kurtosis as well. All together, they form a 4-parameter family of probability distributions. The probability density function of Johnson SB is defined below. It is the best

bounded probability distribution. The Johnson 4-parameter probability density is given by

$$f(x) = \frac{\delta}{\lambda\sqrt{2\pi}z(1-z)} exp(-\frac{1}{2}(\gamma + \delta \ln(\frac{z}{1-z}))^2), \tag{5.3}$$

$$where\ z \equiv \frac{x-\epsilon}{\lambda},\ 0 \leq z \leq 1.$$

In the above pdf, δ and γ are shape parameters and ϵ , λ are location and scale parameters, respectively. The most significant advantage of this pdf is a tight relation with Normal probability distribution. This probability distribution is used as transformation to convert a non-normal data set to normal, that will satisfy the basic assumption of using the appropriate methodology. The Johnson Transformation is extracted from the above pdf and the result will be in the form of:

$$y = \gamma + \delta ln(\frac{x-\epsilon}{\lambda+\epsilon-x}). \tag{5.4}$$

In the defined form, x is the original raw data and y is transformed data. A histogram of the data with the Johnson SB pdf is given by Figure 5.1.

5.3 The QoL Data

The original data which was collected from different regions of Italy by Doxa, the Italian branch of the Gallup International Association, is the combination of two questionnaires. The first is the summary of demographic information of individuals involved in the study as well as health background. The second form summarizes the internal feeling and psychological background of units in the experiment. The latter information is extracted

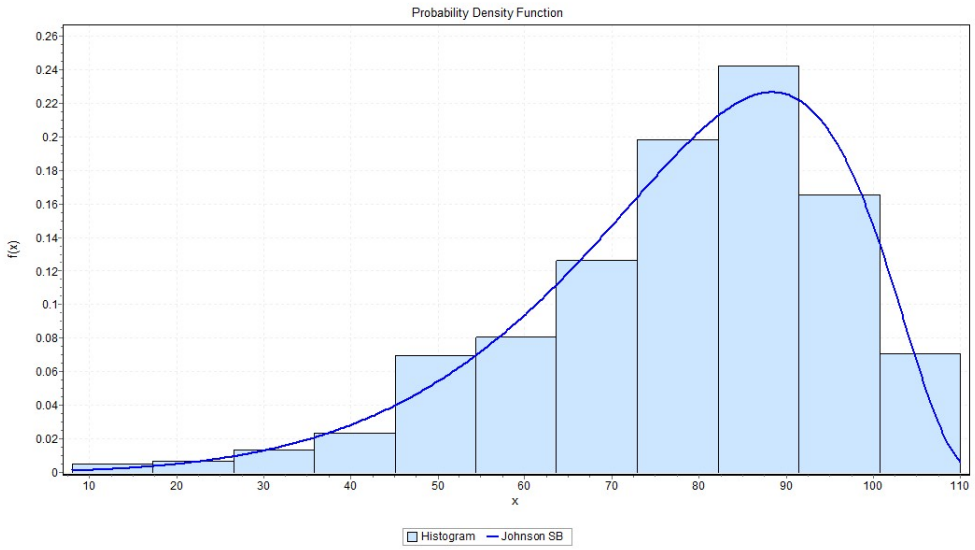

Figure 5.1: Data follow Johnson SB distribution

by asking individuals to fill out PGWBI, Psychological General Well-being Index, questionnaire, a 22-question questionnaire developed by Harold J Dupuy in 1971. PGWBI divides each individual's internal sense into six categories. Anxiety, depression, general health, self-control, vitality, and positive well-being are six categories that constitutes the questionnaire. Table 5.1 summarizes the basic properties of the subject data. In addition to six categories that cab be extracted from the second category, one overall column as PGWBI index is generated by aggregating all of six categories. The range of this column is from 0 to 110. This column is utilized to validate the result of clustering. The Figure 5.1 is the distribution of this column.

Table 5.1: Description of Variables

Variable	**Type**	**Levels**
Region	Nominal	3
Municipal	Nominal	2
Amplitude	Ordinal	5
Gender	Nominal	2
Education	Ordinal	5
Marital Status	Nominal	4
Occupation	Nominal	7
Income	Ordinal	5
Age	Continuous	-
Anxiety	Continuous	-
Depression	Continuous	-
Welfare	Continuous	-
Self Control	Continuous	-
Health	Continuous	-
Vitality	Continuous	-
Diseases	Binary	16 diseases and disabilities
Number of Observations		**1027**
Number of Attributable variables		**31**

In Table 5.2, the basic statistics of all continuous variables is illustrated.

Table 5.2: Basic Statistics of Continuous Variables

Variable	**Mean**	**SD**	**Range**
Age	45.83	16.53	[18, 93]
Anxiety	17.31	4.92	[1, 25]
Depression	12.37	2.64	[0, 15]
Welfare	11.81	3.97	[0, 20]
Self Control	11.89	2.76	[1, 15]
Health	11.05	3.11	[0, 15]
Vitality	13.42	3.97	[0, 20]

5.4 The Analysis

The algorithm which is based on PAM or K-medoids, is implemented for analyzing the subject data. The first step is to determine the number of clusters, K. For this goal, a variety of metrics exist to help choose the number of clusters to be extracted in a cluster analysis. The silhouette width an internal validation metric which is an aggregated measure of how similar an observation is to its own cluster compared with its closest neighboring cluster, is implemented. Figure 5.2 illustrates the relationship between the number of clusters versus the silhouette width. It shows that 3 clusters is the best choice for our dataset. The result of cluster analysis using three clusters is visualized in Figure 5.3.

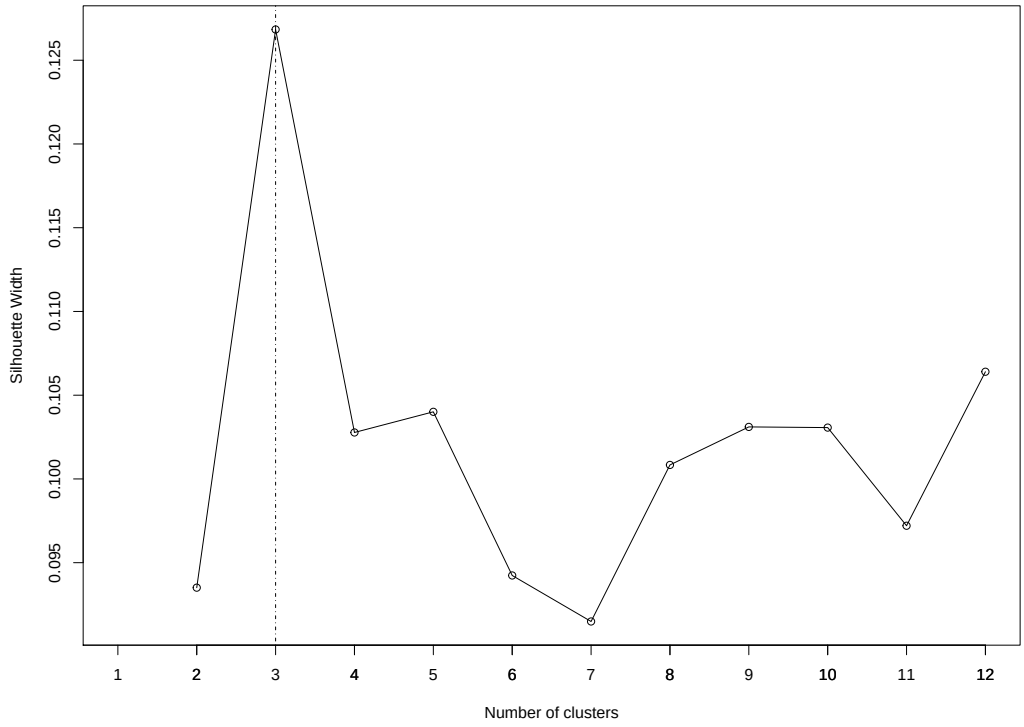

Figure 5.2: Determination of the number of cluster using Silhouette Width

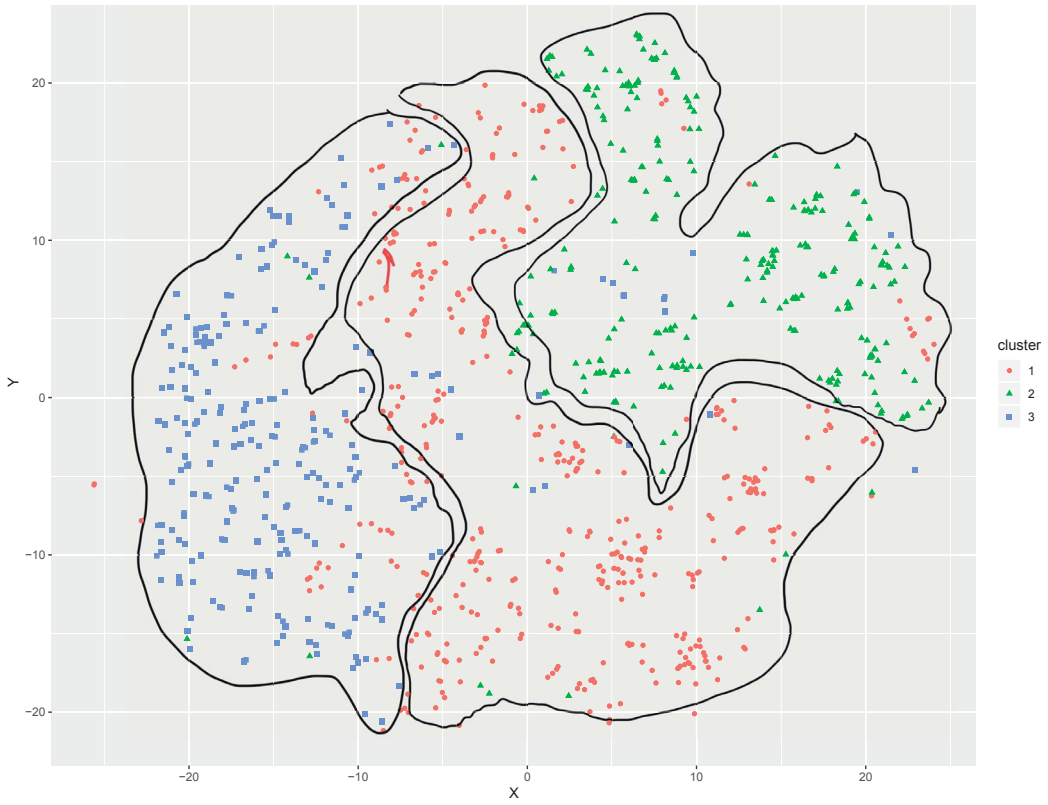

Figure 5.3: Result of Cluster Analysis of the PGWBI Data

There are some overlapping among the three clusters in Figure 5.3 which is confirmed by the Silhouette graph plotted in Figure 5.2. The Silhouette coefficient is closer to zero than one which it validates the result of cluster analysis but it shows some overlapping among clusters. As it is observable the overlapping is negligible and the clusters are fairly distinguishable. The results in the study have been programmed in R 3.5.0. PAM algorithm can be implemented using the Daisy function (Maechler et al. [82]). This function can accept all types of variables including continuous, nominal, and ordinal. As previously explained, this function accepts the dissimilarity matrix using Gower distance.

Table 5.3: Three Medoids of K-Medoids Clustering of the PGWBI Data

Risk factors/Medoids	**164/cluster 1**	**618/ cluster 2**	**803/ cluster 3**
Region	1	3	1
Municipal	2	2	2
Gender	2	1	1
Marital.status	2	1	2
Occupation	3	3	5
Amplitude	2	2	2
Education	2	2	4
Income	4	4	3
Hypertension	0	0	0
Heart attack	0	0	0
Heart failure	0	0	0
Diabetes	0	0	0
Angina	0	0	0
Cancer	0	0	0
Allergy	0	0	0
Arthritis	0	0	1
Sciatica	0	0	0
Blindness	0	0	0
Lungs	0	0	0
Dermatitis	0	0	0
Deafness	0	0	0
Weakness in arms	0	0	0
Depression	0	0	0
Mental disorder	0	0	0
Age	50	28	61
Anxiety	19	15	17
Depression	12	13	12
Welfare	13	14	11
Self control	13	13	14
Health	14	12	12
Vitality	20	12	15

5.5 Interpretation and discussion

This section delves into each cluster in detail and investigate the contribution of each risk factor introduced in Table 5.1. Table 5.3 shows three final and stable medoids, in which all observations accumulated around them.

5.5.1 Cluster 1

Tables 5.4 and 5.5 and Figures 5.4, 5.5 describe the details of the cluster 1. Most of the people are middle-age with average age of 44.57 year and standard deviation of 13.28. Anxiety, depression and self-control are not major problems in this cluster. Welfare and Health are two more important problems stated by individuals themselves. About 75% have a feeling of welfare less than 15 out of 20.

Table 5.4: Basic Statistics of Continuous Variables of cluster 1

Variable	**Mean**	**SD**	**Range**	Q_1	Q_3
Age	44.57	13.28	[18, 82]	34	53
Anxiety	16.95	4.91	[1, 25]	14	20
Depression	12.31	2.61	[1, 15]	11	14
Welfare	11.75	4.01	[0, 20]	9	15
Self Control	11.84	2.72	[1, 15]	10	14
Health	11.15	3.00	[0, 15]	9	13
Vitality	13.33	3.98	[0, 20]	11	16

The cluster 1 is the biggest of all clusters with 481 members that the majority, 60%,originates from the North. About 60% of the cluster originates from the North and 25% of the people come from the South Italy. The majority of individuals in this cluster are female(86%), married(83%), employee(85%), earn below average salary(84%), and 90% hold a degree less than high school. Hypertension (15%), arthritis (26%), allergy (15%), sciatica (22%),and dermatitis (7%) are major health problems in this group. Heart failure, blindness, deafness, and weakness in arms are other minor issues.

Table 5.5: Summary of Categorical Variables in Cluster 1

Size of Cluster 1 = 481								
Variables	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
region	-	288	74	119	-	-	-	-
municipal	-	183	298	-	-	-	-	-
gender	-	68	413	-	-	-	-	-
marital.status	-	47	397	21	16	-	-	-
occupation	-	24	93	318	4	20	10	12
amplitude	59	147	146	66	63	-	-	-
education	-	49	185	142	104	1	-	-
income	-	18	60	125	180	98	-	-
hypertension	409	72	-	-	-	-	-	-
heart attack	479	2	-	-	-	-	-	-
heart failure	476	14	-	-	-	-	-	-
diabetes	469	12	-	-	-	-	-	-
angina	479	2	-	-	-	-	-	-
cancer	471	10	-	-	-	-	-	-
allergy	411	70	-	-	-	-	-	-
arthritis	357	124	-	-	-	-	-	-
sciatica	375	106	-	-	-	-	-	-
blindness	463	18	-	-	-	-	-	-
lungs	460	21	-	-	-	-	-	-
dermatitis	448	33	-	-	-	-	-	-
deafness	463	18	-	-	-	-	-	-
Weakness in arms	464	17	-	-	-	-	-	-
depression	464	17	-	-	-	-	-	-
mental disorder	474	7	-	-	-	-	-	-

h

Finally, this cluster is a collection of all sampled regions where married females are the most questioned individuals. People are middle-age with some health problems specifically arthritis, sciatica, and hypertension.

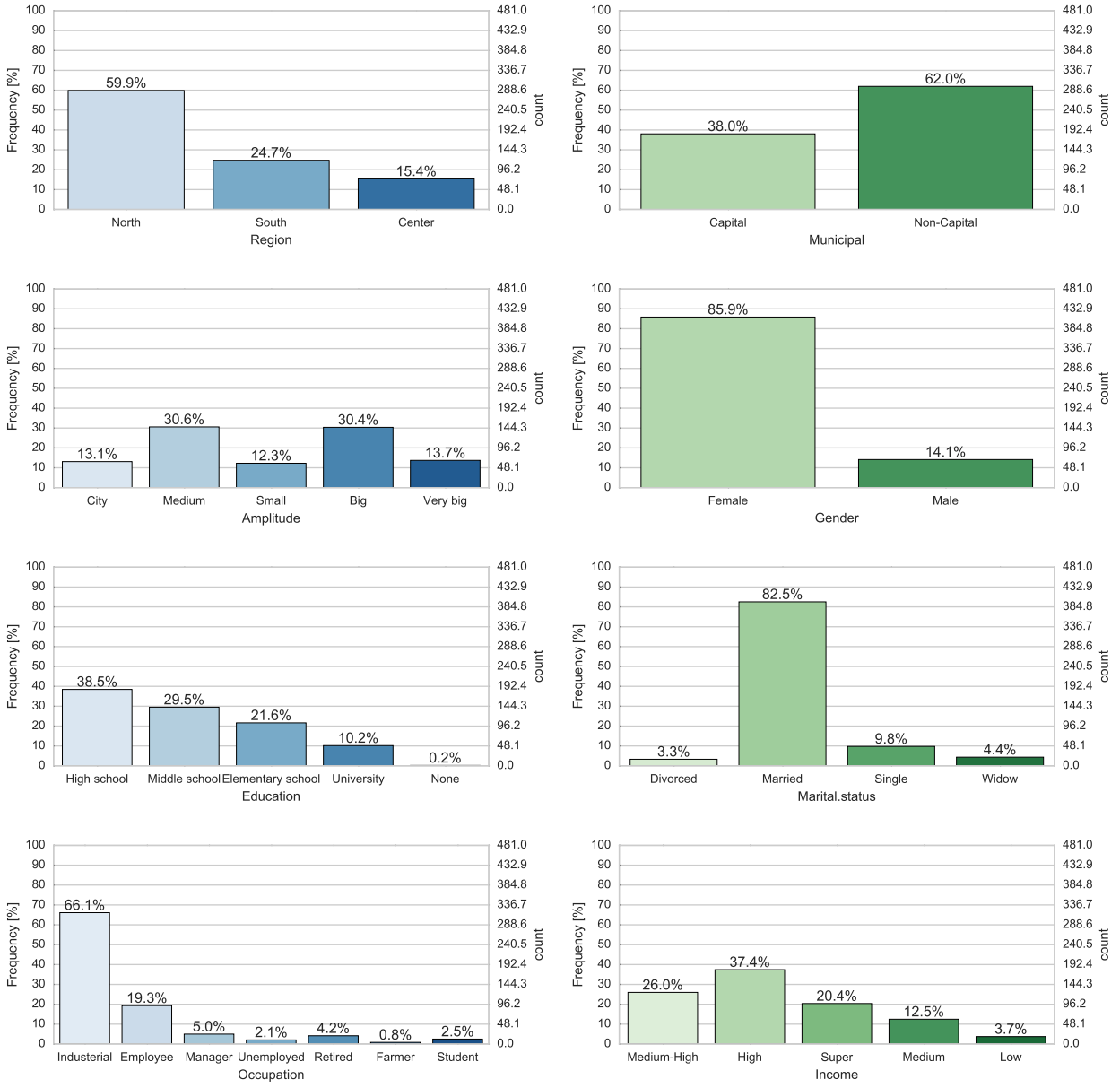

Figure 5.4: Cluster 1 Demographic Distribution

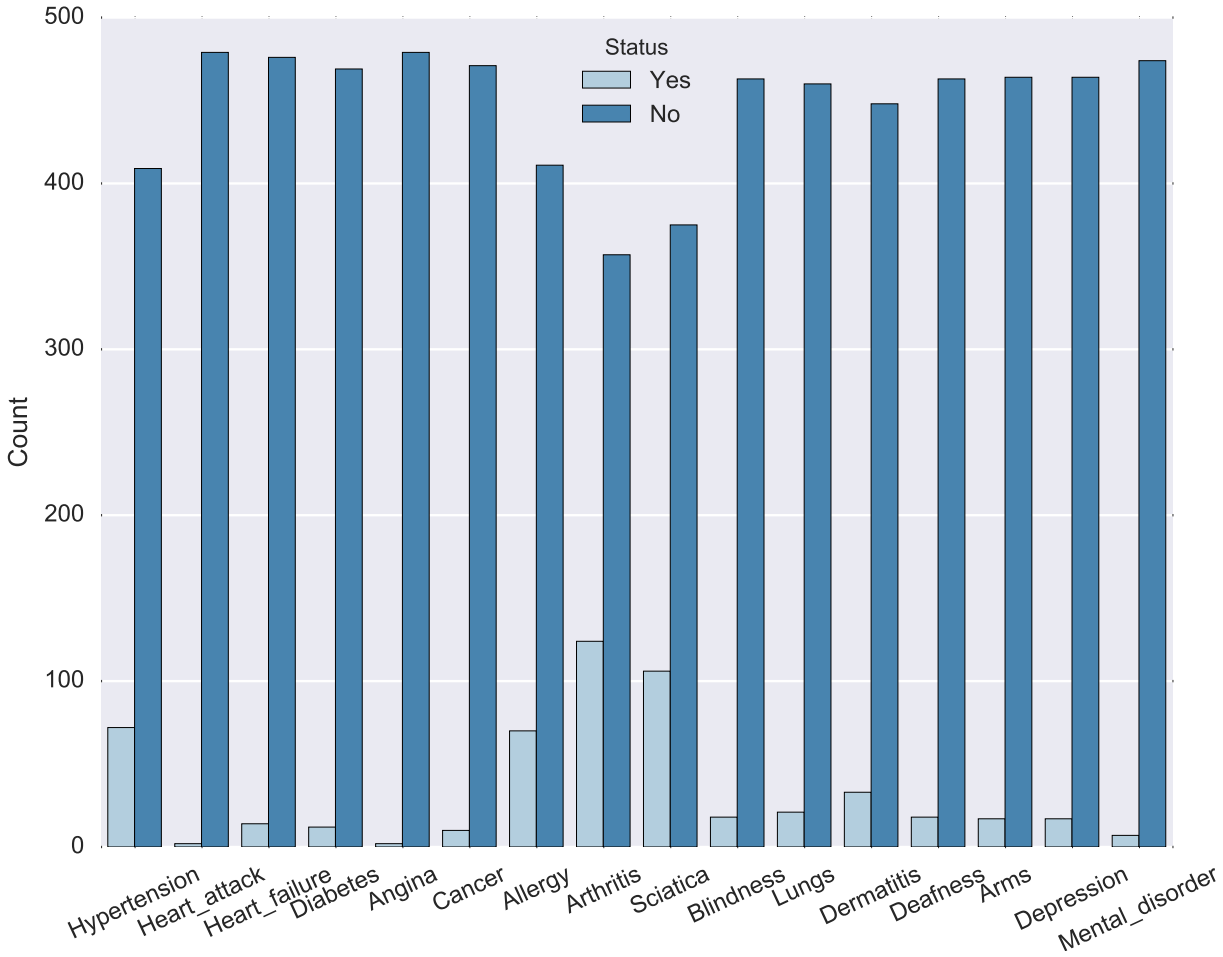

Figure 5.5: Cluster 1 Health Major Problem Distributions

5.5.2 Cluster 2

Tables 5.7, 5.6 and Figures 5.6, 5.7 summarize the properties of categorical and continuous variables in cluster 2 respectively. Generally this cluster is the youngest among all clusters. 75% of people are under 39 years old. From Table 5.6, we can observe that cluster 2 is the healthiest cluster psychologically. Anxiety has an average of 17.59 which is indication of calmness and a good internal feeling. The majority of people have not experienced any level of depression. As we mentioned above, this cluster is the healthiest one physically and

psychologically. High vitality is another good aspect of this cluster.

Table 5.6: Basic Statistics of Continuous Variables of cluster 2

Variable	**Mean**	**SD**	**Range**	Q_1	Q_3
Age	32.31	11.19	[18, 70]	23	39
Anxiety	17.59	4.77	[3, 25]	15	21
Depression	12.69	2.40	[3, 15]	12	14
Welfare	12.37	3.89	[3, 20]	9	16
Self Control	12.2	2.62	[3, 15]	11	14
Health	12.17	2.41	[4, 15]	11	14
Vitality	14.12	3.62	[1, 20]	12	17

283 people (27%) out of 1027 of people in this experiment fall into cluster 2. Almost 60% of people come from south of Italy. The size of the cities where they live is mainly medium or large. The majority of of people are single and male with at least high school degree. About 75% of experimental units claimed that they have a salary higher than the average. Major health problems are almost rare in this cluster. Only 8% reported hypertension, 13% allergy, 5% arthritis, and 6% dermatitis.

Table 5.7: Summary of Categorical Variables in Cluster 2

Size of Cluster 2 = 283								
Variables	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
region	-	69	44	170	-	-	-	-
municipal	-	100	183	-	-	-	-	-
gender	-	227	56	-	-	-	-	-
marital.status	-	208	70	2	3	-	-	-
occupation	-	21	68	96	4	7	27	60
amplitude	32	84	99	40	28	-	-	-
education	-	78	117	75	13	0	-	-
income	-	22	44	90	82	45	-	-
hypertension	260	23	-	-	-	-	-	-
heart attack	279	4	-	-	-	-	-	-
heart failure	279	4	-	-	-	-	-	-
diabetes	279	4	-	-	-	-	-	-
angina	280	3	-	-	-	-	-	-
cancer	283	0	-	-	-	-	-	-
allergy	246	37	-	-	-	-	-	-
arthritis	269	14	-	-	-	-	-	-
sciatica	261	22	-	-	-	-	-	-
blindness	277	6	-	-	-	-	-	-
lungs	279	4	-	-	-	-	-	-
dermatitis	267	16	-	-	-	-	-	-
deafness	279	4	-	-	-	-	-	-
Weakness in arms	281	2	-	-	-	-	-	-
depression	280	3	-	-	-	-	-	-
mental disorder	280	3	-	-	-	-	-	-

All in all, this cluster in the youngest, healthiest, and well-educated group with a high quality of life, and most of them are male originates from the south of Italy.

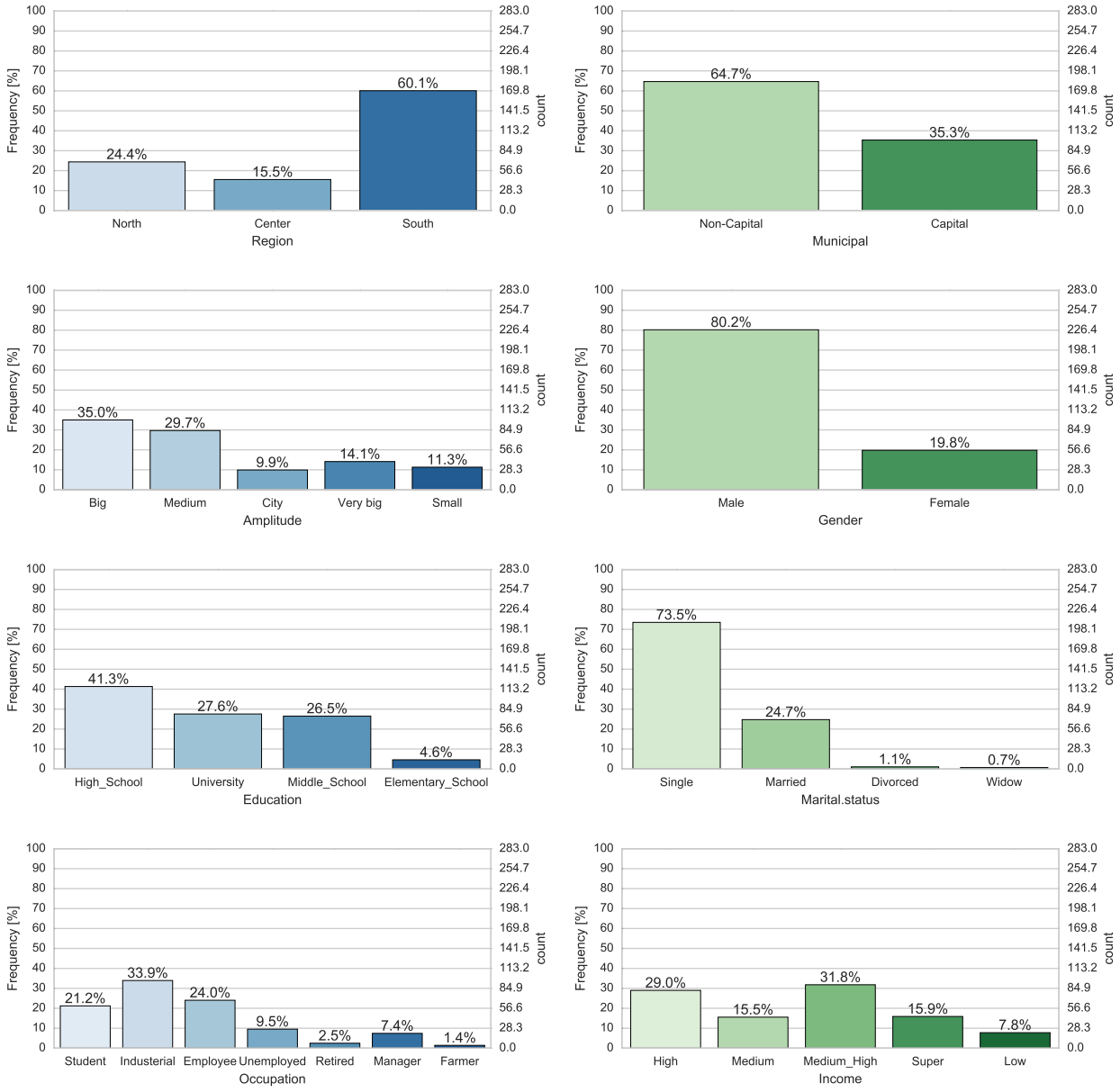

Figure 5.6: Cluster 2 Demographic Distribution

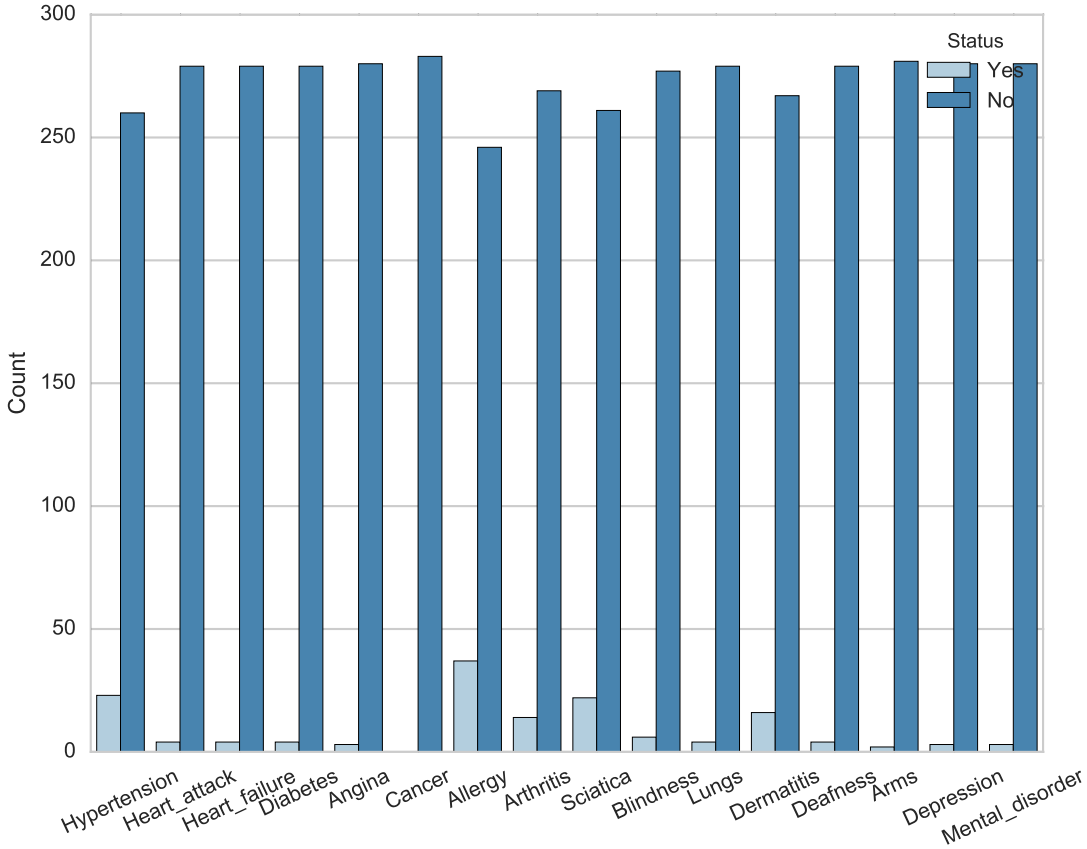

Figure 5.7: Cluster 2 Health Major Problem Distributions

5.5.3 Cluster 3

Tables 5.9, 5.10 and Figures 5.8, 5.9 present some statistical information regarding cluster 3. On average, this cluster is the oldest cluster. The average age is 62.68 with standard deviation of 11.24. The most self-confessed psychological problems in this cluster are welfare, health, vitality. Almost 25% have a welfare level of less than 8 out of 20. On the other hand, 50% of individuals have health status of less than 10 out of 15. Furthermore, about 25% of people are not very energetic since they have a level of vitality less than 10 out of 20.

Table 5.8: Basic Statistics of Continuous Variables of cluster 3

Variable	**Mean**	**SD**	**Range**	Q_1	Q_3
Age	62.68	11.24	[25, 93]	57	70
Anxiety	17.66	5.08	[2, 25]	14	22
Depression	12.16	2.92	[0, 15]	11	14
Welfare	11.3	3.92	[0, 20]	8	15
Self Control	11.63	2.97	[1, 15]	10	14
Health	9.67	3.42	[0, 15]	7	13
Vitality	12.83	4.20	[1, 20]	10	16

This cluster is the smallest group among all the clusters. The majority of people originated from non-capital cities(62%) of northern(57%) Italy which is medium size in population. Almost 76% of people are male. On the other hand, almost 76% of individuals are retired. The big portion of this cluster(75%) consists of married male people. About 51% of individuals have only elementary school education. Salary-wise, this cluster has the majority of the people below average(71%).

The least healthy cluster is cluster 3. 72% of people suffer from arthritis and 40% tolerate sciatica. Moreover, at the same level of sciatica, people should control their hypertension. Heart failure (16%), deafness(18%), allergy(14%), diabetes(13%), lungs' problem (12%), weakness in arms(11%) are other health problems, but they are not as common as the aforementioned issues.

Table 5.9: Summary of Categorical Variables in Cluster 3

Size of Cluster 3 = 263								
Variables	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
region	-	149	49	65	-	-	-	-
municipal	-	101	162	-	-	-	-	-
gender	-	200	63	-	-	-	-	-
marital.status	-	20	198	40	5	-	-	-
occupation	-	8	16	27	7	199	6	0
amplitude	24	89	75	45	30	-	-	-
education	-	22	50	58	131	2	-	-
income	-	32	45	91	64	31	-	-
hypertension	157	106	-	-	-	-	-	-
heart attack	242	21	-	-	-	-	-	-
heart failure	220	43	-	-	-	-	-	-
diabetes	230	33	-	-	-	-	-	-
angina	244	19	-	-	-	-	-	-
cancer	253	10	-	-	-	-	-	-
allergy	227	36	-	-	-	-	-	-
arthritis	76	187	-	-	-	-	-	-
sciatica	157	106	-	-	-	-	-	-
blindness	240	23	-	-	-	-	-	-
lungs	232	31	-	-	-	-	-	-
dermatitis	249	14	-	-	-	-	-	-
deafness	217	46	-	-	-	-	-	-
Weakness in arms	235	28	-	-	-	-	-	-
depression	253	10	-	-	-	-	-	-
mental disorder	260	3	-	-	-	-	-	-

Altogether, although this cluster is the smallest group, it is the least healthy and oldest group of the people in this study. The majority of them originated from the north with a medium level of salary. Most of the people are male and married.

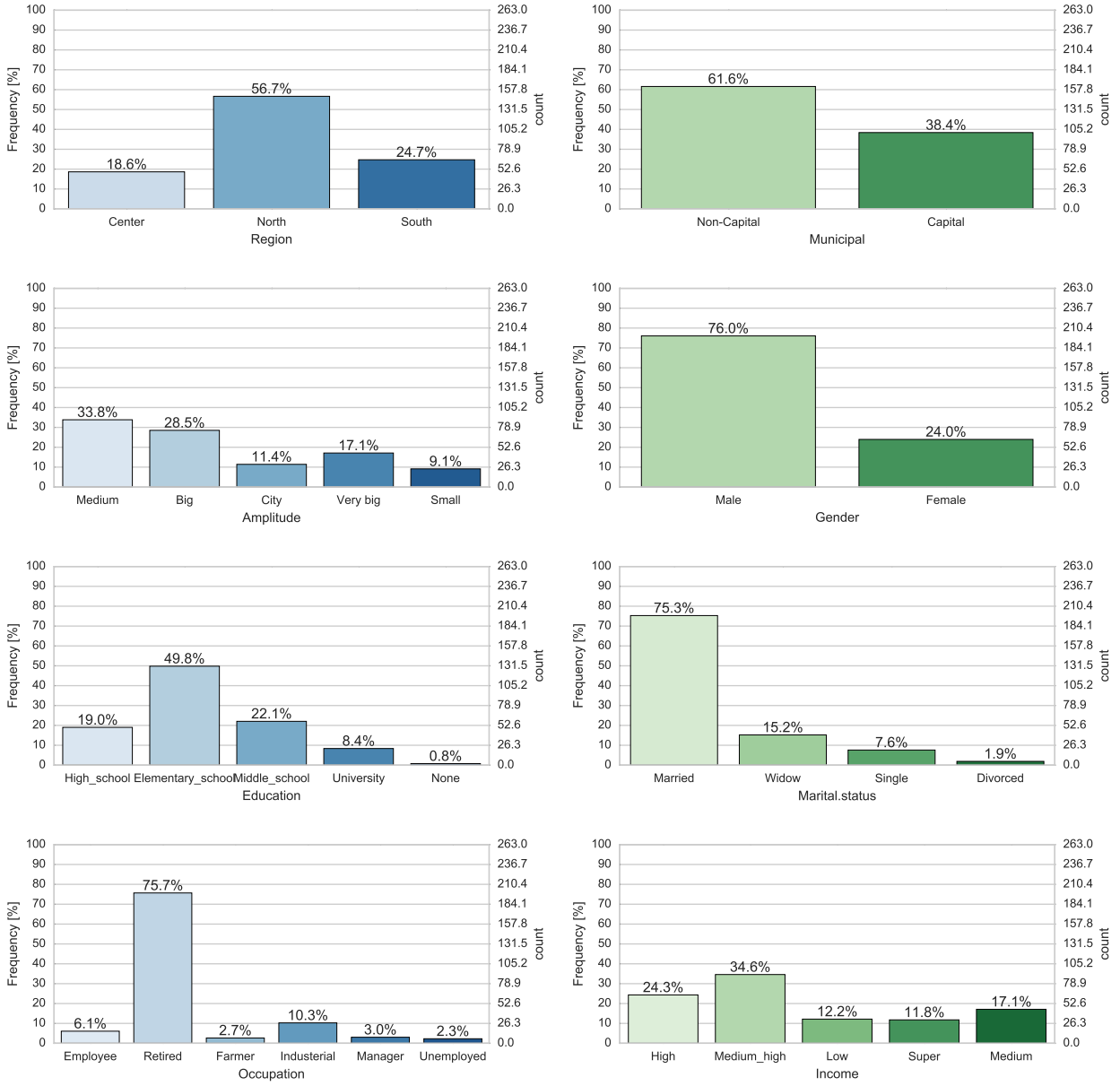

Figure 5.8: Cluster 3 Demographic Distribution

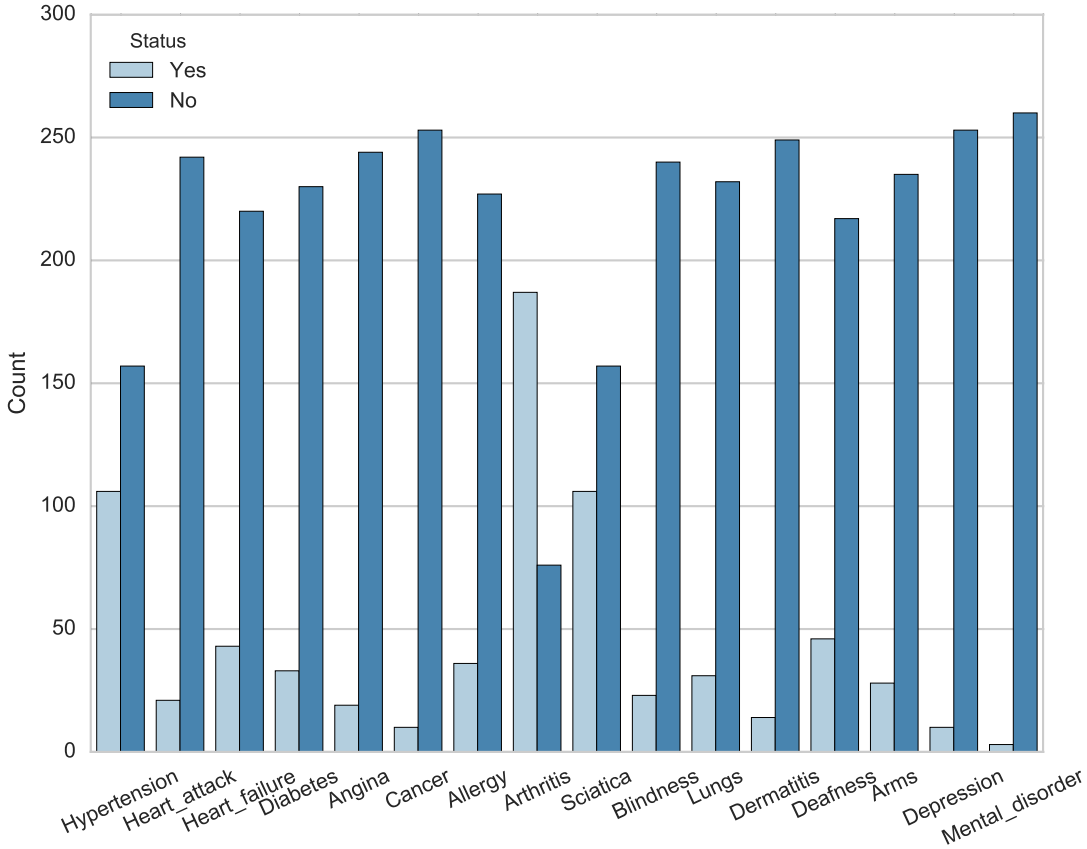

Figure 5.9: Cluster 3 Health Major Problem Distributions

5.6 Validation of Cluster Analysis

The silhouette method was the first step of validation of the cluster analysis. In this section, two more validation methods will be applied to support the result of silhouette distance, supervised learning which consider clusters as labels and parametric analysis of

Qol(quality of life index).

5.6.1 Parametric analysis of the Qol index

The Qol index was the final aggregation of six psychological categories extracted from the PGWBI questionnaire. In the preprocessing phase, this index was excluded from clustering analysis to be utilized for validation of cluster analysis.

In this section, Kruskal-Wallis test is applied to show that three clusters are originated from three different populations. Then Kolmogorov-Smirinov goodness-of-fit tests is utilized to find the best three distributions which fit to each cluster.

Figure 5.10 illustrates the result of Kruskal-Wallis and Kolmogorov-Smirinov goodness-of-fit test. The p-value of Kruskal-Wallis test is 0.001893 and at the level of 0.05, the null hypothesis is rejected and it is deduced that all clusters come from different distributions. As a result, the Kolmogorov-Smirinov(KS) test uses to find the best three distributions fitted to each clusters. Wakeby (Houghton [83]), Dagum (Dagum [84]), Johnson SB described in section 5.2, and Generalized Extreme Value distributions are different distributions fitted to clusters and the calculated p-value of KS test is greater than 0.05 and therefore the null hypothesis, the distribution is fitted to data, is not rejected. It is observable from the Figure 5.10 that all the clusters follow different distributions or the same distribution, but different parameters.

Kruskal-Wallis test Result (Null Hypothesis : all samples originated from the same population)		
Kruskal-Wallis test	Degree of Freedom	P-value
12.539	2	0.001893

	Distribution 1	Distribution 2	Distribution 3
Cluster 1	Wakeby (5 parameters) $\alpha = 299.64,$ $\beta = 7.41,$ $\gamma = 25.07,$ $\delta = -0.633,$ $\zeta = 25.994$	Dagum (4 parameters) $\kappa = 0.13272,$ $\alpha = 27.357,$ $\beta = 97.495,$ $\gamma = 0$	Johnson SB(4 parameters) $\gamma = -2.0337,$ $\delta = 1.4713,$ $\lambda = 159.9,$ $\zeta = -47.299$
Cluster 2	Wakeby (5 parameters) $\alpha = 271.72,$ $\beta = 9.892,$ $\gamma = 35.999,$ $\delta = -0.83053,$ $\zeta = 36.534$	Gen. Extreme Value $\kappa = -.52432,$ $\sigma = 17.539,$ $\mu = 77.37$	Johnson SB(4 parameters) $\gamma = -1.4511,$ $\delta = 1.4726,$ $\lambda = 123.68,$ $\zeta = -6.6856$
Cluster 3	Johnson SB(4 parameters) $\gamma = -1.1585,$ $\delta = 1.1782,$ $\lambda = 119.6,$ $\zeta = -8.6375$	Gen. Extreme Value $\kappa = -.57175,$ $\sigma = 20.968,$ $\mu = 71.256$	Wakeby (5 parameters) $\alpha = 215.09,$ $\beta = 6.6046,$ $\gamma = 38.394,$ $\delta = -0.83171,$ $\zeta = 26.026$

Figure 5.10: Kruskal-Wallis test result and Kolmogorov-Smirinov Goodness-of-fit test of clusters

5.6.2 Supervised learning: Classification

The second method for cluster analysis validation is classification. Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. The predictive classification modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). The accuracy

of a classification model is the fraction of corrected predicted labels to the size of dataset. This section is considered for two purposes, prediction of new data points and validation of the cluster analysis. All the predictors considered for cluster analysis, are used as input variables(X) and extracted clusters as labels or discrete outputs (y). Table 5.10 shows the accuracy of the classifiers used to the aforementioned purposes. Random Forest, AdaBoost , and Naive Bayes (Friedman et al. [85]) are applied as classifiers. The high accuracy of classifiers shows the homogeneity and low variance within clusters. It shows that most of similar units are in the same cluster.

Table 5.10: 10-fold cross validation Classification based on cluster labels

Classifier	**Accuracy**	**F-measure**
Random Forest	95.11	95.1
AdaBoost	79.84	79.1
Naive Bayes	82.08	82.1

5.7 Contribution

In present explanatory analysis, an investigation has been done to find similar groups of individuals. In this study, the main target is to homogenize a group of people to have better of understanding of different aspects of each individual's life. The second objective is to develop a predictive model according to the primary findings of this study to judge about a new individual as well as a method to evaluate the result of the first part of the

study. 1027 people have been involved in this study and we have demonstrated that three clusters are able to explain the similarity and nature of the dataset with a good accuracy using Silhouette width measurement.

First, the youngest and healthiest group is the first cluster. The majority of individuals have a university degree and the number of males is greater than females. Second, the third cluster is the smallest and contains the least healthy people in the experiment. The majority of them are married males with some major health problems originated from south and north. Almost seventy percent of individuals in this cluster suffer arthritis. Finally, the second cluster is a group between the first and third.

Grouping people into similar clusters facilitates making decision. For instance, administrative units like governments can allocate their limited source of funds with almost accurate priority that they can extract from each cluster. Moreover, individuals can predict their Qol by investigation of the cluster which is more similar to their characteristics.

The method proposed here is K-medoids method using Gower distance which can handle mixed data appropriately. This method is validated with parametric analysis of Qol index and several predictive models with the best one has accuracy of 95%. In our future work, we are trying to define the new distance that can outline social big data with more flexibility. The distance can measure and classify new data with more accuracy than Gower distance. The result of clustering was validated using three different methods which all of them verified the result of clustering and its applicability.

CHAPTER 6 : FUTURE RESEARCH WORKS

6.1 Application of the developed model on Heart signal, ECG

My statistical method on analyzing the non-stationary brain signals will have a direct applicability into the signals that result in heart abnormality detection and heart testing, both regular tests and stress testing. The resulting signals need to be classified as normal response, presence of disease A, presence of disease B, or both. Therefore the main goal is to generalize this method to a problem of multi-class classification. The classifier assist doctors and expert in this field to identify the problem easier and in earlier time. We are in the process of obtaining such a information from medical clinical and other institute that are working in the subject area. However, up to now we have not been very successful.

6.2 Monitoring Health using Qol as a time series

This study is the continuation of the quality of life prediction and monitoring. In this study, we generate a huge number of signals from age of individuals. The quality of life of a specific individual is not possible to be monitored for all of their life. In this study we look at age in stochastic perspective. we consider target value as a random variable over age. It

means, we assume that there is an abstract individual that we are going to find the best signal, the rate of changing to the quality of life index, to this new individual. In the next step, these short signals are clustered into several groups. By clustering, we can observe approximately the change of the pattern of quality of life over age. An expert by looking at this patterns can give more accurate advice to their patients how to change risk factor of the patient life to be in the better cluster. The big challenge that we have in this process is the number of generated signals. We need to design some filters in the preprocessing phase to reduce the dimension of data to be suitable for further analysis.

REFERENCES

[1] Kadir Sabancı and Murat Koklu. The classification of eye state by using knn and mlp classification models according to the eeg signals. *International Journal of Intelligent Systems and Applications in Engineering*, 3(4):127–130, 2015.

[2] Abolfazl Saghafi, Chris P Tsokos, Mahdi Goudarzi, and Hamidreza Farhidzadeh. Random eye state change detection in real-time using eeg signals. *Expert Systems With Applications*, 72:42–48, 2017.

[3] Oliver Rösler and David Suendermann. A first step towards eye state prediction using eeg. *Proc. of the AIHLS*, 2013.

[4] Mahnaz Arvaneh, Cuntai Guan, Kai Keng Ang, and Chai Quek. Optimizing the channel selection and classification accuracy in eeg-based bci. *IEEE Transactions on Biomedical Engineering*, 58(6):1865–1873, 2011.

[5] Thomas B McKee, Nolan J Doesken, John Kleist, et al. The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, volume 17, pages 179–183. American Meteorological Society Boston, MA, 1993.

[6] Raphael M Wambua, Benedict M Mutua, and James M Raude. Spatio-temporal drought characterization for the upper tana river basin, kenya using standardized precipitation index. *World Journal of Environmental Engineering*, 3(4):111–120, 2015.

[7] George Townsend, Bernhard Graimann, and Gert Pfurtscheller. Continuous eeg classification during motor imagery-simulation of an asynchronous bci. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 12(2):258–265, 2004.

[8] Samanwoy Ghosh-Dastidar and Hojjat Adeli. Improved spiking neural networks for eeg classification and epilepsy and seizure detection. *Integrated Computer-Aided Engineering*, 14(3):187–212, 2007.

[9] Deng Wang, Duoqian Miao, and Chen Xie. Best basis-based wavelet packet entropy feature extraction and hierarchical eeg classification for epileptic detection. *Expert Systems with Applications*, 38(11):14314–14320, 2011.

[10] Abdulhamit Subasi and Ergun Ercelebi. Classification of eeg signals using neural network and logistic regression. *Computer methods and programs in biomedicine*, 78(2): 87–99, 2005.

[11] Abdulhamit Subasi and M Ismail Gursoy. Eeg signal classification using pca, ica, lda and support vector machines. *Expert systems with applications*, 37(12):8659–8666, 2010.

[12] Qi Xu, Hui Zhou, Yongji Wang, and Jian Huang. Fuzzy support vector machine for classification of eeg signals using wavelet-based features. *Medical engineering & physics*, 31(7):858–865, 2009.

[13] Gtinter Edlinger, Paul Wach, and Gert Pfurtscheller. On the realization of an analytic high-resolution eeg. *IEEE transactions on biomedical engineering*, 45(6):736–745, 1998.

[14] Boris Burle, Laure Spieser, Clémence Roger, Laurence Casini, Thierry Hasbroucq, and Franck Vidal. Spatial and temporal resolutions of eeg: Is it really black and white? a scalp current density view. *International Journal of Psychophysiology*, 97(3):210–220, 2015.

[15] Ling Guo, Daniel Rivero, Julián Dorado, Cristian R Munteanu, and Alejandro Pazos. Automatic feature extraction using genetic programming: An application to epileptic eeg classification. *Expert Systems with Applications*, 38(8):10425–10436, 2011.

[16] Robert D Vincent, Joelle Pineau, Philip De Guzman, and Massimo Avoli. Recurrent boosting for classification of natural and synthetic time-series data. In *Advances in Artificial Intelligence*, pages 192–203. Springer, 2007.

[17] Andrew Frank. Uci machine learning repository. *http://archive. ics. uci. edu/ml*, 2010.

[18] Whoqol Group et al. The world health organization quality of life assessment (whoqol): position paper from the world health organization. *Social science & medicine*, 41(10): 1403–1409, 1995.

[19] Ed Diener. Subjective well-being: The science of happiness and a proposal for a national index. *American psychologist*, 55(1):34, 2000.

[20] Rick Harrington and Donald A Loffredo. Insight, rumination, and self-reflection as predictors of well-being. *The Journal of psychology*, 145(1):39–57, 2010.

[21] Neil K Aaronson. Quantitative issues in health-related quality of life assessment. *Health Policy*, 10(3):217–230, 1988.

[22] Francesc Casellas, Josefa López-Vivancos, Xavier Badia, Jaime Vilaseca, and Juan-Ramon Malagelada. Impact of surgery for crohn's disease on health-related quality of life. *The American journal of gastroenterology*, 95(1):177, 2000.

[23] Enzo Grossi, Nicola Groth, Paola Mosconi, Renata Cerutti, Fabio Pace, Angelo Compare, and Giovanni Apolone. Development and validation of the short version of the psychological general well-being index (pgwb-s). *Health and quality of life outcomes*, 4 (1):88, 2006.

[24] Clairice T Veit and John E Ware. The structure of psychological distress and well-being in general populations. *Journal of consulting and clinical psychology*, 51(5):730, 1983.

[25] Ed Diener, Derrick Wirtz, William Tov, Chu Kim-Prieto, Dong-won Choi, Shigehiro Oishi, and Robert Biswas-Diener. New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, 97(2):143–156, 2010.

[26] Åsa Lundgren-Nilsson, Ingibjörg H Jonsdottir, Gunnar Ahlborg, and Alan Tennant. Construct validity of the psychological general well being index (pgwbi) in a sample of patients undergoing treatment for stress-related exhaustion: a rasch analysis. *Health and quality of life outcomes*, 11(1):2, 2013.

[27] Angelo Compare, Elena Germani, Riccardo Proietti, and David Janeway. Clinical psychology and cardiovascular disease: an up-to-date clinical practice review for assessment and treatment of anxiety and depression. *Clinical practice and epidemiology in mental health: CP & EMH*, 7:148, 2011.

[28] Saeid Yazdi-Ravandi, Zahra Taslimi, Narges Jamshidian, Hayede Saberi, Jamal Shams, and Abbas Haghparast. Prediction of quality of life by self-efficacy, pain intensity and pain duration in patient with pain disorders. *Basic and clinical neuroscience*, 4(2):117, 2013.

[29] RJJ Gobbens and MALM Van Assen. The prediction of quality of life by physical, psychological and social components of frailty in community-dwelling older people. *Quality of Life Research*, 23(8):2289–2300, 2014.

[30] Anette Schrag, Marjan Jahanshahi, and Niall Quinn. What contributes to quality of life in patients with parkinson's disease? *Journal of Neurology, Neurosurgery & Psychiatry*, 69(3):308–312, 2000.

[31] Karen Herlofson Karlsen, Jan P Larsen, Elise Tandberg, and John G Mæland. Influence of clinical and demographic variables on quality of life in patients with parkinsonâĂŹs disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 66(4):431–435, 1999.

[32] S D'alisa, G Miscio, S Baudo, A Simone, L Tesio, and Alessandro Mauro. Depression is the main determinant of quality of life in multiple sclerosis: a classification-regression (cart) study. *Disability and rehabilitation*, 28(5):307–314, 2006.

[33] Rebecca G Logsdon, Laura E Gibbons, Susan M McCurry, Linda Teri, et al. Quality of life in alzheimer's disease: patient and caregiver reports. *Journal of Mental health and Aging*, 5:21–32, 1999.

[34] Kathleen M Fenn, Suzanne B Evans, Ruth McCorkle, Michael P DiGiovanna, Lajos Pusztai, Tara Sanft, Erin W Hofstatter, Brigid K Killelea, M Tish Knobf, Donald R Lannin, et al. Impact of financial burden of cancer on survivors' quality of life. *Journal of Oncology Practice*, 10(5):332–338, 2014.

[35] G Bianchi, G Marchesini, F Nicolino, R Graziani, D Sgarbi, Carmelina Loguercio, R Abbiati, and M Zoli. Psychological status and depression in patients with liver cirrhosis. *Digestive and Liver Disease*, 37(8):593–600, 2005.

[36] Anna Carotenuto, Angiola M Fasanaro, Ivana Molino, Fabio Sibilio, Andrea Saturnino, Enea Traini, and Francesco Amenta. The psychological general well-being index (pgwbi) for assessing stress of seafarers on board merchant ships. *International maritime health*, 64(4):215–220, 2013.

[37] Cynthia Rudin. Can machine learning be useful for social science?. *In: The Cities: An essay collection from the Decent City initiative.*, 9(1):86–90, 2015.

[38] Camila Maione, Donald R Nelson, and Rommel Melgaço Barbosa. Research on social data by means of cluster analysis. *Applied Computing and Informatics*, 2018.

[39] Kuldeep Singh, Harish Kumar Shakya, and Bhaskar Biswas. Clustering of people in social network based on textual similarity. *Perspectives in Science*, 8:570–573, 2016.

[40] Emilio Ferrara, Mohsen JafariAsbagh, Onur Varol, Vahed Qazvinian, Filippo Menczer, and Alessandro Flammini. Clustering memes in social media. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 548–555. IEEE, 2013.

[41] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.

[42] T Velmurugan and T Santhanam. Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points. *Journal of computer science*, 6(3):363, 2010.

[43] Preeti Arora, Shipra Varshney, et al. Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, 78:507–512, 2016.

[44] Frederico AC Azevedo, Ludmila RB Carvalho, Lea T Grinberg, José Marcelo Farfel, Renata EL Ferretti, Renata EP Leite, Wilson Jacob Filho, Roberto Lent, and Suzana Herculano-Houzel. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513 (5):532–541, 2009.

[45] Jan Pieter Pijn, Jan Van Neerven, Andre Noest, and Fernando H Lopes da Silva. Chaos or noise in eeg signals; dependence on state and brain site. *Electroencephalography and clinical Neurophysiology*, 79(5):371–381, 1991.

[46] Justin Dauwels, Francois Vialatte, and Andrzej Cichocki. Diagnosis of alzheimer's disease from eeg signals: where are we standing? *Current Alzheimer Research*, 7(6): 487–505, 2010.

[47] William Dement and Nathaniel Kleitman. Cyclic variations in eeg during sleep and their relation to eye movements, body motility, and dreaming. *Electroencephalography and clinical neurophysiology*, 9(4):673–690, 1957.

[48] Wu Ting, Yan Guo-zheng, Yang Bang-hua, and Sun Hong. Eeg feature extraction based on wavelet packet decomposition for brain computer interface. *Measurement*, 41 (6):618–625, 2008.

[49] Robert Jenke, Angelika Peer, and Martin Buss. Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions on Affective Computing*, 5(3):327–339, 2014.

[50] Huabiao Qin, Jun Liu, and Tianyi Hong. An eye state identification method based on the embedded hidden markov model. In *2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2012)*, pages 255–260. IEEE, 2012.

[51] Ting Wang, Sheng Uei Guan, Ka Lok Man, and TO Ting. Time series classification for eeg eye state identification based on incremental attribute learning. In *2014 International Symposium on Computer, Consumer and Control*, pages 158–161. IEEE, 2014.

[52] PA Estévez, CM Held, CA Holzmann, CA Perez, JP Pérez, J Heiss, M Garrido, and P Peirano. Polysomnographic pattern recognition for automated classification of sleep-waking states in infants. *Medical and Biological Engineering and Computing*, 40(1): 105–113, 2002.

[53] John D Musa and Kazuhira Okumoto. A logarithmic poisson execution time model for software reliability measurement. In *Proceedings of the 7th international conference on Software engineering*, pages 230–238. IEEE Press, 1984.

[54] AL Goel and K Okumoto. An analysis of recurrent software failures on a real-time control system. In *Proceedings of ACM Conference*, pages 496–500, 1978.

[55] DR Cox and PAWL Lewis. The statistical analysis of series of events. 1966.

[56] Chris P Tsokos. *Reliability Growth: Non-homogeneous Poisson Process*. CRD Press, 1995.

[57] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[58] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[59] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

[60] Ashutosh Patri and Yugesh Patnaik. Random forest and stochastic gradient tree boosting based approach for the prediction of airfoil self-noise. *Procedia Computer Science*, 46:109–121, 2015.

[61] Simone Grebner, Norbert K Semmer, and Achim Elfering. Working conditions and three types of well-being: a longitudinal study with self-report and rating data. *Journal of Occupational Health Psychology*, 10(1):31, 2005.

[62] M Powell Lawton. A multidimensional view of quality of life in frail elders. In *The concept and measurement of quality of life in the frail elderly*, pages 3–27. Elsevier, 1991.

[63] Carol D Ryff. Happiness is everything, or is it? explorations on the meaning of psychological well-being. *Journal of personality and social psychology*, 57(6):1069, 1989.

[64] James Griffin. Well-being: Its meaning, measurement and moral importance. 1986.

[65] E Rabiei, E López Droguett, M Modarres, and M Amiri. Damage precursor based structural health monitoring and damage prognosis framework. *Safety and Reliability of Complex Engineered Systems*, pages 2441–2449, 2015.

[66] Elaheh Rabiei, Enrique Lopez Droguett, and Mohammad Modarres. A prognostics approach based on the evolution of damage precursors using dynamic bayesian networks. *Advances in Mechanical Engineering*, 8(9):1687814016666747, 2016.

[67] Elaheh Rabiei. *Damage Precursor Based Structural Health Monitoring and Prognostic Framework Using Dynamic Bayesian Network*. PhD thesis, 2016.

[68] Elaheh Rabiei, Enrique Lopez Droguett, and Mohammad Modarres. Damage monitoring and prognostics in composites via dynamic bayesian networks. In *2017 Annual Reliability and Maintainability Symposium (RAMS)*, pages 1–7. IEEE, 2017.

[69] Elaheh Rabiei, Enrique Droguett, and Mohammad Modarres. Fully adaptive particle filtering algorithm for damage diagnosis and prognosis. *Entropy*, 20(2):100, 2018.

[70] Kamal Jafarian, Mohammadsadegh Mobin, Ruholla Jafari-Marandi, and Elaheh Rabiei. Misfire and valve clearance faults detection in the combustion engines based on a multi-sensor vibration signal monitoring. *Measurement*, 128:527–536, 2018.

[71] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: data mining, inference, and prediction, springer series in statistics, 2009.

[72] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[73] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

[74] Alexander Statnikov, Lily Wang, and Constantin F Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1):319, 2008.

[75] Nicolás L Gutiérrez, Ray Hilborn, and Omar Defeo. Leadership, social capital and incentives promote successful fisheries. *Nature*, 470(7334):386, 2011.

[76] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3:2522, 2013.

[77] Tessa K Anderson. Kernel density estimation and k-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41(3):359–364, 2009.

[78] Thomas Liebig. *Relating mobility patterns to socio-demographic profiles*. PhD thesis, PhD Thesis, ULB Bonn, 2013.

[79] Amin Khatami, Saeed Mirghasemi, Abbas Khosravi, Chee Peng Lim, and Saeid Nahavandi. A new pso-based approach to fire flame detection using k-medoids clustering. *Expert Systems with Applications*, 68:69–80, 2017.

[80] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.

[81] Norman L Johnson. Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2):149–176, 1949.

[82] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. Package âĂŸclusterâĂŹ. *Dosegljivo na*, 2013.

[83] John C Houghton. Birth of a parent: The wakeby distribution for modeling flood flows. *Water Resources Research*, 14(6):1105–1109, 1978.

[84] Camilo Dagum. A model of income distribution and the conditions of existence of moments of finite order. *Bulletin of the International Statistical Institute*, 46:199–205, 1975.

[85] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.