
# Davranışsal Risk Faktörleri İzleme Sistemi'nin 2023 Yılına Ait Verilerinden Kalp Krizi Tahmini

*İlayda Ayvat, Ravi Memmedov, Elif Dilşah Bahçeci, Burak Dal, Büşra Küçük*

# 📖 *Veri Seti Hikayesi & Problem Tanımı*

## Veri Seti Hikayesi

2023 yılına ait Davranışsal Risk Faktörleri İzleme Sistemi (BRFSS) veri seti, Amerika Birleşik Devletleri'nde yetişkin nüfusun sağlıkla ilgili davranışlarını ve risk faktörlerini izlemek amacıyla yapılan en kapsamlı telefon anketlerinden biridir.

- **Toplanan Veriler:** Anket, sağlık durumu, egzersiz alışkanlıkları, hipertansiyon, kolesterol, kronik sağlık koşulları, demografik bilgiler, engellilik durumu, düşmeler, tütün ve alkol kullanımı, bağışıklama, HIV/AIDS, emniyet kemeri kullanımı, araç kullanırken alkol alma ve uzun süreli COVID etkileri gibi konuları kapsamaktadır.

## Problem Tanımı

Kalp krizi, dünya genelinde en yaygın ve ölümcül sağlık sorunlarından biridir. Erken teşhis ve risk altındaki bireylerin belirlenmesi, bu hastalığın önlenmesinde kritik bir rol oynamaktadır. Bu çalışmanın amacı, **Davranışsal Risk Faktörleri İzleme Sistemi (BRFSS) 2023 verileri** kullanılarak bireylerin kalp krizi geçirme riskinin tahmin edilmesidir.

Makine öğrenmesi yöntemleriyle, bireylerin demografik bilgileri, sağlık durumu ve yaşam tarzı alışkanlıklarına dayalı olarak kalp krizi riski öngörülmeye çalışılacaktır. Bu doğrultuda geliştirilecek model, sağlık politikaları oluşturulurken önleyici müdahalelerde rehber olabilecek potansiyele sahiptir.

# 🔍 *Keşifsel Veri Analizi*

**Çalışmamıza temel oluşturan veri seti, 433.323 gözlem ve 350 değişkenden oluşmaktadır. Anket yanıtları sayısal olarak kodlandığı için veri seti tamamen sayısal değişkenlerden oluşmaktadır.**

Bazı sorulara yanıt verilmemesi veya katılımcının cevabı bilmemesi durumunda değerler 7-9, 77-99 gibi sayılarla kodlanmıştır. Bu tür kodlamalar, veri setinde aykırı değerlerin ortaya çıkmasına neden olabilmektedir.

Ayrıca, bazı eksik değerlerin aslında belirli durumları temsil ettiği gözlemlenmiştir. Örneğin, diyabet hastası olmayan bireylerin ilaç kullanımıyla ilgili verileri eksik (missing) olarak kodlanmıştır. Bu tür eksiklikler, bilinçli ve bağlama özgü eksik veri olarak değerlendirilebilir.

# 🔧 *Veri Ön İşleme ve Özellik Mühendisliği*

## Veri Ön İşleme

Veri ön işleme sürecinde;

- Görüşmeyi tamamlamayan ve 18 yaş altı katılımcıların gözlemleri,
- Analiz açısından anlamlı bilgi taşımayan bazı değişkenler,
- "_" karakteriyle türetilmiş olan değişkenler,
- %60'ın üzerinde eksik veriye sahip değişkenler analiz dışı bırakılmıştır.

Bazı değişkenlerde "bilmiyorum", "cevap vermek istemiyorum" gibi anlamlara gelen 7–9, 77–99, 777–999 gibi kodlamalar analiz açısından geçerli veri sunmadığı için temizlenmiştir.

> ### Neden 18 Yaşının Altındaki Gözlemleri Çıkarttık?
>
> Bu çalışmada kalp krizi (miyokard enfarktüsü) tahmini yapılmaktadır. Kalp krizi, genellikle yetişkin bireylerde görülen bir sağlık sorunudur ve 18 yaş altı bireylerde çok nadiren rastlanır. Ayrıca, çocuklar ve ergenler için kalp-damar sistemi, risk faktörleri ve hastalık gelişim süreçleri yetişkinlerden önemli ölçüde farklıdır. Bu nedenle, istatistiksel analizlerin doğruluğunu artırmak ve modelin genelleme yeteneğini iyileştirmek amacıyla, 18 yaş altındaki gözlemler veri setinden çıkarılmıştır.

🔪 **Birim Tutarsızlıklarının Giderilmesi:**

Veri setinde "Kilo", "Boy" ve "Aktivite Sıklığı" gibi bazı değişkenlerde birim tutarsızlıkları tespit edilmiştir. Bu durum, analiz sonuçlarının güvenilirliğini olumsuz etkileyebileceğinden, ilgili değişkenler *codebook* referans alınarak standart birimlere dönüştürülmüştür.

### ❓ Eksik Değerlerin İşlenmesi:

*Codebook* referans alınarak, belirli durumları temsil eden eksik veriler anlamlı karşılıklarıyla yeniden kodlanmıştır.

**missing_report**

| Aa Name | # Missing Percentage | # Missing Count |
|---|---|---|
| EXEROFT1 | 23.822% | 59535 |
| EXERHMM1 | 23.823% | 59537 |
| BPMEDS1 | 57.279% | 143149 |
| TOLDHI3 | 5.357% | 13387 |
| CHOLMED3 | 5.777% | 14438 |
| HEIGHT3 | 0% | 1 |
| ALCDAY4 | 0% | 1 |
| AVEDRNK3 | 45.525% | 113773 |
| SHINGLE2 | 31.475% | 78660 |
| COVIDSM1 | 48.846% | 122073 |
| RACE | 0.019% | 47 |

**gerekli atamalardan sonra missing_report**

| Aa Name | # Missing Percentage | # Missing Count |
|---|---|---|
| CHOLMED3 | 0.421% | 1051 |
| EXRACTM | 23.823% | 1407 |

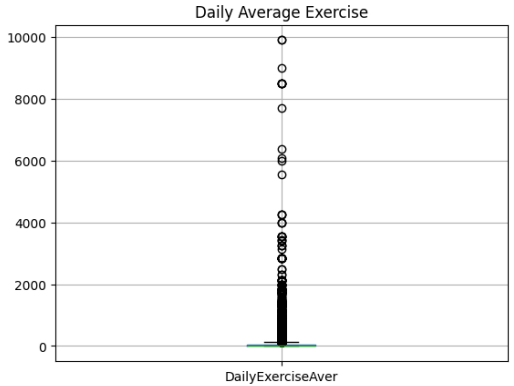
Eksik değerlere yönelik bu işlemler sonrasında, `CHOLMED3` ve `EXRACTM` değişkenlerinde *codebook* tarafından tanımlanmamış, dolayısıyla anlamı belirsiz olan bazı değerlerin veri setinde yer aldığı görülmüştür. Analizin doğruluğunu korumak amacıyla bu değişkenlerdeki eksik gözlemler veri setinden çıkarılmıştır.

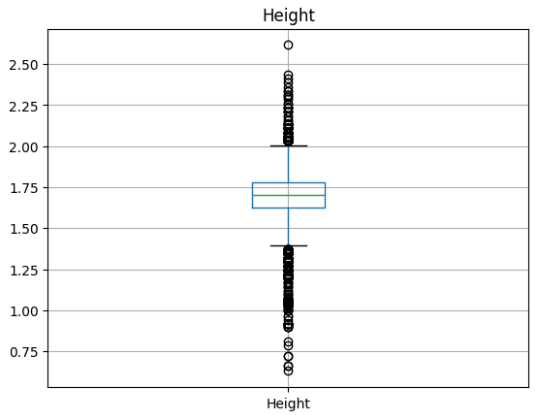
### 📊 Kategorik ve Sayısal Değişkenlerin Yakalanması:

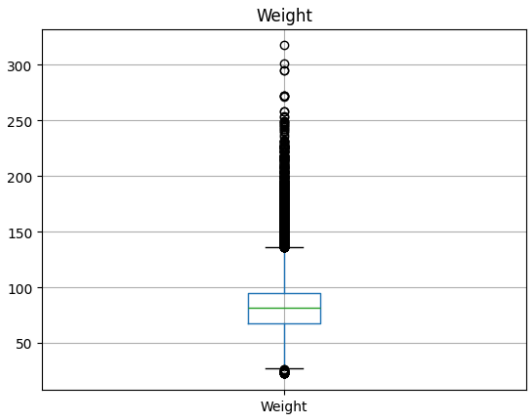
Öncelikle okunabilirliği artırmak amacıyla değişken isimleri düzenlenmiştir. Ardından veri setindeki değişkenler kategorik ve numerik olarak sınıflandırılmıştır.

Observation: 247463
Variables: 36
cat_cols: 30
num_cols: 6

### 📊Sayısal Değişkenlerin Analizi:




| | count | mean | std | min | 5% | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 95% | 99% | max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PhysicalHealth | 242438.000 | 4.334 | 8.643 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 2.000 | 5.000 | 19.000 | 30.000 | 30.000 | 30.000 |
| MentalHealth | 242438.000 | 4.059 | 7.976 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 2.000 | 5.000 | 15.000 | 30.000 | 30.000 | 30.000 |
| AlcoholUsingAver | 242438.000 | 1.144 | 1.812 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 | 2.000 | 3.000 | 4.000 | 8.000 | 76.000 |
| Height | 242438.000 | 1.702 | 0.104 | 1.400 | 1.549 | 1.575 | 1.600 | 1.626 | 1.676 | 1.702 | 1.727 | 1.753 | 1.803 | 1.829 | 1.880 | 1.930 | 2.083 |
| Weight | 242438.000 | 83.444 | 21.180 | 40.000 | 54.431 | 58.967 | 65.771 | 70.760 | 76.204 | 81.647 | 86.183 | 90.718 | 99.790 | 111.130 | 122.470 | 147.418 | 317.515 |
| DailyExerciseAver | 242438.000 | 38.199 | 52.178 | 0.000 | 0.000 | 0.000 | 0.000 | 6.430 | 12.860 | 19.290 | 28.570 | 42.860 | 65.710 | 100.000 | 142.860 | 257.140 | 300.000 |

## Özellik Mühendisliği

❓ **Aykırı Değerler:**

“Height”, “Weight” ve DailyExerciseAver” sayısal değişkenlerinde bulunan aykırı değerler literatür araştırması yaparak veri setinden çıkartıldı.

➕ **Yeni değişkenler oluşturduk:**

- BMI_Category
- BMI_
- Cronic_Count
- Vaccine_Score
- Health_Tracking
- Drug_Tracking
- Old_Risk
- Nicotin_Score
- Risk_Score
- Risk_Level

🔢 **Encoding işlemleri:**

İki kategorili değişkenler için “Label Encoder”, daha fazla kategorili değişkenler için “On-Hot Encoder” uyguladık.

# 🤖 *Modelleme*

### 🧠 Model Seçimi ve Tercih Nedeni

Kalp krizi riskini tahmin edebilmek için birden fazla makine öğrenmesi algoritması denenmiştir. Bu süreçte **CatBoost**, **XGBoost** ve **LightGBM** gibi gelişmiş sınıflandırma modelleri değerlendirilmiştir.

Problemimizin doğası gereği, yüksek doğruluk oranı ve işlem verimliliği sunan **LightGBM** algoritması tercih edilmiştir. LightGBM, karar ağaçlarına dayalı, özellikle büyük veri setlerinde hızlı ve etkili çalışabilen bir sınıflandırma yöntemidir.

Veri setimizin yüz binlerce satırdan oluşması nedeniyle, modelin işlem süresi ve bellek kullanımı kritik öneme sahiptir. LightGBM bu noktada sağladığı hız ve performans ile en uygun model olarak öne çıkmıştır.

### ⚖️ Sınıf Dengesizliği Problemi ile Nasıl Başa Çıktık?

Veri setimizde **kalp krizi geçiren bireyler**, toplam örneklem içinde oldukça düşük bir orana sahipti. Bu durum, modelin azınlık sınıfı doğru tahmin etmekte zorlanmasına ve bazı performans metriklerinin düşük çıkmasına neden oldu.

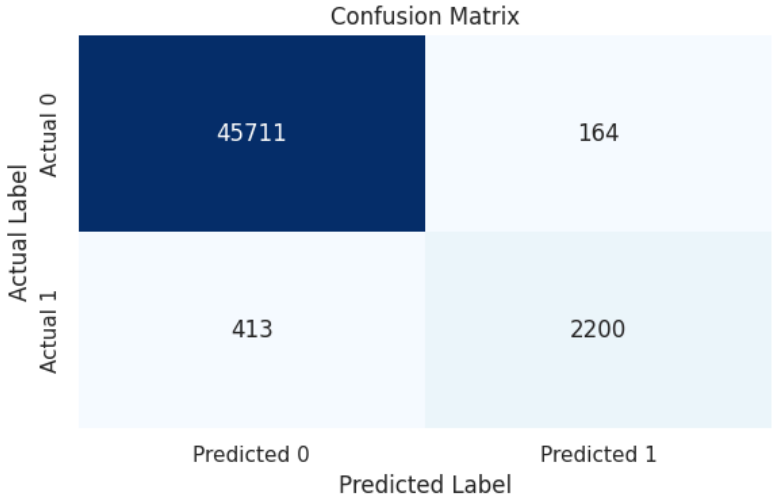
Bu sorunu çözmek için **SMOTE (Synthetic Minority Over-sampling Technique)** yöntemi kullanıldı. SMOTE, azınlık sınıfına ait örneklerin sayısını artırarak veri kümesinde daha dengeli bir sınıf dağılımı sağlar. Böylece model, her iki sınıfı da daha iyi öğrenme şansı elde eder.

SMOTE uygulandıktan sonra, modelin performansında özellikle **recall** ve **F1-score** gibi metriklerde anlamlı iyileşmeler gözlemlenmiştir.

# ✅ Bulgular & İş Önerileri

## Bulgular

### 📊 Model Başarı Değerlendirme: Classification Report & Confusion Matrix




```
Classification Report:

              precision    recall  f1-score   support
Class 0           0.990     1.000     0.990 45875.000
Class 1           0.930     0.840     0.880  2613.000
accuracy          0.990     0.990     0.990     0.990
macro avg         0.960     0.920     0.940 48488.000
weighted avg      0.990     0.990     0.990 48488.000
```

```
best_params = {
    'learning_rate': 0.05,
    'max_depth': 7,
    'n_estimators': 500,
    "subsample": 0.8,
    "colsample_bytree": 0.8}
```

### Kalp krizi olmama durumu:

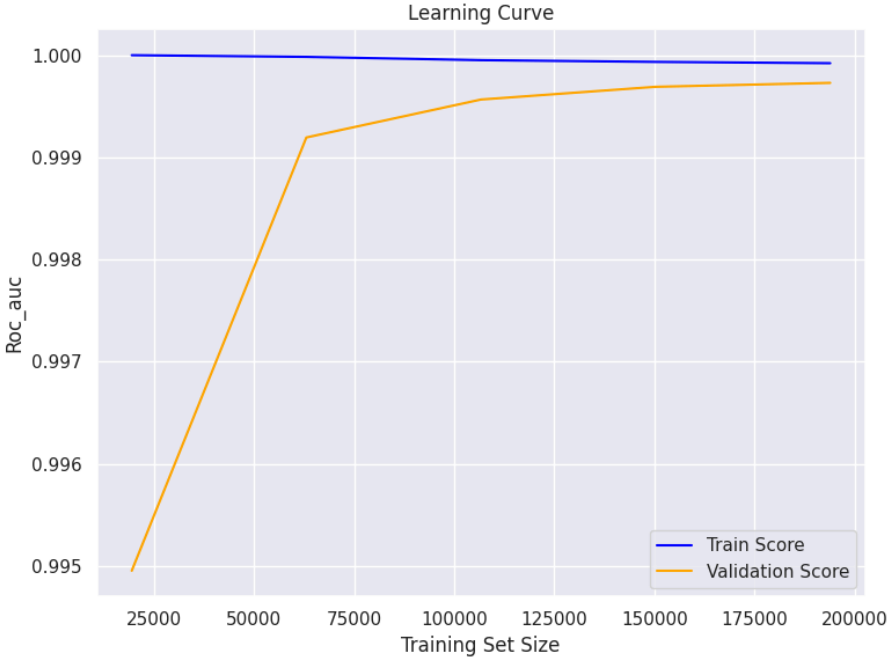
- **Precision:** 0.990 → Modelin "0" dediği örneklerin %99'u gerçekten "0".
- **Recall:** 1.000 → Gerçekten "0" olan tüm örneklerin %100'ü doğru tahmin edilmiş.
- **F1-Score:** 0.990 → Precision ve recall'un harmonik ortalaması. Gayet yüksek.

### Kalp krizi olma durumu:

- **Precision:** 0.930 → Modelin "1" dediği örneklerin %93'ü gerçekten "1".
- **Recall:** 0.840 → Gerçek "1" örneklerinin sadece %84'ü doğru şekilde tahmin edilmiş.
- **F1-Score:** 0.880 → Daha düşük ama hâlâ iyi seviyede.

Model, yukarıdaki parametreler ile başarılı olmuş sayılabilir. Ancak, **Sınıf 1'i tespit etmede biraz gelişmeye ihtiyaç vardır.**

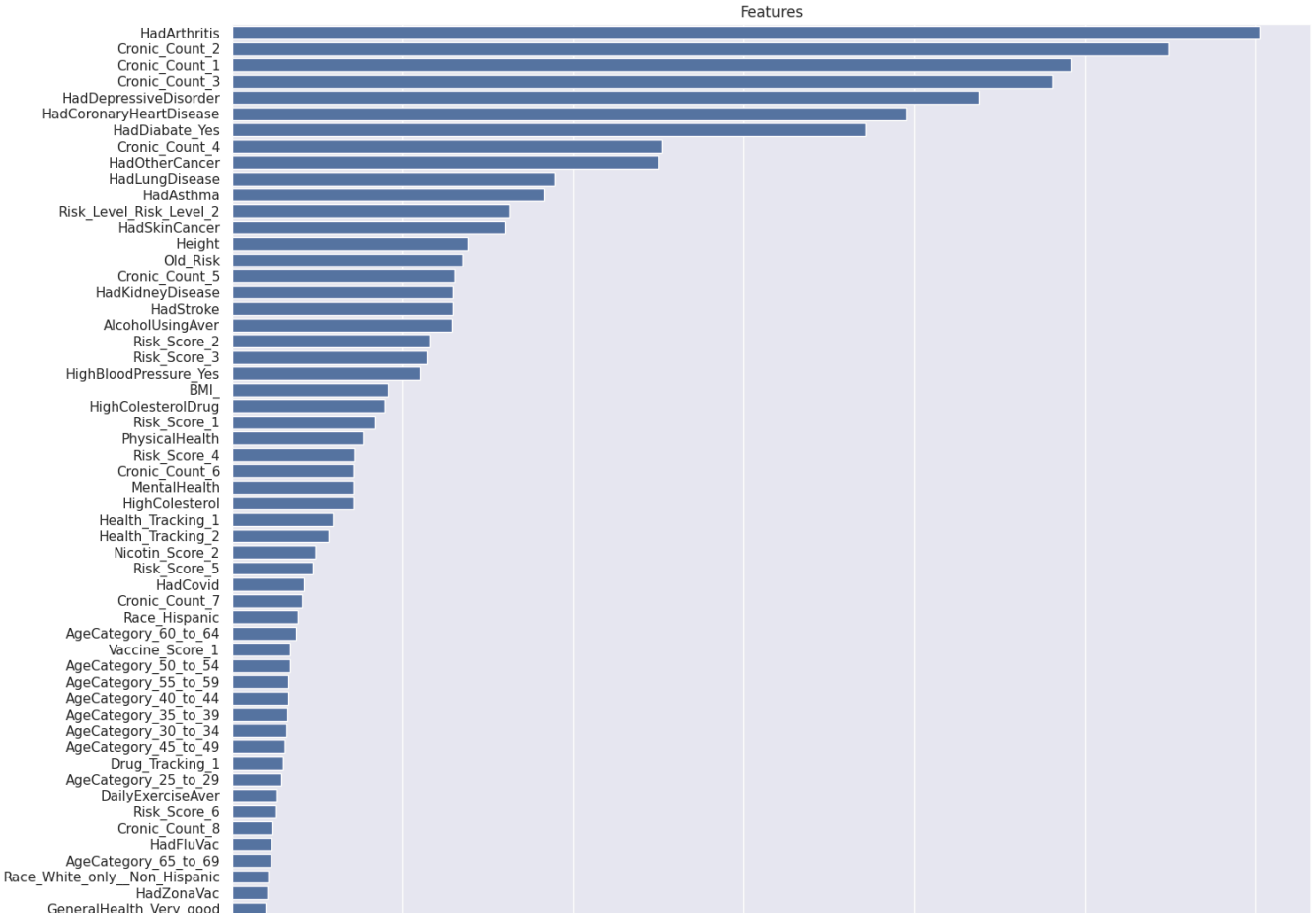
## 📊 *Öğrenme Eğrileri*





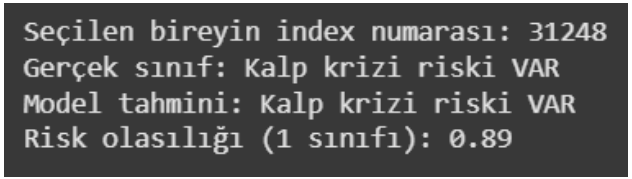
Modelin öğrenme eğrileri incelendiğinde, **aşırı öğrenme (overfitting)** belirtilerine rastlanmamıştır. Eğitim ve doğrulama hatalarının benzer seviyelerde seyretmesi, modelin veriye iyi genelleme yapabildiğini göstermektedir.

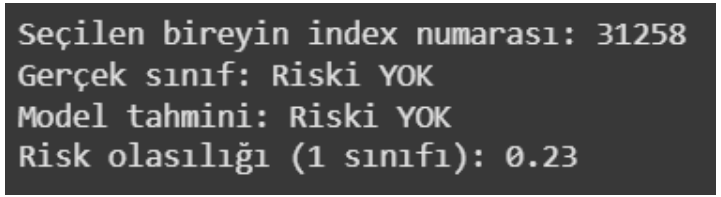
## 📊 *Model Başarı Değerlendirme: Feature Importance*




Modelin değişken önem grafiği incelendiğinde, **özellik mühendisliği sürecinde oluşturulan yeni değişkenlerin**, modelin performansına anlamlı katkı sağladığı görülmektedir. Bu yeni değişkenler, modelin öğrenme kapasitesini artırarak tahmin doğruluğunu olumlu yönde etkilemiştir.

### ✅Model Testi




Bu aşamada, modelin gerçek dünyadaki performansını gözlemlemek amacıyla **rastgele seçilen örnek gözlemler** üzerinde testler gerçekleştirilmiştir. Görsellerde görüldüğü üzere, model tahminleri ile gerçek değerler karşılaştırılarak modelin sınıflandırma başarısı değerlendirilmektedir.

Bu testler, modelin bireysel bazda ne kadar doğru tahmin yapabildiğini ve genelleme yeteneğini gözlemlemek açısından önemli bir adımdır.

### 📌Alternatif Yaklaşım Önerisi

Modelin daha başarılı ve gerçekçi sonuçlar verebilmesi için, **SMOTE** yerine alternatif bir veri dengeleme yöntemi uygulanabilir. Bu doğrultuda, aynı veri tabanının diğer yıllarına ait kayıtlar kullanılarak, **kalp krizi geçiren ve geçirmeyen bireylerden dengeli sayıda gözlem** içeren yeni bir veri seti oluşturulabilir. Böylece, sentetik veri üretimi yerine gerçek gözlemlerle çalışılarak modelin genelleme yeteneği artırılabilir.

## İş Önerileri

### 💻*Dijital Müdahale Sistemleri:*

Geliştirilen model, dijital bir sağlık platformuna entegre edilerek bireylerin kendi risk düzeylerini hesaplamalarına olanak tanıyabilir. Vatandaşlar, kişisel verilerini sisteme girerek **kişiselleştirilmiş sağlık tavsiyeleri** alabilir.

Ayrıca, model **erken uyarı sistemi** olarak da kullanılabilir. Yüksek riskli bireyler tespit edilerek, aile hekimlerine önceden bilgi verilebilir. Böylece, birey henüz hastalık belirtisi göstermeden **önleyici sağlık hizmetleri** devreye alınabilir.