Kalp Sağlığı Yardımcı Asistanı

Talha Bacak
Computer Engineering

Yildiz Technical University
talhadhh@gmail.com

***Özet*—Bu çalışmada veri madenciliği için sınıflandırma algoritmalarından logistic regression, decision tree, support vector machine ve random forest kullanılmış ve bunlar karşılaştırılmıştır.**

Keywords—decision tree, logistic regression, support vector machine, random forest, data mining, cardiovascular-diseases-analysis

I. GİRİŞ

Bu çalışmada kardiyovasküler hastalık ihtimalinin veri analiz yöntemleriyle tahmin edilmesi öngörülmüştür. Bu yöntemler kullanılarak doktorlara yardımcı olabilecek asistan yapıldı. Günden güne artan kalp hastalıkları riskinin olasılığını gösteren kullanıcı ara yüzü oluşturuldu. Risk analizi için doktorlara fikir verecektir.

II. VERI KEŞFI

Bu çalışmada kullanılan veri kaggle'dan [1] elde edilmiştir. Veri 12 özellikten oluşmaktadır. Bunlar:

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm)
3. Weight | Objective Feature | weight | float (kg)
4. Gender | Objective Feature | gender | categorical code
5. Systolic blood pressure | Examination Feature | ap_hi | int
6. Diastolic blood pressure | Examination Feature | ap_lo | int
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal
9. Smoking | Subjective Feature | smoke | binary
10. Alcohol intake | Subjective Feature | alco | binary
11. Physical activity | Subjective Feature | active | binary
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary

Bu özelliklerde bazı değişiklikler yapıldı. Bunlar:

- Age özelliği gün olarak belirtilmiş, bu değer yıla çevrildi.
- Gender özelliği 1 = kadın, 2 = erkek olarak verilmiş. Bu değerlerden 1 çıkartılarak 0 = kadın, 1 = erkek olarak değiştirildi.
- Weight, Systolic blood pressure ve Diastolic blood pressure özelliklerinde mantık dışı olarak verilen datalar temizlendi.

Veriler düzenlendikten sonra veri setinin boyutu 69977 olmuştur.

Özelliklerin histogramları ise şu şekildedir:

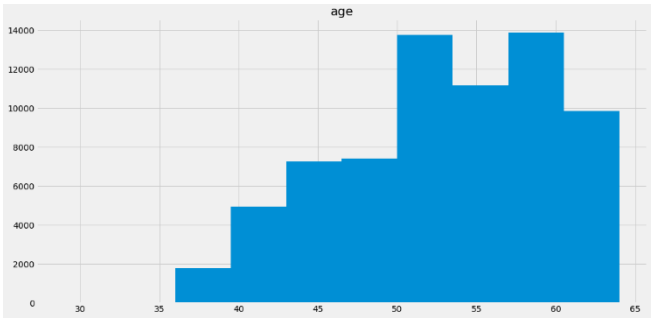

Fig. 1. Age

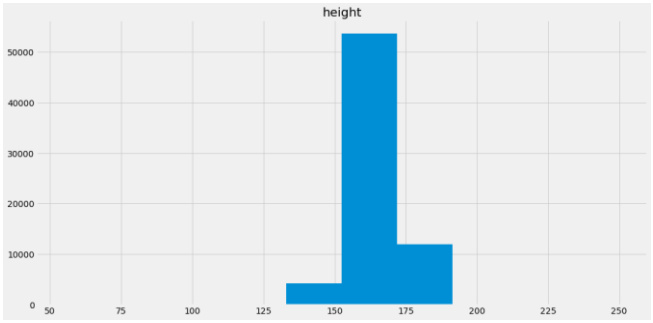

Fig. 2. Height

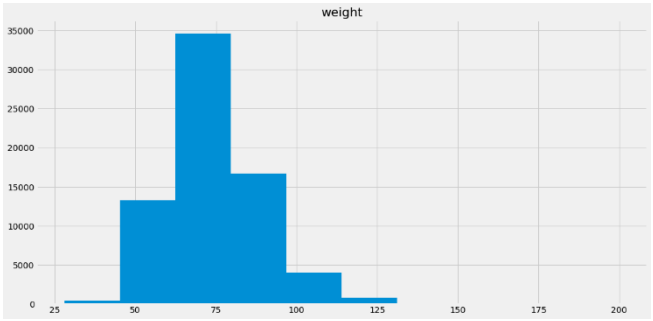

Fig. 3. Weight

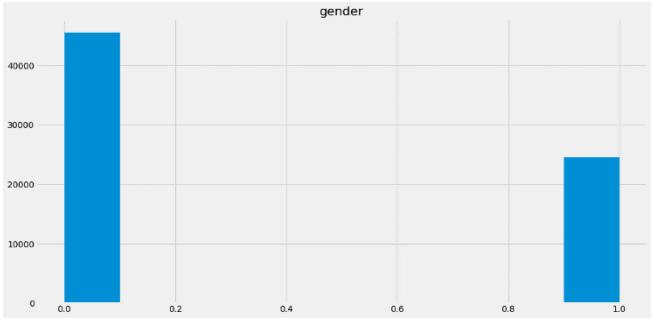

Fig. 4. Gender

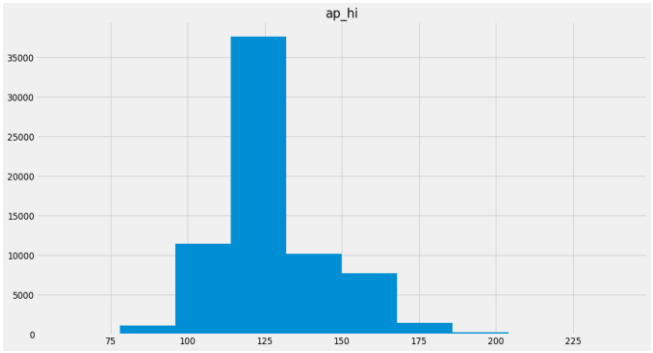

Fig. 5. Systolic blood pressure

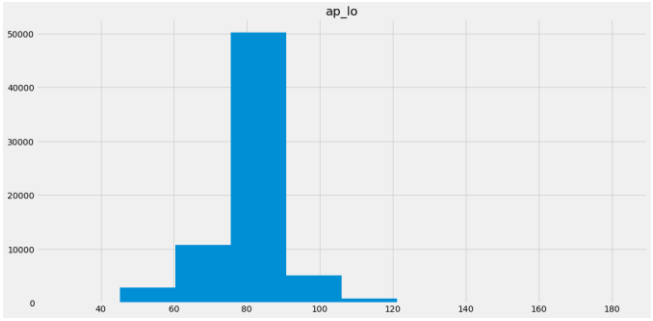

Fig. 6. Diastolic blood pressure

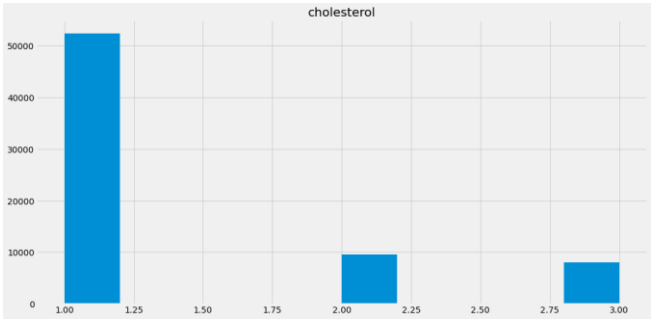

Fig. 7. Cholesterol

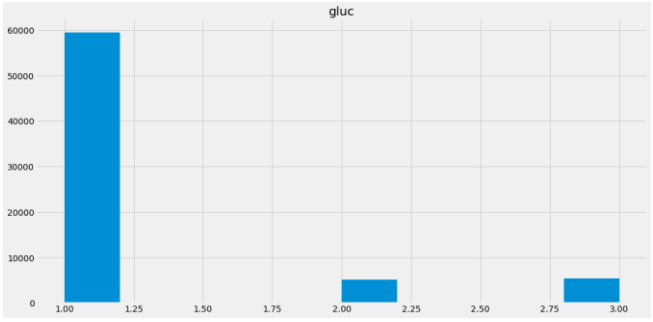

Fig. 8. Glucose

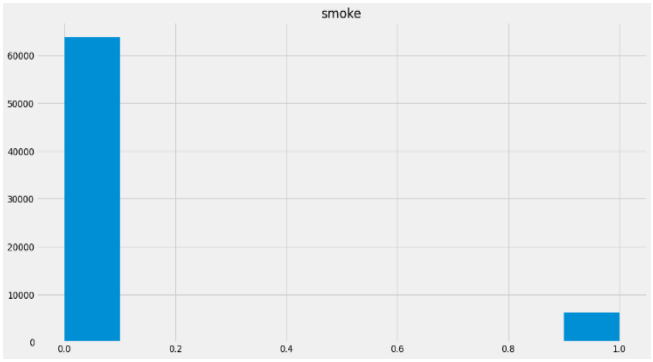

Fig. 9. Smoking

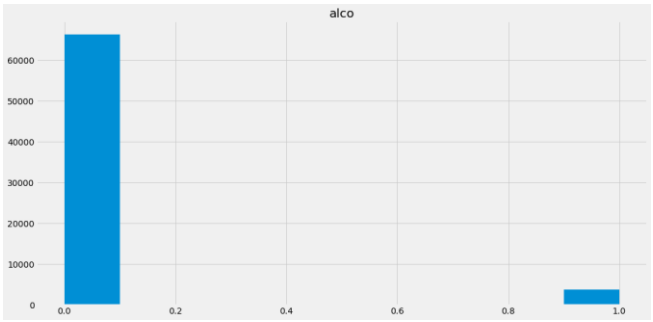

Fig. 10. Alcohol intake

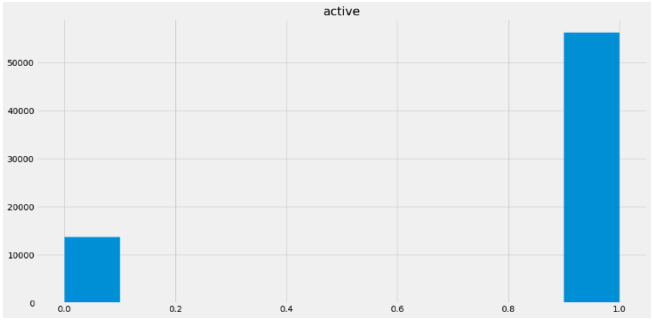

Fig. 11. Physical activity

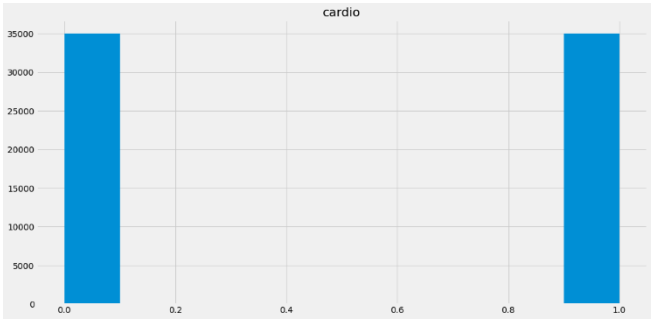

Fig. 12. Cardio

Bu özelliklerin korelasyonu ise şu şekildedir:

Fig. 13. Korelasyon

Korelasyon değerlerine baktığımızda kardiyovasküler hastalıkla en fazla ilişkili olandan en az ilişkili olana özellikler şu sıra ile tespit edildi:

1. Yüksek tansiyon (Çok Yüksek)
2. Düşük tansiyon (Çok Yüksek)
3. Yaş (Yüksek)
4. Kolesterol (Yüksek)
5. Kilo (Alakalı)
6. Glukoz (Orta üstü)
7. Fiziksel aktivite (Orta)
8. Sigara (Az)
9. Boy (Neredeyse alakasız)
10. Alkol (Neredeyse alakasız)
11. Cinsiyet (Neredeyse alakasız)

Korelasyon değerlerine baktığımızda cinsiyet, boy, alkol değerlerinin ilgisi hemen hemen bulunmadığı ve sigaranın ilgisinin çok az olduğu tespit edilmiştir. Sigara ve alkolün etkisinin bu kadar az olmasının sebebi olarak veride sigara ve alkol kullanan insan oranının çok az olduğu varsayılabilir.

III. Model

Veri setine bakarak geliştirilecek modellerin classification olduğu tespit edilerek buna uygun modeller geliştirilmiştir. Bu modeller logistic regression, decision tree, random forest ve support vector machine olarak karar verildi. Logistic regression maksimum iterasyon sayısı 1000 olarak belirlendi. Decision tree maksimum ağaç derinliği 3 olarak belirlendi ve entropy ile gini karşılaştırlmasında entropy daha başarılı sonuçlar verdiği için entropy seçildi. Random forest modelinde ise 100 ağaç kullanıldı. Support vector machine modelinde linear kernel tercih edildi.

PCA yöntemi kullanılarak daha doğru sonuçlar elde edilmiştir.

IV. Sonuçlar

Modellerin eğitilme zamanına bakıldığında ciddi farklar olduğu gözlemlenmiştir. Fig. 14'e bakıldığında support vector machine algoritmasının çok uzun sürdüğü ve başarı olarak Linear regression ile aynı olduğu görülmüştür. Decision tree ve logistic regression en hızlı çalışan algoritmalar olmuştur. Random Forest algoritması ise support vector machine algoritmasından çok hızlı olmasına rağmen logistic regression ve decision tree'ye göre yavaş çalıştığı görülmektedir.

F1 skorlarına baktığımızda (fig. 14) ise yaklaşık benzer sonuçlar göstermişlerdir. En başarılı sonuçları logistic regresiion ve support vector machine algoritmaları sağlamıştır ve % 73'lük bir başarı elde edildi. En düşük başarı ise decision tree algoritması ile % 71 olarak elde edildi. Random forest'ta ise % 72'lik bir skor elde edilmiştir.

Veri seti büyüklüğü 69977'dir. Veri setinin % 20'si test verisi, % 80'i train verisi olarak ayrılmıştır.

Model Sonuçları

Model	F1 Score	Zaman (sn)
Logistic Regression	0.73	0.42
Decision Tree	0.71	0.34
Random Forest	0.72	57.66
SVM	0.73	8369.82

Fig. 14. Model karşılaştırmaları

V. Uygulama

Bu algoritmaların kullanımı için PyQt5 ile ara yüz geliştirilmiştir. Bu ara yüz aşağıdaki şekildedir:

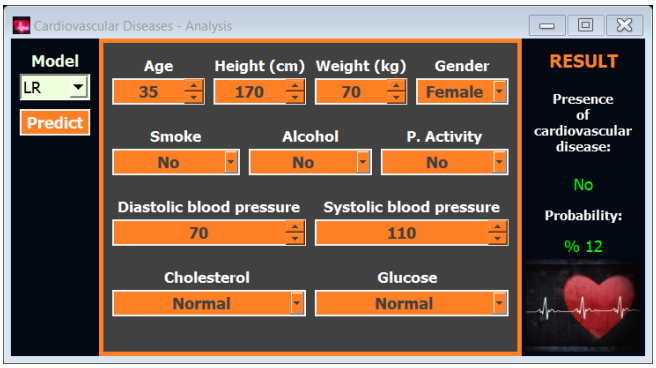

Fig. 15. GUI (1)

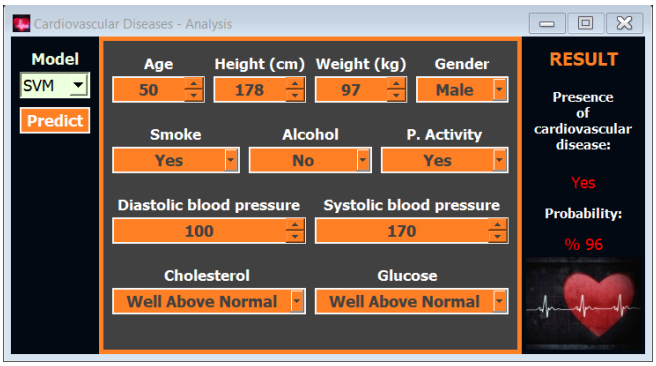

Fig. 16. GUI (2)

VI. İLIŞKILI ÇALIŞMALAR

GuangliNie ve arkadaşları kayıp kredi kartlarının tespiti için decision tree ve logistic regression modelleri ile yaptığı çalışmalarda genel olarak logistic regression modellerinin daha doğru çalıştığını tespit etmişlerdir. Bu çalışmada çoklu logistic regression modellerinin 2'li modellere göre daha başarılı olduğu fark edilmiştir. Çalışmalarında error değerlerine daha fazla ağırlık vermişler ve bankalar için anomali tespit modeli oluşturmuşlardır. [2]

VII. ÇALIŞMANIN KODLARI

Bu çalışma esnasında kullanılan bazı kodlara aşağıdaki Github linki üzerinden erişebilirsiniz:

https://github.com/talhabacak/4.class/tree/main/Data_Mining/project

REFERENCES

[1] https://www.kaggle.com/code/ntltam/cardiovascular-diseases-analysis-w-data-mining/data

[2] Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, Yong Shi, Credit card churn forecasting by logistic regression and decision tree, Expert Systems with Applications, Volume 38, Issue 12, 2011, Pages 15273-15285, ISSN 0957-4174