KÜTAHYA
SAĞLIK BİLİMLERİ
ÜNİVERSİTESİ

Tip 2 Diyabet Teşhisi

İlyas Cemal Erginli

Ocak 2025

Özet

Bu projede, diyabet teşhisini desteklemek amacıyla LiNGAM algoritması ile veri setindeki nedensel ilişkiler analiz edilmiş ve bu ilişkiler bir nedensel grafikle görselleştirilmiştir. Elde edilen analiz sonuçları, lojistik regresyon modeliyle tahmin sürecine entegre edilmiştir. Örnek insanlar yaratılarak test edilmiş ve doğruluk oranı hesaplanmıştır.

1 Hipoglisemi

Hipoglisemi[1], kan şekerinin olması gerektiğinden daha düşük olması durumudur. Hipoglisemik reaksiyonun başlangıcında bulanık görme, baş ağrısı, baş dönmesi, terleme ve bayılma hissi görülür. Uygun müdahale edilmediği takdirde bilinç kaybı yaşanır. Diyabetli hastaların sıklıkla karşılaştığı bir rahatsızlıktır. Vücudun ihtiyaç duyduğu insülin alınan gıda miktarına, yenen yemeğin çeşidine, ne kadar egzersiz yapıldığına, insülin enjekte edilen bölgeye, vücutta başka hastalık olup olmamasına ve içinde bulunulan stres oranına bağlıdır.

Diyabetli kişilerde hipoglisemiye neden olabilecek durumlar[2]:

- Diyabet İlaçlarını Kullanmada Yapılan Hatalar: Diyabet tedavisinde kullanılan insülinin aşırı doz alımı veya yanlış kullanımı veya insülini yağ yerine kana enjekte etmek.
- Yemek Düzenindeki Değişiklikler: Karbonhidrat alımını doğru sınırlandırmamak, gerektiğinden daha az almak. Özellikle egzersizden önce veya egzersiz sırasında yeteri kadar karbonhidrat tüketmemek. Öğünleri atlamak, geciktirmek veya öğünlerde normalden daha az yemek de ani şeker düşmesi belirtilerine neden olabilmektedir. Yemek yemeden alkol tüketmek de diğer riskli davranışlardandır.
- Egzersiz: Kan şekerinin düşmesinde yapılan egzersizler de etkili olabiliyor. Diyabet ilaçları kullanan kişiler, egzersiz yaparken kan şekeri düzeylerini kontrol etmeli ve gerekirse atıştırmalık tüketmelidirler. Aşırı düzeyde egzersiz yapmak da nedenlerinden biridir.
- Alkol: Karaciğerin glikoz üretmesi ve salgılamasında alkol etkili olmaktadır. Bu yüzden kan şekerinin düşmesinde etkisi olabilir. Diyabetli kişilerin alkol tüketmesi riskli olabilmektedir.

Hipoglisemi[3] yaşanması durumunda hızlıca şekerli bir yiyecek tüketilmelidir; aksi takdirde bilinç kaybı yaşanabilir.

Hipoglisemiyi düzeltmeye yardımcı olacak bir şeyi her zaman diyabet hastalarının yanında bulundurması önemlidir. Meyve suyu, meşrubat, tatlılar veya glukoz tabletleri gibi hızla emilen ve 15 gram karbonhidrat içeren ürünlerin tüketilmesi önerilir.

2 Diyabet Şeker Ölçüm Cihazları

Diyabet şeker ölçüm cihazları[4], diyabet hastalarının kan şeker seviyelerini düzenli olarak takip etmelerini sağlayan teknolojik cihazlardır. Kan şekeri (glukoz) seviyelerinin sürekli kontrol edilmesi, diyabetin yönetimi ve komplikasyonların önlenmesi açısından kritik öneme sahiptir. Geleneksel olarak, glukoz seviyesi parmak delme yöntemiyle alınan bir kan damlası kullanılarak ölçülür. Ancak teknolojideki ilerlemeler sayesinde artık sürekli glukoz ölçüm sistemleri (Continuous Glucose Monitoring - CGM) gibi daha gelişmiş cihazlar da mevcuttur.

2.1 Parmak Delerek Ölçüm Yapan Cihazlar

2.2 Nasıl Çalışır?

- Kullanıcı, parmağını delerek bir damla kan alır.
- Kan, test şeridine yerleştirilir ve cihaz tarafından analiz edilir.
- Kan şekeri seviyesi cihaz ekranında gösterilir.

2.3 Avantajları

- Taşınabilir ve genellikle ucuzdur.
- Kullanımı kolaydır ve anlık ölçüm sağlar.

2.4 Dezavantajları

- Günde birden fazla kez kan alınması gerektiğinden, rahatsızlık ve cilt tahrişi yaratabilir.
- Anlık ölçüm yapar, bu nedenle kan şekeri dalgalanmalarını sürekli takip etmek zordur.

2.5 Sürekli Glukoz Ölçüm Sistemleri (CGM)

Bu cihazlar, deri altına yerleştirilen küçük bir sensör aracılığıyla kan şekeri seviyelerini sürekli olarak ölçer ve verileri bir alıcıya ya da akıllı telefona gönderir.

2.6 Nasıl Çalışır?

- Cihazdaki sensör, deri altına yerleştirilir ve burada dokuların glukoz seviyelerini ölçer.
- Sensör, ölçümleri bir alıcıya veya mobil uygulamaya kablosuz olarak iletir.
- Kan şekerini 24 saat boyunca düzenli aralıklarla ölçer.

2.7 Avantajları

- Sürekli Takip: Kan şeker seviyesini sürekli izler, böylece dalgalanmalar anlık olarak görülebilir.
- Hipoglisemi ve Hiperglisemi Uyarısı: Tehlikeli kan şekeri seviyelerinde alarm verir.
- Azalmış Rahatsızlık: Parmak delme gereksinimi büyük ölçüde azalır (genellikle haftalık bir kalibrasyon için gerekebilir).
- Uzun Vadeli Veriler: Kullanıcıya geçmiş kan şekeri verilerini ve eğilimlerini gösterir, bu da tedavi planlamasında yardımcı olur.
- Tedaviye Entegre: Bazı sistemler insülin pompalarıyla entegre çalışabilir, böylece kan şekeri düzeyine bağlı otomatik insülin dozlaması yapılabilir.

2.8 Dezavantajları

- Maliyetleri daha yüksektir.
- Sensörün yerleştirilmesi bazı kullanıcılar için başlangıçta zor olabilir.
- Su veya diğer çevresel faktörlerle kullanım dikkat gerektirebilir.

3 Kişisel Yorumlar ve Fikirler

Bu kısımda geçen günlerde başımdan geçen bir olay ve sonrasında projeye ekleyebileceğimi düşündüğüm bir fikir hakkında açıklamalarda bulunacağım.

Birkaç gün önce geç saatlerde(23-24 gibi) yani normal yemek yeme saatim dışında bir zamanda günün tüm ana öğünlerini tamamlamışken ekstra bir insülin dozuyla tatlı yiyecektim, insülinin dozunu normalde doğru yaptım yani yiyeceğim şeye göre ama yiyeceğim şeyi tamamen bitiremedim çünkü midem bulandı, bu olaylar sonucunda yediğim şeye göre insülin dozunu fazla yapmış oldum bunu hemen farketmedim ve yaptığım dozda az değildi. Yiyeceğim şeyi dışarıda yemiştim ve yemekten hemen sonra eve dönmek için yürümeye başladım, normalde kan şekeri düştüğü zaman vücudun verdiği tepkiler sayesinde anlayabiliyorum ama o gün düştüğünü tam anlayamadım çünkü benim tahminim çok hızlı düşüyordu tüm bu faktörler birleşince, gözlerimin kararmaya başladığını farkettiğim zaman çok geçti bir şeyler yemeye çalıştıysamda yiyemeden bayıldım sokak ortasında yoldan geçen birileri ambulansı çağırmış herhalde sonra gözümü ambulansta açtım.

Şimdi fikirden bahsedeceğim. Bayılma durumu genelde kan şekeri 40mg/dl'nin altına düştüğü ve o seviyeden düşme eğiliminde devam ettiği durumlarda olabiliyor. Benim gibi çoğunlukla tek yaşayan ve bu hastalığa sahip olan hastların bu gibi acil durumlarda çaresizce beklememesi için aklıma bir şeyler geldi. Çünkü yolda değilde evde olsa yapabileceğim hiçbir şey olmayacaktı. Fikir aslında basit ama kullanımda olması gerektiğini düşündüğüm bir şey. Sürekli ölçüm yapan sensör cihazlar halihazırda şeker seviyesi çok düştüğü zaman uyarı veriyor ama en ekstrem durumlarda örneğin uyarı verdiği zaman hastanın yanında yiyecek bir şeyin olmaması ya da bağlı olduğu telefonun şarjı bittiği zaman bayılma gibi durumlar gerçekleşirse insan yardımına bakılmaksızın 112'ye bildirim göndermeli. Bu olayı yaşadıktan sonra ciddiyetini anladığım için bunu bilen biri olarak bir şeyler yapmak istiyorum. Yani kısaca bir uygulama. Fikrin detayları ileriki zamanlarda netlik kazanacaktır.

4 Nedensel Keşif Algoritmaları

Nedensel keşif algoritmaları (Causal Discovery Algorithms)[5], verilerdeki nedensel ilişkileri keşfetmek ve bu ilişkileri modellemek için kullanılan yöntemlerdir. Bu algoritmalar, gözlemlerden nedensel bağlantılar çıkarma sürecine yardımcı olur. Geleneksel istatistiksel yöntemler korelasyonları keşfetmeye odaklanırken, nedensel keşif algoritmaları, bir değişkenin diğerini nasıl etkilediğini ve bu etkilerin hangi koşullarda ortaya çıktığını anlamaya çalışır.

4.1 PC Algoritması (Peter-Clark Algorithm)

PC Algoritması (Peter-Clark Algorithm)[6], nedensel keşif algoritmalarından biridir ve değişkenler arasındaki nedensel yapıyı öğrenmeye yönelik bir yöntemdir. PC algoritması, özellikle büyük veri setlerinde ve karmaşık ilişkilerde kullanılabilir, çünkü bir değişkenin diğerine olan nedensel etkilerini ortaya çıkarmak için istatistiksel bağımlılıkları test eder. PC algoritması, temel olarak üç ana adımdan oluşur:

- Bağımsızlık Testleri: Algoritma, değişkenler arasındaki ilişkileri bulmaya başlamak için "bağımsızlık testleri" yapar. Yani, hangi değişkenlerin birbirinden bağımsız olduğunu anlamaya çalışır. Bağımsızlık testlerinde, genellikle koşullu bağımsızlık testleri kullanılır. Bu, belirli bir değişkenin, başka bir değişkenle ilişki kurmayan diğer değişkenler üzerinde kontrol edilmesi anlamına gelir.
- Bağımsızlık İlişkilerini Çıkarma: İlk adımda yapılan bağımsızlık testlerine göre, PC algoritması birbirine bağımsız olan değişkenler arasındaki bağlantıları (kenarları) kaldırır. Bu, veri setinin temel yapısını (graf yapısı) oluşturan ilk adımdır.

- Ağaç Yapısını İnşa Etme (Acyclic Graph): Son adımda, PC algoritması bir dönüşümsüz grafik (Acyclic Directed Graph) oluşturur. Bu, değişkenler arasındaki nedensel ilişkileri (neden-sonuç bağlantıları) temsil eder. Algoritma, daha önce çıkarılmış bağımsızlık ilişkilerine dayanarak, her iki değişkenin nedensel olarak birbirine bağlı olup olmadığını anlamaya çalışır.

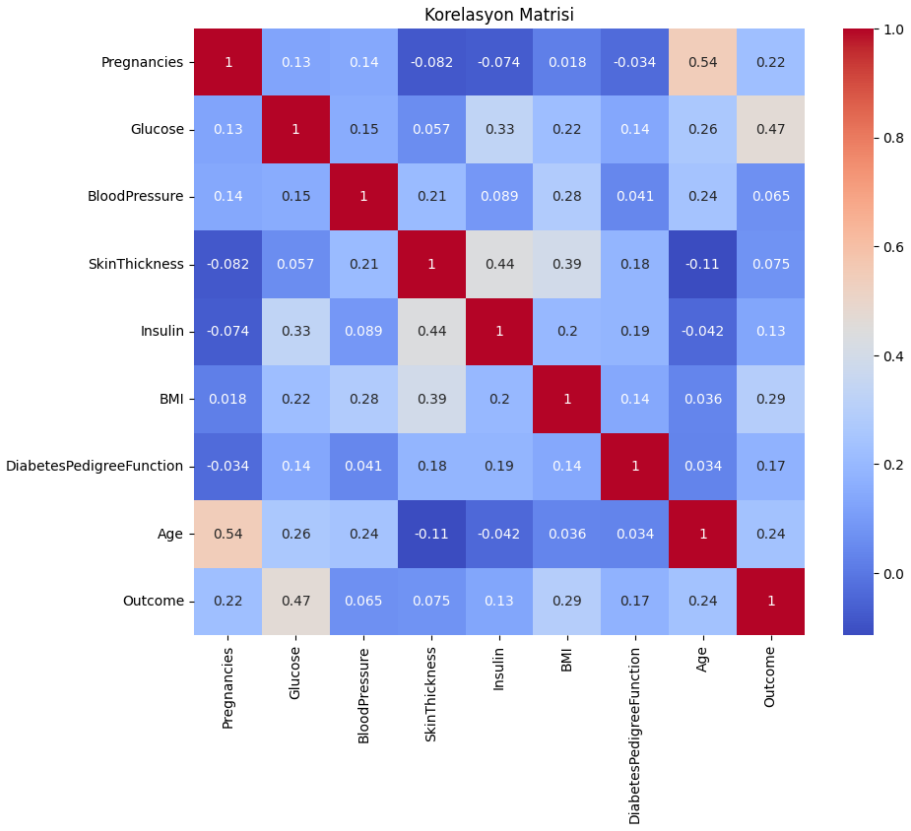
5 Örnek Uygulama

Bu örnek uygulamada kaggle sitesindeki "Diabetes Dataset"[7] veri seti kullanılacaktır. Veri setinin amacı, veri setinde yer alan belirli tanı ölçümlerine dayanarak bir hastanın diyabet olup olmadığını tanısal olarak tahmin etmektir.Veri kümesi hakkında:

- Pregnancies: Gebelik sayısını ifade etmek için
- Glucose: Kandaki Glikoz seviyesini ifade etmek için
- BloodPressure: Kan basıncı ölçümünü ifade etmek için
- SkinThickness: Cildin kalınlığını ifade etmek için
- Insulin: Kandaki insülin seviyesini ifade etmek için
- BMI: Vücut kitle indeksini ifade etmek için
- DiabetesPedigreeFunction: Diyabet yüzdesini ifade etmek için
- Age: Yaşı ifade etmek için
- Outcome: Sonucu ifade etmek için 1 Evet, 0 ise Hayır'dır.

Bu örnek, bir diyabet veri seti üzerinde Nedensel Keşif yapmak için kullanılır. Örneğin amacı, PC algoritması (Peter-Clark) kullanarak veriler arasındaki nedensel ilişkileri belirlemek ve bu ilişkileri görselleştiren bir Yönlendirilmiş Acyclic Graph (DAG) oluşturmak ve görselleştirmektir.

İlk olarak, veri setinin temel bilgileri ve korelasyon matrisi görüntülenir. Ardından, PC algoritması kullanılarak değişkenler arasındaki nedensel ilişkiler tespit edilir ve bir DAG oluşturulur. Son olarak, bu DAG görselleştirilir ve görsel olarak değişkenler arasındaki ilişkiler incelenebilir.

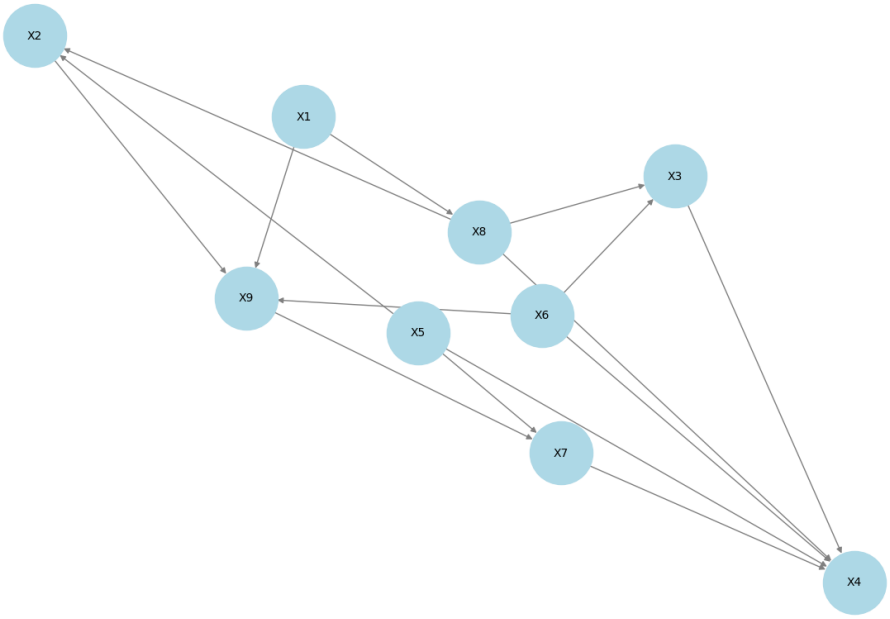


Şekil 1: Veri Setinin Korelasyon Matrisi

Korelasyon matrisi, iki veya daha fazla değişken arasındaki doğrusal ilişkiyi ölçen bir tablodur. Bu matris, her bir değişkenin diğerleriyle olan ilişkisini korelasyon katsayısı olarak gösterir. Korelasyon katsayısı genellikle -1 ile +1 arasında bir değere sahiptir:

- +1: Tam pozitif korelasyon (Bir değişken arttıkça diğeri de artar).
- -1: Tam negatif korelasyon (Bir değişken arttıkça diğeri azalır).
- 0: Hiçbir ilişki yok (Değişkenler birbirinden bağımsızdır).

- 'x1': 'Pregnancies'
- 'x2': 'Glucose'
- 'x3': 'BloodPressure'
- 'x4': 'SkinThickness'
- 'x5': 'Insulin'
- 'x6': 'BMI'
- 'x7': 'DiabetesPedigreeFunction'
- 'x8': 'Age'
- 'x9': 'Outcome'



Şekil 2: PC Algoritmasının Yönlendirilmiş Döngüsüz Grafiği

Bu grafiğin açıklaması şu şekildedir:

x2 (Glucose) → x9 (Outcome)
Yüksek Glukoz seviyesinin diyabet sonucunu (Outcome) doğrudan etkileyebileceğini gösterir. Bu bağlantı, glukoz seviyesinin, bireyin diyabet tanısı alıp almamasında önemli bir rol oynadığı anlamına gelir.

x1 (Pregnancies) → x9 (Outcome)

Hamilelik sayısının (Pregnancies), diyabet sonucu üzerinde etkisi olduğunu gösterir. Özellikle gestasyonel diyabet açısından bu ilişki anlamlı olabilir.

x8 (Age) → x9 (Outcome)

Yaş ilerledikçe diyabet riskinin arttığına dair bir nedensel bağ bulunmuştur.

x2 (Glucose) → x8 (Age)

Bu bağlantı, yaşın glukoz seviyesini etkileyebileceğini göstermektedir. Yaşlanma ile birlikte metabolizma değişiklikleri glukoz seviyesini etkileyebilir.

x3 (BloodPressure) → x6 (BMI)

Kan basıncının vücut kitle indeksini etkileyebileceği yönünde bir ilişki gösterir. Yüksek kan basıncı, genellikle obezite ile ilişkili olabilir.

x6 (BMI) → x9 (Outcome)

BMI'nin diyabet sonucu üzerinde etkisi olduğunu belirtir. Fazla kilo, diyabet riskini artıran önemli bir faktördür.

x7 (DiabetesPedigreeFunction) → x9 (Outcome)

Ailede diyabet geçmişinin, bireyin diyabet riskini etkilediğini belirtir. Bu genetik yatkınlıkları ifade eder.

x5 (Insulin) → x9 (Outcome)

İnsülin seviyesinin, diyabet sonucu üzerinde doğrudan bir etkisi olduğu görülmektedir. İnsülin direnci veya eksikliği, diyabetin temel nedenlerinden biridir.

x4 (SkinThickness) → x6 (BMI)

Cilt kalınlığının BMI üzerinde etkili olduğunu gösterir. Cilt kalınlığı genelde vücut yağ oranıyla ilişkilidir ve bu da BMI ile ilişkilendirilebilir.

x8 (Age) → x6 (BMI)

Yaşın BMI üzerindeki etkisini gösterir. İlerleyen yaşlarda metabolizmanın yavaşlaması BMI artışına yol açabilir.

x5 (Insulin) → x6 (BMI)

İnsülin seviyelerinin BMI üzerinde etkili olduğunu ifade eder. İnsülin direnci kilo alımına yol açabilir.

x3 (BloodPressure) → x7 (DiabetesPedigreeFunction)

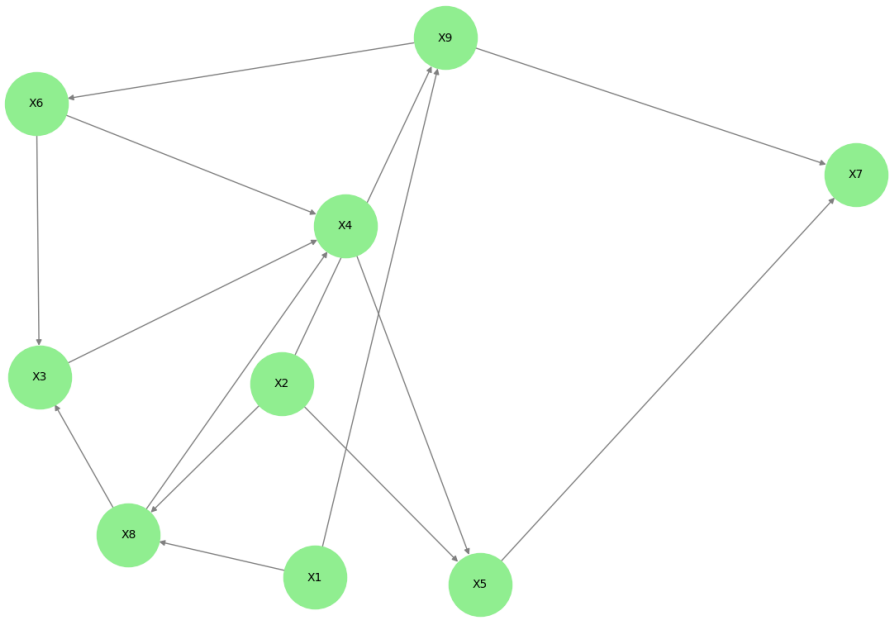
Kan basıncının, ailede diyabet geçmişiyle bağlantılı olabileceğini gösterir. Bu ilişki çevresel ve genetik faktörlerle açıklanabilir.

6 Hibrit Modeller

6.1 GES + Random Forest

GES[8], bir nedensel keşif algoritmasıdır. Veriler arasındaki nedensel ilişkileri öğrenmek amacıyla, özellikle değişkenler arasındaki bağımlılıkları ve etkileşimleri keşfetmek için kullanılır. GES, Bayes ağları kullanarak verilerdeki olası neden-sonuç ilişkilerini belirlemeye çalışır.
Random Forest[9], karar ağaçları tabanlı bir ansamble öğrenme yöntemidir. Birçok karar ağacının çıktılarının birleştirilmesiyle güçlü bir sınıflandırma modeli oluşturur. Overfitting (aşırı öğrenme)'e karşı dayanıklıdır ve veri setindeki karmaşıklıkları iyi öğrenebilir.
Daha önce yapılan örnekteki veri setini[10] kullanarak bu sefer hibrit modelle bir çalışma yapıldı ancak hibrit modelin sonucunda yüzde yüz doğruluk veriyor yani 'Overfitting' durumu oluyor bunun sebebini veri setinin yeterli büyüklükte olmayışı olabilir çünkü tek başına GES algoritması çalıştırıldığı zaman çıktı olarak alınan graph doğru.

- 'x1': 'Pregnancies' 'x2': 'Glucose' 'x3': 'BloodPressure' 'x4': 'SkinThickness' 'x5': 'Insulin' 'x6': 'BMI' 'x7': 'DiabetesPedigreeFunction' 'x8': 'Age' 'x9': 'Outcome'



Şekil 3: GES Algoritmasının Yönlendirilmiş Döngüsüz Grafiği

Random Forest Sınıflandırma Raporu:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	99
1	1.00	1.00	1.00	55
accuracy			1.00	154
macro avg	1.00	1.00	1.00	154
weighted avg	1.00	1.00	1.00	154

Şekil 4: Random Forest Sınıflandırma Raporu

Bu sonuçlar modelin aşırı uyum yaptığını gösteriyor, random forest modelini daha az karmaşık hale getirip düzenledikten sonra yine sonuç değişmedi sonra koda KFold Cross Validation eklenip düzenleme yapıldıktan sonra da model yine aynı sonuçları vermeye devam etti yani aşırı uyum.

6.2 GES + XGBoost

XGBoost[11] yüksek performanslı bir makine öğrenimi modelidir ve özellikle büyük veri ve karmaşık sınıflandırma/regresyon problemlerinde başarıyla kullanılır. İlk örnekten sonra bu model denendi ancak sonuç değişmedi yine overfitting durumu var. Bu durumda sorun veri setinde olabilir.

XGBoost Sınıflandırma Raporu:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	99
1	1.00	1.00	1.00	55
accuracy			1.00	154
macro avg	1.00	1.00	1.00	154
weighted avg	1.00	1.00	1.00	154

Şekil 5: XGBoost Sınıflandırma Raporu

Sorunu çözmek için veri setine sentetik veri ekleme çalışması yapılacak ve başka modellere bakılacaktır.

7 Sentetik Veri

Sentetik veri[12], gerçek dünya verilerinden türetilen ancak orijinal verilere benzer özellikler taşıyan yapay verilerdir. Yapılan uygulamalarda model sonucu yüzde yüz doğruluk verdiğinden veri setine sentetik veriler ekleyerek veri setini genişletmek ve overfitting durumunu çözmek amaçlanmıştır. Sentetik veri üretmek için birden fazla method vardır, bunlardan en çok kullanılanlarından olan SMOTE ile çalışılacaktır.

7.1 SMOTE(Synthetic Minority Over-sampling TEchnique)

SMOTE[13], dengesiz veri setlerinde, özellikle azınlık sınıfının yetersiz olduğu durumlarda, azınlık sınıfındaki örneklerin sayısını artırmak için kullanılan bir tekniktir.

Normalde veri setinde Outcome değeri 0 yani sağlıklı olan veri sayısı 500, Outcome değeri 1 yani hasta olan veri sayısı 268 kişidir. İki sınıf arasında bir dengesizlik var. Öncelikle SMOTE ile bu sınıfları dengelemek için Outcome değeri 1 olan 232 yeni veri üretildi. Oluşturulan yeni veri seti ile önceden yapılan çalışmalar tekrar yapıldı ve Nedensel Çizgeler ile oluşturulan graphlar aynı değil.

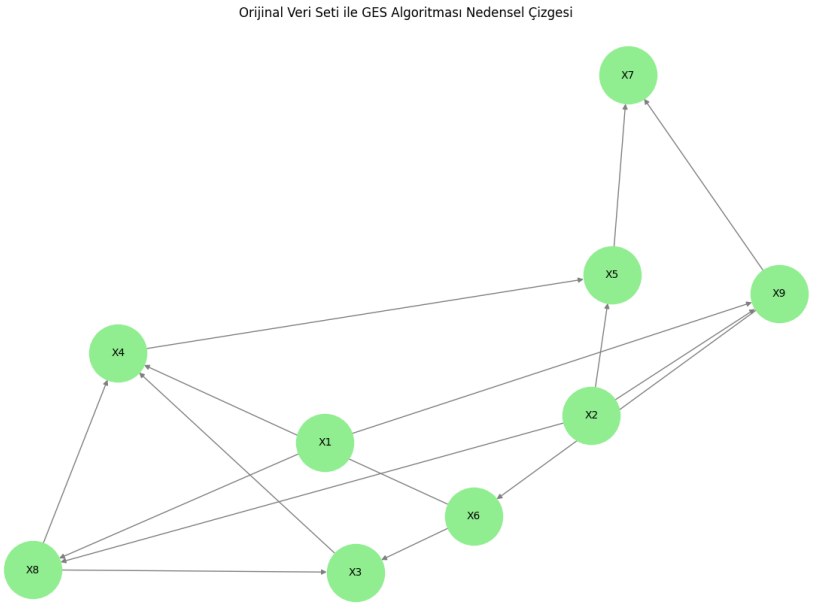
Ancak daha önce kullanılan hibrit model kullanıldığı zaman çıkan sonuçlar yine bir aşırı uyum yani Overfitting olduğunu gösteriyor.

Random Forest Sınıflandırma Raporu:

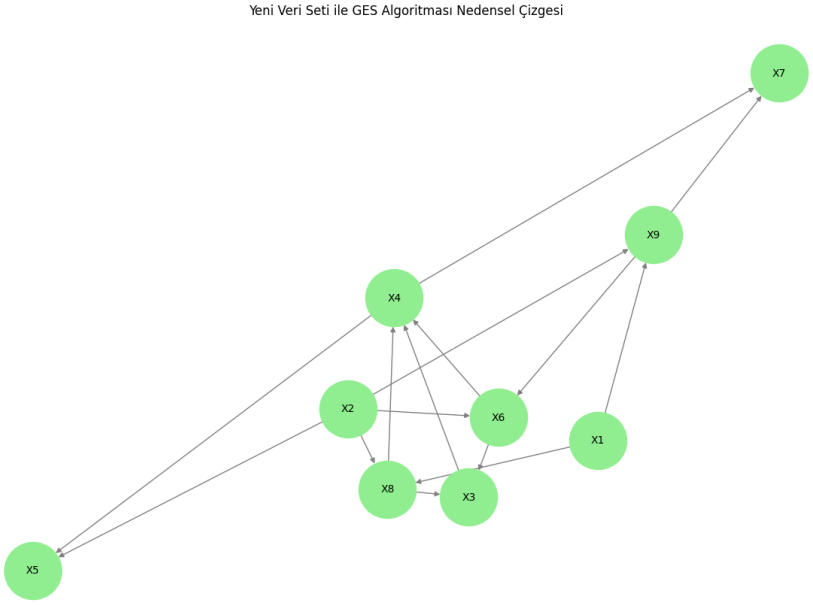
	precision	recall	f1-score	support
0	1.00	1.00	1.00	100
1	1.00	1.00	1.00	100
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Şekil 6: Yeni veri setiyle yapılan çalışmanın sonuçları

Karşılaştırdığımız zaman bazı nedensel çizgelerin diğer graphda bulunmadığını anlayabiliyoruz. Bu çizgeler şöyle:



Şekil 7:



Şekil 8:

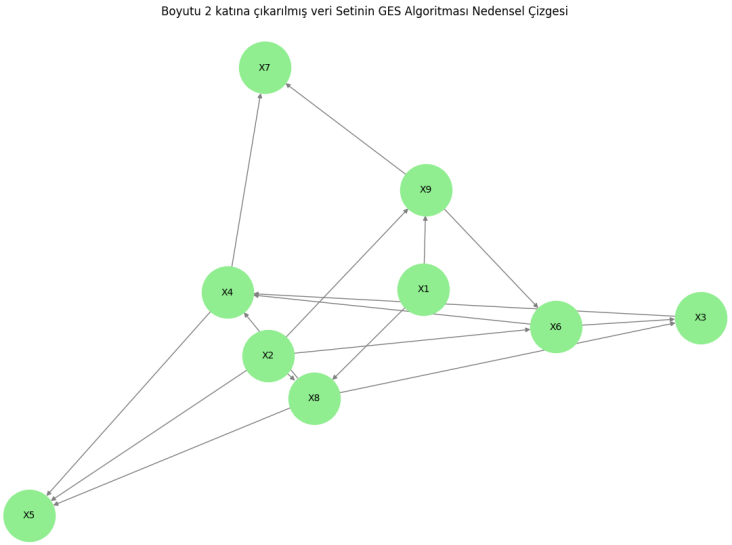
- 'x1': 'Pregnancies' 'x2': 'Glucose' 'x3': 'BloodPressure' 'x4': 'SkinThickness' 'x5': 'Insulin' 'x6': 'BMI' 'x7': 'DiabetesPedigreeFunction' 'x8': 'Age' 'x9': 'Outcome'

- X1-X4, X1-X6, X5-X7 Bu çizgeler orijinal veri setiyle yapılan çalışmada var yeni veri setiyle yapılan çalışmada yok.

- X2-X6, X4-X7 Bu çizgeler yeni veri setiyle yapılan çalışmada var orijinal veri setiyle yapılan çalışmada yok.

Azınlık sınıfı dengelemek bu çalışma için pek de mantıklı olmayabilir çünkü veri setinin kendine has olan oranını bozmuş oluyoruz. Bunun yerine sınıflar arasındaki dengeyi bozmadan veri setinin boyutunu arttırmak daha mantıklı olabilir. Yine SMOTE ile aynı oranda veri setinin boytunu 2 katına çıkarıldı yani sağlıklı insan sayısı 1000 olurken hasta insan sayısı 536 oldu.

Bu işlem sonucu oluşturulan yeni veri setiyle aynı graph çizdirildiği zaman çıktısı şöyle:



Şekil 9:

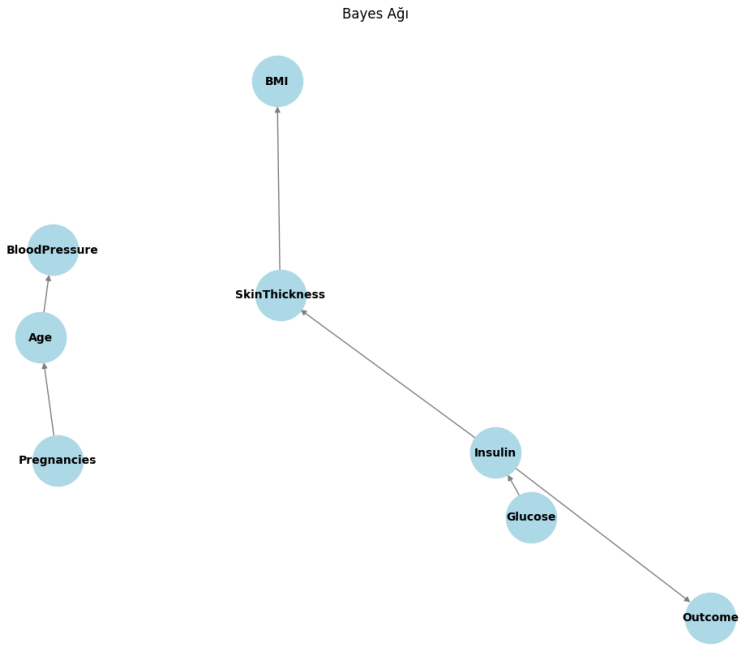
- X1-X4, X1-X6, X5-X7 Bu çizgeler orijinal veri setiyle yapılan çalışmada var yeni veri setiyle yapılan çalışmada yok.

- X2-X6, X4-X7, X8-X5 Bu çizgeler yeni veri setiyle yapılan çalışmada var orijinal veri setiyle yapılan çalışmada yok.

Boyutu 2 katına çıkarılan bu veri setiyle hibrit modeli çalıştırdığımız zaman yine sonuç değişmiyor ve overfitting durumu oluyor. Buradan çıkarılacak sonuç hibrit model bu veri seti için uygun değil. Ayrıca sentetik veriyle oluşturulan yeni veri setlerinin nedensel çizgelerde vermiş olduğu bağlantılar orijinal veri setinin vermiş olduğu bağlantılardan daha mantıksız yani orijinal veri seti nedensel çizgelerde daha mantıklı.

8 Bayes Ağı

Nedensel Bayes ağları[14], değişkenler arasındaki neden-sonuç ilişkilerini modellemek için kullanılır. Bu ağlar, olasılıksal temellere dayanmasının yanı sıra, değişkenler arasındaki nedensel ilişkileri de kullanarak analizlerin ve oluşturulan senaryoların daha açıklayıcı olmasını sağlar.

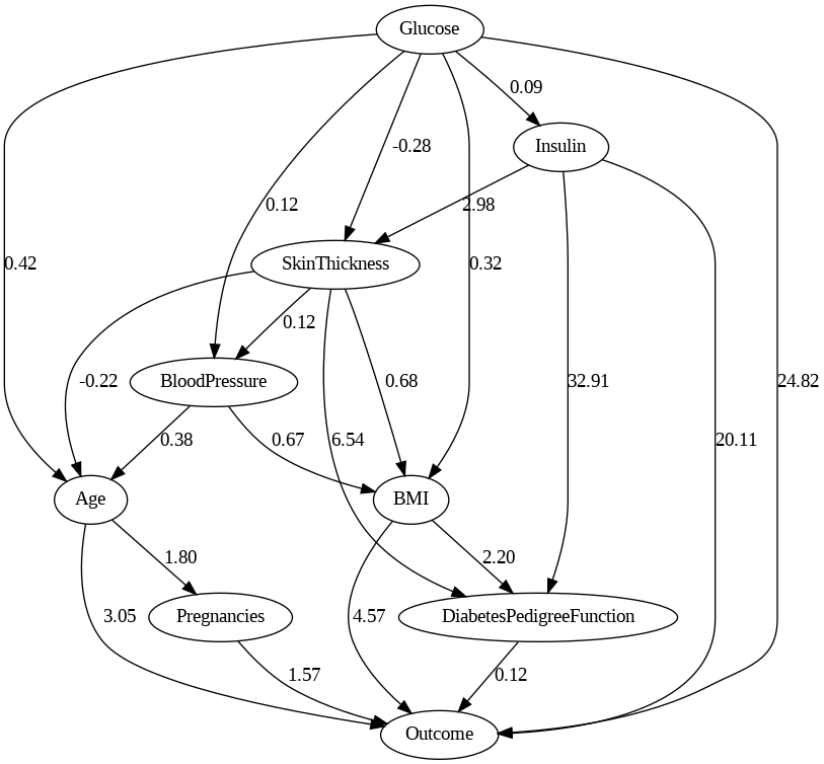


Şekil 10: Bayes Ağı

Veri setindeki bazı değişkenler arasında yeterince güçlü bağımlılıklar tespit edilememiş olabilir. Bu durumda, bu değişkenler ağın dışında kalır. Hill Climb Search gibi algoritmalar, yalnızca anlamlı bağlantıları belirler. Eğer bazı değişkenler diğer değişkenlere bağımlı değilse, ağda yer almazlar. Ağda bazı değişkenler arasında bağımsızlık ilişkisi olduğu için birden fazla bağımsız alt grafik oluşmuş olabilir.

9 LiNGAM (Linear Non-Gaussian Acyclic Model)

Lingam[15], nedensel ilişkileri belirlemek için kullanılan bir algoritmadır. Bu yöntem, özellikle doğrusal ve doğrusal olmayan ilişkileri tespit etmede etkilidir. LiNGAM, veriler arasındaki nedensel yönleri belirlemek için bağımsız bileşen analizi (ICA) kullanır.



Şekil 11: Lingam Ağı

Bağlantılar üzerindeki sayılar, doğrusal regresyon katsayılarını temsil eder. Bu katsayılar, iki değişken arasındaki doğrudan ilişkinin büyüklüğünü ve yönünü ifade eder:

- Pozitif değerler: Artan bir ilişkiyi ifade eder.
- Negatif değerler: Azalan bir ilişkiyi ifade eder.
- **Glucose → Insulin (0.09):** Glikoz seviyesi, insülin seviyesini pozitif yönde etkiler.
- **Glucose → Outcome (24.82):** Glikoz seviyesi, diyabet sonucunu (Outcome) güçlü bir şekilde etkiler.
- **Glucose → SkinThickness (-0.28):** Glikoz seviyesi, deri kalınlığını negatif yönde etkiler.
- **Glucose → BloodPressure (0.12):** Glikoz seviyesi, kan basıncını pozitif yönde etkiler.
- **Glucose → Age (0.42):** Glikoz seviyesi, yaşı pozitif yönde etkiler.
- **Glucose → BMI (0.32):** Glikoz seviyesi, vücut kitle indeksini (BMI) pozitif yönde etkiler.
- **Insulin → Outcome (20.11):** İnsülin seviyesi, diyabet sonucunu (Outcome) güçlü bir şekilde etkiler.
- **Insulin → DiabetesPedigreeFunction (32.91):** İnsülin seviyesi, genetik yatkınlık (DiabetesPedigreeFunction) üzerinde çok güçlü bir etkiye sahiptir.
- **Insulin → SkinThickness (2.98):** İnsülin seviyesi, deri kalınlığını pozitif yönde etkiler.
- **SkinThickness → BMI (0.68):** Deri kalınlığı, vücut kitle indeksini (BMI) pozitif yönde etkiler.
- **SkinThickness → DiabetesPedigreeFunction (6.54):** Deri kalınlığı, genetik yatkınlık (DiabetesPedigreeFunction) üzerinde çok güçlü bir etkiye sahiptir.
- **SkinThickness → BloodPressure (0.12):** Deri kalınlığı, kan basıncını pozitif yönde etkiler.
- **SkinThickness → Age (-0.22):** Deri kalınlığı, yaş üzerinde negatif bir etkiye sahiptir.
- **BloodPressure → BMI (0.67):** Kan basıncı, vücut kitle indeksini (BMI) pozitif yönde etkiler.
- **BloodPressure → Age (0.38):** Kan basıncı, yaş üzerinde pozitif bir etkiye sahiptir.
- **Age → Pregnancies (1.80):** Yaş, hamilelik sayısını pozitif yönde etkiler.

- **Age → Outcome (3.05):** Yaş, diyabet sonucunu (Outcome) pozitif yönde etkiler.
- **BMI → DiabetesPedigreeFunction (2.20):** BMI (vücut kitle indeksi), genetik yatkınlık (DiabetesPedigreeFunction) üzerinde pozitif bir etkiye sahiptir.
- **BMI → Outcome (4.57):** BMI, diyabet sonucunu (Outcome) pozitif yönde etkiler.
- **Pregnancies → Outcome (1.57):** Hamilelik sayısı, diyabet sonucunu (Outcome) pozitif yönde etkiler.
- **DiabetesPedigreeFunction → Outcome (0.12):** Genetik yatkınlık (DiabetesPedigreeFunction), diyabet sonucunu (Outcome) pozitif yönde etkiler.

10 Threshold Değeri

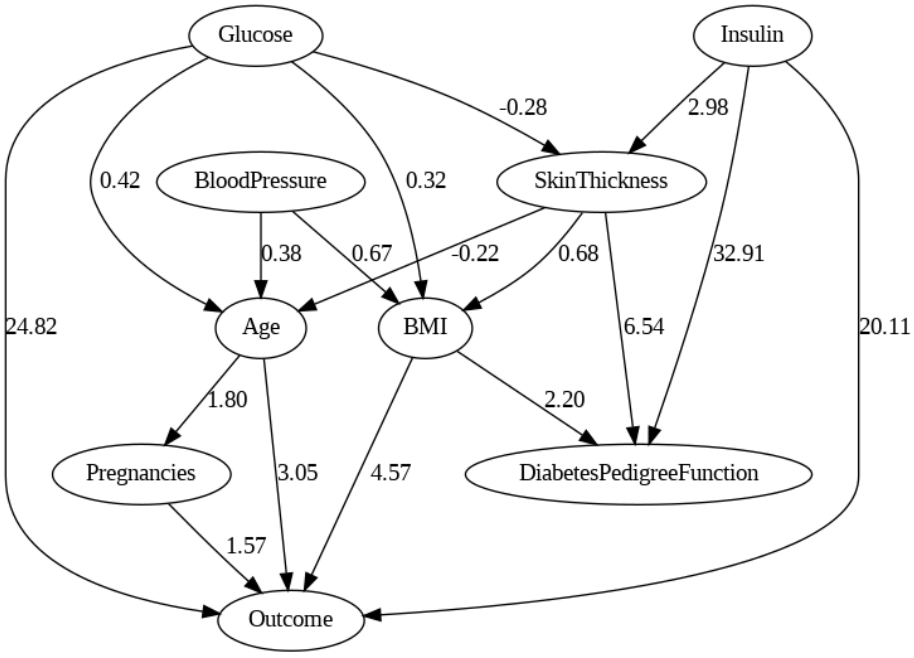
Threshold (Eşik Değeri), nedensel bağların modelden alınan ağırlıklarına göre filtrelenmesi için kullanılan bir kriterdir. Bu değer, hangi bağların dikkate alınacağını belirler. Oluşturulan graphın karmaşıklığını azaltmak ve daha önemli bağları dikkate almak için kullanıldı. Thershold değeri olarak 0.2 seçildi.

Eşik Değeri	**Kalan Kenar Sayısı**
0.0	21
0.1	20
0.2	17
0.3	15

Table 1: Farklı eşik değerleri için modelde kalan kenar sayıları.

Nedensel ağı sadeleştirdikten sonra şimdiki işlem nedensel bağların ağırlıkları üzerinden outcome değerine ulaşma yani örnek olarak girilen verilerin diyabet hastası olup olmadığını anlamak.

Buradaki diyabet türü "Tip 2 Diyabet" çünkü veri setinde "C-Peptid" gibi Tip 1 Diyabet tanısı koymada ana eteknlerden biri olan pankreasın insülin üretme kapasitesini ölçen kritik verilerin hepsi yok.



Şekil 12: Threshold değeri "0.2" seçildikten sonra Lingam Ağı

11 Örnek İnsanların Verilerine Göre Outcome Değeri Hesaplama

Yapılan işlem daha önce bulunan ağırlıkların bir takım işlemlerle outcome değeri oluşturmak.

11.1 1. Ortalama ve Standart Sapma Hesaplama

Her bir değişken için veri setinden ortalama (μ) ve standart sapma (σ) değerleri hesaplanır:

$$\mu = \frac{\sum X}{n}, \quad \sigma = \sqrt{\frac{\sum (X - \mu)^2}{n}}$$

Burada X, değişkenin tüm bireylerdeki değerlerini ifade eder.

11.2 2. Standartlaştırma (Z-Skoru) Hesaplama

Her bir değişken için bireyin değeri şu şekilde standartlaştırılır:[16]

$$z = \frac{X - \mu}{\sigma}$$

Burada:

- X: Bireyin değişken için değeri,
- μ: Veri setindeki o değişkenin ortalaması,
- σ: Veri setindeki o değişkenin standart sapması.

11.3 3. Nedensel Ağırlıklarla Etki Hesaplama

Her bir değişkenin Outcome üzerindeki etkisi şu şekilde hesaplanır:

$$\text{Etkiler} = z \times \text{Ağırlık}$$

Burada:

- z: Değişkenin standartlaştırılmış değeri,
- Ağırlık: Nedensel bağlantı ağırlığı.

Tüm etkiler toplandıktan sonra toplam Outcome değeri hesaplanır:

$$\text{Toplam Outcome Değeri} = \sum(\text{Etkiler})$$

11.4 4. Sigmoid Fonksiyonu ile Normalize Etme

Hesaplanan Outcome değeri sigmoid fonksiyonuna uygulanarak normalize edilir:[17]

$$S(x) = \frac{1}{1 + e^{-x}}$$

Burada $S(x)$, Outcome değerinin 0 ile 1 arasında normalize edilmiş halini ifade eder.

11.5 5. Diyabet Sınıflandırması

Sigmoid fonksiyonundan elde edilen değer, bir eşik değeri ile karşılaştırılarak bireyin durumu belirlenir:

- Eğer $S(x) \geq 0.5$ ise birey **Diyabet** olarak sınıflandırılır.
- Eğer $S(x) < 0.5$ ise birey **Diyabet Değil** olarak sınıflandırılır.

Önce veri setindeki hem hasta hem hasta olmayan kişilerin değerleri girilince çıkan sonuçların veri setindekiyle uyumlu olup olmadığına yani yapılan çalışmanın doğruluğuna bakalım. Daha sonra örnek kişiler oluşturup bu kişilerin hastalık durumunun ne olduğunu ve bu sonucun mantıklı olup olmadığına bakalım.

Değişken	**Değer**
Pregnancies	0
Glucose	137
BloodPressure	40
SkinThickness	35
Insulin	168
BMI	43.1
DiabetesPedigreeFunction	2.288
Age	33
Normalize Edilmiş Outcome (Sigmoid)	**0.65**
Diyabet Durumu	**Diyabet**

Table 2: Veri setindeki hasta biri için tahmin sonuçları.

Değişken	**Değer**
Pregnancies	1
Glucose	103
BloodPressure	30
SkinThickness	38
Insulin	83
BMI	43.3
DiabetesPedigreeFunction	0.183
Age	33
Normalize Edilmiş Outcome (Sigmoid)	**0.46**
Diyabet Durumu	**Diyabet Değil**

Table 3: Veri setindeki hasta olmayan biri için tahmin sonuçları.

Değişken	**Değer**
Pregnancies	0
Glucose	120
BloodPressure	70
SkinThickness	25
Insulin	80
BMI	28.0
DiabetesPedigreeFunction	0.8
Age	40
Normalize Edilmiş Outcome (Sigmoid)	**0.49**
Diyabet Durumu	**Diyabet Değil**

Table 4: 1.Örnek birey için tahmin sonuçları.

Yapılan uygulamada veri seti değerleri ve örnek değerlerle oluşturulan outcome değerleri mantıklıdır.

Değişken	**Değer**
Pregnancies	0
Glucose	140
BloodPressure	70
SkinThickness	25
Insulin	80
BMI	28.0
DiabetesPedigreeFunction	0.8
Age	40
Normalize Edilmiş Outcome (Sigmoid)	**0.56**
Diyabet Durumu	**Diyabet**

Table 5: 2.Örnek birey için tahmin sonuçları.

12 Doğruluk Hesaplama

En son eklenen çalışmada, örnek insanlar yaratarak bu kişilerin diyabet olup olmadığını kontrol etme işlemi gerçekleştirilmiş ve doğruluk oranı %34.63 olarak kaydedilmiştir. Modelin doğruluğunu artırmak için lojistik regresyon modeli[18] eklenmiş ve LiNGAM[19] algoritmasıyla elde edilen nedensel bağlantılar kullanılmıştır.

LiNGAM modeli, veri setindeki değişkenler arasındaki nedensel ilişkileri analiz ederek bu ilişkilerden elde edilen ağırlıkları belirlemiştir. Bu ağırlıklar, lojistik regresyon modeline ek özellikler olarak dahil edilmiştir. Lojistik regresyon, veriyi doğrusal bir sınır ile sınıflandırarak, belirli bir kişinin diyabetli olma olasılığını hesaplamak için sigmoid fonksiyonunu kullanır. Model, girdilerdeki her bir özelliğe bir katsayı atar ve bu katsayılar, LiNGAM'den gelen nedensel ağırlıklarla zenginleştirilmiştir. Sonuçta, tahmin doğruluğunu artırmak için veri setindeki nedensel bilgi, lojistik regresyonun tahmin yeteneğine entegre edilmiştir.

Eklenen lojistik regresyon ile önce çalışmada kullanılan veri setini rastgele bir şekilde %70 eğitim %30 test olarak ayırıp çalıştırıldığı zaman modelin doğruluk oranı %73.59 oluyor.

Daha sonra ise eğitim verisi olarak, kullanılan veri setini ayırmadan bütün veri seti , test verisi için ise sentetik veri seti oluşturulup o veri seti kullanıldı. Normal veri setinde 768 değer olduğu için yeni veri setinde 330 değer oluşturuldu böylece eğitim ve test verisi oranı korunmuş oldu. Sentetik veri seti, orijinal veri setinin ortalama vektörü ve kovaryans matrisine dayalı olarak çok değişkenli normal dağılımdan üretilerek oluşturulmuştur. Lojistik regresyon modeli, sentetik veriler için tahmin olasılıklarını hesaplamış ve Bernoulli dağılımı kullanılarak sınıf etiketleri atanmıştır. Veri seti oluşturulduktan sonra birkaç eksi değer üretti bu değerler 0 olarak değiştirildi. Yapılan işlemler sonrasında model doğruluğu %79.09 oldu.

13 Veri Setindeki Mantıksız Değerleri Düzeltme

Önceki raporlarda da bahsedildiği şekilde normal veri setindeki bazı mantıksız veriler var, daha önce bu veriler düzeltilmeye çalışıldığı zaman çalışmada mantıksız sonuçlar ortaya çıkarmıştı. Şimdi ise bu verileri lojistik regresyon ile düzeltildi ama nedensel bağların arasındaki ağırlıkları mantıksız hale getirdi örneğin insülinle outcome değerleri arasındaki ağırlık -30.94 gibi gerçekle uyuşmayan bir değer ortaya çıkardı. Yani veri setindeki değerlerle oynamak mantıklı sonuçlar çıkarmıyor bundan dolayı verilerle oynamadan veri seti olduğu gibi kullanılacaktır.

14 Sonuç

Modelin ulaştığı en yüksek doğruluk oranı %79.09 olup çeşitli parametreleri değiştirererk modelin doğruluk oranında değişiklik sağlanamamıştır. Aynı şekilde lojistik regresyon yerine farklı modeller kullanılsa da %79.09'u geçen model olmamıştır. Bunun sebebi lingam ağından çıkan nedensel bağ ve ağırlıkların bu modellere girmesi olabilir çünkü bu ağlar ne kadar anlamlıysa modeller de ona göre işler. Yani %79.09 yüksek bir doğruluk oranı olarak görünmeyebilir ancak azımsanmayacak kadar da iyi bir oran. Özellikle bu veri seti için çünkü veri setinin doğruluk ve gerçekliği sorgulanabilecek düzeyde. Ayrıyeten lingam ağıyla oluşturulan nedensel ağın bağları büyük ölçüde biyolojik olarak mantıklı. Yaratılan örnek insan değişkenlerinin hasta olup olmama durumuyla ilgili yapılan çalışmalar da modelin doğruluğunu gösteriyor.

Kaynakça

[1] Wikipedia, "https://tr.wikipedia.org/wiki/Hipoglisemi," 2024.

[2] medicalpark, "https://www.medicalpark.com.tr/seker-dusmesi-belirtileri/hg-4402," 2024.

[3] diyabet, "https://www.diyabet.com/tr/diyabet-hakkinda/Tip-1-diyabet-nedir-Tip-1-diyabette-hiperglisemi-ve-hipoglisemi.html," 2024.

[4] deremedikal, "https://www.deremedikal.com/blog/icerik/seker-olcum-cihazi-nasil-kullanilir: :text=

[5] causaLens, "https://causalens.com/resources/white-papers/how-can-ai-discover-cause-and-effect/: :text=Causal

[6] salesforce, "https://opensource.salesforce.com/causalai/latest/tutorials/PC$_{A}lgorithm_{T}abular.html$," 202

[7] Kaggle, "https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset/data," 2024.

[8] frontiersin, "https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2019.00524/full," 2024.

[9] analyticsvidhya, "https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/," 2024.

[10] Kaggle, "https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset/data," 2024.

[11] simplilearn, "https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article: :text=XgBoost

[12] wikipedia, "https://en.wikipedia.org/wiki/Synthetic$_{d}ata$: : $text$ = $Synthetic$

[13] microsoft, "https://learn.microsoft.com/tr-tr/azure/machine-learning/component-reference/smote?view=azureml-api-2," 2024.

[14] E. N. ÇİNİCİOĞLU, Ş. Ö. EKİCİ, and F. ÜLENGİN, "Bayes ağ yapisinin oluşturulmasinda farkli yaklaşimlar: Nedensel bayes ağlari ve veriden ağ öğrenme,"

[15] S. Shimizu, "Lingam: Non-gaussian methods for estimating causal structures," *Behaviormetrika*, vol. 41, no. 1, pp. 65–98, 2014.

[16] wikipedia, "https://tr.wikipedia.org/wiki/Standartla

[17] wikipedia, "https://tr.wikipedia.org/wiki/Sigmoid$_{i}$

[18] medium-towardsdatascience, "https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc," 2025.

[19] lingam, "https://lingam.readthedocs.io/en/stable/tutorial/lingam.html," 2025.