Ertek, G., Tokdil, B., Günaydın, İ. "Risk Factors and Identifiers for Alzheimer's Disease: A Data Mining Analysis". In Proceedings of Industrial Conference on Data Mining (ICDM 2014), Springer. Ed: Petra Perner (2014)

Risk Factors and Identifiers for Alzheimer's Disease: A Data Mining Analysis

Gürdal Ertek, Bengi Tokdil, İbrahim Günaydın

Sabanci University, Faculty of Engineering and Natural Sciences, Istanbul, Turkey

Abstract. The topic of this paper is the Alzheimer's Disease (AD), with the goal being the analysis of risk factors and identifying tests that can help diagnose AD. While there exists multiple studies that analyze the factors that can help diagnose or predict AD, this is the first study that considers only non-image data, while using a multitude of techniques from machine learning and data mining. The applied methods include classification tree analysis, cluster analysis, data visualization, and classification analysis. All the analysis, except classification analysis, resulted in insights that eventually lead to the construction of a risk table for AD. The study contributes to the literature not only with new insights, but also by demonstrating a framework for analysis of such data. The insights obtained in this study can be used by individuals and health professionals to assess possible risks, and take preventive measures.

1 Introduction

The topic of this paper is the Alzheimer's Disease (AD) and the analysis of risk factors and identifying tests that can help diagnose AD. AD, a type of dementia disease, involves the irreversible degeneration of the brain which gradually ends up with the complete brain failure.

According to a 2012 report of the World Health Organization (WHO), 35.6 million people throughout the world are suffering from dementia diseases (Alzheimer Canada, 2013). Moreover, WHO projects that the total population of sufferers will double by 2030 and triple by 2050. It is also crucial to mention that, AD is the most common type of dementia disease. According to the statistics of Alzheimer's Association, AD accounts for 60 to 80 percent of the dementia cases (Alzheimer.org, 2013).

Neurodegeneration, progressive loss of neurons, increases due to aging and other factors, and these factors can lead to AD. On the other hand, neurodegenerative diseases such as AD cannot be diagnosed and treated fully due to the lack of treatment methods (Unay et al, 2010).

Besides the current statistics and forecasted spread of AD, the lack of a proven treatment method is another significant fact about this disease. Especially after the age of 65, AD generates a high risk to the population. A great percentage of the population suffers from this cureless disease, which eventually leads to death. Therefore, analysis of AD and insights based on available data are significant in understanding, alleviating the effects of, and paving the way to curing the disease.

Our study aims at generating a risk map of having AD after the age of 60. The probability of having AD will be analyzed in terms of age, social & economic status, gender, medical tests, and other factors, based on data coming from a field study. A detailed review of the literature on the factors that cause dementia and Alzheimer's disease can be found in a supplementary document (Supplement), and will not be included in this paper. Instead, we will focus on the work that we performed.

2 Data and Model

Our study uses data obtained from the Open Access Series of Imaging Studies (Marcus et al., 2010). The dataset consists of a collection of 354 observations for 142 subjects aged 60 to 96. Each patient may appear in more than one row. The subjects are all right-handed and include both men and women. The data also includes the education level and socio-economic status of the subjects. Moreover, some other medical statistics exist in the dataset, including intracranial volumes and brain volumes of the subjects. Summary statistics on the data, as well as some exploratory data analysis are presented in Marcus et al. (2010). We analyze the dataset using various visualization methodologies and a create risk map of the disease based on the given factors using classification trees.

Demented and *non-demented* are the classes in which the patient has the AD or not, respectively. *Converted* is the class that refers to the patients that develop the AD during the tests. The class *converted* was included in the classification tree analysis and cluster analysis, but removed from the dataset during the classification analysis. In classification tree and classification analyses, non-demented was selected as the target class, namely, the class that is predicted by the predictor attributes.

Table 1 presents the attributes (factors) in the analyzed dataset, explaining their meanings and providing their respective value ranges. Figure 1 presents the data mining process followed in the study. Figure 2 presents the roles assigned to attributes in the process ("Select Attributes" block of Figure 1). The attributes listed inside the "Attributes" box are the predictor attributes, whereas the attribute listed inside the "Class" box is the predicted attribute.

In Table 1, Clinical Dementia Rating is abbreviated "CDR". CDR can only take values 0, 0.5, 1 and 2. CDR being equal to 0 corresponds to non-demented subject CDR being equal to 0.5 corresponds to very mild dementia and CDR above 1 corresponds to moderate dementia. This medical test carries significance to entitle a subject as Alzheimer patient. The mini–mental state examination (MMSE) is a questionnaire test that has 30 questions. The goal of this test is to examine the cognitive situations of individuals. The questions of MMSE cover arithmetic, memory, and orientation. MR delay refers to the number of days between two medical visits. Other than those parameters, there is also information about age of the subjects in the classification tree. As mentioned earlier, the range of the age of the subjects is 60 to 96.

Classification tree analysis has been conducted with respect to the classes that the observed subjects belong to, namely demented, non-demented, and converted. In the data mining process (Figure 1), there are three main types of analysis: classification tree (decision tree) analysis, hierarchical clustering, and classification analysis. The data mining process begins with reading of the data from file (File block), and the validation of the data by displaying it in a data table, as well as observing the histogram (Distributions block), scatter plot (Scatterplot block), and attribute statistics (Attribute Statistics block). Then, each of the attributes is specified either as the class attribute or one of the predictor attributes (Select Attributes block). The roles specified for the attributes are given Figure 2.

The class label is "Group" and the key attribute is "MRIID". The attributes under the Attributes list box are predicator / factors in the classification tree analysis and classification analysis. In the clustering analysis, the Available Attributes "Visit", "MR Delay" and "CDR" are also included. The classification tree algorithm used is C4.5 and the created classification tree is visualized as a graph (Classification Tree Graph block). The visualized classification trees are displayed in Figures 3 and 4.

Hierarchical clustering analysis begins with the calculation of the attribute distances and storing these distances in a matrix (Attribute Distance block). Then hierarchical

clustering is carried out (Hierarchical Clustering block). The visualization of the clusters is displayed in Figure 5.

Table 1. The explanation and the value ranges of the attributes of the OASIS dataset.

Attribute	**Explanation and Value Range**
Group	The class label. Demented, non-demented, or converted.
MRIID	The test ID. Unique for each row. 1 to 354.
SubjectID	The subject's ID. 1 to 142. A subject may be visiting more th once, so the number of rows (354) is larger than the numbe subjects (142).
Visit	Visit of the subject. 1 to 5.
MRDelay	The delay of a subject since the last visit.
CDR	Clinical Dementia Rating. 0 = no dementia, 0.5 = very mild AD, 1 = mild AD, 2 = moderate AD. (Morris, 1993)
Gender	Male (M) or Female (F)
Age	The age of the subject at the time of observation
EDUC	Education level
SES	Socioeconomic status, which is assessed by the Hollingshea Index of Social Position. 1 (highest status) to 5 (lowest statu (Hollingshead, 1957)
MMSE	Mini-Mental State Examination value. 0 (worst value) to 3((best value). (Folstein, Folstein, & McHugh, 1975)
eTIV	Estimated total intracranial volume (cm3) (Buckner et al., 2004)
nWBV	Normalized whole-brain volume, expressed as a percent of voxels (Fotenos et al., 2005)
ASF	Atlas Scale Factor; volume scaling factor for brain size.

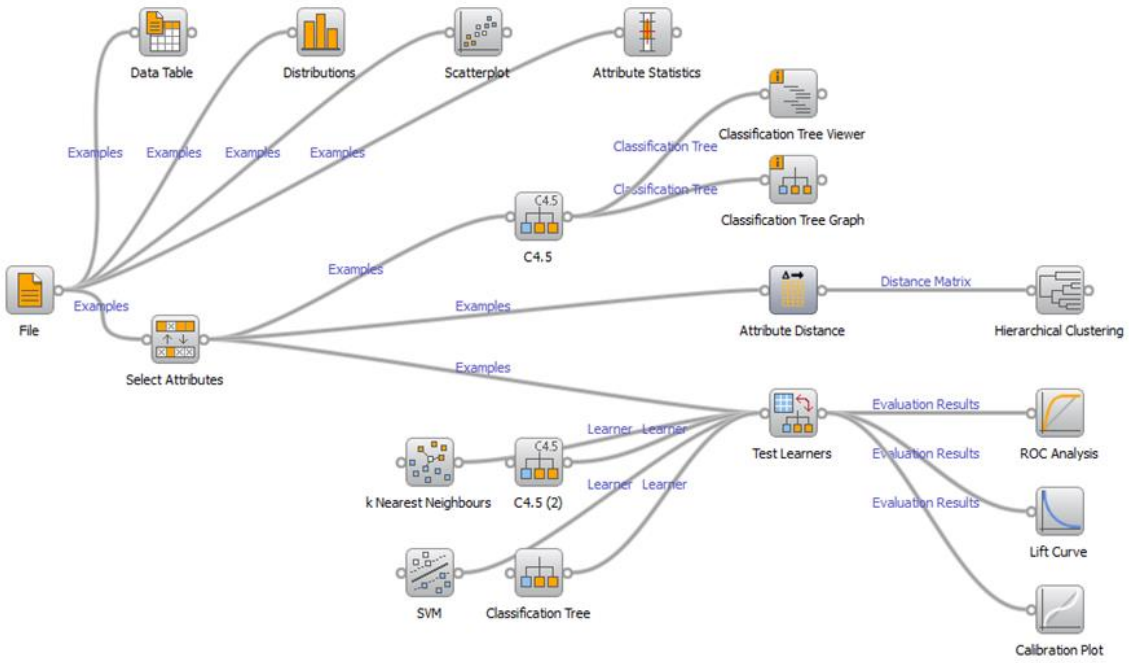

Fig. 1. The data mining process followed in the study.

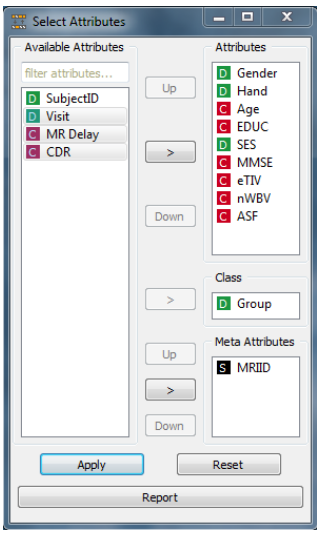

Fig. 2. The key attribute (Meta Attributes), class label (Class), and the attributes used for prediction (Attributes). The clustering analysis also includes the grey-shaded attributes within Available Attributes.

The classification analysis involves four classification algorithms (learners), namely k Nearest Neighbors, C4.5, SVM, and Classification tree. The performances of these four learners were compared (Test Learners block) with respect to classification accuracy, using a 5-fold design.

3 Analysis and Results

In this section, we present the data mining results and the insights that we obtain through these results. The analysis has been carried out using Orange (Orange) and Tableau (Tableau) data mining software. The analysis results are presented as a list of insights, and are later summarized in Table 2.

The preliminary classification tree constructed considered MRIID as the key attribute, Group as the class attribute, and included all the other attributes (except SubjectID) as factors. However, this resulted in a tree where the first split based on CDR perfectly distinguished the demented patients (CDR=1) from other subjects (non-demented and converted). This showed that CDR was too good of a factor to include in the analysis.

In the preliminary analysis, the next split in the tree was based on the attribute “MR Delay”. However, using this attribute also had an inherent flaw: The demented subjects need to be under control with frequent medical tests. Most of the potential Alzheimer patients take the MR tests earlier than 675 days. Therefore “MR Delay” is dependent on the “CDR” score, and the probability of being converted. The subjects whose “CDR” values are greater than 0, and additionally if “MR delay” periods of these patients are smaller than 675 days, with 97.9% probability these subjects are either now or eventually became converted Alzheimer patients. The attribute “Visit” (number of visits) is also dependent on the “CDR” results.

Observing the “too perfect” results in the preliminary classification tree analysis due to “CDR” and the inherent dependency problem of “MR Delay” and “Visit”, we decided to carry out our analysis by excluding these three attributes from the list of factors, as given in Figure 2.

Figures 3 and 4 show the graph visualizations of the classification tree after the attributes were selected as in Figure 2. In each pie, the light-colored slice represents the non-demented observations, darker slice represents demented observations and the darkest slice represents the converted observations (subjects who were observed not to have AD at that observation, but later possessed the disease).

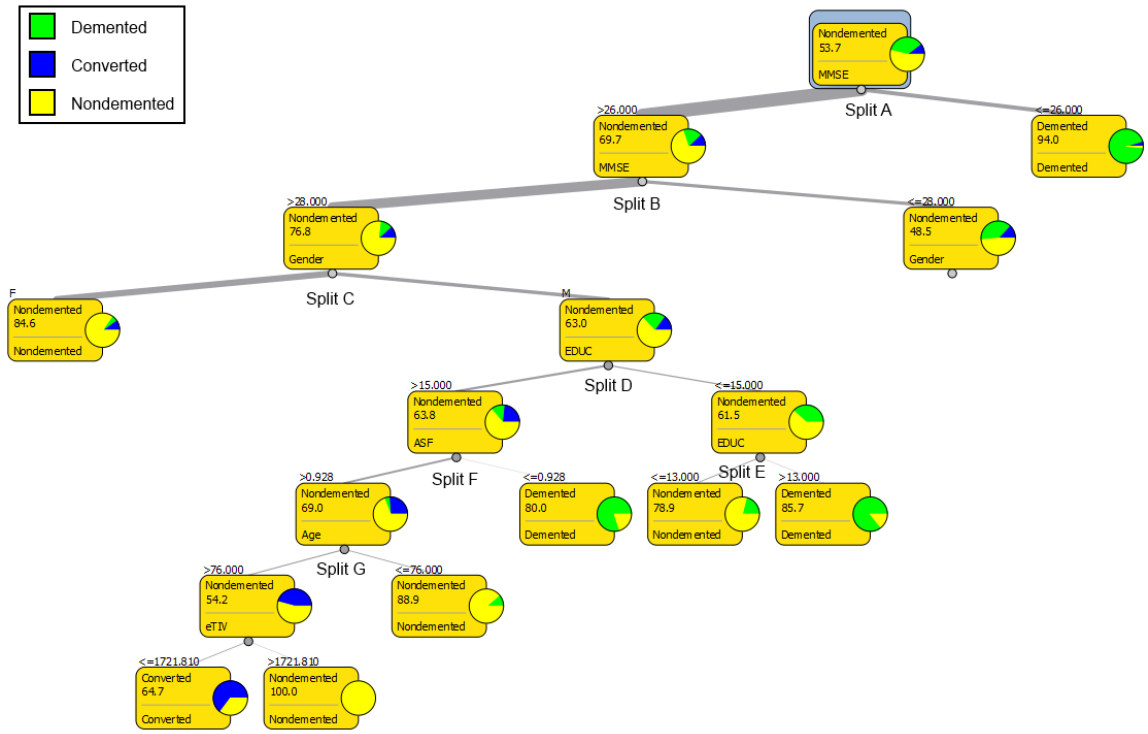

Fig. 3. The expansion of the classification tree for CDR >0.250.

In analyzing the classification tree graph, as visualized in Figures 3 and 4, we will be especially interested in two types of observations: 1) The deviations from the original distribution of the class labels (root of the tree), 2) The significant deviations between the parent and children nodes after a split is made.

The insights obtained from the classification tree analysis are now presented, following the observations that lead to those insights. Insights 1 through 6 are based on the expansion of the left mode (Figure 3), whereas insights 7 and 8 are based on the expansion of the right mode (Figure 4).

In Figure 3, the branches of the classification tree are split firstly (Split A) with respect to the values of MMSE. MMSE is thus a high-ranking indicator of AD. As it can be observed seen from the right branch of Figure 3, if MMSE<26, then the patient is demented at the time of the observation with a very high probability (94%). However, the left branch needs a further analysis.

Insight 1: *If the MMSE value is smaller than 26, the risk of AD increases considerably to 94%.*

In Figure 3, in the left branch, the tree is first split based on MMSE again (Split B), and then based on gender (Split C). The MMSE values are greater than 28. In the female gender side of the branch, there is 84.6% probability of being non-demented.

Insight 2: *If a woman has MMSE value greater than 28, than she has a probability of being non-demented with a probability of 84.6% at the time of the observation.*

The branch of men is split further to make more analysis. The next split (Split D) is with respect to education values of the male people. Education level is divided into two. In the left branch, there are males with education level EDUC≤15 while in the right branch the education level EDUC≥15.

Insight 3: *Among the men who have MMSE*>28*, those who have an education level EDUC*>15 *have* 63.8% *chance of being non-demented at the time of observation, and those with EDUC*≤15 *have* 61.5%. *Yet there are no converted among those on the right branch; they are all demented. Therefore, less educated subjects show signs of dementia early on, whereas more educated convert later, summing to similar percentage of AD in the long run.*

Even though Insight 3 says that the percentage of demented plus converted is very close for Split D, Insight 4 goes into the detail, based on Split E.

Insight 4: *Among the men whose MMSE*>28*, those who have an education level in the range* (13, 15] *have much higher chance of having dementia, compared to those in other value ranges. Thus, the most risky range of education level for males who have MMSE*>28 *is the interval* (13, 15]*, which refers to Bachelor's diploma at a university.*

When the branch of education level is EDUC>15 (left branch below Split D) is considered, there are again two other branches. These branches split according to their ASF values. ASF is the abbreviation of Atlas Scale Factor. This is a clinical term, which is the result of the MRI scans, and explained in Table 1.

Insight 5: *Among the men who have EDUC*>15 *and MMSE*>28*, those who have the ASF*≤0.928*,* 80% *are demented at the time of observation (right branch under Split F). Therefore, for men in this group, ASF is a major identifier of AD.*

The branch where the ASF value is greater than 0.928 is divided into two, according to Age (Split G). The left branch is where the age is greater than 76 while the right branch is the age is equal to or less than 76. In this study, the ages were between 60 and 96, and the age 76 seems to be the threshold age for men where significant changes take place.

Insight 6: *For men with EDUC*>15*, MMSE*>28*, and ASF*>0.928*, the age is equal to* 76 *or less than* 76*, there is* 88.9% *conditional probability that they are non-demented.*

This insight can also be expressed as follows: If a man with more than 15 years of education has MMSE>28 and ASF>0.928 when he is older than 76, then he will most probably *not* have AD.

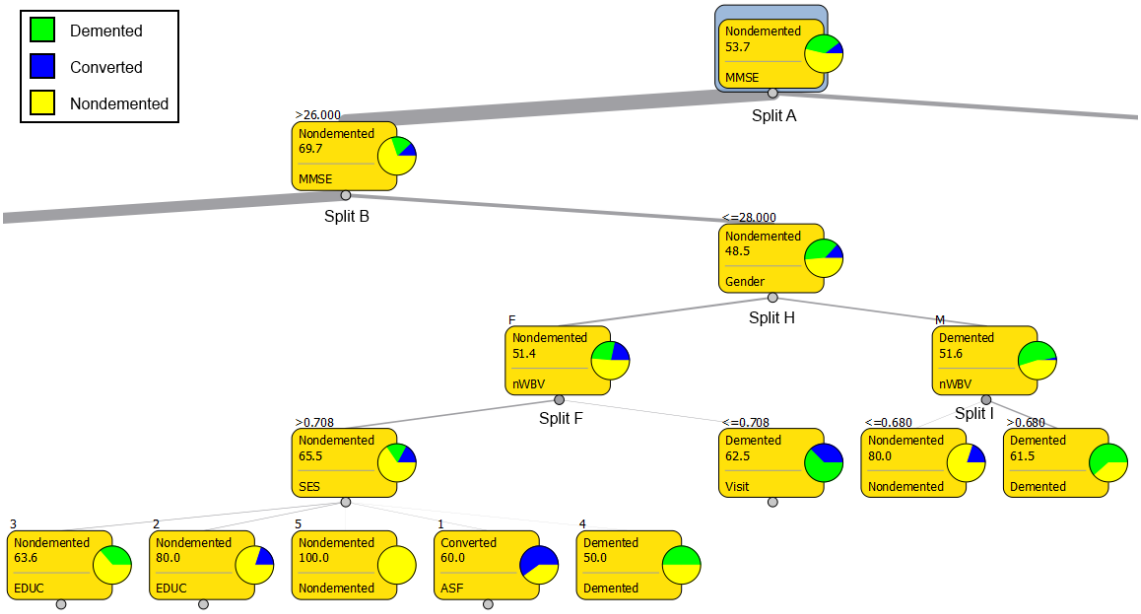

Fig. 4. The expansion of the non-demented branch of the classification tree.

So far, the branch of the classification tree for the subjects who have MMSE value greater than 28 has been analyzed by observing Figure 3. Now the subjects who have MMSE value between 26 and 28 will be analyzed through Figure 4 (the right branch under Split B). This branch of the tree contains a greater portion of demented and converted subjects compared to the other branch. As the effect of the MMSE value has been indicated, the same effect can be observed in Figure 4. The smaller MMSE value results in higher

risk of having the disease. The effect of other factors such as gender, SES and nWBV will be explored through Figure 4.

Gender could be an indicator for the statistical studies. When the non-demented percentages (under Split H) are compared with respect to gender, it can be seen that both genders have more or less the same percentage of non-demented subjects. Specifically, the female branch has 51.4% non-demented and male branch has 51.6% demented proportions. Therefore, there is not a clear distinction between these two values in terms of reasoning a differentiation. However, when the composition of the remaining portion of the pie is analyzed, it is observed that the remaining men (M) are almost all demented, whereas about half of the women (F) are converted later.

Insight 7: *For subjects that have MMSE in the range* (26, 28]*, men and women exhibit similar percentages of non-demented, but almost all the remaining exhibit dementia at the time of observation, whereas nearly half of the women develop dementia later (being converted).*

The next important factor according to the classification tree graph is nWBV, which is an abbreviation for "Normalized whole brain volume". nWBV is the next splitting attribute for both men (M) and women (F), as can be seen in Splits I and F, respectively. For men, nWBV>0.680 signals a big risk factor, since 61.5% of the men under Split I, who have nWBV>0.680 are demented. For women, Split F tells that having nBWV≤0.708 completely guarantees being demented or being converted. There is a significant deduction from these observations, as given in Insight 8:

Insight 8: *When men and women with MMSE in the range* (26, 28] *are considered, larger nWBV values with nWBV*>0.680 *(larger brain volumes) are more risky for men, whereas smaller brain volumes (nWBV*≤0.708*) are more risky for women.*

The other factor that has an impact on having AD is socioeconomic status of the subjects, SES value. There is an increase in the converted ratio if the SES=1. While one might hypothesize that "People with the highest socioeconomic status are more likely to develop AD over time", this may not be true. It may be the case that the people with the largest income are those that continue to come to future MR tests, until they develop AD. There was not enough data in our sample (with SES=1 and multiple visits) to test whether this was the case or not.

The next analysis carried out was the hierarchical cluster analysis, whose results are displayed in Figure 5 as a dendrogram. In the dendrogram, attributes that are in neighboring branches, or from the same parent branch are related to each other. There is no

input/output or cause/effect relation in the clustering analysis and the construction of the dendrogram; therefore, “CDR”, “MR Delay” and “Visit” have been included among the attributes. The combination of several factors is more conclusive in terms of the risk map. The proximity of the attributes to each other can be seen through clustering. For instance, the education level of the subject is closely related with the socio-economic status of the subject (SES). In addition, both of SES and education are related to the result of mini-mental state examination (MMSE) of the subjects. Based on this observation from Figure 5, the relation of education to Alzheimer’s deserves further investigation. Education directly influences the social status of the individuals in real life. Therefore, those two factors are connected to each other and the dataset of OASIS specifies these two data have similar effects. For further insights, the values for the education attribute “EDUC” can be discretized to take the following categorical values:

1: less than high school degree

2: high school degree

3: some college

4: college degree

5: beyond college

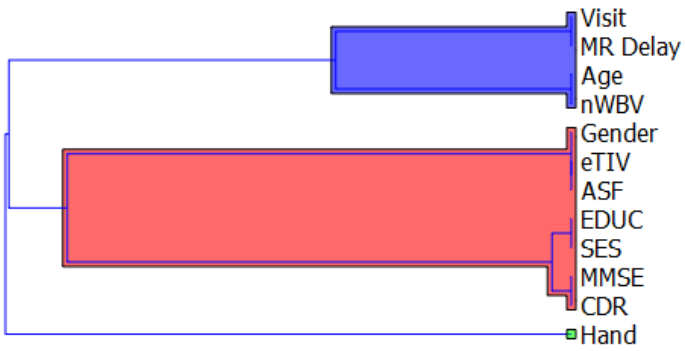

Fig. 5. Dendrogram for the attributes, showing their proximity to each other based on the sample data.

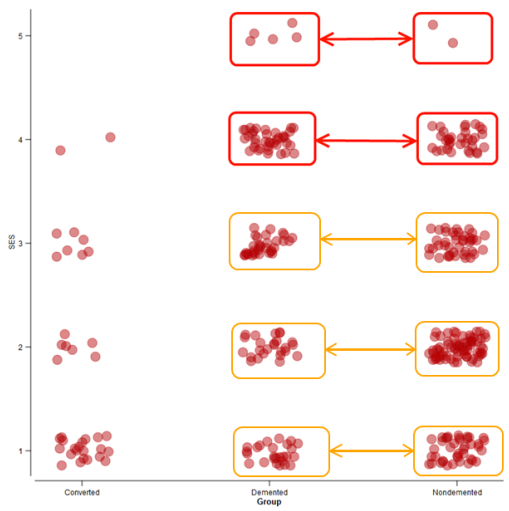

Fig. 6. Scatter plot of education effect.

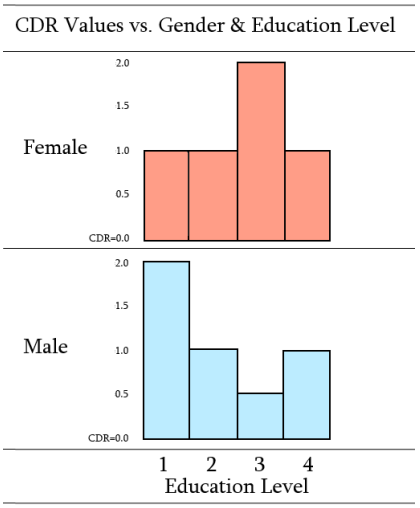

Fig. 7. The relation between gender, education, and CDR.

In Figure 6, an analysis is performed to reveal the possible relation between the education levels and the risk of having Alzheimer. For this visualization, the density of points for non-demented and the density points for other values can be compared. For education level taking values of 1, 2 or 3, the density of non-demented subjects is greater than the others, meaning that the risk of the disease is lower. However, for the education level values of 4 and 5, it can be seen that the number of demented subjects are greater than the non-demented subjects. For better illustration, the comparisons were highlighted in Figure 6. The insight obtained is the following:

Insight 9: *The risk of Alzheimer is higher for people with college degree or higher.*

Next, the effect of education level and the effect of gender were considered together, as shown in Figure 7. CDR is a crucial factor to identify AD. As it is evaluated before, if CDR is greater than 0.5, the possibility of having the disease increases considerably. The important observation for the Figure 7 is that the risk for women is greater for the education level 4. A similar observation can be done for the men for the education level 1. On the other hand, for the education level 4 the opposite observation can be made. Hence, it can be summarized that women with college degrees are in a riskier position than men with college degrees, in terms of being an Alzheimer patient.

Insight 10: *For higher education levels, especially for women college graduates are in a riskier position to having AD.*

The final analysis carried out was classification analysis, where the predictive power of the attributes has been tested. Unfortunately, the classification accuracies came out to be too low.

Insight 11: *The listed attributes cannot predict the risk of AD accurately.*

Therefore, as a deduction of Insight 11, one should rather focus on exploratory data mining for the given data, rather than predictive data mining.

4 Conclusions

It is expected that the number of Alzheimer patients will increase in upcoming years. Apart from this projection, the lack of a precise medical treatment method for this disease will also continue to increase the possibility of deaths due to Alzheimer. Due to these facts, the understanding of AD risks is crucial.

Table 2. The summary of insights on the risk of AD.

Risky ranges	**Related Insight(s**
MMSE≤26	Insight #1 & 2
EDUC∈(13,15] *(for men with MMSE*>28*)*	Insight #3 & 4
ASF≤0.928 *(for men with EDUC*>15 *and MMSE*>28*)*	Insight # 5
MMSE∈(26,28]	Insight # 7
nWBV>0.680 for men (M), nWBV≤0.708 for women (F) *(for MMSE*∈(26,28]*)*	Insight # 8
College degree or higher	Insight # 9
College degree for women, less than high school degree for me	Insight # 10

As a contribution to the previous literature on AD, in this study, the effects of the factors are examined from a broader perspective through data visualization and mining methods. Rather analyzing brain images, the demographics and test statistics for the subjects have been examined. In terms of presenting the risk map of AD, the riskier ranges of each crucial factor can be summarized as in Table 2. As a distinctive factor from other studies of AD, our study is based on a recent dataset that includes not only demographic attributes, but also test results as attributes.

The study contributes to the literature not only with new insights, but also by demonstrating a framework for analysis of such data. Individuals and health professionals to

assess possible risks, and take preventive measures can use the insights obtained in this study. The insights can also be used by health institutions, pharmaceutical companies, insurance companies, government institutions for planning their strategies for the current and the future.

5 Acknowledgements

The authors thank Precious Joy Balmaceda for her help in proofreading the paper.

6 References

1. Alzheimer Canada, (2013), available under http://www.alzheimer.ca/en/Get-involved/Raise-your-voice/WHO-report-dementia-2012. Accessed on January 24, 2013.

2. Alzheimer.Org (2013), available under http://www.alz.org/dementia/types-of-dementia.asp. Accessed on January 24, 2013.

3. Marcus, D. S., Fotenos A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L. (2010), Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults, *Journal of Cognitive Neuroscience*, 22, 2677-2684.

4. Supplement for "Risk Factors and Identifiers for Alzheimer's Disease: A Data Mining Analysis". Available under http://people.sabanciuniv.edu/ertekg/papers/supp/11.pdf.

5. Unay, D., Chen, X., Erçil, A., Çetin, M., Jasinschi, R., van Buchem, M.A., & Ekin, A. (2009), Binary and nonbinary description of hypointensity for search and retrieval of brain MR images, *IS&T/SPIE Electronic Imaging, Multimedia Content Access: Algorithms and Systems III*, San Jose, California, USA, January 2009.

6. WHO (2912), Dementia: a public health priority. World Health Organization and Alzheimer's Disease International. Available under http://www.who.int/mental_health/publications/dementia_report_2012/en/. Accessed on November 13, 2013.