CLUSTERING OF DIVERS USING DATA MINING TECHNIQUES

(DALGIÇLARIN VERİ MADENCİLİĞİ TEKNİKLERİ İLE SINIFLANDIRILMASI)

by

A. CÜNEYT YAVUZ, M.S.

Thesis

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

in the

INSTITUTE OF SCIENCE AND ENGINEERING

of

GALATASARAY UNIVERSITY

Supervisor: Assoc. Prof. Dr. S. Murat Egi

July 2018

ACKNOWLEDGEMENTS

I would like to thank my supervisors Assoc. Prof. Dr. Murat Egi and Dr. Tamer Özyiğit for their kind support during the development of the ideas of this thesis.

July 2018

A. Cüneyt Yavuz

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Divers Alert Network (DAN) created a database (DB) with a big amount of dive related data which has been collected since 1994 within the scope of Dive Safety Laboratory project. The aim of this study is to analyse DB using data mining techniques. The clustering of divers by their health and demographic information and revealing significant differences between diver groups are the main objectives of this study.

To eliminate time effect of age, divers who participated to only one dive and one dive event were included in the study. The number of one-dive and one-event divers are 874 and 1669 respectively. Before applying clustering methods, data cleaning was performed to eliminate the potential mistakes resulting from inconsistencies, inaccuracies and missing information. TwoStep and K-means clustering methods were performed on DB to find the naturally associated clusters. Conventional statistical analysis was performed to understand differences in clusters and between male and female divers.

As the result, divers were separated into 3 groups and distinguishing variables of these clusters were revealed. One dive and one event results were similar for all clusters. In order to analyse the dive-related variables with nonrecurring data, we focused on one-dive divers. The reason of this is to avoid inconsistencies in the data which changes in time. As TwoStep is suitable for categorical variables, age and dive activity years were distributed in 3 categories. For K-Means Clustering, original numerical values of these variables was used for clustering. The most distinct clusters were formed by TwoStep Clustering. The middle aged male divers without any health problem are in Cluster 1. Male and female divers with health problems and high rate of cigarette smoking are in the Cluster 2 and old divers with many dive activity years are in the Cluster 3. The search for significant differences in dive-related variables was performed based on the TwoStep Clustering results and separating male and female divers.

Keywords: Data Mining, Clustering Analysis, Diver Classification

ÖZET

Divers Alert Network (DAN), dalış güvenlik laboratuarı projesi kapsamında 1994 yılından beri topladığı dalış ile ilgili verilerle büyük bir veritabanı oluşturmuştur. Bu çalışmanın amacı, bu veritabanını veri madenciliği teknikleri kullanarak analiz etmektir. Dalıcıların sağlık ve demografik bilgilerine göre kümelenmesi ve dalıcı grupları arasındaki anlamlı farkların bulunması çalışmanın ana hedefleridir.

Değişkenlerin zaman etkilerinden arındırılması için DAN Avrupa veritabanında kayıtlı toplam 3097 dalıcıdan sadece tek dalışı olanlar ve tek dalış etkinliğine katılmış olanlar seçilmiş ve sırasıyla 874 (tek dalışı olan) ve 1669 (tek etkinliğe katılmış) dalıcının verileri kullanılarak, İki Adımlı Analiz ve K-Ortalamaları yöntemleri ile kümeleme analizleri gerçekleştirilmiştir. Kümeleme analizi sonrası, kadın ve erkek dalıcıların ve elde edilen kümelerin arasındaki farklılıkları anlamak için istatistiksel analizler yapılmıştır.

Bu analizler sonucu her dalıcılar 3 farklı gruba ayrılmış ve bu grupların ayırt edici özellikleri ortaya koyulmuştur. Tek dalışı olan ve tek etkinliğe katılmış dalıcı gruplarının özellikleri birbirleriyle benzerdir. İki Adımlı Analiz yöntemi kategorik değişkenler için uygun olduğu için, yaş ve dalış tecrübesi değişkenleri 3 kategoriye dağıtılmıştır. Sayısal veriler için etkin olan K-Ortalamaları yöntemi için ise bu değişkenlerin sayısal değerleri kullanılmıştır. En ayırt edici kümeler iki adımlı kümeleme yöntemi ile oluşturulmuştur. Orta yaşlı ve sağlıklı dalıcılar birinci kümede toplanmıştır. Sağlık problemi olan ve yüksek sigara içme oranına sahip kadın ve erkek dalıcılar ikinci kümede, uzun süredir dalış yapan dalıcılar ise üçüncü kümede toplanmıştır. Dalış ile ilgili değişkenlerin arasındaki farklılıkların araştırılması ise İki Adımlı Analiz sonuçlarına ve dalıcıların cinsiyetlerine göre yapılmıştır.

Anahtar Kelimeler : Veri Madenciliği, Kümeleme Analizi, Dalıcı Sınıflandırması

RESUMÉ

Divers Alert Network (DAN) a créé une base de données (DB) avec une grande quantité de données de plongée liées qui ont été collectées depuis 1994 dans le cadre du projet de laboratoire de sécurité de plongée. Le but de cette étude est d'analyser DB en utilisant des techniques d'exploration de données. Le regroupement des plongeurs par leur état de santé et des informations démographiques et de révéler des différences significatives entre les groupes de plongeurs sont les principaux objectifs de cette étude.
Pour éliminer l'effet du temps de l'âge, les plongeurs qui ont participé à une seule plongée et un événement de plongée ont été inclus dans l'étude. Le nombre d'une plongée et un événement plongeurs sont 874 et 1669 respectivement. Avant d'appliquer les méthodes de classification, le nettoyage des données a été effectuée pour éliminer les erreurs potentielles résultant d'incohérences, des inexactitudes et des informations manquantes. TwoStep et K-means méthodes ont été effectuées sur DB pour trouver les grappes naturellement associés. analyse statistique conventionnelle a été réalisée pour comprendre les différences dans les grappes et entre hommes et femmes plongeuses.
À la suite de ces analyses, les plongeurs ont été séparés en 3 groupes et les variables distinctives de ces groupes ont été révélés. Une plongée et un événement résultats étaient similaires pour tous les groupes. Afin d'analyser les variables liées à la plongée avec des données non récurrents, nous nous sommes concentrés sur les plongeurs d'une plongée. La raison en est d'éviter les incohérences dans les données qui changent dans le temps. Comme TwoStep est adapté pour les variables, l'âge et l'activité de plongée années ont été distribués en 3 catégories. Pour K-Means Clustering, les valeurs numériques originales de ces variables ont été utilisées pour la classification. Les groupes les plus distincts ont été formés par TwoStep Clustering. Les plongeurs mâles d'âge moyen sans aucun problème de santé sont en cluster 1.. Homme et plongeuses avec des problèmes de santé et le taux élevé de la cigarette sont dans le groupe 2 et anciens plongeurs avec de nombreuses années d'activité de plongée sont dans le cluster 3. La recherche des différences significatives dans les variables liées à la plongée a été

effectuée sur la base des résultats TwoStep Clustering et séparer les plongeurs masculins et féminins.

Mots-clés: Forage de Données, l'analyse Regroupement, classification des plongeurs

1. INTRODUCTION

Divers Alert Network (DAN) is collecting data in scope of dive safety research since 1994. The aim of collecting these data is to create a database consisting of reliable and large number of data in order to support dive related scientific research. Collected by volunteer divers, this database is the first example of research applications that spread to community in dive related research.

Originally, this project is named Safe Dive and the data is collected by taking notes manually. These notes have been replaced by dive computers since 2002 and the project has begun to be called Dive Safety Laboratory since then. In 2013, a system is implemented which allows sending divers' profiles and other information to DAN online. DAN is also publishing yearly reports about dive accidents and decompression sickness.

At the beginning of this study, there were 3.108 divers and 50.151 dive records in the DB and these numbers are growing day by day. To obtain useful information about divers, dives and dive safety, the use of data mining techniques was necessary. The primary objective of the study is to reveal demographic/physical profiles of divers and clustering these profiles by using data mining techniques to investigate if there are different groups consisting of distinct properties.

Multivariate statistical studies and data mining applications are quite new in diving. In past studies, Ozyigit et. al., implemented clustering of decompression sickness using dive accident data provided by DAN America [1]. Another study is the use of multi criteria decision making systems for selecting personal for manned underwater operations [2].

Data regarding to divers consists of divers demographical information and disease history denoting that if the diver have had any symptom relating to any disease (e.g

Allergy). For the first step of the study, demographical and sickness information of divers are checked to see if they include empty or bizarre data. If so those records are omitted from the data. Also if variables expressing any symptom occurred before have an insufficient amount of *True* value are also omitted as they are non-significant. After that, divers are separated into different groups using 2 different cluster analysis methods and the distinguishing variables of diver clusters are revealed in cluster tables. Clustering produced groups that are significantly different form each other. Finally the study focused on dive-related data of one-dive divers to uncover significant differences between diver groups and male-female divers.

By analysing dive relating data for each cluster, we inferred meaningful information which reveals divers' diving habits. We performed statistical tests for each variable to see if their varieties between clusters are significant or not. By looking at these results, we deduced that some variables are relating to the clusters very strongly. We explained why they relation is strong and what are these tests reveals about dive habits of divers.

In the second section, we gave information about related studies which are performing data mining techniques to similar data. In the third section, we described the methods we used which include the two clustering techniques, k-means and two-step clustering and statistical tests which are chi-square test, student's t-test and ANOVA test. The fourth section, we presented the results of clustering in two tables. After that, we showed graphs for each dive related variable, demonstrating their percentages for each cluster and gender. We explain why these differences are important and what information they reveal. In the final section, we summarise our work, give information about how this work could be improved and discuss possible future research directions.

2. RELATED WORK

Similar studies exist which applies data mining techniques to data about healthcare. In these studies, data mining techniques are used to assess the performance of medical treatments. In [3], an overview of applications of data mining in healthcare are reviewed. In these procedures, causes and symptoms are compared and data mining presents an effective analysis [4]. For example, by comparing the results of diseases treated with different drugs, the best treatment can be determined [5].

Also, another study is conducted by United HealthCare in which new ways to calculate costs are explored by applying data mining to treatment data [6].

Similarly, by the employment of data mining, successful standardised treatments for distinct diseases can be identified. For example, in [7] the processes involved in mining a clinical database about obstetrical patients including data warehousing, data query and cleaning, and data analysis.

Other data mining applications related to treatments consists of associating the various side-effects of treatment, comparing common symptoms to help diagnosis, determining the most effective drug for a treatment [8].

Another aspect of data mining of medical data is studied in [9]. In this study, a broad range of ethical, legal and technical issues about mining of medical data is discussed.

3. DATA MINING

Data mining is an area of computer science that is used for searching and discovering patterns in large sets of data. Useful and implicit information is extracted from data in databases through data mining. [3]. There are also other terms for defining data mining such as data archeology, knowledge extraction, knowledge mining from databases, etc. The information is extracted through data mining is then transformed into an comprehensible format for further use. Beside elementary analysis, data mining is composed of database and data management aspects, data pre-processing, model and inference concerns, interestingness measures, complexity estimation, data visualisation. With data mining, useful knowledge can be uncovered from the relevant data sets and be examined from different point of views, and thereby large data stores aid as comprehensive and reliable sources for information production and validation. The discovered knowledge can be implemented to decision making, query processing, information management and many other applications. Researchers in diverse disciplines, including machine learning, artificial intelligence, knowledge-base systems, database systems, spatial databases, statistics, knowledge acquisition and data visualisation have also studied data mining.

3.1 KNOWLEDGE DISCOVERY

The Knowledge Discovery process is defined with these phases:

- Data Selection
- Preprocessing
- Transformation
- Data Mining
- Evaluation

3.1.1 Preprocessing

Before data mining algorithms can be applied, a set of relevant data must be compiled. The target data set should be large enough to make it possible to extract patterns and remain small enough to not exceed a certain time limit when processed. Data is generally compiled from different sources or the existing pre-compiled data in data warehouses is used. Before data mining, pre-processing must be applied to analyse the multivariate data sets. After that data cleaning is applied. Missing data and observations which involve noisy data are removed with data cleaning methods.

3.1.2 Data Mining

Data mining involves six basic classes of tasks: [4]

- Outlier detection – The recognition of extraordinary data records or bizarre data that require more examination.
- Association rule learning – The process of searching for associations between variables. For example, association rule learning can be applied on a data collected by a supermarket on customer purchasing habits, the supermarket can find out which products are often bought together and use this information for marketing.
- Clustering – is the task of revealing useful information from the data by grouping data that are similar in some aspects.
- Classification – is the task of calculating a model from the data in order to be used for applying to new data. The calculated model is then used to classify the new data. For example, an e-mail service may classify an e-mail as "reliable" or as “spam".
- Regression – Similar to the classification, attempts to calculate a model, which is used as a function for estimation purposes.
- Summarisation – By visualising and generating report provides a more solid representation of the data set.

3.1.3 Results Validation

With an unintentional misuse of data mining, misleading results which seem to be significant may be produced. These results actually do not predict any feature behaviour. Also it is impossible to reproduce these results on a new sample of data. This is often caused by having too many hypotheses but not performing convenient hypotheses testing. This problem is called overfitting in machine learning.

To overcome overfitting, in the final stage of knowledge discovery process, it is validated that the patterns extracted in data mining step exist in the wider data set. Because data mining algorithms could have produced results from the training data set which does not exist in the general data set. To test this, the algorithm is evaluated by using a set of test data which is different from the data that is used in the training. This is performed in two phases. In the first phase, patterns are learned from the training data. In the second phase, these learned patterns are applied to the test data set and the results are compared to the real output. For example, spam detection algorithms used by email providers use a set of emails for training. After training, the learned patterns are applied to a test set of emails which is different from the set which is used in the training. The accuracy of the algorithm can then be decided by applying test statistics to the results.

If the results show that the learned patterns causes overfitting, than the pre-processing and data mining steps should be reapplied to extract a new set of learned patterns. Otherwise, the process is proceeded to the next step which is to interpret the learned patterns and obtain knowledge using them.

3.2 CLUSTER ANALYSIS

Clustering or unsupervised classification is the process of grouping objects together into classes. Clustering analysis is used to partition a large set of data using a “divide and conquer” method. The main advantage of this method is that it decomposes a large system into smaller elements. By this way, the design and implementation of the algorithm is simplified.

Data clustering recognises clusters in a multidimensional data set according to some distance metrics. In a big multidimensional data, the data points are not usually uniformly distributed. Data clustering discriminates sparse and dense groups of data and therefore detects the overall distribution patterns in the data.

Clustering analysis has been studied broadly for many years. Distance-based clustering analysis is the most popular of the clustering analysis methods. In the distance-based methods, all the data points are assumed to be present. They are global or semi-global approaches. This means that when deciding the cluster of each point, all data points or all currently existing clusters are examined by ignoring the distance between the points and global distance measures which require iterating all data points or all clusters. Therefore, it is not possible to scale the algorithm linearly to the data with these approaches.

Clustering analysis is referred to as unsupervised learning in machine learning, since the information of which classes an object belong to is not supplied to the algorithm, or it is called conceptual clustering, because the distance measurement may be based on conceptual classes and clustering algorithm is performed on these conceptual classes and decides whether an object belongs to a class or not. A similarity measure should be defined and then applied to the data to determine the classes of the objects.

3.2.1 K-Means Clustering

Local search is used in k-means to divide points into k groups. Firstly, k initial cluster is selected randomly. After that, based on those clusters' centroids the resulting clusters are composed [5]. K-means has a very extensive demand in the areas such as biology and computer graphics [6].

The algorithm works on a set of vectors, $D = \{x_i \mid i = 1, \ldots, N\}$, where $x_i \in R_d$ designates the i_{th} data point. K points are picked in R_d as the preliminary k clusters. This selection might be performed by selecting random points from the dataset or disrupting the global mean of the data k times. Then the algorithm iterates between two steps until it converges:

Step 1: Data Assignment. Each point is joined into the cluster of which centroid is closest to that point.

Step 2: Relocation of "means". Each cluster's centroid is recomputed by taking into account of all newly joined data points. If a probability measure is given in the data, then the relocation of the centroids is computed probabilistically.

The algorithm ends when no assignment is performed. N × k comparisons are performed in each iteration. The total number of iterations depends on N but the algorithm is considered to be linear in the data size.

The essential issue is how to measure the distance between data points in the assignment step. This is called closeness measure and the most popular of these measures is the Euclidean distance which is applied in the cost function as

$$\sum_{i=1}^{N} \left(argmin \parallel x_i - c_j \parallel_2^2 \right) \tag{1}$$

The value computed here will decrease whenever a new assignment is performed and therefore the algorithm will definitely converge in a finite number of iterations. This non-convex cost function implies that the algorithm will only converge to a local optimum and the algorithm is very sensitive to the preliminary centroid points. A solution to the local minima problem might be to run the algorithm multiple times with different preliminary points or by limiting the local search.

3.2.2 Two-Step Clustering

Two-Step clustering results with a cluster tree which is obtained by establishing preliminary clusters and re-clustering of these clusters [7,8]. It is designed to manage large data sets and scale to even larger or smaller data sets. The algorithm can handle both continuous and categorical variables. In the first step, data is pre-clustered into

small sub-groups. After that, these sub-clusters are again clustered into the desired number of clusters.

Step 1: Pre-cluster the data. In this step data is clustered sequentially. It iterates over the objects and checks if the object should merge with any of the existing clusters. If not a new cluster is formed based on the distance metric. A modified cluster feature (CF) tree is built throughout the process.

A leaf entry (an entry in the leaf node) represents a sub-cluster. The non-leaf node entries directs a new object into the correct leaf node. For example, if the CF-tree can have a maximum three levels of nodes and a maximum of eight entries per node then the number of entries can be maximum 512 therefore 512 sub-clusters.

A CF entry holds the number of objects in the cluster, the mean and variance of each continuous variable and the counts for each category of each categorical variable. For each record, starting from the root node, the distance of the object to the each cluster entry in the node is computed and the closest entry directs the object to the closest child node. The object is pointed downward with this method. After reaching a leaf node it finds the closest leaf entry in the leaf node. If the distance between the object and the CF entry is below a pre-specified threshold then the object is assigned to that cluster. Otherwise a new leaf entry in that leaf node is created. If there is no space left in the leaf node than the leaf node splits in two.

Step 2: Group the data into sub-clusters. The resulting sub-clusters are taken as input in the second step and grouped together into the desired number of clusters. The number of sub-clusters obtained in the first step is much less than the number of the objects in the data. Hence, any traditional clustering methods can be used effectively.

3.3 RESULTS ANALYSIS

After clustering, statistical hypothesis tests were applied to variables on obtained results. There are many tests which can be applied to clustering results. In this thesis chi-squared test, t-test and analysis of variance (ANOVA) tests are mentioned.

3.3.1 Student's T-Test

A t-test checks if a test statistic follows a Student's t-distribution if the null hypotheses is true. The test sample is drawn from a normal distributed population and the standard deviation is assumed to be unknown. It can be applied to two data sets to decide whether they differ from each other in some particular way.

T-Tests are performed by computing the confidence interval around the difference of two observations ($d = \theta_1 - \theta_2$). s_1^2 and s_2^2 denotes the sampling variances and $cov_{1,2}$ denotes the covariance between the estimates, the sampling error (standard error) of the difference is computed by $se(d) = \sqrt{s_1^2 + s_2^2 - 2\,cov_{1,2}}$. The T-statistic is therefore given by $t = \frac{d}{se(d)}$. Commonly, it is tested if there is any real difference exists or not which is assumed to be the null hypothesis. Therefore, the confidence intervals are computed by assuming a distribution which has a mean of zero and a standard deviation of $se(d)$.

3.3.2 Chi-Square Test

A chi-squared test is any hypothesis test which is used to decide if the data distribution of the test statistic follows a chi-square distribution when the null hypothesis is true. Chi-squared tests are formed from a sum of squared errors and are also referred to as χ^2 test. Chi-squared tests assumes an independent normally distributed data. A chi-squared test is used to reject the hypothesis that the data are independent. A greater value of the test statistics means a bigger evidence against the null hypotheses. An example of a chi-squared distribution is shown in figure 2.1.

The formula for calculating chi-square statistics is given in Eq. 2. From the formula, we can conclude that if the observed counts are close to expected then test statistic will be low, if the observed counts are far from expected then test statistic will be high. Calculated value is named as p-value. This p-value is going to be the probability under

the null hypothesis of getting the observed value of the the test statistic or something even larger. In other words, the area to the right of the observed test statistics.

$$\sum_{all\ cells} \frac{(Observed - Expected)^2}{Expected} \tag{2}$$

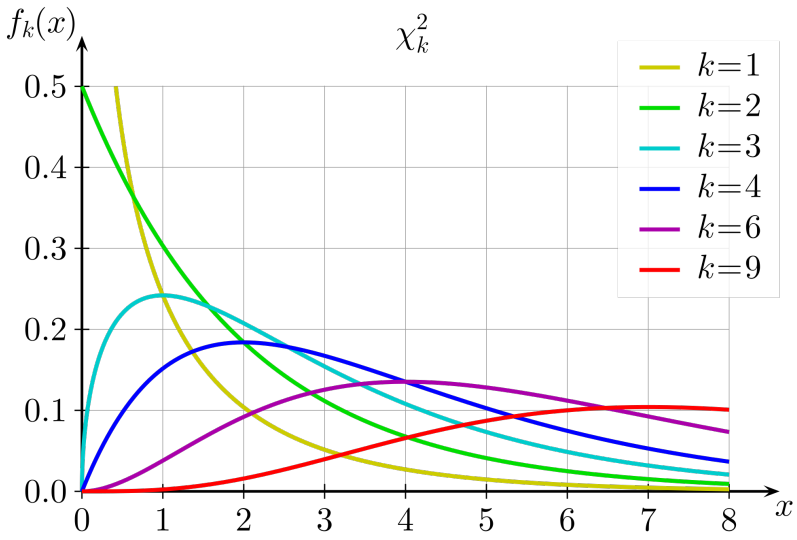

Figure 2.1: Chi-squared distribution (k is the degrees of freedom)

3.3.3 ANOVA Test

Analysis of variance (ANOVA) is used to investigate the varieties between group means and also their associated functions such as variation within and between groups.The measured variance in a specific variable is divided into parts In the ANOVA setting, the observed variance in a particular variable is partitioned into components traceable to diverse sources of variation. ANOVA presents a statistical test for deciding if the means of groups are equal or not. It can be applied to more than two groups. ANOVA tests are used to compare three or more variable means for statistical significance.

4. CLUSTERING OF DIVERS

In this section, we describe our work in four parts. In the first part we gave information about the software we have used during implementation. In the second part we gave information about how we obtained the data, the data model and the statistics of data before cleaning. In the third part we evaluated the results obtained from clustering. And lastly, in the fourth part we analyse the results using statistical methods mentioned in section 2.2.

4.1 SOFTWARE

We have used R language and programming environment as data mining tool. It provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form [9].

We also used IBM SPSS for Two-Step clustering. SPSS Statistics is a software package used for statistical analysis. First, it was produced by SPSS Inc., then it was acquired by IBM in 2009. It is a widely used program for statistical analysis in social science. It is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organisations, data miners, and others [10].

4.2 DATA

The data is stored in MSSQL which is a relational database. Data is separated into two groups. First one consists of demographical and anamnesis information of divers. Latter consists of pre-dive and post-dive information collected from hardcopy forms filled by divers and digital data collected from dive computers.

Basically, records are kept in tables consisting of diver, dive and dive event variables. There is an ID variable (DiverID and DiveID respectively) for each diver and dive. Similarly, there is also an ID variable (EventID) for dive events. A dive event consists of dives performed in 48 hours.

Data repetition is caused by the fact that a diver may have more than one dive in one event and forms being refilled for every dive. Keeping records in this fashion created several issues for arranging data for analysis. Most important of those is the necessity to eliminate this repetition effect.

Since there are more than one dive and event records for a diver in different dates and diver data is collected for each dive, change of repeated data about diver in time caused another problem. For example, a diver might have no allergic symptom in the dive he participated earlier than a dive which he have encountered allergic symptom in the meantime. This creates an inconsistency of data. We avoid this by taking divers who participated only one dive or one event. However, taking divers who participated only one event has another problem. A diver may participate to an event at the age of 30 and another one later at the age of 35. Considering divers' age is also an important variable for clustering, it is required to eliminate time effect of variable values so divers who participated to only one dive or one event are included in analysis. Because of this, we used divers data who participated only dive.

13 tables are included in database where diver, dive and dive event data are recorded. These tables are dive, event, record header, diver identification and demographics, dive header at start, dive profile, dive log details, dive safety report, diver additional anamnesis, dive additional data, dive additional problems, dive additional medications.

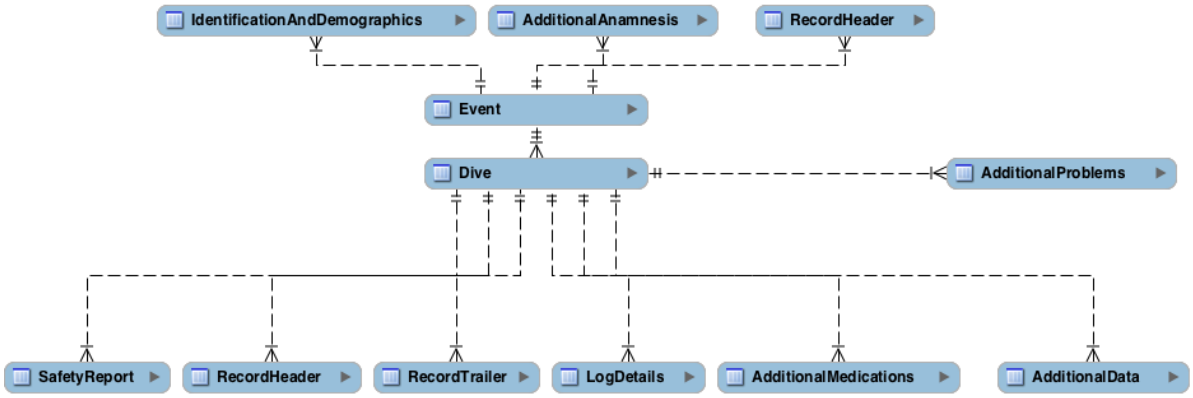

Figure 2.2: Database Diagram

Because data used for analysis is separated into these tables, it was required to aggregate selected variables into one table. Database diagram is shown in Figure 2.2.

Primarily these tables and their variables were examined and variables which will be used for cluster analysis were selected. The criteria for data selection is to choose divers' personal data (e.g. dive activity years) that may have effect on dive-related data. Variables which have a large number of missing data (e.g. height and weight) were eliminated even if they are very important since they reduce the number of data and analysis' reliability.

Variables used for analysis are age, sex, dive activity years, allergy, asthma, back pain, back surgery, cigarette smoking, diabetes, ear-sinus problems, hearth problems, muscle pain, nervous system disorder, vascular diseases and sea sickness. Age and dive activity years were numeric variables and others were binary (0-1) variables. As TwoStep is suitable for categorical data, age and dive activity years were distributed in 3 categories

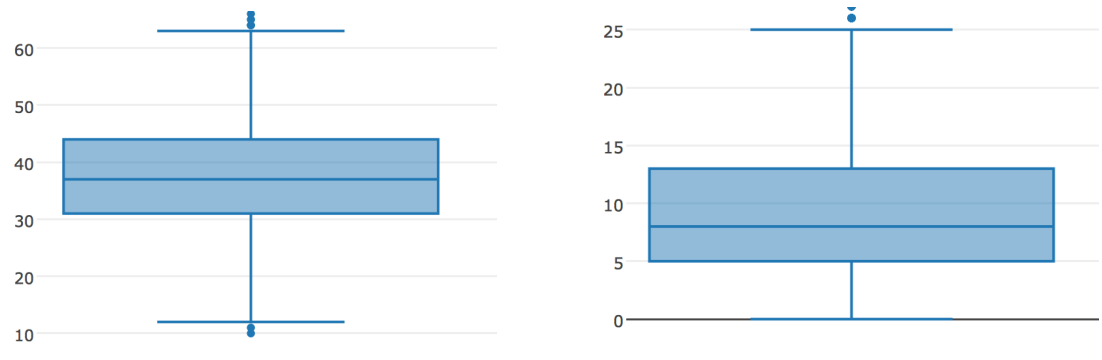

Figure 2.3: Box plots for Age (left) and Activity Years (right)

to be included into analysis. For K-Means Clustering, original numerical values of these variables are used for clustering.

Past and present anamnesis information of divers was transformed into one variable which denotes if that diver has that disease or not. For combining these variables into one table, tables acquired in MS Excel format were transferred into MySQL Database. Two tables used for analysis consisting of divers who participated only one dive and one event were obtained based on DiverID variable.

In order to be able to perform clustering with categorical variables, age and activity year variables were separated into 3 categories by examining the box plots shown in Figure 2.3.

Observing the box plots above, the optimum intervals age and dive activity years categories were built by minimum to 1st quarter, 1st quarter to 3rd quarter and 3rd quarter to maximum. These categories are shown in Table 2.1 and Table 2.2.

Table 2.1: Activity Years Diving Categories

Interval	Activity Years Category
0 < 5	Few
5 < 13	Average
< 13	Many

Table 2.2: Age Categories

Interval	Age Category
12 < 30	Young
30 < 45	Middle Aged
45 < 70	Old

The study on dive-related data was based on investigating differences between clusters and male and female divers. Examined variables were alcohol before dive, exercise before dive, state of rest before dive, minimum water temperature, thermal comfort, diving platform, maximum depth, workload, breathing gas, apparatus, equipment malfunctions and any symptoms. For categorical dive-related variables, chi-square test

for dependency on diver clusters and male/female divers were applied. For numerical variables one-way ANOVA test for multiple groups and student t tests for pairwise comparisons were applied to find significant discrepancies for each combination of cluster and for male-female divers.

4.3 CLUSTERING

Before cleaning of the data, there were 3108 divers and 50151 dives performed by these divers. Total event count was 20052. There were 16.13 dives and 6.45 events per diver. Dive count per event was 2.5.

After cleaning, the number of divers who participated to only one dive was 874 (120 female, 754 male) and the number of divers who participated to only one event was 1669 (286 female, 1382 male). Divers who are aged between 12 and 70 are included in the analysis and average age for divers who participated to only one dive and event are 37.52 and 37.28 respectively. Box plots for number of dives per diver and number of events per diver are shown in Figure 3.1.

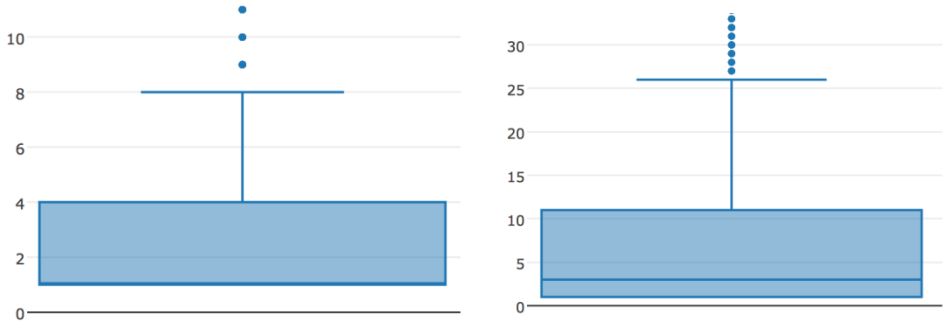

Figure 3.1: Box plots for Age (left) and Activity Years (right)

Observing the box plots in Figure 3.1, we can say that most of the divers participated to one or few dives or events. On the other hand, we can see by the outliers that there are divers who participated to a great number of dives and events which dramatically increase the average number of dives and events per diver.

To clear the data, inconsistent and inaccurate values were omitted (e.g divers who aged 150, whose dive activity years is more than 70 years and whose gender is undefined in database). Before omitting there were 930 divers who participated only one dive and 1774 divers who participated only one event. After filtering these number became 874 and 1669 respectively. Number of filtered divers is shown in Table 4. Finally, clustering

analysis was performed on cleaned data which has divers aged between 12 and 70. The number of filtered divers is given in Table 3.1

Clustering is performed using Two-Step method which is effective with categorical data. To provide a different perspective to the analysis K-means method which is more effective on numerical data was carried out as well.

Divers are separated into 3 clusters using 3 clustering analysis methods. The distinguishing variables for each group obtained from TwoStep Clustering are shown in bold in Table 3.2.

Table 3.1: Number of filtered divers by variables.

Interval	Value	One Dive	One Event
Age	< 12 and > 70	53	95
Activity Years	> 70	0	1
Sex	Undefined	3	3

4.3.1 Two-Step Clustering

When we observe Table 3.2, we can say that the first cluster is formed by middle aged male divers who have no health problem.

The second cluster is formed by male and female divers (all female divers are in the Cluster 2) with high rate of cigarette smoking. Especially allergy, ear/sinus problems and sea sickness have high rates of occurrence in this group. But other health problems have also higher rates than other 2 clusters. We can name this group as the group with health problems or unhealthy divers.

The Cluster 3 is formed by older male divers. The diving activity in years is greater than other groups and most of the heart problems are observed in this group. Only one outlier is observed in this group which is a diver who has a nervous system disorder.

The chi-square tests show highly significant specific distribution of variables into groups except diabetes, nervous system disorder and vascular disease which are rarely seen. This means that the one-dive diver clusters are formed by significantly different groups of divers.

Table 3.2: TwoStep Clustering Results

TwoStep		Clusters			Chi-Square p value
		1	2	3	
		Count	Count	Count	
DiverSex	Male	**345**	238	171	0.000
	Female	0	120	0	
Age Category	Young	0	165	0	0.000
	Middle	**345**	174	3	
	Old	0	19	**168**	
Activity Years Category	Few	57	115	15	0.000
	Average	175	197	64	
	Many	113	46	**92**	
Allergy	1	0	**48**	3	0.000
Asthma	1	0	**7**	0	0.006
Backpain	1	0	**12**	3	0.003
Backsurgery	1	0	2	3	0.045
Cigarette Smoking	1	0	**105**	20	0.000
Diabetes	1	0	0	1	0.128
Ear/Sinus Problem	1	0	**36**	0	0.000
Hearth Problem	1	0	2	**11**	0.000
Muscle Pain	1	0	9	3	0.015
Nerveus System Disorder	1	1	1	0	0.783
Vascular Disease	1	0	3	0	0.114
Sea Sickness	1	0	**18**	3	0.000

4.3.2 K-Means Clustering

As mentioned before, in K-Means clustering, the numerical variables are taken into account rather than categorical variable. We used the original numerical data for age and dive activity years when performing clustering using this method and as expected, the distinguishing variables were these two.

Observing the Table 3.3, we can say that younger divers with few dive activity years are in the Cluster 1, older divers with many diving activity years are in the Cluster 2 and middle aged divers with average dive activity years are in the Cluster 3.

Table 3.3: K-Means Clustering Results

K-Means Clustering		Clusters						Chi-Square p value
		1		2		3		
		Count	Mean	Count	Mean	Count	Mean	
DiverSex	Male	321		84		349		0.000
	Female	77		2		41		
DiverAgeYears			**30**		**50**		**43**	0.000*
DiverActivityYears			**6**		**27**		**10**	0.000*
Allergy	1	30		2		19		0.096
Asthma	1	6		0		1		0.098
Backpain	1	7		2		6		0.875
Backsurgery	1	1		1		3		0.469
CigaretteSmoking	1	73		8		44		0.007
Diabetes	1	0		0		1		0.537
SinusProblem	1	24		1		11		0.027
HearthProblem	1	0		6		7		0.000
MusclePain	1	4		2		6		0.591
NerveusSystemDisorder	1	1		0		1		0.896
VascularDisease	1	2		0		1		0.713
SeaSickness	1	10		0		11		0.297

Another remark is that, there are few female divers in Cluster 3, the group of old and experienced divers.

The categorical variable frequencies are random except heart problems. Heart problems have a high rate of occurrence in the Cluster 2 which was expected as older divers are in this cluster.

4.3.3 Statistical Analysis of Clusters

One-way ANOVA tests were applied to numerical variables (age and dive activity years) to investigate whether the means of all groups are equal or significantly different. However the other (categorical) variables have high p values of chi-square tests indicating these variables are not dependent to K-Means clusters.

As the most distinct clusters were formed by TwoStep Clustering, analysis on dive-related data were performed on the clusters obtained by this method.

In Figure 3.2, alcohols before dive percentages for each cluster are shown. Older (male) divers have the higher percentage. However, the chi-square p value is 0.3471 which means that there is no significant dependency between clusters and alcohol before dive.

To analyse the significant differences between groups deeply, chi-square tests for cluster combinations were performed as well. For Cluster 1 and Cluster 2, the p value is 0.3580. For Cluster 1 and Cluster 3, p value is 0.5156. The most significant difference was found between Cluster 2 and Cluster 3 with the p value of 0.1597 but still less than 0.05.

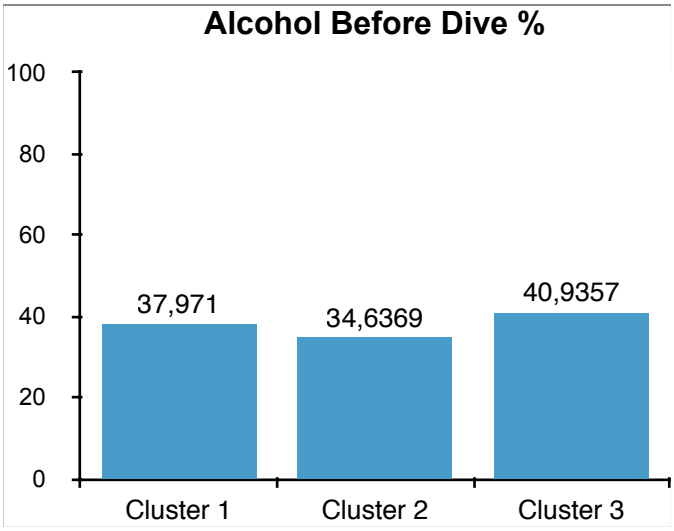

Figure 3.2: Chi-squared test results for Alcohol Before Dive by clusters

In Figure 3.3, exercise before dive percentages for each group are given. The chi-square p value for all 3 clusters is 0.0004, showing that exercise level before dive differs significantly according to clusters. Most active group before dive is old and

experienced. The percentage of light exercise is high in Cluster 3, group of old and experienced divers. The Cluster 2, group of male and female divers with health problems has the highest percentage of no exercise. These results show that older and experienced divers give more attention for preparing to diving.

After group combinations analysis, p value of chi-square test between Cluster 1 and Cluster 2 is 0.01498, Cluster 1 and Cluster 3 is 0.0760, Cluster 2 and Cluster 3 is 0,0760. The most significant difference is between Cluster 1 and Cluster 2 (healthy and unhealthy divers).

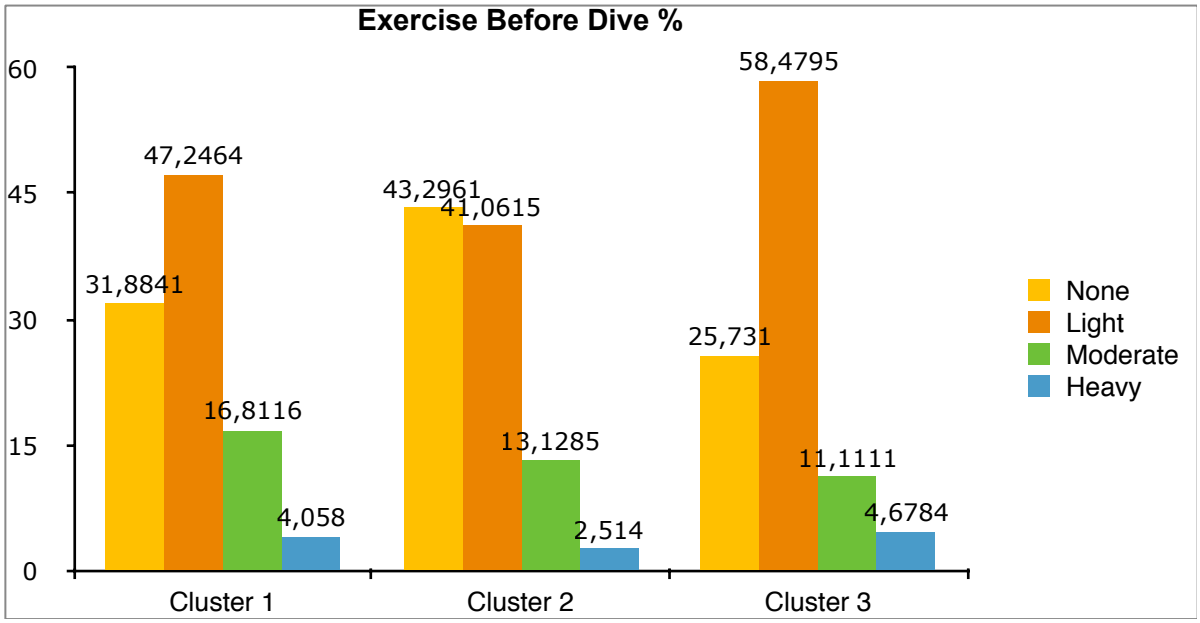

Figure 3.3: Chi-squared test results for Exercise Before Dive by clusters

In Figure 3.4, the percentages for state of rest before dive are shown. The chi-square p value for all 3 clusters is 0.0006, showing that this variable is significantly dependent to clusters. The p value of chi-square test between Cluster 1 and 2 is 0.0069, between Cluster 1 and Cluster 3 is 0.1670 which is not highly significant and between Cluster 2 and Cluster 3 is 0.0002. The biggest difference is between Cluster 2 and Cluster 3. We can say that the Cluster 2 differs from other clusters in state of rest before dive. We can conclude that older and experienced divers taking care of their pre-dive conditions more than other divers.

We didn't observe significant differences between groups in minimum water temperature. The one-way ANOVA test p value is 0.554. The averages of clusters are

given in Figure 3.5. The p value of student t test between Cluster 1 and Cluster 2 is 0.3975, between Cluster 1 and Cluster 3 is 0.7499 and between Cluster 2 and Cluster 3 is 0.3429.

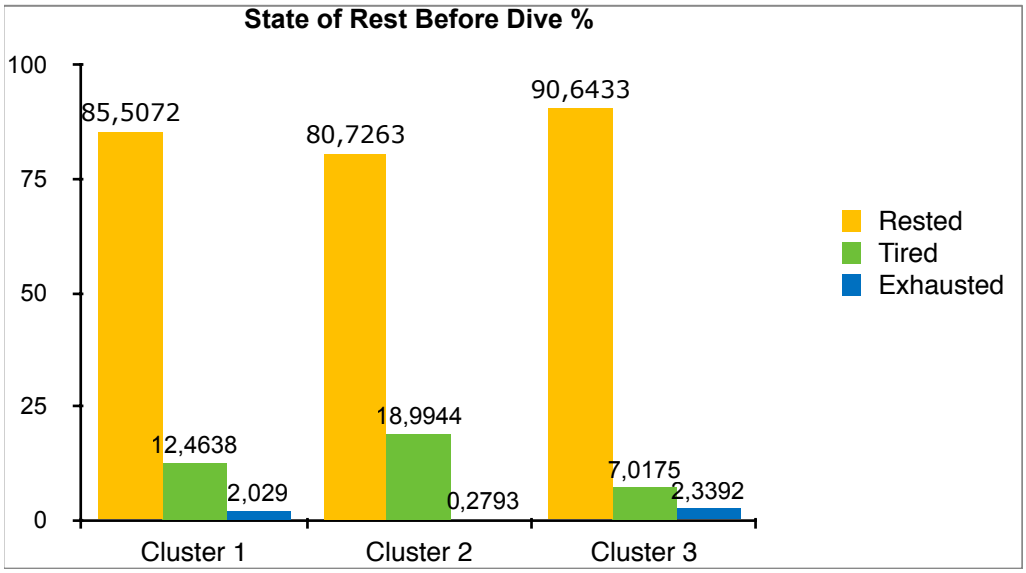

Figure 3.4: Chi-squared test results for State of Rest Before Dive by clusters

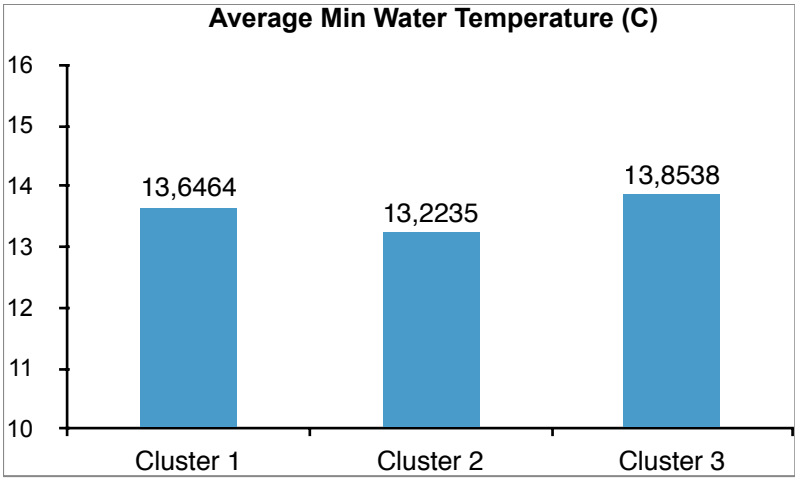

Figure 3.5: Chi-squared test results for Min Water Temperature by clusters

In Figure 3.6, the thermal comfort chart for clusters is shown. The chi-square p value is 0.0183 which means the thermal comfort during dive depend on the clusters. The Cluster 2, the group of male and female divers with health problems complains more about cold water. The chi-square p value performed to Cluster 1 and Cluster 2 is 0.0022. For Cluster 1 and Cluster 3 p value of chi-square is 0.2223 and for Cluster 2 and Cluster 3 is 0.5785. This means that the biggest difference is between Cluster 1 and Cluster 2,

groups of healthy and unhealthy divers. These results mean that health condition significantly effects thermal comfort underwater.

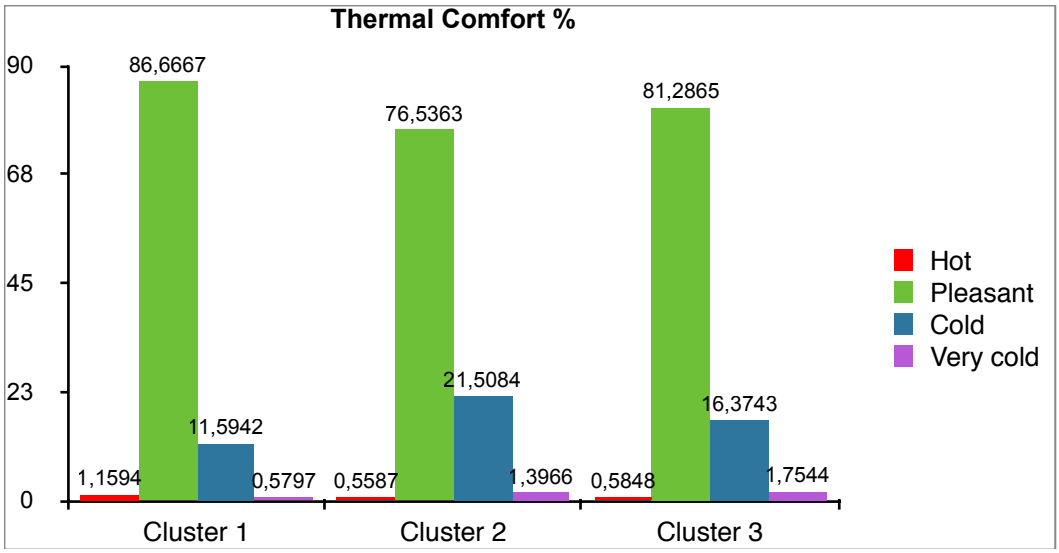

Figure 3.6: Chi-squared test results for Thermal Comfort by clusters

In Figure 3.7, platforms of diving for clusters are given. The chi-square p value is 0.0095 showing that the clusters differ in diving platform. Comparing clusters 1 and 2, the p value of chi-square test is 0.1861 and for 1 and 3 is 0.1812 which are not very significant. The p value of chi-square test between Cluster 2 and Cluster 3 is 0.0007 showing that these two diver groups are very distinct in diving platform. The main difference is seen between divers who dive on a small boat and divers who prefer to dive near the beach. The results show that older and more experienced divers dive on a small boat where as younger divers prefer to dive near the beach.

The average maximum depths of cluster are shown in Figure 3.8. The one-way ANOVA test p value is 0.0034 meaning that there are significant differences between clusters.

The student t test p value for Cluster 1 and Cluster 2 is 0.2620, for Cluster 1 and Cluster 3 is 0.0215 and for Cluster 2 and Cluster 3 is 0.0009. These p values show that Cluster 3, older and experienced divers dive significantly deeper than other groups.

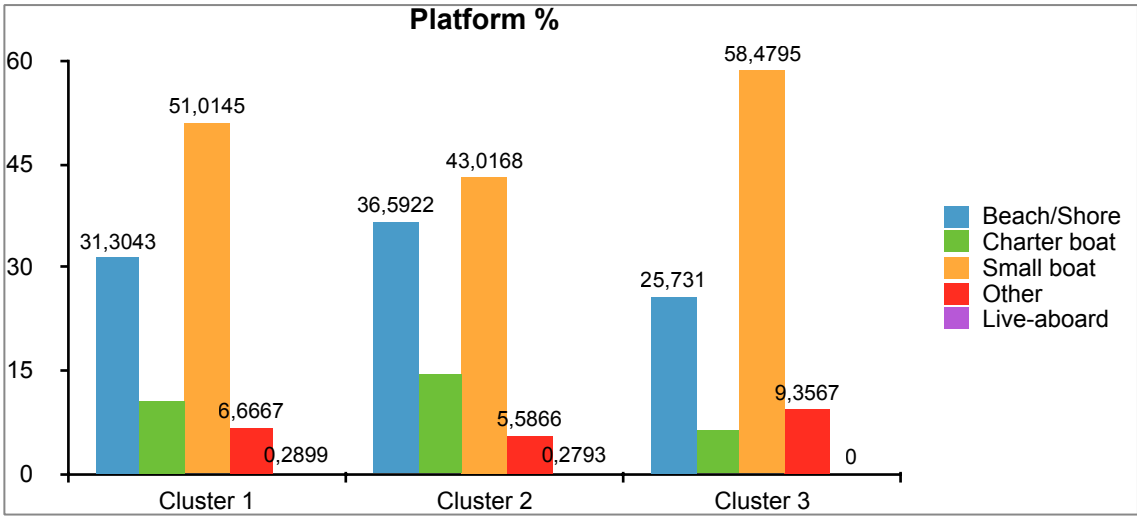

Figure 3.7: Chi-squared test results for Platform by clusters

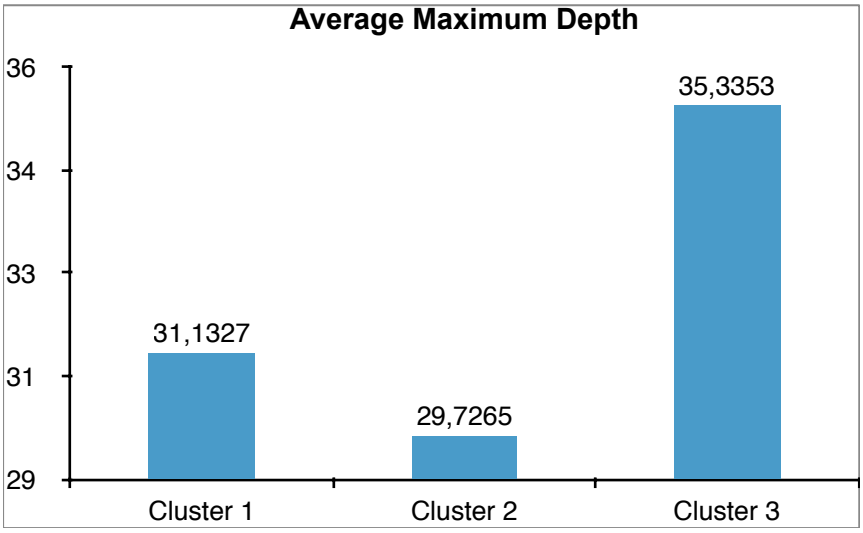

Figure 3.8: Chi-squared test results for Average Maximum Depth by clusters

In Figure 3.9, the workload level percentages are shown. The chi-square p value for all clusters is 0.0350 meaning that there are significant differences between diver groups. Chi-square p value for Cluster 1 and Cluster 2 is 0.1661, for Cluster 1 and Cluster 3 is 0.035 and for Cluster 2 and Cluster 3 is 0.0602. The Cluster 3 differs significantly from other two clusters in workload.

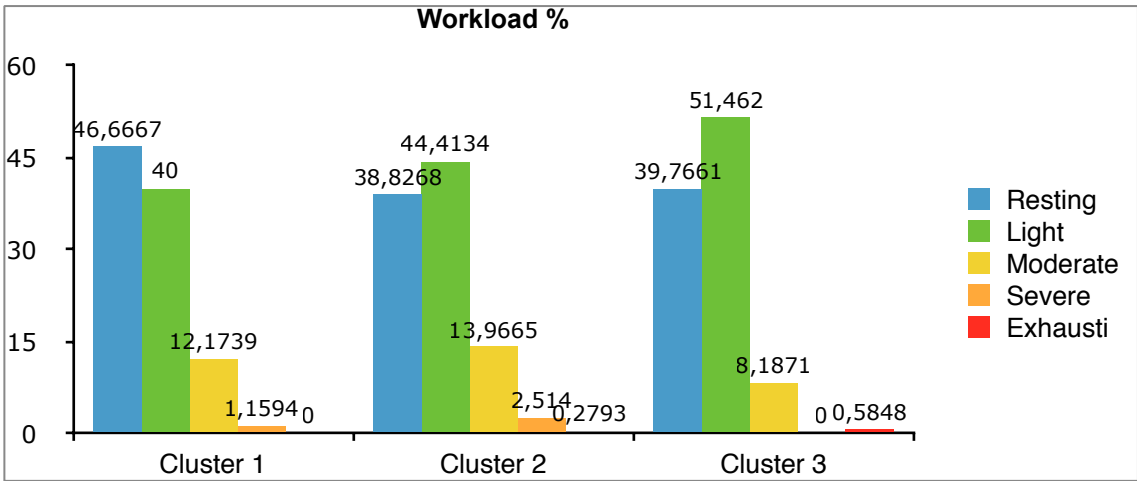

Figure 3.9: Chi-squared test results for Workload by clusters

In Figure 3.10, breathing gas percentages for clusters are shown. The p value of chi-square test is 0.0004, indicating significant differences between clusters. The experienced divers use more nitrox, Trimix and other gases than other groups. Chi square p values of pairwise tests justify this as the value for Cluster 1 and Cluster 2 is 0.353352, for Cluster 1 and Cluster 3 is 0.0037 and for Cluster 2 and Cluster 3 is 0.0001. This results tells us that more experienced divers prefer gases other than air than younger divers.

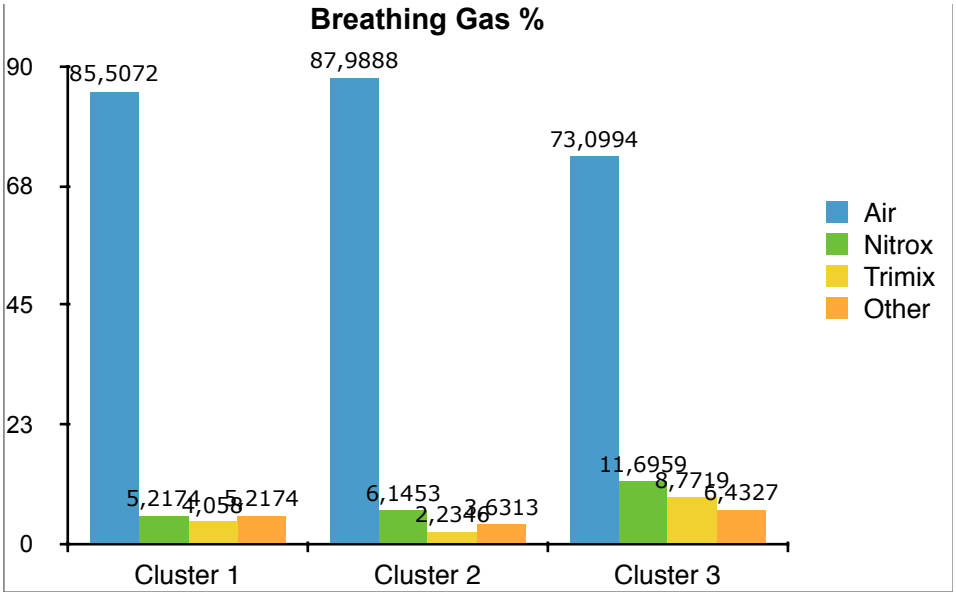

Figure 3.10: Chi-squared test results for Breathing Gas by clusters

For Breathing Apparatus, we didn't find significant dependency on clusters. The chi-square p value is 0.3006. The divers use open circuit scuba with high percentage. The pairwise chi-square tests p value for Cluster 1 and Cluster 2 is 0.1589, for Cluster 1 and

Cluster 3 is 0.8971 and for Cluster 2 and Cluster 3 is 0.1321. The percentages for cluster are given in Figure 3.11.

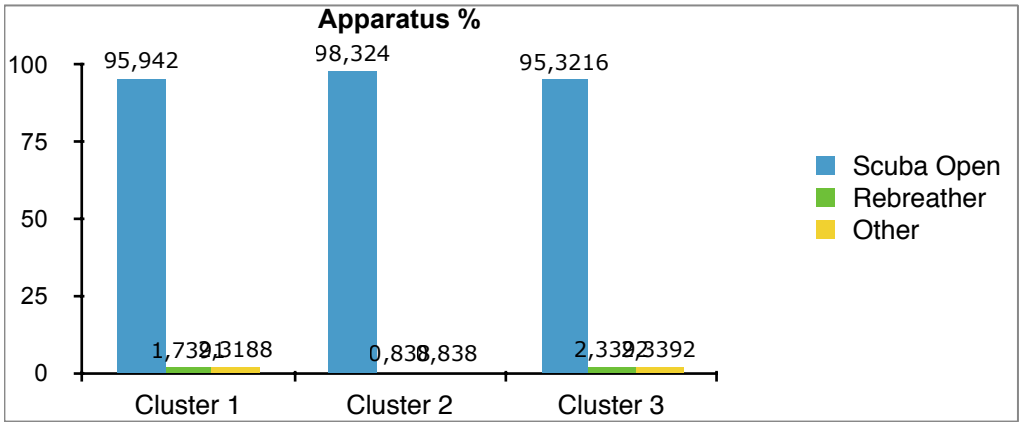

Figure 3.11: Chi-squared test results for Breathing Apparatus by clusters

In figure 3.12, equipment malfunction percentages for each cluster are given. The highest percentages are in the Cluster 2 especially in face mask and thermal protection. The chi-square p value for all clusters is 0.0106 which indicates significant differences between clusters. The p value of chi-square test between Cluster 1 and Cluster 2 is 0.0151, between Cluster 1 and cluster 3 is 0.4141 and between Cluster 2 and Cluster 3 is 0.083588. The most important difference is between Cluster 1 and Cluster 2, groups of healthy and unhealthy divers. We can say that the Cluster 2 is the most distinct group from other clusters. This shows that unhealthy divers are encountered with equipment malfunctions more than other divers. This might be due to disordered body functions such as irregular breathing.

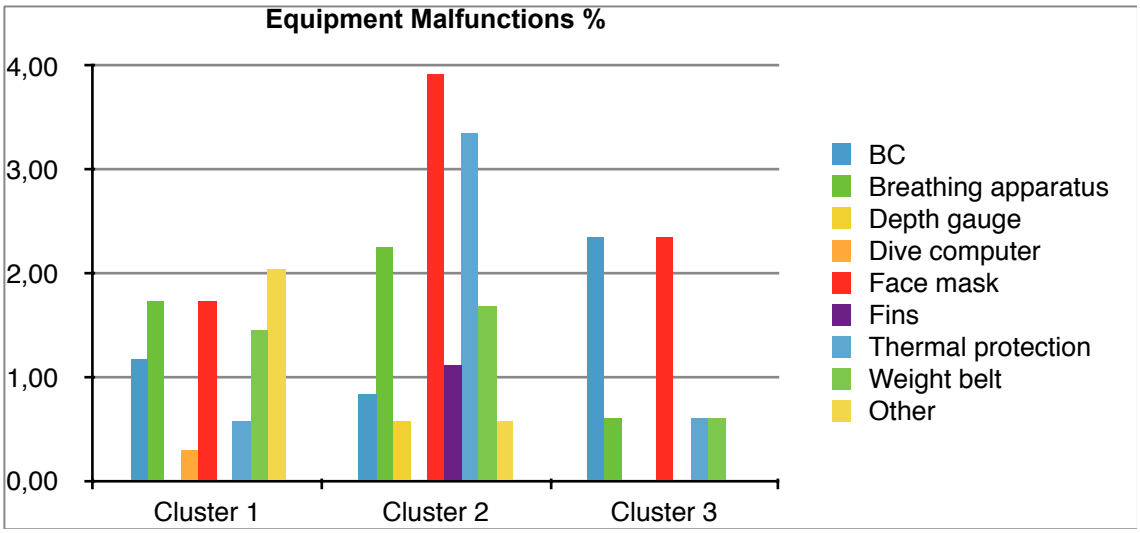

Figure 3.12: Chi-squared test results for Equipment Malfunctions by clusters

In Figure 3.13, percentages for any symptom for clusters are given. The chi-square p value is 0.000008, indicating very significant dependency of this variable to clusters. The test p value between Cluster 1 and 2 is 0.0053, between Cluster 1 and Cluster 3 is 0.0158 and between Cluster 2 and Cluster 3 is 0.000001. The Cluster 3, group of old and experienced (and deep) divers has the highest percentage. However, unhealthy divers are the ones who experiences the least number of symptoms after dives. This might be due to their less healthy metabolism causing them not to notice any differences.

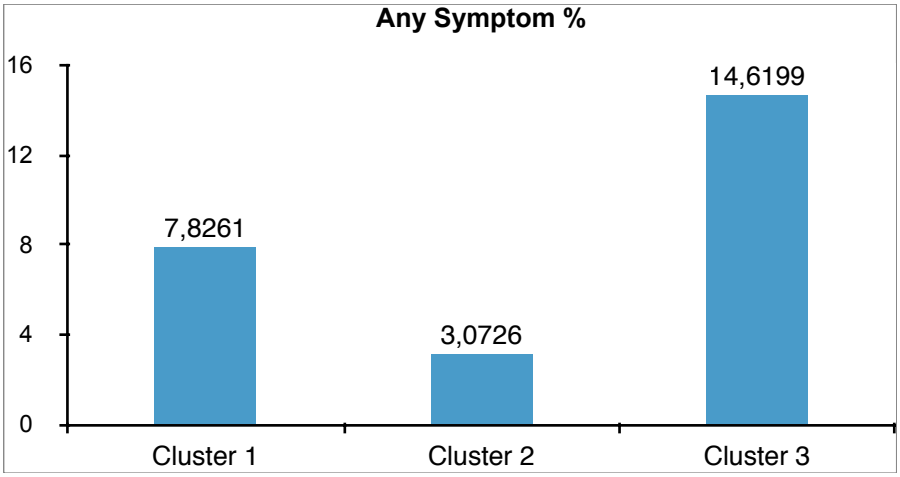

Figure 3.13: Chi-squared test results for Any Symptoms by clusters

In Figure 3.14, we can see the percentages of alcohol before dive for male and female divers. The p value of chi-square test is 0.05037. This is not surprising that male divers use more alcohol than female divers.

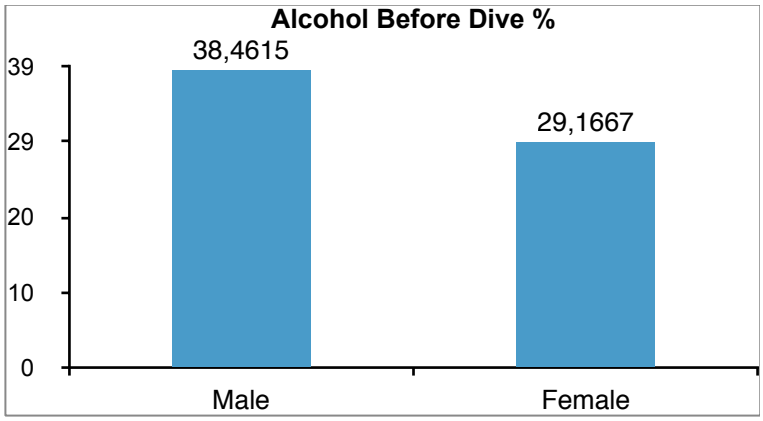

Figure 3.14: Chi-squared test results for Alcohol Before Dive by gender

Figure 3.15 shows that none of the female divers perform heavy exercise before dive. The percentage of "No Exercise" is higher in females while males have higher percentage of moderate exercise. In general, this graph shows that male divers show more attention to their pre-dive body conditions. However the p value of chi-square test is 0.0939 indicating that significance of difference between males and females in terms of exercise before dive is not very high.

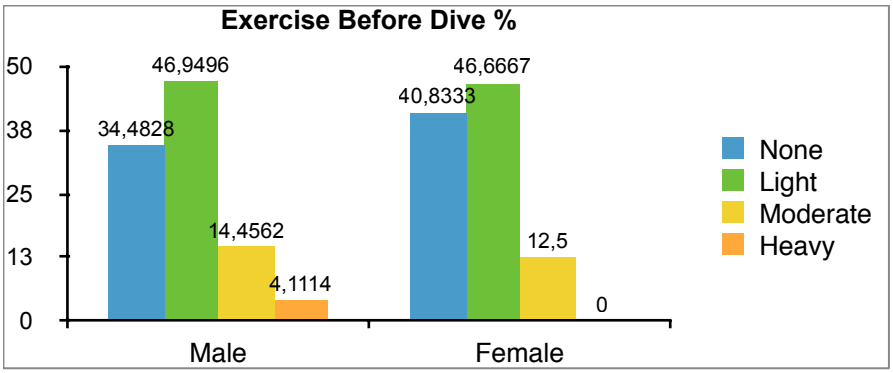

Figure 3.15: Chi-squared test results for Exercise Before Dive by gender

The Figure 3.16, where we can see State of Rest Before Dive. The p value of chi-square test is 0.0155 indicating a significant difference between male and female divers in state of rest before dive. We can say that female divers feel more tired than male divers before dive.

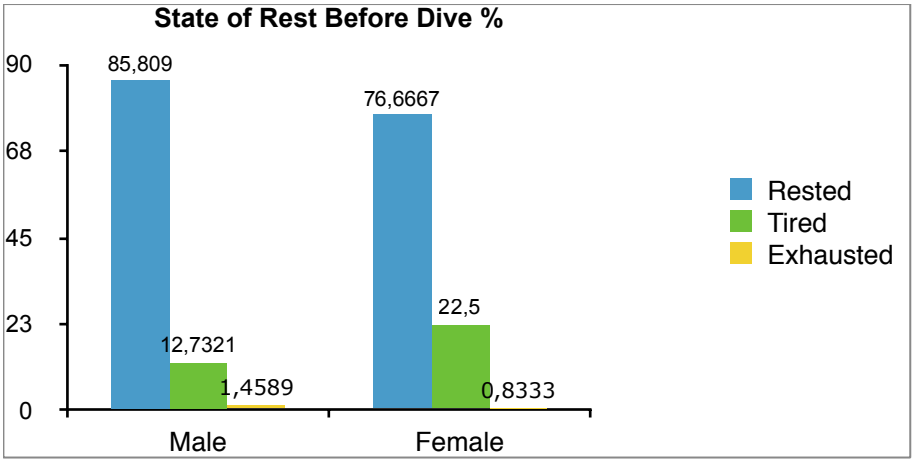

Figure 3.16: Chi-squared test results for State of Rest Before Dive by gender

When it comes to water temperature, we observe that female divers prefer diving in warmer water than male divers as seen in Figure 3.17. This value correlates with average max. depth as female divers prefer to dive less deeper than male divers. The

student t test p value is 0.0028 meaning that the difference between male and female divers is significant.

On the other hand, there is no big difference in thermal comfort. The p value of chi-square test is 0.9380. The thermal comfort percentages are shown is Figure 3.18.

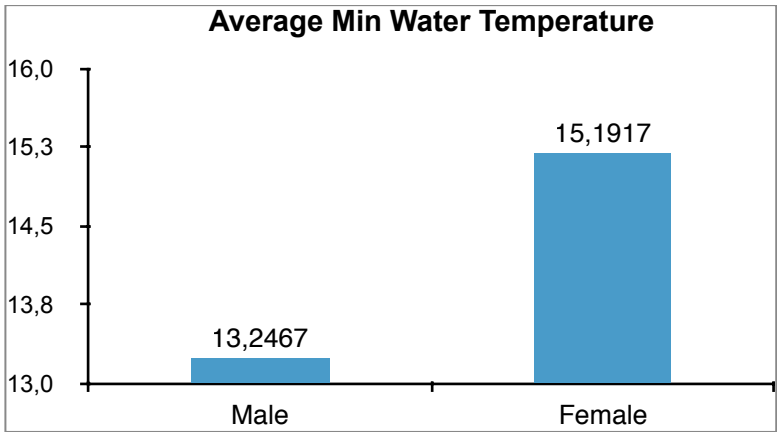

Figure 3.17: Chi-squared test results for Minimum Water Temperature by gender

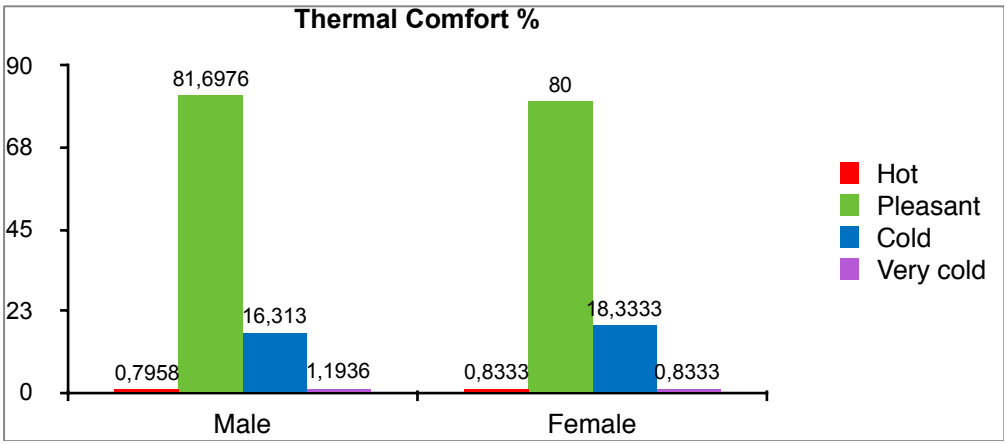

Figure 3.18: Chi-squared test results for Thermal Comfort by gender

In Table 3.19, diving platform percentages for male and female divers are given. The biggest difference is in small boat platform. The p value of chi-square test is 0.04790.

In Figure 3.20, we can see that male divers tend to dive slightly deeper than female divers. This is because male divers are more experienced than female divers. The p value of student t test is 0.0325.

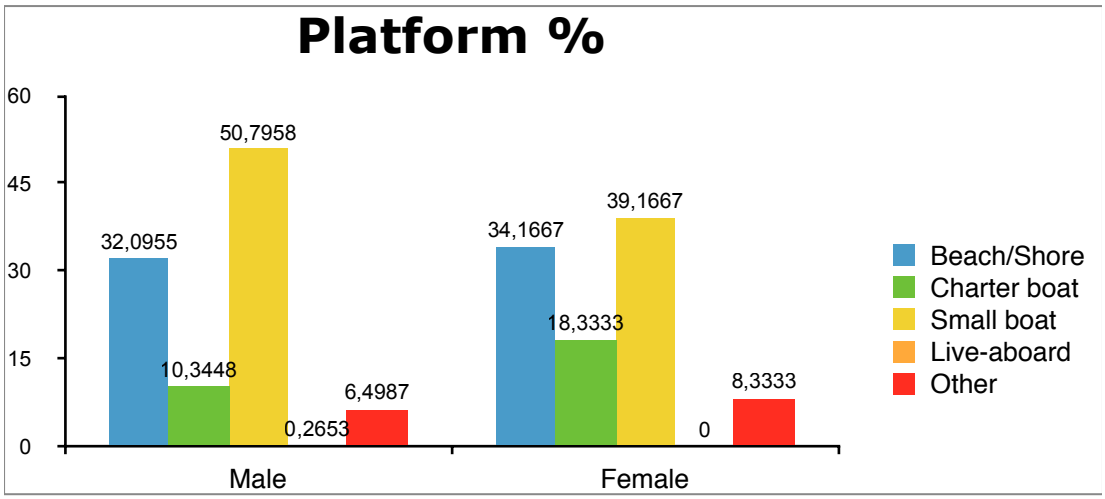

Figure 3.19: Chi-squared test results for Platform by gender

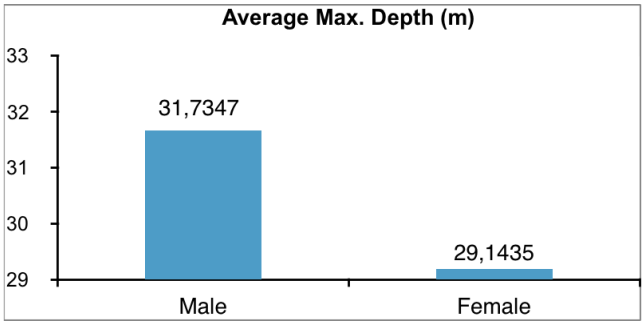

Figure 3.20: Chi-squared test results for Average Maximum Depth by gender

In Figure 3.21, workload percentages of male and female divers can be seen. The p value of chi-square test is 0.03997 meaning that workload depends on sex.

The breathing gas percentages are shown in Figure 3.22. Breathing gas does not depend significantly on sex as p value of chi-square test is 0.2113. However 36 male and only 1 female diver used Trimix. When we focus on Trimix specifically, the difference is significant as p value is 0.0464.

All of the female divers used open circuit scuba as it can be seen in Figure 3.23. The p value of chi-square test is 0,1001 which indicates a weak significance for dependancy of diving apparatus to sex.

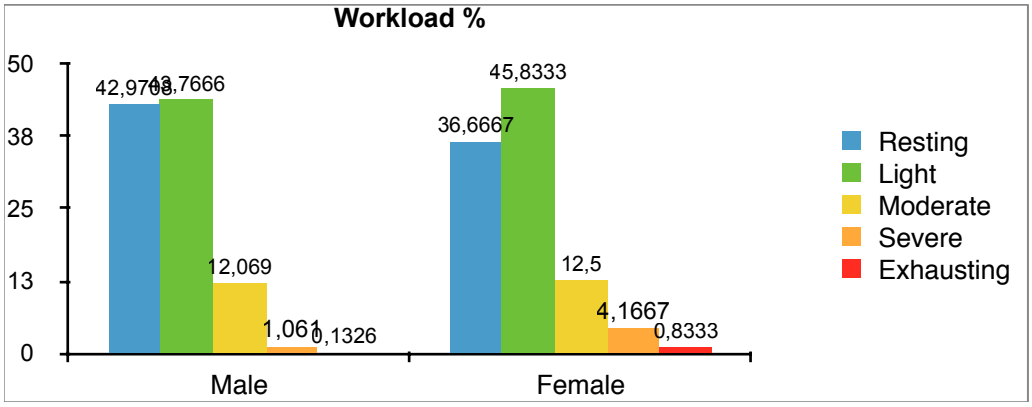

Figure 3.21: Chi-squared test results for Workload by gender

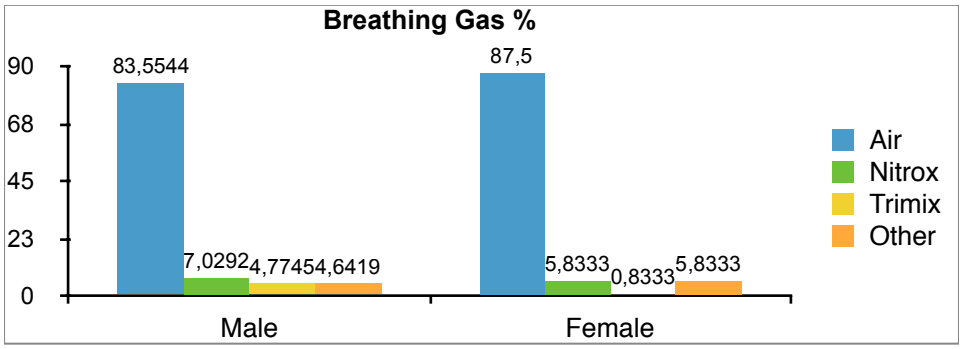

Figure 3.22: Chi-squared test results for Breathing Gas by gender

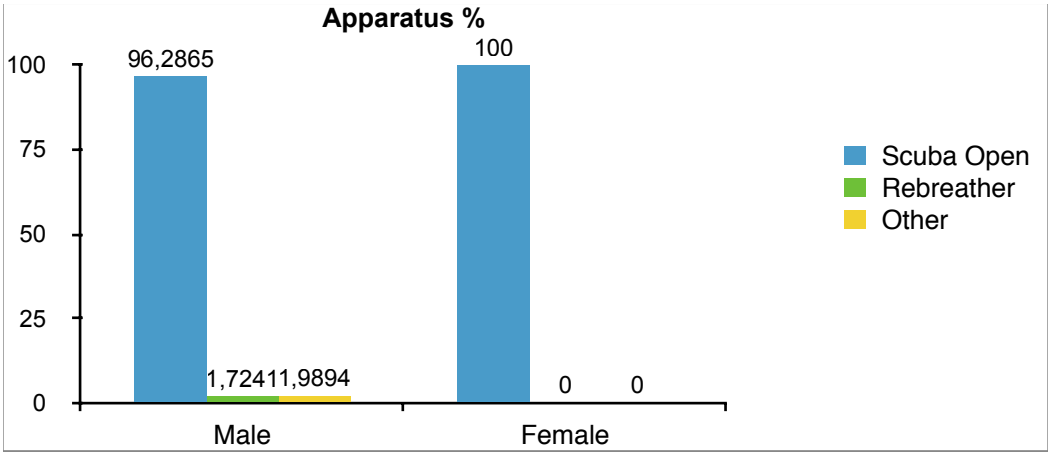

Figure 3.23: Chi-squared test results for Apparatus by gender

In Figure 3.24, the equipment malfunction percentages for male and female divers are shown. Male divers experienced equipment malfunction problem in 9.68 % of dives while this rate is 16.67 %. We can see that the face mask problems rate is very high in female divers compared to male divers. The p value of chi-square test is 0.0022 meaning that there is a significant difference between males and females.

There is not significant difference in symptom occurrence between male and female divers as seen in Figure 3.25. The p value of chi-square test is 0.8942.

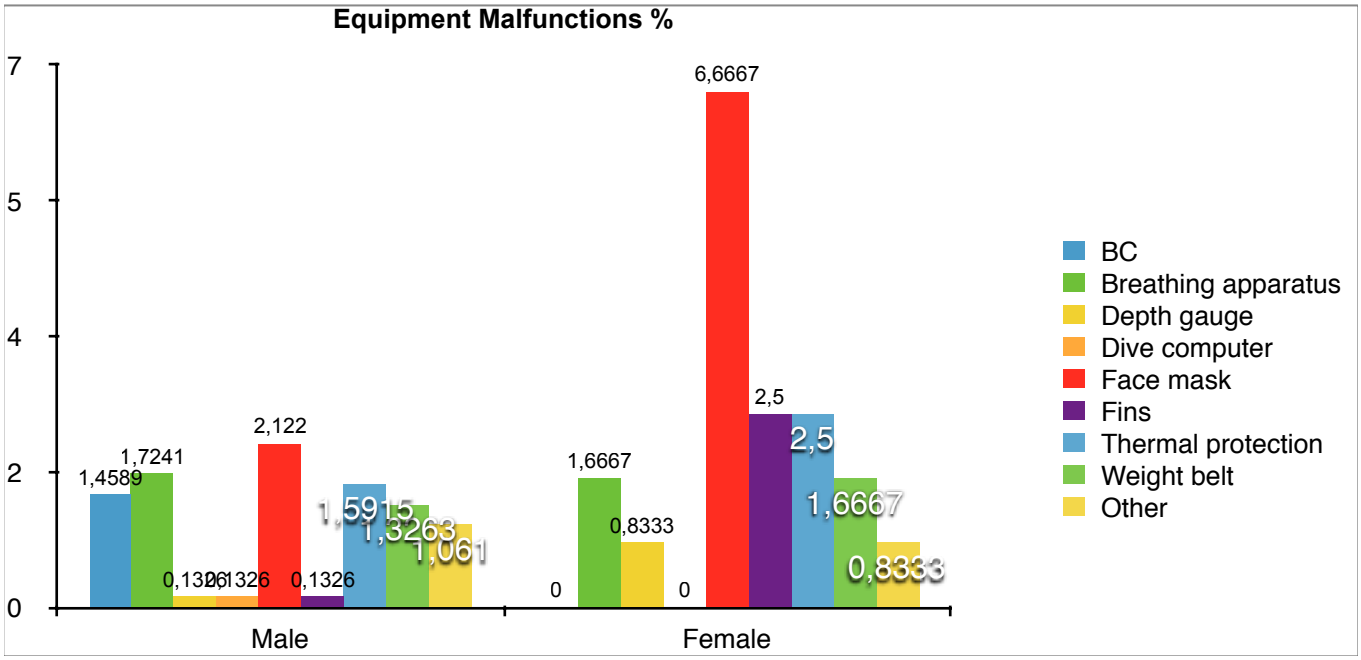

Figure 3.24: Chi-squared test results for Equipment Malfunctions by gender

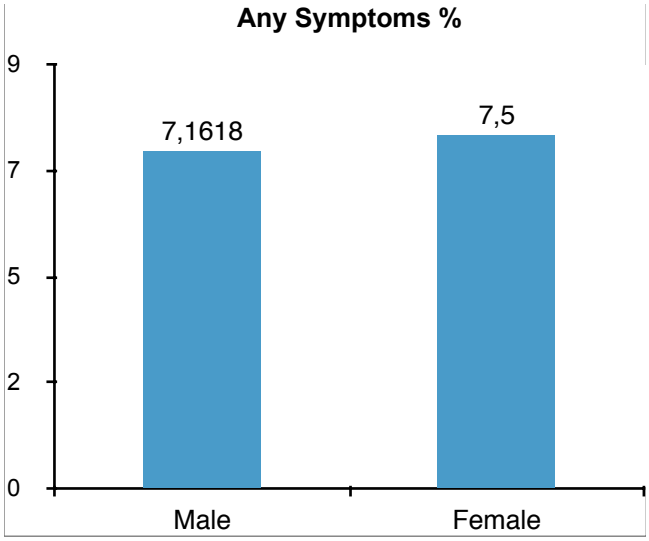

Figure 3.25: Chi-squared test results for Any Symptoms by gender

Overall, analysis of dive related variables regarding to clusters show some significant features about diving habits of divers. These can be summarised as follows :

- Exercise level before dive differs significantly according to clusters. Middle aged healthy divers is the most active group before dive.
- Older and experienced divers dive significantly deeper than other groups.

- The group of male and female divers with health problems complains more about cold water.
- Old and experienced divers dive into more depths whereas unhealthy divers dive into the least depth.
- The highest percentages of equipment malfunctions are in the unhealthy divers group especially in face mask and thermal protection.
- Group of old and experienced (and deep) divers has the highest percentage.
- The percentage of “No Exercise” is higher in females while males have higher percentage of moderate exercise.

5. CONCLUSION

We observed that two-step clustering produced clusters which are significantly different from each other. TwoStep method clusters data based on binary variables, whereas K-means creates clusters based on numerical variables (age and dive activity years).

Considering that similar results are obtained from one-dive and one-event divers' data, we focused on one-dive divers when performing statistical analysis on dive-related data to avoid repetitive information.

TwoStep clustering was the most suitable method for clustering given the type of diver's data especially with age and dive activity years transformed to categorical variables.

Obtained groups have meaningful characteristics. One group consists of middle aged male divers who have no health problems. Another group consists of male and female divers who have higher rates than other two clusters in following problems: cigarette smoking, allergy, asthma, back pain, ear/sinus problems and sea sickness. We named this group as divers who have health problems. Another group consists of old and experienced divers. Hearth problems in this group have higher rate than other groups.

For dive-related data, the group of unhealthy divers differs from others in exercise before dive (highest rate of no exercise), state of rest before dive (highest rate of tired), equipment malfunctions (especially in face mask and thermal protection problems) and any symptoms (lowest rate). The group of older and experienced divers differs from other groups in maximum depth (deeper than other groups), workload (highest rate in resting), breathing gas (highest rates of nitrox and Trimix) and any symptom (highest rate). The group of middle aged male divers with no health problems has significantly higher rates than the unhealthy divers group exercise before dive thermal comfort, lower rate in equipment malfunctions.

Male and female divers are significantly different in state of rest before dive, minimum water temperature, maximum depth and equipment malfunctions.

To avoid problems mentioned in method section prospective data collection studies, data should be recorded in a single table and keeping one record for each diver especially when recording demographical data.

Important data such as weight and height should be collected more carefully for all divers. Furthermore, current database mostly consists of data about health information. For much more comprehensive and effective data mining studies, collecting data about socio-economic status about divers is also important (having their own diving equipment, diving outside country, etc.)

Diving experience of divers is also an essential data for this study. The variable named DiveActivityYears indicates since when the diver is diving. However this variable doesn't provide accurate information about experience of divers. Therefore, variables expressing number of dives a diver has participated in his/her lifetime should be included in the database. Likewise, certificate degree information and diving courses taken by divers also give information about divers' experience hence these variables should also be included in the database for providing a much better analysis.

Physiological data about pre-dive and post-dive is highly important in terms of dive health research. To be able to carry out detailed studies about dive profile and bubble measurement, collecting these data should be considered crucial. Besides, a more detailed physiological statistics like water loss measurements, vascular measurements can be extracted even if they require devices which may be hard to keep in dive sites.

Further analysis can be carried out for observing the data of the "aging" divers who participate to multiple events over the life time of the DB.

The characteristics of diver clusters can be compared with the similar statistics of general population to analyse if divers differentiate from non-diver population in terms of variables like health, smoking, drug use, alcohol use.

Another future research direction is investigating multi-variable relationship between diver profile, dive profile - bubble measurement relation. By this way, risky diver groups, risky dive profiles and the possible results of combinations of dive and dive profiles can be examined.

REFERENCES

Ozyigit, T., Egi, S.M., Denoble, P., Balestra C., Aydin, S., Vann, V., Marroni, A., (2010). Decompression Illness Medically Reported by Hyperbaric Treatment Facilities: Cluster Analysis of 1929 Cases., *Aviation, Space, and Environmental Medicine*, 81(1):1-5.

Ozyigit, T., Egi, S.M., (2014). Commercial Diver Selection Using Multiple-Criteria Decision-Making Methods., *Undersea and Hyperbaric Medicine*, 41(6): 565-572.

Milley, A., (2000). Healthcare and data mining. Health Management Technology, 21(8), pp.44-45.

Kincade, K., (1998). Data mining: digging for healthcare gold. Insurance & Technology, 23(2), pp.2-7.

Young, J. and Pitta, J., 1997. Wal-Mart or Western Union? United HealthCare Corp.

Kolar, H.R., 2001. Caring for healthcare. Health management technology, 22(4), pp. 46-47.

Prather, J.C., Lobach, D.F., Goodwin, L.K., Hales, J.W., Hage, M.L. and Hammond, W.E., (1997). Medical data mining: knowledge discovery in a clinical data warehouse. In Proceedings of the AMIA annual fall symposium (p. 101). American Medical Informatics Association.

Gillespie, G. (2000). There's gold in them thar' databases. Health Data Management, 8(11), pp.40-52.

Veletsos, A. (2003). Getting to the bottom of hospital finances. Health Management Technology, 24(8), pp.30-31.

Jensen, P.B., Jensen, L.J. and Brunak, S., (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), p.395.

G. Piatetsky-Shapiro and W.J. Frawley (1991), Knowledge Discovery in Databases. AAAI/MIT Press.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth (1996). From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, 17(3):37.

Lloyd., S.P., (1982). Least Squares Quantization in PCM., *IEEE Transactions on Information Theory*, 28(2):129-137.

Aggarwal, C.C., Yu, P.S., (1999). Data Mining Techniques for Associations, Clustering and Classification, *Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining*. Springer-Verlag, Germany, pp. 13-23.

Chang, H.L, Yeh, T.H., (2007). Motorcyclist Accident Involment by Age, Gender and Risky Behaviors in Taipei.i Taiwan*., Transportation Research Part F*, 10(2):109-22.

Chiu T, Fang D, Chen J, Wang Y, Jeris C. (2001). A robust and scalable clustering algorithm for mixed type of attributes in large database environment, *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Fransisco, California, USA, pp. 263-8.

“The R Project for Statistical Computing.” [Online]. Available: https://www.r-project.org. [Accessed: 10-Jun-2016].

"SPSS Statistics.” [Online]. Available: http://www-03.ibm.com/software/products/en/spss-statistics [Accessed: 10-Jun-2016].

BIOGRAPHICAL SKETCH

PERSONAL INFORMATION

Name	Ahmet Cüneyt Yavuz
E-mail	cuneytyvz@gmail.com
Date of Birth	20.12.1990
Place of Birth	Turkey-Sakarya
Driving Licence	Yes
Marital Status	Single
Nationality	Turkey
Personal ID No	21686059670
Military Service	Postponed (20.09.2016)

EDUCATION

09.2014 - …	Galatasaray Üniversitesi, İstanbul Graduate, Computer Engineering
09.2008-01.2013	Yıldız Teknik Üniversitesi, İstanbul Undergraduate, Computer Engineering
09.2004-06.2008	Sakarya Figen Sakallıoğlu Anatolian High School Science

OCCUPATION

10.2014 - 01.2016 — İstanbul İletişim
Web Developer

- A Web application (http://showroomist.co) is developed using Spring framework on the backend and angular.js on front-end.

05.2014 - 08.2014 — OBSS Bilişim Hizmetleri Dan. San. A.Ş
Java Software Developer

- Customization of Atlassian products (JIRA, Bamboo, etc.) using Java and groovy.
- A web service project is implemented which provides communication interface for different Atlassian products using Spring and Hibernate.

04.2013 - 04.2014 — OBSS Bilişim Hizmetleri Dan. San. A.Ş (Turkcell Consultance)
Java Software Developer

- Architecture and programming of backend application which provides services for mobile applications is implemented using Spring and Hibernate.
- Database design and implementation.

02.2012 -12.2012 Shopamani Bilişim Hizmetleri Sanayi ve Ticaret A.Ş, İstanbul
Java Software Developer

- Implementation of server which provides services for mobile applications and cash registry systems.

PUBLICATIONS

- Ozyigit, T., Yavuz, C., Pieri, M., Egi, S.M., Egi, B., Altepe, C., Cialoni, D. and Marroni, A., 2016, July. Data Mining on Divers Alert Network DSL Database: Classification of Divers. In Industrial Conference on Data Mining (pp. 96-109). Springer, Cham.