Research Article

Framingham Risk Score by Data Mining Method

Şükrü Kitiş [a],*

ªControl Engineering, Kütahya Dumlupınar University, Simav-Kütahya, Turkey

ABSTRACT

There are cleaning, integration, reduction, conversion, algorithm implementation and evaluation stages in data mining meaning finding necessary data from a wide variety of variables and data. It is important to create a data warehouse to realize these steps. Data randomly selected from data warehouse is evaluated with certain algorithms. While deaths resulting from heart diseases in our country are 37% according to 2016 data, 420-440 thousand people are diagnosed as heart patients each year and the number of deaths per year can reach 340 thousand people. These values correspond to approximately three times of Europe. In this study, risk of heart attack is calculated by data mining method by taking advantage of Framingham risk score. In order to determine this risk factor; 10-year risk is calculated by looking at sex, age, total cholesterol, HDL cholesterol, blood pressure, diabetes and smoking. While the effects of the ages for men starts -9 points, ends with +13 points and for women starts -7 points, ends with +16 points. While the effects of the total cholesterol for men starts 0 points, ends with +11 points and for women starts 0 points, ends with +13 points. Total scores are between 0-17 and over in men, and scores between 0-25 and over in women. There are risk values ranging from 1% to 30%.

1. Introduction

Data mining, which expresses relations and rules providing estimation from large amounts of data by using computer programs [1], is also evaluated as a data analysis process [2] that reveals information that raw data cannot present alone. It is also defined as revealing previously unknown, implicit, unclear but useful information from available data [3]. According to Fayyad [4], it is the emergence of valid, reliable, potentially useful, previously unknown and understandable patterns from databases. Data mining is the process of converting huge data into information by saving time. In the future, there will be huge increases in data in many fields such as education, medicine, telecommunications, banking and industry. By using data mining techniques, costs will decrease, revenues will increase, productivity will increase, new opportunities will arise, new discoveries will be made, and frauds will emerge [5].

Cardiovascular diseases include coronary heart diseases, rheumatic heart diseases, hypertension, peripheral artery disease, cerebrovascular diseases, congenital heart diseases and heart failure. In the development of heart and vascular diseases; physical inactivity, hypertension, unhealthy diet, diabetes, and tobacco use play an important role [6].

In 2012, 46.2% (17.5 million) of deaths in the world were caused by cardiovascular diseases, 7.4 million of these deaths were due to heart attack and 6.7 million due to stroke. 37% of deaths under the age of 70 are due to cardiovascular diseases. Deaths related to cardiovascular diseases are estimated to be approximately 22.2 million in 2030 [7].

According to Turkey Statistical Institute (TUIK) mortality data, share of heart attacks among total deaths has gradually increased. Heart diseases rank first among causes of death with 40% in 1989, 45% in 1993, 40% in 2009, 39.6% in 2013, 40.4% in 2014. 39.6% of deaths due to circulatory system diseases are caused by ischemic heart disease, 24.7% from cerebrovascular disease, 18.8% from other heart diseases and 11.6% from hypertensive disease.

When analyzed as age groups, most cause of death was circulatory system diseases in the 75-84 age group. When deaths due to circulatory system diseases are analyzed by residence, the first five provinces with a high rate are Denizli, Kırklareli, Yozgat, Samsun and Artvin [6].

Currently, cardiovascular diseases are the leading causes of morbidity and mortality. This situation is more evident in developing countries like our country. Atherosclerotic heart disease occurs as a result of multiple risk factors. Underlying disease develops over the years, and it often becomes much more advanced when symptoms appear. Benefit of treatment interventions to be made after this stage will be limited [8].

2. Material and Method

In order that a risk calculation method is to be useful, it must be easy to use and based on strong and up-to-date information. Such methods generally depend on factors that cannot be changed, such as age, gender, and changeable factors such as smoking, blood lipids, and blood pressure [8].

There are still many risk calculation methods. The oldest and most preferred of these is the Framingham system. Other known systems include SCORE, PROCAM, QRISK, WH /ISH, Reynolds Risk Score [8].

While circulatory system diseases have the highest rate with 39.7% in distribution of causes of deaths distribution according to 2017 data, nervous system and sensory organs have the lowest rate with 4.9% (Table 1). In the distribution of circulatory system diseases, ischemic heart diseases ranked first with 39.7%, and hypertensive diseases ranked last with 8.9% (Table 2).

Distribution of causes of death was as in Table 1 in 2016 and 2017 according to TUIK data [9].

Table 1. Distribution of causes of death [9]

Distribution of causes of death (2016-2017)	2016		2017	
	Number	%	Number	%
Total	420.189	100.0	416.881	100
Circulatory system diseases	166.069	39,5	165.323	39,7
Benign and malignant tumors	81.647	19,4	81.527	19,6
Respiratory system diseases	49.295	11,7	49.855	12,0
External laceration causes and poisoning	21.473	5,1	18.901	4,5
Diseases related to endocrine, nutrition and metabolism	20.731	4,9	20.110	4,8
Nervous system and sensory organs diseases	20.220	4,8	20.504	4,9
Other	60.754	14,5	60.661	14,6

According to TÜİK [9] data, distribution of circulatory system diseases in 2016 and 2017 was as in Table 2.

Table 2. Distribution of circulatory system diseases [9]

Distribution of deaths from circulatory system diseases by subgroups(2016,2017)	2016		2017	
	Number	%	Number	%
Circulatory system diseases	166.069	100,0	165.323	100,0
Ischemic heart disease	67.412	40,6	65.666	39,7
Cerebro-vascular disease	38.959	23,5	37.885	22,9
Other heart disease	37.077	22,3	38.606	23,4
Hypertensive diseases	14.537	8,8	14.706	8,9
Other	8.084	4,9	8.460	5,1

Based on Framingham risk score data, American Heart Association (AHA) developed a risk assessment system. In this system, risk of heart attack within 10 years calculates with gender, age, smoking, family history, presence of cardiovascular disease, presence of diabetes, elevated blood glucose (> 100 mg), height, weight, waist circumference, systolic and diastolic blood pressure, antihypertensive therapy, total cholesterol, HDL-Cholesterol, LDL-cholesterol, triglyceride parameters [8].

In our country, Framingham study is also involved in e-pulse application.

According to this system, when information for risk of heart attack is matched, it is assumed that if 10-year risk is more than 20%, patients should be treated very seriously, and if risk is between 10-20%, they should receive a moderate treatment. <10% shows low risk, 10-20% medium risk, and > 20% high risk [10].

Framingham Risk Scoring calculates 10-year cardiovascular risk in percentages corresponding to question such as:

Gender

Age

Total Cholesterol (mg / dL)

HDL Cholesterol (mg / dL)

Blood Pressure (mm Hg)

Diabetes

Cigarette [11]

In this study, regression tree model was used to calculate the risk of heart attack with data mining method. The algorithm prepared for this method differs for men and women. For groupings and scoring used in this tree model (Figure 1-10 and Table 3). It was taken advantage of articled performed by Peter W.F.W. et al. [12].

Considering age for men, the relevant part of the regression tree is as given in Figure 1.
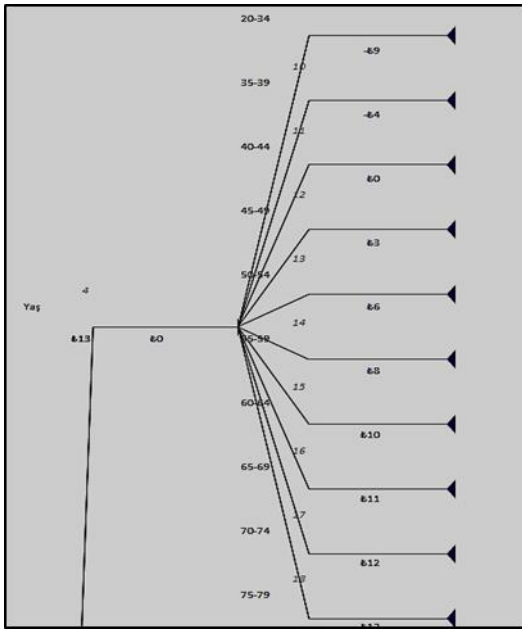
Figure 1. Age factor affecting heart attack in men

As seen in Figure 1 in scoring, it starts with -9 points between the ages of 20-34 and ends with +13 points between the ages of 75-79. The risk of heart attack tends to increase with increasing age. The values of -9 and -4 used here affect 20-34 age group and 35-39 age group among other risk groups. For this reason, abnormalities in the values for future risk calculations arise from these -9 and -4 scores.
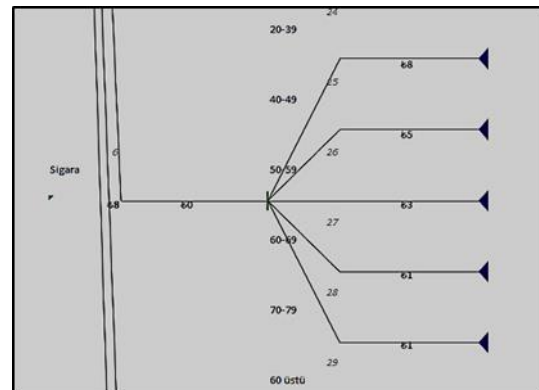
The section for total cholesterol (mg / dl) is as given in Figure 2.
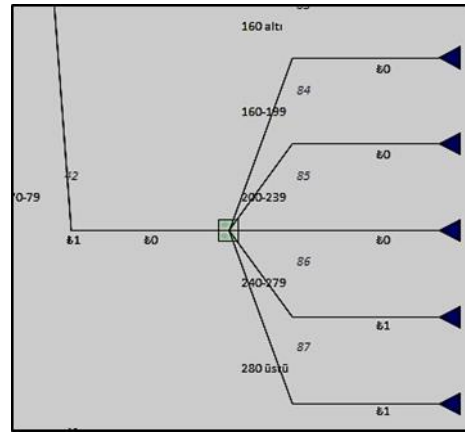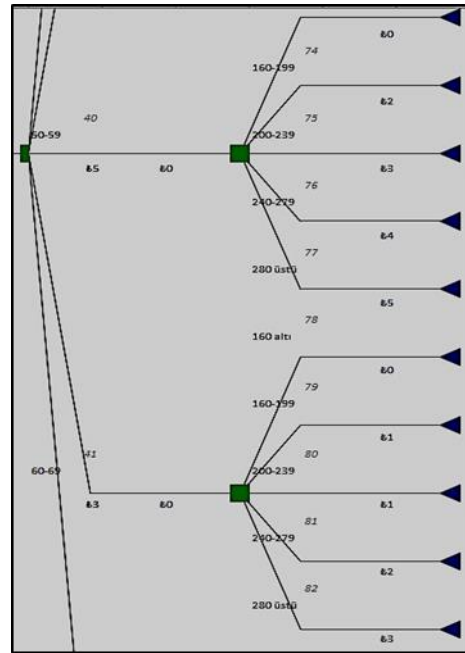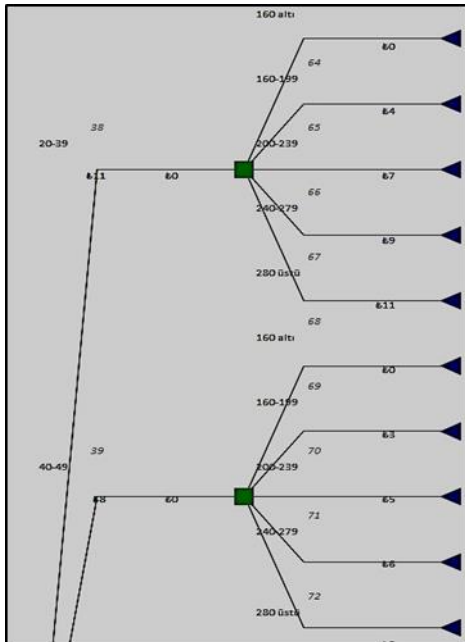
Figure 2. Total cholesterol factor affecting heart attack in men.

In the scoring, as seen in Figure 2, total cholesterol values affect each age group separately. It takes values as minimum 0 and maximum 11.
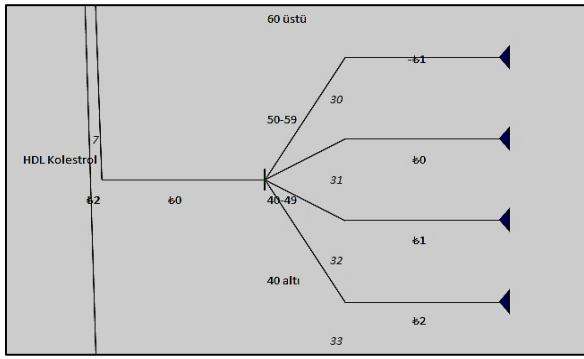
The relevant section for smoking is as given in Figure 3.

Figure 3. Smoking factor that affects heart attack in men

As seen in Figure 3, smoking shows a tendency to increase with age (values of -9 and -4 in Figure 1 should not be forgotten).
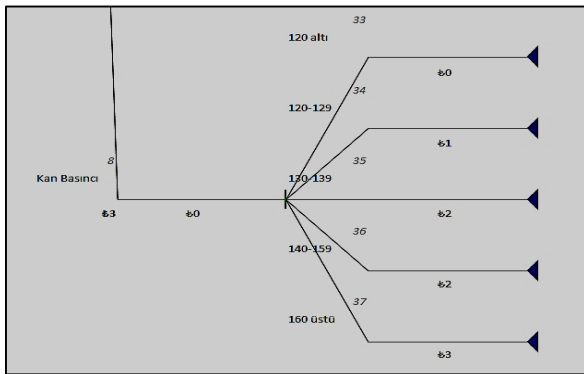
The relevant part for HDL cholesterol (mg / dl) is as given in Figure 4.

Figure 4. HDL cholesterol factor affecting heart attack in men

As seen in Figure 4, as HDL cholesterol decreases, the risk of heart attack increases.
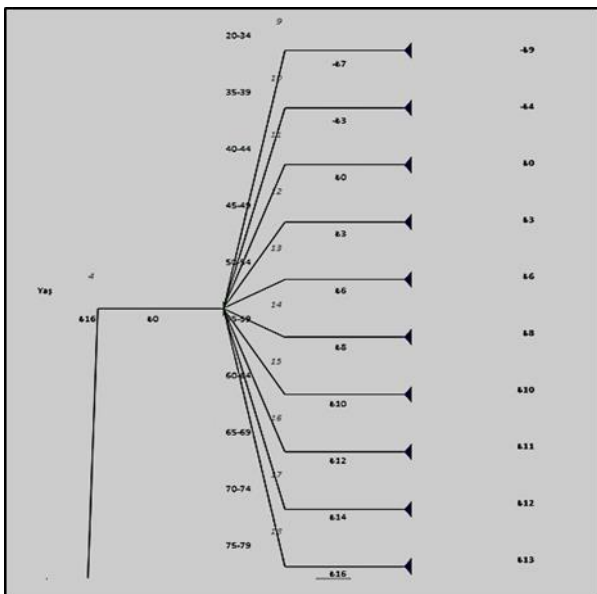
The relevant part for Blood Pressure is as given in Figure 5.

Figure 5. Blood pressure factor affecting heart attack in men

As seen in Figure 5, the risk of heart attack increases as blood pressure value/ tension value increases.

Considering age for women, the section in the regression tree is as given in Figure 6.
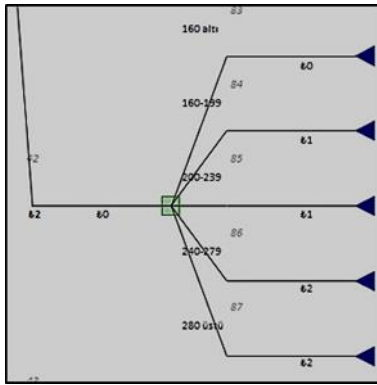

Figure 6. Age factor affecting heart attack in women

As seen in Figure 6 in scoring, it starts with -7 points between the ages 20-34 and ends with +16 points between the ages of 75-79. The risk of heart attack tends to increase with increasing age. The values of -7 and -3 used here also affect the 20-34 age group and 35-39 age group among other risk groups. For this reason, abnormalities in the values in the future risk calculations result from these -7 and -3 scores.

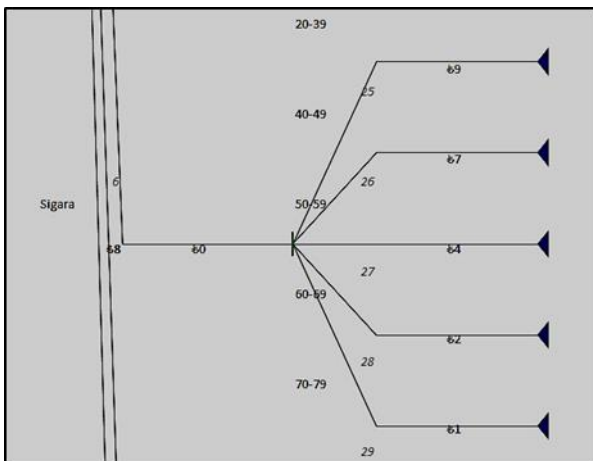The relevant part for total cholesterol (mg / dl) is as given in Figure 7.
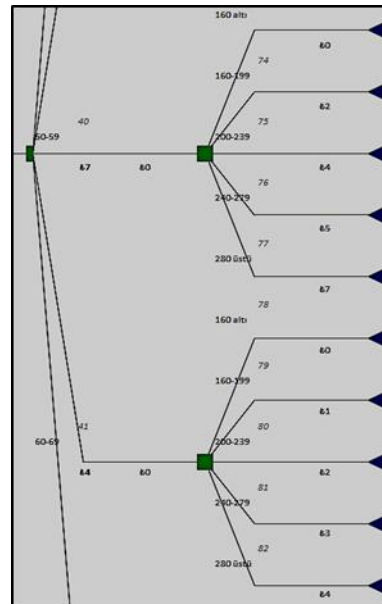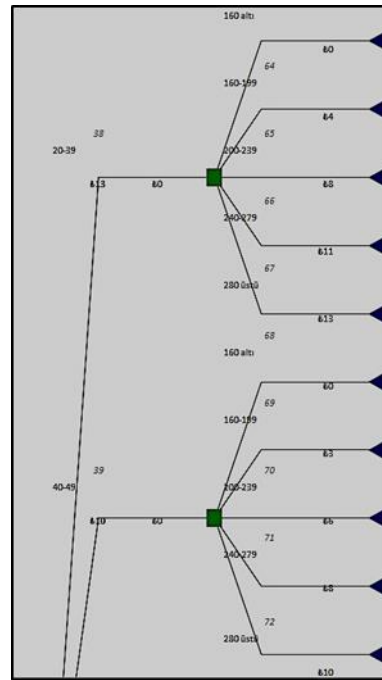
Figure 7. Total cholesterol factor affecting heart attack in women.

In the scoring, as seen in Figure 7, total cholesterol values affect each age group separately. It has an effect of minimum 0 maximum 13.
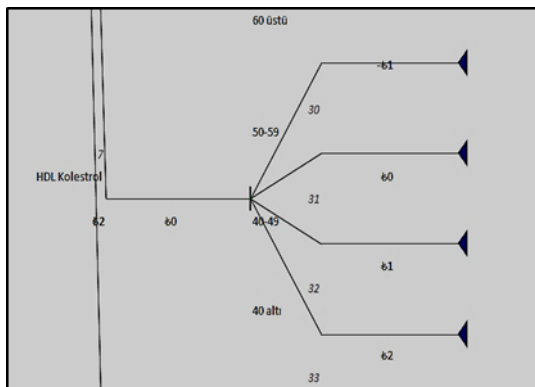
The relevant section for smoking is as given in Figure 8.

Figure 8. Smoking factor affecting heart attack in women.

As seen in Figure 8, cigarette smoking tends to increase with age (values of -7 and -3 in Figure 6 should not be forgotten).
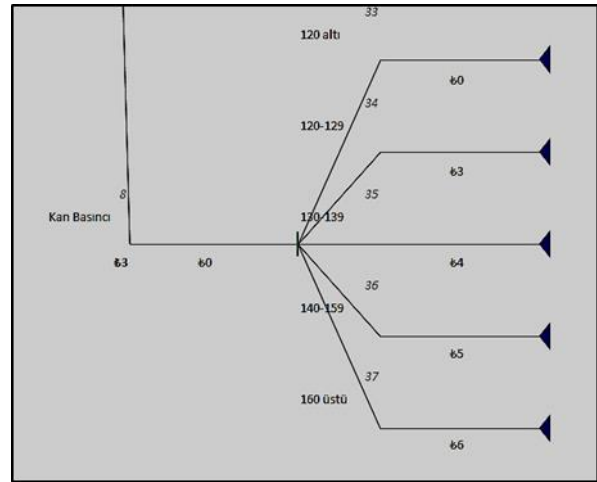
The relevant part for HDL cholesterol is as given in Figure 9.

Figure 9. HDL cholesterol factor affecting heart attack in women.

As seen in Figure 9, the risk of heart attack increases as HDL cholesterol decreases. The relevant part for blood pressure is as given in Figure 10.

Figure 10. Blood pressure factor affecting heart attack in women.

As seen in Figure 10, the risk of heart attack increases as blood pressure value/tension value increases.

3. Research-Results and Discussion

Framingham risk score reveals different risk results for men and women. As can be seen from Table 3, there are risk values ranging from 1% to 30%. These values vary according to scores between 0-17 and over in men, and scores between 0-25 and over in women. While the effects of the ages for men starts -9 points, ends with +13 points and for women starts -7 points, ends with +16 points. While the effects of the total cholesterol for men starts 0 points, ends with +11 points and for women starts 0 points, ends with +13 points. While the effects of the HDL cholesterol for men starts -1 points, ends with +2 points and for women starts -1 points, ends with +2 points. While the effects the blood pressure for men starts 0 points, ends with +3 points and for women starts 0 points, ends with +6 points. While the effects of the smoke cigarette for men starts +8 points, ends with +1 points and for women starts +9 points, ends with +1 points.

As a result of scoring the values that age factor, total cholesterol factor and smoke cigarette factor are important for heart attack risk.

As a result of scoring the values that possible patients will get from routine blood tests results and daily measurements, the risk factor of heart attack is revealed in percent. This system is also included in e-pulse application.

4. Conclusion

Considering the mentioned criteria, 10-year risk of women and men according to the total score as % emerges as given in Table 3. By summing the effects of all factors, the heart attack risk of the people was obtained in terms of percentage. While the score of men starts 0 and ends 17, the score of women starts 0 and ends 25. While value of 0-

4 corresponds to % 1 for men, 0-9 corresponds to % 1 over for women. While value of 17 and over corresponds to % 30 for men, 25 and over corresponds to % 30 over for women.

Table 3. Scoring Result Table

Men Score			Women Score	
Value	%		Value	%
0-4	1		0-9	1 lower
5-6	2		10-12	1
7	3		13-14	2
8	4		15	3
9	5		16	4
10	6		17	5
11	8		18	6
12	10		19	8
13	12		20	11
14	16		21	14
15	20		22	17
16	25		23	22
17 over	30		24	27
			25 over	30

In this study; Framingham risk score was used to calculate the risk of heart attack. Regression tree model was used in algorithm created for this. The effects of five factors on heart attack risk were evaluated. These factors can be multiplied and the risk of heart attack can be examined in more detail. Data warehouse was created, and algorithm implementation and evaluation steps were carried out. In addition, it has continued studies for cleaning, integration, reduction and conversion stages.

With this study, it is planned to create regional, city-city heart attack maps and create awareness of heart attacks and to reduce the probability of heart attacks to some extent.

Author's Note

Abstract version of this paper was presented at 9th International Conference on Advanced Technologies (ICAT'20), 10-12 August 2020, Istanbul, Turkey with the title of "Framingham Risk Score by Data Mining Method" [13].

References

[1] Babadağ, K. (2006). Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri. Industrial Application Software, 85-87

[2] Jacobs, P. (1999). Data Mining: What General Managers Need To Know. Harvard Management Update, 4 (10): 8.

[3] Alataş, B. ve Akın, E. (2004). Veri Madenciliğinde Yeni Yaklaşımlar. Ya/Em-2004- Yöneylem Araştırması/Endüstri Mühendisliği XXIV Ulusal Kongresi, 15-18 Haziran, Gaziantep-Adana.

[4] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. AKDDM, AAAI/MIT Press, pp. 1-30.

[5] Bırtıl, F. S. (2011). Kız Meslek Lisesi Öğrencilerinin Akademik Başarısızlık Nedenlerinin Veri Madenciliği Tekniği İle Analizi. Afyon Kocatepe Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi

[6] Ministry of Health. (2015). Türkiye Halk Sağlığı Kurumu, Türkiye Kalp ve Damar Hastalıkları Önleme ve Kontrol Programı Eylem Planı 2015-2020, Ankara 2015.

[7] WHO. (2014). Global Status Report on Noncommunicable Diseases. https://www.who.int/nmh/publications/ncd-status-report-2014/en/ (Erişim tarihi: Temmuz 2019)

[8] Kültürsay, H. (2011). Kardiyovasküler Hastalık Riski Hesaplama Yöntemleri. Türk Kardiyol Dern Arş - Arch Turk Soc Cardiol 2011, 39 Suppl 4:6-13 doi: 10.5543/tkda.2011.kultursay

[9] TUİK Sayı: 27620. (Erişim tarihi: Temmuz 2019) http://www.tuik.gov.tr/PreHaberBultenleri.do?id=27620

[10] Tekkeşin, N., Kılınç, C. ve Ökmen, A.Ş. (2011). Türk Erişkinlerde Framingham Risk Faktörlerinin Araştırılması. Klinik ve Deneysel Araştırmalar Dergisi / 2011; 2 (1): 42-49

[11] Framingham Risk Skorlaması. (2020). http://www.hipertansiyonmd.com/hesaplama/hesaplama_Framingham_risk_skorlamasi.htm (Erişim Tarihi: Şubat 2020)

[12] Peter, W. F. W., Ralph, B. D., Daniel L., Albert M. B., Halit S. ve William B. K. (1998). Prediction of Coronary Heart Disease Using Risk Factor Categories, Circulation, Vol. 97, No. 18, 12 May 1998, 1998;97:1837–1847 https://doi.org/10.1161/01.CIR.97.18.1837.

[13] Kitiş, Ş. "Framingham Risk Score By Data Mining Method", 9th International Conference on Advanced Technologies 2020, Online İstanbul, 10-12 August 2020, ISBN:978-625-44427-0-4.