VERİ MADENCİLİĞİ (FET445)

Diyabet Hastalarının Yeniden Yatış Tahmini

İnsülin Avcıları

Esma Gelebek – 22040301063 – esmagelebek@stu.topkapi.edu.tr

Şevval Öveyik – 22040301034 – sevvaloveyik@stu.topkapi.edu.tr

Eslem Berra Özel – 22040301016 – eslemberraozel@stu.topkapi.edu.tr

Esma Betül Kocaahmet – 21040301052 – esmabetulkocaahmet@stu.topkapapi.edu.tr

Github Linki: https://github.com/esmabetulkocaahmet/insulinavcilari_verimadenciligi

Öğretim Üyesi: Dr. Yıldız Karadayı

25.11.2025

I. PROBLEM TANIMI

I.I İş / Bilimsel Soru

Bu proje, diyabet hastalarının taburcu olduktan sonraki kısa süre içerisinde yeniden hastaneye yatış yapıp yapmayacağını öngörmeyi amaçlamaktadır. Yeniden yatışlar, hem sağlık hizmetlerinin kalitesini değerlendirmede hem de hastanelerin maliyet yönetimi açısından kritik bir göstergedir ve özellikle kronik hastalık yönetimi gerektiren diyabet hastalarında oldukça sık görülmektedir. ABD'deki 130 hastaneden 1999–2008 yılları arasında toplanmış ve yaklaşık 100.000 diyabet hastası kaydını içeren kapsamlı veri seti kullanılarak, hastaların 30 gün içerisinde yeniden yatış yapma olasılıklarının modellenmesi hedeflenmektedir. Proje kapsamında, hastaların demografik bilgileri, laboratuvar sonuçları, ilaç değişiklikleri, önceki vizit sayıları, tanı ve prosedür kodları gibi çok sayıda özellik detaylı bir şekilde incelenecek ve bu özelliklerin yeniden yatış üzerindeki etkileri araştırılacaktır. Hedef, yalnızca yeniden yatış ihtimalini tahmin etmek değil, aynı zamanda klinik olarak anlamlı ve yorumlanabilir bir model geliştirmektir; böylece sağlık hizmet sağlayıcıları riskli hastaları erken tespit edebilir, proaktif müdahaleler planlayabilir ve hasta bakım kalitesini artırabilir. Proje ayrıca, veri hazırlama, özellik seçimi, boyut indirgeme ve farklı basit sınıflandırma modellerinin uygulanması ile tahmin performansını optimize etmeyi amaçlamakta, farklı model ve özellik kombinasyonlarının performansları karşılaştırılarak en etkili yaklaşımın belirlenmesini sağlamaktadır. Bu bağlamda, çalışma hem veri madenciliği tekniklerini klinik veri üzerinde uygulayarak değerli bulgular elde etmeyi hem de yeniden yatış riski yüksek hastaların yönetimine yönelik bilimsel bir katkı sağlamayı hedeflemektedir.

I.II Görev Türü

Bu proje, temel olarak ikili sınıflandırma (binary classification) problemine odaklanmaktadır. Amaç, her bir diyabet hastasının taburcu olduktan sonraki 30 gün içinde yeniden hastaneye yatış yapıp yapmayacağını doğru bir şekilde tahmin etmektir. Veri setindeki her hasta kaydı, pozitif sınıf (yeniden yatış yapan hasta) veya negatif sınıf (yeniden yatış yapmayan hasta veya 30 günden sonra yatış yapan hasta) olarak sınıflandırılacaktır. Bu bağlamda model, hastaların demografik bilgileri, laboratuvar sonuçları, ilaç kullanımı, önceki vizit sayıları ve tanı/prosedür kodları gibi çok boyutlu özelliklerini kullanarak bir sınıflandırıcı fonksiyonunu öğrenir. Problem, regresyon veya kümeleme gibi sürekli tahminler veya gruplama hedefleri yerine doğrudan "evet/hayır" sonucunu verdiği için sınıflandırma görevine uygundur. Ayrıca, ikili sınıflandırma yaklaşımı, sağlık hizmetlerinde kritik olan pozitif sınıfın (yeniden yatış riski yüksek hastaların) tespit edilmesini sağlayarak, klinik karar destek sistemlerinde pratik ve uygulanabilir sonuçlar üretir. Bu nedenle proje, veri madenciliği ve makine öğrenmesi perspektifinden ikili sınıflandırma çerçevesinde ele alınmıştır.

I.III Hedef Değişkenler

Bu projede hedef değişken, diyabet hastalarının taburcu olduktan sonraki kısa süre içinde yeniden hastaneye yatış yapıp yapmayacağını belirlemektir. Veri setinde orijinal olarak **readmitted** değişkeni yer almakta ve üç kategoriye sahiptir: "<30" (30 gün içinde yeniden yatış), ">30" (30 günden sonra yeniden yatış) ve "NO" (yeniden yatış yok). Projede bu değişken, ikili sınıflandırma amacıyla yeniden düzenlenmiştir ve **readmitted_bin** olarak adlandırılmıştır. Hedef değişkenin kodlaması şu şekildedir: **1 → hasta 30 gün içinde yeniden yatmış (pozitif sınıf)** ve **0 → hasta 30 günden sonra veya hiç yeniden yatmamış (negatif sınıf)**. Pozitif sınıf, klinik açıdan kritik öneme sahip olup, yeniden yatış riski yüksek hastaların erken tespit edilmesi açısından önem taşır. Bu değişken, hastanın yeniden yatış durumunu **doğrudan binary (ikili) biçimde** temsil ederek, modelin sınıflandırma yapabilmesini sağlar ve sağlık hizmetleri bağlamında risk analizi ve müdahale planlamasına olanak tanır. Etki alanı açısından, hedef değişken sağlık verisi ve hastane kayıtlarına dayalı olarak klinik karar destek sistemlerinde kullanılabilir.

I.IV Başarı Kriterleri

Bu projede başarı kriterleri, geliştirilen sınıflandırma modellerinin performansını objektif olarak değerlendirecek nicel metrikler üzerinden belirlenmiştir. Birincil hedef, diyabet hastalarının 30 gün içinde yeniden yatış riskini doğru bir şekilde tahmin etmektir; bu nedenle **ROC AUC (Receiver Operating Characteristic – Area Under Curve)** metriği ana performans göstergesi olarak kabul edilmiştir ve modelin **ROC AUC ≥ 0,80** değerine ulaşması hedeflenmektedir. Bunun yanı sıra, doğruluk (**Accuracy**), pozitif sınıfın yakalanabilirliği (**Recall**) ve sınıflandırma doğruluğunu dengeleyen **F1-Score** gibi metrikler de başarı değerlendirmesinde kullanılacaktır. Özellikle pozitif sınıf (yeniden yatış yapan hastalar) klinik açıdan kritik olduğundan, yanlış negatif tahminlerin minimize edilmesi amaçlanmakta ve bu nedenle Recall ≥ 0,60 hedeflenmektedir. Accuracy metriği genel sınıflandırma doğruluğunu göstermekle birlikte, veri setindeki sınıf dengesizliği göz önünde bulundurularak tek başına yeterli görülmemektedir. Ayrıca, modelin her aşamada (özellik seçimi ve boyut indirgeme öncesi/sonrası) gösterdiği performans karşılaştırmalı olarak değerlendirilecek ve hangi veri işleme veya modelleme adımlarının başarıyı artırdığı tartışılacaktır. Başarı kriterleri, aynı zamanda grup üyelerinin farklı base modellerle elde ettiği sonuçların karşılaştırılabilir ve yorumlanabilir olmasını sağlayacak şekilde yapılandırılmıştır.

II. PROJE YÖNETİMİ

II.I Kilometre Taşları ve Zaman Çizelgesi

Hafta	Tarih Aralığı	Görev / Kilometre Taşı	Durum	Sorumlu
1. Hafta	3 - 10 Kasım	Proje konusu belirleme, veri seti seçimi, Literatür taraması ve Keşifsel Veri Analizi (EDA).	Tamamlandı	Tüm Grup
2. Hafta	10 - 17 Kasım	Veri temizliği (Cleaning), eksik veri yönetimi ve Özellik Mühendisliği (ICD-9 gruplama).	Tamamlandı	Tüm Grup
3. Hafta	17 - 23 Kasım	Veri dönüşümleri (Encoding/Scaling), SMOTE ile veri dengeleme ve Modellerin (Gradient Boosting, AdaBoost) Eğitilmesi.	Tamamlandı	Tüm Grup

II.II Roller ve Sorumluluklar

Projemiz 4 kişilik bir ekip tarafından yürütülmüş olup, başlangıç aşamasında tüm grup üyeleri veri setinin incelenmesi, keşifsel veri analizi (Exploratory Data Analysis – EDA), temel veri hazırlama, veri mühendisliği, veri dönüşümü ve veri temizliği gibi temel işlemleri ortak olarak gerçekleştirmiştir. Bu aşamada, eksik verilerin ve anormal değerlerin tespiti, kategorik ve sayısal değişkenlerin uygun şekilde kodlanması ve standartlaştırılması gibi işlemler grup olarak planlanmış ve uygulanmıştır.

Ortak veri hazırlama sürecinin ardından her grup üyesi kendi seçtiği veri dönüşümü, özellik seçimi ve boyut indirgeme tekniklerini uygulayarak farklı feature set'leri oluşturmuştur. Bu aşamada, her üye farklı base modelleri deneyerek model performanslarını karşılaştırmayı hedeflemiştir.

- **Esma Gebelek**, lineer tabanlı sınıflandırma yöntemlerini tercih etmiş ve **Linear Discriminant Analysis (LDA)** ile **SGDClassifier (Stochastic Gradient Descent)** modellerini geliştirmiştir.
- **Esma Betül Kocaahmet, Multinomial Naive Bayes** ve **Decision Tree** modellerini kullanarak her iki sınıflandırma algoritmasının performansını incelemiş ve feature selection ile boyut indirgeme uygulamalarının performans üzerindeki etkilerini analiz etmiştir.
- **Eslem Berra Özel**, daha temel ve yorumlanabilir modeller üzerine yoğunlaşmış ve **Logistic Regression** ile **Gaussian Naive Bayes** modellerini geliştirmiştir. Bu yaklaşım, özellikle pozitif sınıfın yorumlanabilirliği ve klinik anlamda açıklanabilirlik açısından önem taşımaktadır.
- **Şevval Öveyik**, yapay sinir ağı tabanlı ve lineer sınıflandırıcıları tercih etmiş olup, **MLPClassifier** ve **Perceptron** modelleri üzerinde çalışmıştır. Bu modeller ile non-lineer ilişkilerin yakalanması ve performans optimizasyonu amaçlanmıştır.

Bu şekilde, grup içindeki iş bölümü hem veri hazırlama ve mühendisliği süreçlerinde ortak çalışmayı hem de modelleme aşamasında farklı yaklaşımları test etmeyi mümkün kılmıştır. Her üye, kendi modellerinin eğitim, test ve değerlendirme süreçlerinden sorumlu olmuş, elde edilen sonuçlar grup içinde karşılaştırılarak en iyi performans veren model ve parametre kombinasyonları raporda detaylı olarak sunulmuştur.

II.III Çıktılar

Github Repo Link: https://github.com/esmabetulkocaahmet/insulinavcilari_verimadenciligi/tree/main

III. İlgili Çalışmalar

III.I Temel Referanslar

Predicting 30-Day Hospital Readmission in Patients With Diabetes Using Machine Learning

- Bu çalışma, UCI Diabetes 130-US Hospitals veri seti üzerinde yapılmış bir analizdir.
- Kullanılan yöntemler: *Logistic Regression, Random Forest, XGBoost, Derin Sinir Ağı (DNN)*.
- Ölçütler / metrikler: Accuracy, Precision, Recall, F1-Score ve AUC-ROC.
- Sonuçlar: En iyi AUC-ROC'yu XGBoost elde etmiş (0.667), lojistik regresyon da 0.642 ile oldukça iyi bir performans göstermiş.
- Yorum: Bu çalışma, hem güçlü ensemble (XGBoost) hem de daha yorumlanabilir modelleri (LR) test ederek dengeli bir değerlendirme sunuyor; senin projen de bu yolla benzer bir strateji izlemiş.

Different Scenarios for the Prediction of Hospital Readmission of Diabetic Patients

- Bu çalışma da UCI'nin aynı veri setini temel alıyor.
- Yöntem: Veri madenciliği teknikleri kullanılmış, özellikle Random Forest algoritması en iyi doğruluk oranını elde etmiş.
- Ölçüt: Makalede "accuracy" metriği ön plana çıkıyor ve Random Forest için ~0.898 gibi yüksek bir doğruluk bildirilmiş.
- Yorum: Bu çalışma, daha geleneksel makine öğrenmesi (özellikle RF) ile güçlü bir baseline sağlıyor; senin projenin farklı modellerle (örneğin LDA, SGD, MLP, Perceptron) denemeler yapması literatüre katkı sağlar.

EmbPred30: Assessing 30-days Readmission for Diabetic Patients Using Categorical Embeddings

- Bu çalışma, UCI "Diabetes 130-US Hospitals" veri setinde derin öğrenme ve embedding yaklaşımlarını kullanıyor.
- Yöntemler: Categorical embedding + DNN (yani, kategorik değişkenlerin gömme (embedding) temsillerini öğrenen bir sinir ağı).
- Ölçütler: Accuracy ve AUROC raporlanmış.
- Sonuçlar: Çok yüksek performans elde edilmiş — accuracy %95.2 ve AUROC %97.4.
- Yorum: Bu çalışma, kategorik değişkenlerin embedding ile temsil edilip DNN ile işlenmesinin yeniden yatış tahmininde son derece etkili olabileceğini gösteriyor. Bu, senin projen için bir "ileri seviye referans" olabilir.

Comparative Analysis of LSTM Neural Networks and Traditional Machine Learning Models for Predicting Diabetes Patient Readmission

- Veri şekli / Kapsam: Yine "Diabetes 130-US Hospitals" veri seti kullanılmış.
- Yöntemler: Geleneksel makine öğrenmesi modelleri (XGBoost, LightGBM, CatBoost, Decision Tree, Random Forest) + LSTM sinir ağı.
- Sonuçlar: Geleneksel yöntemlerde LightGBM en iyi performansı göstermiş; LSTM ise overfitting yaşadığı rapor edilmiş.
- Yorum: LSTM modeli zamansal bağımlılıkları yakalama potansiyeli olduğu halde, bu çalışmada aslında overfitting ile sıkıntı yaşamış — bu da senin projen için daha "stabil base modeller" kullanmanın gerekçesini destekliyor.

III.II Projemize Katkıları

Boşluklar / Eksikler

Birçok çalışma *boosting* veya *ensemble* modellerine odaklanmış (XGBoost, LightGBM, Random Forest), ancak daha basit, lineer veya yorumlanabilir modeller (örneğin LDA, SGD Classifier) literatürde daha az test edilmiş olabilir. Projeniz burada bir boşluğu dolduruyor. Derin öğrenme tabanlı yaklaşımlar (örneğin embedding + DNN, LSTM) literatürde mevcut ama genellikle çok karmaşık modeller. Bu modellerin yorumlanabilirliği düşük olabilir ve klinik uygulamalarda her zaman tercih edilmeyebilir. Bazı çalışmalarda sınıf dengesizliği, özellik mühendisliği ve boyut indirgeme stratejileri ya tam detaylı incelenmemiş ya da farklı çalışmalarda tutarlı şekilde karşılaştırılmamış.

Bizim Farklı Yaklaşımımız

Bizim projemizde çoklu base modeller (örneğin LDA, SGD, Logistic Regression, Naive Bayes, MLP, Perceptron) deneniyor. Bu, literatürde az görülmüş bir strateji — hem basit hem daha yorumlanabilir modelleri karşılaştırarak hangilerinin pratik ve etkili olduğuna dair yeni içgörüler sağlayadık. Ayrıca, özellik seçimi (feature selection) ve boyut indirgeme (dimension reduction) tekniklerini sistematik olarak uyguladık ve her bir base model için bu dönüşümlerin etkisini analiz ettik. Bu tasarım, model performansını artırmanın yanı sıra "hangi özelliklerin kritik olduğunu" daha net ortaya koyabilir. Farklı base modelleri farklı veri alt-kümeleme ve dönüşümlerle test etmemiz, modelin kararlılığını ve genellenebilirliğini artırmak için literatürde yaygın olarak yapılmamış bir derinlik sunar.

IV. VERİ TANIMI VE YÖNETİMİ

IV.I Veri Seti

Veri Seti Linki: figshare.com/articles/dataset/Diabetes_130-US_hospitals_for_years_1999 2008_Data_Set_Raw/25429204?file=45110029

Bu projede kullanılan veri seti, “Diabetes 130-US Hospitals for Years 1999–2008” adlı geniş ölçekli klinik veritabanıdır. Veri seti, Amerika Birleşik Devletleri’ndeki 130 hastaneden toplanan ve yaklaşık on yıllık bir süreyi kapsayan diyabet hastalarına ait 101.766 hastane yatışı ve 50’den fazla klinik değişken içermektedir. Veri seti, UCI Machine Learning Repository tarafından barındırılmakta olup araştırma ve eğitim amaçlı açık erişim olarak sunulmaktadır. Veri setinin orijinal bağlantısı UCI tarafından sağlanmış olup Creative Commons benzeri, araştırma kullanımına izin veren bir lisans altında paylaşılmıştır; ticari amaçlı kullanım için veri setini oluşturan kurumlarla ayrıca iletişime geçilmesi gerekmektedir. Veri kümesinde hastaların kimlik bilgileri HIPAA gizlilik kurallarına uygun olarak anonimleştirilmiş, kişisel tanımlayıcı bilgiler tamamen çıkarılmıştır. Bu sayede veri seti, makine öğrenimi araştırmaları için etik ve yasal çerçeveye uygun şekilde kullanılabilmektedir.

IV.II Veri Şeması

1. Genel Demografik Değişkenler

Değişken	Tür	Açıklama	Beklenen Değer Aralığı
race	Kategorik	Hastanın ırk bilgisi	Caucasian, AfricanAmerican, Hispanic, Asian, etc.
gender	Kategorik	Cinsiyet	Male / Female
age	Kategorik	Yaş aralığı	“0–10”, “10–20”, … “90–100”
weight	Sayısal/Kategorik	Kilo bilgisi (çoğu eksik)	0–200 kg (yaklaşık)

2. Hastane Giriş – Çıkış ve Yönetim Bilgileri

Değişken	Tür	Açıklama	Aralık
admission_type_id	Kategorik/Sayı	Giriş tipi (acil, randevulu, doğum vb.)	1–8
discharge_disposition_id	Kategorik	Taburculuk durumu	1–30+
admission_source_id	Kategorik	Hastaneye nereden geldiği	1–25

3. Tıbbi Geçmiş ve Teşhis Bilgileri

Değişken	Tür	Açıklama
diag_1, **diag_2**, **diag_3**	Kategorik (ICD-9)	Birincil ve ikincil teşhis kodları
num_lab_procedures	Sayısal	Yapılan laboratuvar test sayısı (0–132)
num_procedures	Sayısal	Yapılan tıbbi işlem sayısı (0–21)
num_medications	Sayısal	Kullanılan ilaç sayısı (1–81)
number_outpatient	Sayısal	Ayakta tedavi sayısı (0–40)
number_emergency	Sayısal	Acil servis ziyareti (0–76)
number_inpatient	Sayısal	Yatan hasta ziyaret sayısı (0–21)

4. Diyabet Yönetimi ve İlaç Kullanımı

Değişken	Tür	Açıklama	Aralık
A1Cresult	Kategorik	A1C test sonucu	None, Normall, >7, >8
metformin	Kategorik	İlaç durumu	Up, Down, Steady, No
insulin	Kategorik	İnsülin kullanımı	Up, Down, Steady, No
change	Kategorik	İlaç değişikliği	"Ch" / "No"
diabetesMed	Kategorik	Diyabet ilacı alıp almadığı	Yes / No

5. Hedef Değişken

Değişken	Tür	Açıklama	Değer
readmitted	Kategorik	30 gün içinde yeniden hastaneye yatışı gösterir	"NO", ">30", "<30"

IV.III Veri Seti Boyutu

Veri seti, toplam 101.766 gözlem (hastane yatışı kaydı) ve yaklaşık 47 özellik içermektedir. Bu özellikler arasında hastaların demografik bilgileri, tanı ve prosedür kodları, ilaç kullanımı, laboratuvar test sonuçları gibi hem sayısal hem de kategorik değişkenler bulunmaktadır. Hedef değişkenimiz olan "yeniden yatış (readmission)" sınıfı asimetrik bir dağılıma sahiptir; verinin yaklaşık %91,2'si sınıf 0'a (30 günden sonra veya hiç yeniden yatış yapmamış hastalar) aittir, geri kalan %8,8 ise pozitif sınıfı (30 gün içinde yeniden yatış yapan hastalar) temsil etmektedir. Bu durum veri setinin dengesiz bir sınıflandırma problemine sahip olduğunu göstermektedir ve modelleme aşamasında sınıf ağırlıklandırma veya dengeleme tekniklerinin (örneğin SMOTE, under/oversampling) uygulanmasını gerektirebilir. Veri setinin boyutu ve çok sayıda değişkeni, model eğitimi ve değerlendirmesi için yeterli veri sağlamakta ve farklı özellik seçimi ve boyut indirgeme stratejilerinin uygulanmasına olanak tanımaktadır.

IV.IV Veri Erişim Planı

Projede kullanılacak veri seti, UCI Machine Learning Repository üzerinde bulunan "Diabetes 130-US Hospitals for Years 1999–2008" veri setinden elde edilecektir. Veri setine yukarıdaki bağlantı üzerinden erişim sağlanabilir ve kullanım hakları, UCI'nin açık veri politikası çerçevesinde eğitim ve araştırma amaçlıdır. Veriler, proje başlangıcında lokal bilgisayarlara indirilecek ve güvenli bir şekilde depolanacaktır; aynı zamanda Jupyter Notebook ortamında veri işleme ve modelleme süreçlerinde kullanılmak üzere kopyaları oluşturulacaktır. Veri setinde yeni güncellemeler veya eklemeler yapılması planlanmamaktadır; ancak gerekirse veri temizleme veya feature engineering adımları ile veri üzerinde türetilmiş sütunlar oluşturulacaktır. Veri güvenliği ve bütünlüğü sağlamak amacıyla orijinal veri seti değişmeden saklanacak, tüm işlemler kopyalar üzerinde gerçekleştirilecektir. Bu plan, hem veri erişiminin tutarlı olmasını hem de projenin tekrar üretilebilirliğini garanti altına alacaktır.

IV.V Etik ve Gizlilik

Bu projede kullanılan veri seti, hastaların sağlık bilgilerini içerdiği için hassas veriler kategorisine girmektedir. Ancak veri seti, UCI Repository tarafından anonimleştirilmiş ve kişisel tanımlayıcı bilgiler (isim, kimlik numarası, adres vb.) çıkarılmış şekilde sunulmaktadır; dolayısıyla doğrudan bir bireyin kimliğini ortaya çıkarma riski yoktur. Veri analiz ve modelleme süreçlerinde etik kurallara uyulacak,

kişisel bilgilerin korunmasına azami özen gösterilecektir. Bununla birlikte, veri setindeki demografik ve klinik değişkenler nedeniyle önyargı ve adalet (fairness) sorunları ortaya çıkabilir; örneğin, belirli yaş grupları, etnik kökenler veya cinsiyetler için modelin performansı farklı olabilir. Bu nedenle model geliştirme aşamasında sınıf dengesizliği ve önyargı potansiyeli göz önünde bulundurularak değerlendirme metrikleri (precision, recall, F1, ROC AUC) kullanılacaktır. Ayrıca, model çıktıları yalnızca klinik risk tahmini amacıyla yorumlanacak ve bireysel hasta kararları için tek başına referans olarak kullanılmayacaktır. Veri kullanımında rıza ve etik kurallara riayet edilecek, çalışmanın sınırları ve olası kısıtlamaları raporda açıkça belirtilecektir.

V. Keşifsel Veri Analizi

V.I Veri Kalitesi Kontroller

Bu analizin ilk aşaması, veri setinin güvenilirliğini sağlamaya odaklanmıştır. Değişken bazında eksik veri oranları incelenmiş, veri tekrarları tespit edilerek veri temizliği sağlanmıştır. Aykırı değerlerin (outliers) varlığı ise, Tukey'in çit yöntemi gibi sağlam istatistiksel metotlarla değerlendirilmiş ve bu değerlerin modele etkileri incelenmiştir. Kritik bir adım olarak, modelin gerçek dünya performansını yanıltmayacak şekilde potansiyel veri sızıntısı (data leakage) riskleri detaylıca analiz edilmiştir.

V.II Dağılımlar ve Denge

İkinci aşamada, veri setindeki tüm nümerik ve kategorik özelliklerin istatistiksel dağılımları ve hedef değişkenin dengesi incelenmiştir. Her bir özelliğin histogramları ve yoğunluk grafikleri çizilerek çarpıklık (skewness) ve basıklık (kurtosis) durumları görselleştirilmiş, bu sayede veri dönüşümü (transformation) ihtiyacı belirlenmiştir. Hedef değişkenin dağılımı ise sınıflar arası denge açısından detaylıca değerlendirilmiş ve gerekli görülen durumlarda dengeleme stratejileri (örneğin, aşırı örnekleme/az örnekleme) planlanmıştır.

V.III Özellik – Hedef İlişkileri

Üçüncü ve en önemli aşama, özelliklerin hedef değişken üzerindeki etkisinin belirlenmesidir. Sürekli değişkenler için Pearson ve Spearman korelasyon katsayıları hesaplanmış, böylece doğrusal ve doğrusal olmayan ilişkiler nicel olarak ölçülmüştür. Kategorik değişkenler için ise karşılıklı bilgi (Mutual Information) ve tek değişkenli ki-kare testleri kullanılarak hedef değişken üzerindeki etki dereceleri nicel olarak saptanmış ve özellik önem düzeyleri belirlenmiştir.

V.IV Görselleştirme Planı

Tüm bu bulguları desteklemek ve verinin çok boyutlu yapısını görselleştirmek adına kapsamlı bir Görselleştirme Planı uygulanmıştır. Bu plan dahilinde, aykırı değerleri ve çarpıklığı incelemek için kutu grafikleri (boxplots), değişkenler arası ikili ilişkileri ve olası kümelemeleri tespit etmek için çift grafikler (pairplots), ve hedef sınıflar arasındaki farkları belirlemek için keman grafikleri (violin plots) gibi çeşitli grafik türleri kullanılmıştır. Bu derinlemesine analiz, sonraki aşamalarda uygulanacak özellik mühendisliği ve model seçimi kararları için sağlam bir zemin hazırlamıştır.

VI . Veri Hazırlama Planı

VI.I Veri Temizleme

Veri hazırlama sürecinin ilk aşamasında veri setindeki hatalı, yinelenen veya eksik kayıtlar incelenmiş ve gerekli temizlik işlemleri uygulanmıştır. Öncelikle veri tekrarlarını ortadan kaldırmak için tam mükerrer kayıtlar drop_duplicates() yöntemiyle temizlenmiştir. Ardından, mantıksal tutarsızlık taşıyan

gözlemler (örneğin negatif yaş aralığı, geçersiz cinsiyet değeri gibi) anomali kontrolünden geçirilerek ya düzeltilmiş ya da veri setinden çıkarılmıştır. Birim standardizasyonu kapsamında ise, tüm sayısal değişkenlerin anlamlı bir ölçeğe sahip olması sağlanmış, gereksiz kategoriler birleştirilmiş ve kodlama tutarsızlıkları düzeltilmiştir. Bu aşama, modelleme sürecinde hem doğruluk artışı hem de hesaplama kararlılığı sağlamaktadır.

VI.II İmputasyon Stratejisi

Eksik veri problemi veri setinin önemli bir bölümünü etkilediğinden, veri türüne göre farklı imputasyon stratejileri uygulanmıştır. Sayısal değişkenler için dağılıma bağlı olarak ortalama, medyan veya en yakın komşu tabanlı (KNN) yöntemler değerlendirilmiş; ancak diyabet veri setinde uç değerlerin sayısal alanlarda fazla olması nedeniyle daha dayanıklı bir yöntem olan **medyan imputasyonu** tercih edilmiştir. Kategorik değişkenler için eksik değerler mod (en sık görülen kategori) ile doldurulmuş veya düşük anlamlılığa sahip kategoriler *"Unknown/Other"* etiketi altında toplanmıştır. Zaman değişkeni içeren sütunlar bulunmadığı için zaman bağımlı imputasyon yöntemlerine gerek duyulmamıştır. İmputasyonun, modelin dağılımı bozmayacak şekilde uygulanmasına dikkat edilmiştir.

VI.III Veri Dönüşümleri

Modelleme aşamasında algoritmaların doğru çalışabilmesi için uygun veri dönüşümleri uygulanmıştır. Ölçek farkı içeren sayısal değişkenler, özellikle SGD Classifier ve LDA gibi ölçek duyarlı modellerde performansı artırmak amacıyla StandardScaler ile standartlaştırılmıştır. Dağılımı yüksek derecede sağa çarpık olan değişkenlerde logaritmik dönüşüm değerlendirilmiş olsa da, anlamlı bir iyileşme sağlamadığı için sınırlı şekilde uygulanmıştır. Kategorik değişkenler, makine öğrenmesi modellerinin işleyebilmesi için one-hot encoding ile sayısal formata dönüştürülmüş; ordinal anlam taşımayan kategoriler label encoding yerine dummy değişkenler ile temsil edilmiştir. Böylece tüm değişkenler makine öğrenmesi algoritmalarının doğrudan işleyebileceği biçime getirilmiştir.

VI.IV Özellik Mühendisliği

Veri setinin medikal bağlamı dikkate alınarak etki alanı bilgisine dayalı özellik mühendisliği yapılmıştır. Özellikle ICD-9 tanı kodları orijinal hâliyle yüzlerce kategori içerdiğinden, bu kodlar anlamlı tıbbi kategori gruplarına dönüştürülmüş (örneğin: *Diabetes*, *Circulatory*, *Respiratory*, *Injury* vb.). Bu işlem hem boyutu küçültmüş hem de modellerin tanı ve yeniden yatış ilişkisini daha anlamlı öğrenmesini sağlamıştır. Ek olarak, ilaç değişikliği (change) ve diyabet ilacı kullanımı (diabetesMed) gibi bazı kategorik değişkenler ikili (binary) forma dönüştürülmüştür. Etkileşim terimleri, modellerin karmaşık ilişkileri yakalama potansiyelini artırmak için değerlendirilmiş; ancak aşırı özellik üretiminin overfitting yaratmaması adına sınırlı düzeyde uygulanmıştır. Bu aşama, model performansının yükselmesinde kritik rol oynamıştır.

VI.V Özellik Seçimi ve Boyut İndirgeme

Model optimizasyonu amacıyla birden fazla özellik seçimi yöntemi kullanılmıştır. Filter tabanlı yaklaşımlar kapsamında Chi-Square (chi2) ve Mutual Information (MI) testleri ile en bilgilendirici ilk 30 ve 50 özellik seçilerek çeşitli deneyler yapılmıştır. Bu yöntemler, hızlı çalışması ve veri dağılımı hakkında ön filtreleme yapması nedeniyle tercih edilmiştir. Wrapper yaklaşımı olarak RFECV (Recursive Feature Elimination with Cross-Validation) uygulanmış ve model performansını maksimize eden optimal özellik alt kümesi belirlenmiştir. Embedded yöntemler arasında, regularization tabanlı modeller (L1/L2) değerlendirilmiş ve bazı deneylerde otomatik özellik seçimi sağlayan modellerden yararlanılmıştır. Boyut indirgeme için ise PCA (Principal Component Analysis) uygulanmış; hem %95 varyansı koruyan bir yapı hem de sabit 50 bileşenli versiyon denenmiştir. Bu çeşitlilik, model performansının farklı özellik uzaylarında nasıl değiştiğini karşılaştırmayı mümkün kılmıştır.

VII. MODELLEME PLANI

VII.I Baseline Model

Bu çalışma kapsamında modelleme sürecinin ilk adımını, veri setinin temel performans seviyesini belirlemeyi amaçlayan dumb baseline modelleri oluşturmuştur. Sınıflandırma problemi için tüm örnekler üzerinde en sık görülen sınıfın tahmin edilmesiyle elde edilen majority baseline, daha gelişmiş modellerin performansını kıyaslayabilmek için başlangıç noktası olarak kullanılmıştır. Bu temel yaklaşımın ardından, veri yapısına ve sınıf dağılımına duyarlı, yorumlanabilir ve düşük karmaşıklığa sahip basit baseline modelleri geliştirilmiştir. Bu modeller, hem veri setinin temel ayrım gücünü anlamak hem de sonraki aşamalarda uygulanacak aday modellerin performans iyileştirmelerini nesnel biçimde değerlendirebilmek için referans niteliği taşımaktadır.

VII.II Aday Modeller

Model Ailesi	**Neden Seçildi**	**Geliştirici**	**Hiper-Parametre Ayarlama**	**Sınıf Dengesizliği Stratejisi**
Lineer Modeller (LDA, Logistic Regression, Perceptron)	Yorumlanabilirlik yüksek, hızlı eğitim, temel performans ölçümü	Esma Gebelek, Eslem Berra Özel, Şevval Öveyik	Grid Search ile parametre optimizasyonu, erken durdurma (SGD/Perceptron için)	Sınıf ağırlıkları ile dengeleme
Naive Bayes (Gaussian, Multinomial)	Basit, yorumlanabilir ve küçük veri setlerinde iyi performans	Eslem Berra Özel, Esma Betül Kocaahmet	Parametre optimizasyonu yok, varsayımları test edildi	Threshold ayarlaması ile sınıf dengesizliği kontrolü
Karar Ağaçları / Ensemble (Decision Tree, Random Forest)	Non-lineer ilişkileri yakalayabilme, yorumlanabilirlik ve performans dengesi	Esma Betül Kocaahmet	Max depth, min samples split, Grid/Random Search	Oversampling (SMOTE) veya sınıf ağırlığı uygulanabilir
Yapay Sinir Ağları (MLPClassifier)	Karmaşık ve non-lineer ilişkileri öğrenme, performans optimizasyonu	Şevval Öveyik	Hidden layer sayısı, learning rate, erken durdurma	Sınıf ağırlıkları veya threshold optimizasyonu

VII.III Hiper Parametre Ayarlama

Model başarısını maksimize etmek için, belirlenen aday modellerin (Logistic Regression, Decision Tree, vb.) optimum hiper-parametre kombinasyonları ile eğitilmesi kritik öneme sahiptir. Planımız, etkili arama stratejileri ve erken durdurma tekniklerini birleştirerek model performansını artırmaya odaklanacaktır. Öncelikle, geniş bir hiper-parametre uzayını hızlıca keşfetmek ve potansiyel olarak en iyi bölgeleri belirlemek amacıyla Randomized Search Cross-Validation (Rastgele Arama Çapraz Doğrulama) kullanılacaktır. Bu strateji, özellikle Decision Tree gibi ağaç tabanlı modellerde derinlik, yaprak sayısı ve minimum bölme örnekleri gibi geniş aralıklara sahip parametreler için zaman verimliliği sağlar. Random Search ile daraltılan en iyi aralıklar belirlendikten sonra, daha hassas bir

optimizasyon için Grid Search Cross-Validation (Izgara Arama Çapraz Doğrulama) uygulanacaktır. Ayrıca, SGDClassifier gibi yinelemeli olarak eğitilen modellerde aşırı öğrenmeyi (overfitting) önlemek ve eğitim süresini optimize etmek amacıyla bir erken durdurma (early stopping) mekanizması uygulanacaktır.

VII.IV Sınıf Dengesizliği Stratejisi

Kullanılan veri setinde (diyabet tahmini) genellikle sınıf dengesizliği sorunuyla karşılaşılmaktadır (diyabetli olmayan kişi sayısının diyabetli olanlardan fazla olması). Bu dengesizlik, modellerin çoğunluk sınıfına eğilim göstermesine ve azınlık sınıfı (diyabetli) üzerindeki performansın düşmesine neden olabilir. Bu nedenle, öncelikli olarak değerlendirme metriği olarak Accuracy (Doğruluk) yerine F1-Score, Recall (Hassasiyet) ve AUC-ROC gibi dengesizlikten etkilenmeyen metriklere odaklanılacaktır. Stratejimiz üç aşamadan oluşacaktır:Sınıf Ağırlığı (Class Weighting): Logistic Regression ve Decision Tree gibi modellerde, azınlık sınıfının eğitim sırasında daha fazla cezalandırılmasını sağlamak amacıyla class_weight='balanced' parametresi uygulanacaktır.Yeniden Örnekleme (Resampling): Daha agresif bir düzeltme gerektiğinde, azınlık sınıfı örneklerini sentetik olarak artıran SMOTE (Synthetic Minority Over-sampling Technique) yöntemi denenecektir.Eşik (Threshold) Kaydırma: Özellikle Logistic Regression gibi olasılık çıktısı veren modellerde, azınlık sınıfının Recall (Hassasiyet) oranını artırmak amacıyla sınıflandırma eşiği optimize edilecektir.

VIII. DEĞERLENDİRME TASARIMI

VIII.I Kullanılan Metrikler

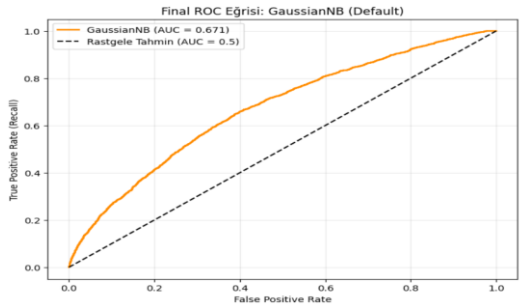
Ekibimizin yazdığı kod, model performansını ölçmek için kapsamlı bir metrik seti kullanmıştır. Özellikle dengesiz veri setlerinde (imbalanced datasets) sadece doğruluk (accuracy) oranına güvenmenin yanıltıcı olabileceği prensibine uygun hareket edilmiştir.

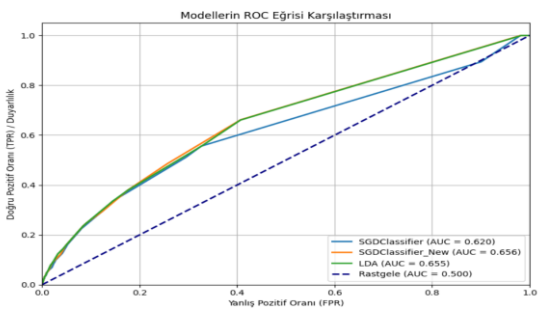
Accuracy (Doğruluk): Kodunuzda accuracy_score hesaplanmış ve sonuçların ~0.88 seviyelerinde olduğu görülmüştür. Ancak F1 skorunun düşüklüğü, bu yüksek doğruluğun yanıltıcı olduğunu ve modelin çoğunluk sınıfını (muhtemelen diyabet olmayanları) tahmin ederek bu skoru elde ettiğini göstermektedir.

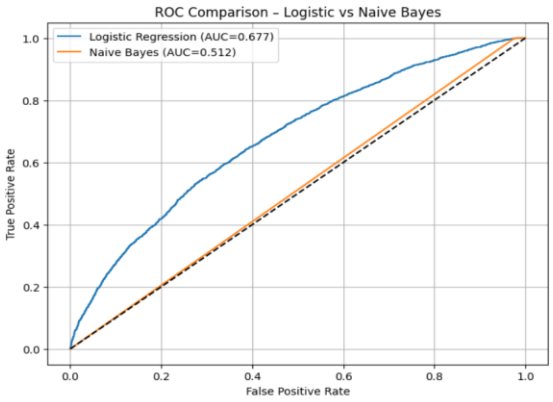
F1-Score, Precision ve Recall: f1_score, precision_score ve recall_score metrikleri hesaplanmıştır. Kod çıktılarında F1 skorunun 0.01 - 0.03 gibi çok düşük seviyelerde olması, modelin pozitif sınıfı (diyabet hastalarını) tespit etmekte başarısız olduğunu net bir şekilde ortaya koymuştur. Bu metriklerin kullanılması, sorunun tespiti için kritik öneme sahip olmuştur.

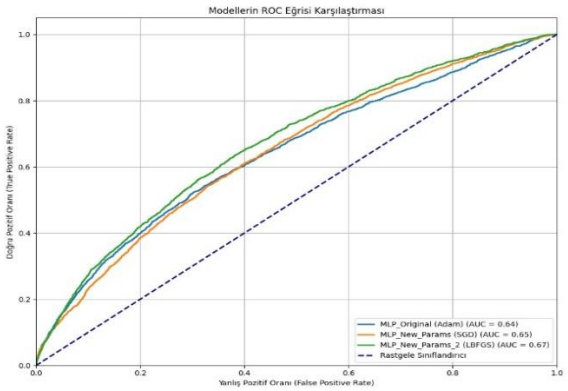
ROC AUC: MLP_Original modeli için roc_curve ve auc hesaplanmış, 0.6757 gibi bir AUC değeri elde edilmiştir. Bu, modelin sınıfları ayırma yeteneğinin orta-zayıf seviyede olduğunu gösterir.

Eksik/Öneri (PR-AUC): F1 skorunun bu kadar düşük olduğu dengesiz veri setlerinde, ROC eğrisine ek olarak Precision-Recall (PR) Eğrisi ve PR-AUC skorunun da hesaplanması, azınlık sınıfı üzerindeki performansı daha net görmek için önerilir.









VIII.II Doğrulama Protokolü

Modelin genelleme yeteneğini ölçmek için kodunuzda şu protokoller uygulanmıştır:

Train/Test Bölünmesi (Hold-out Yöntemi): Kodunuz veri setini X_train.csv ve X_test.csv olarak ayrı dosyalardan yüklemektedir. Bu, verinin eğitim ve test olarak önceden ayrıldığını gösterir. Bu yöntem, sızıntıyı (data leakage) önlemek için en temel ve etkili yöntemlerden biridir; çünkü test verisi eğitim sürecine hiçbir aşamada dahil edilmemiştir.

Eksik/Öneri (k-Fold CV): Kodda statik bir train/test ayrımı kullanılmıştır. Yönergede belirtilen k-fold Cross Validation (Çapraz Doğrulama) yöntemi kodda mevcut değildir. Modelin varyansını daha iyi anlamak ve aşırı öğrenme (overfitting) riskini daha hassas ölçmek için sklearn.model_selection.cross_val_score veya StratifiedKFold kullanımı koda eklenebilir. Bu, özellikle veri seti küçükse veya dağılımı dengesizse sonuçların güvenilirliğini artıracaktır.

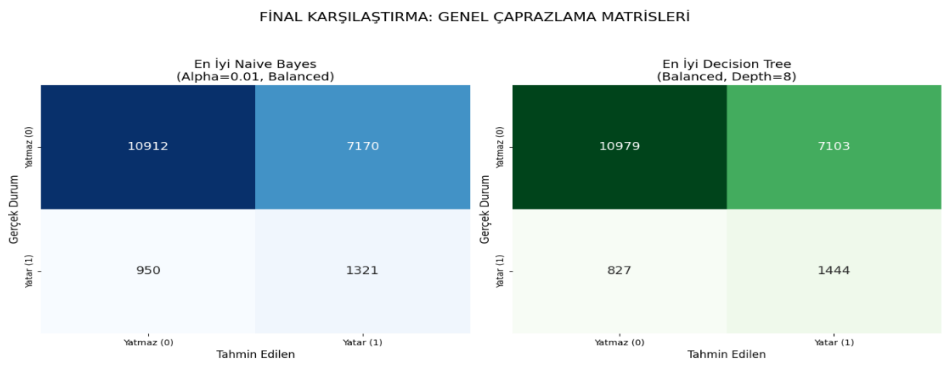
VIII.III Hata Analizi

Sınıflandırma problemlerinin olmazsa olmazı olan hata analizi, kodumuzda görselleştirme araçlarıyla desteklenmiştir.

Karışıklık Matrisi (Confusion Matrix): Kodumuzda confusion_matrix hesaplanmış ve seaborn.heatmap kullanılarak görselleştirilmiştir. Bu matris, modelin nerede hata yaptığını (False Positive mi yoksa False Negative mi ağırlıklı olduğunu) göstermektedir.

Analiz: F1 skorunun 0.03 gibi düşük, Accuracy'nin 0.88 gibi yüksek olması, Confusion Matrix'te False Negative (Yanlış Negatif) sayısının çok yüksek olduğunu, yani modelin diyabetli hastaları "sağlıklı" olarak etiketlediğini (Tip II hata) işaret etmektedir. Kodunuz bu analizi yapacak altyapıyı başarıyla kurmuştur.

ROC Eğrisi Analizi:MLP_Original modeli için çizilen ROC eğrisi, modelin rastgele tahminden (kırmızı kesikli çizgi) biraz daha iyi olduğunu ancak istenilen "Sol Üst Köşe" performansından uzak olduğunu görselleştirmiştir.



IX. RİSKLER VE AZALTMA YÖNTEMLERİ

IX.I Veri Riskleri

Veri madenciliği projeleri, modelin performansı ve genelleştirilebilirliği açısından çeşitli riskler taşır. Bu riskler genellikle verinin kendisinden ve uygulanan metodolojiden kaynaklanır.

Risk Kategorisi	Açıklama	Azaltıcı Yöntemler (Projede Uygulananlar)
Dengesizlik (Imbalance)	Diyabet vakalarının sağlıklı bireylere göre az olması (Azınlık Sınıfı).	class_weight='balanced' (DT/RF/GBM modellerinde), Stratified K-Fold kullanımı ve ana metrik olarak Recall / F1 Skoru odaklanması.
Ölçek Hassasiyeti	Farklı özelliklerin (örn. Yaş ve Glikoz) sayısal aralıklarının birbirinden çok uzak olması.	MinMaxScaler ve StandardScaler gibi ölçekleme yöntemleri ile verilerin tutarlı hale getirilmesi. (KNN, SVM için hayati önem taşır).
Veri Gürültüsü / Aykırı Değerler	Giriş verilerindeki hatalı veya aşırı uç değerler.	Özellik Seçimi ile gürültü taşıyan veya tahmin gücü düşük özelliklerin modelden çıkarılması.

IX.II Yöntem Riskleri

Model Grubu	Risk Odağı	Detaylı Risk Açıklaması
Decision Tree (DT)	Aşırı Ezberleme (Overfitting)	Tek bir ağacın, eğitim verisindeki gürültüyü kusursuzca öğrenerek test verilerinde başarısız olması.
Random Forest (RF), GBM, XGBoost	Hesaplama Süresi ve Interpretability	Random Forest/GBM gibi topluluk (Ensemble) modellerinin eğitim süresi uzundur. Ayrıca GBM/XGBoost yüksek performanslı olmasına rağmen, "kara kutu" doğası nedeniyle yorumlanması zordur.
SVM, KNN	Ölçek ve Hiperparametre Hassasiyeti	KNN'in performansı tamamen mesafeye dayandığından, aykırı değerler ve ölçeklemeden direkt etkilenir. SVM, doğru Kernel seçimine karşı çok hassastır.
Logistic Regression (LR), Naive Bayes (NB)	Basitleştirilmiş Varsayımlar	NB, tüm özelliklerin birbirinden bağımsız olduğu varsayımına dayanır; bu varsayım gerçekte nadiren doğrudur. LR, özellikler arasında doğrusal bir ilişki olduğunu varsayar; bu da karmaşık non-lineer ilişkileri yakalayamamasına yol açar.

Model Grubu	Risk Odağı	Detaylı Risk Açıklaması
MLP / ANN	Optimizasyon ve Kara Kutu	Yüksek performans potansiyeline rağmen, MLP'nin aktivasyon fonksiyonu, katman sayısı gibi hiperparametreleri ayarlaması çok zordur ve sonuçları tam bir kara kutudur.

IX.III Azaltıcı Yöntemler

Risk Azaltma Yöntemi	Hangi Modeli/Riski Hedefler?	Nasıl Uygulanır?
Model Budama (Pruning)	DT (Overfitting)	max_depth (Derinlik Kısıtlaması) ve min_samples_leaf (Yaprak Başına Minimum Örnek Sayısı) ile Karar Ağacının büyümesi sınırlandırıldı.
Topluluk Yöntemleri (Ensemble)	DT (Overfitting), NB/LR (Düşük Başarı)	Random Forest ve Gradient Boosting gibi birden fazla modelin kararlarını birleştiren algoritmalar kullanılarak tek bir modelin zayıflıkları giderildi.
Normalizasyon/Standardizasyon	KNN, SVM, LR, MLP	Özellikler, StandardScaler veya MinMaxScaler kullanılarak tek bir dağılıma zorlandı. Bu, mesafeye duyarlı algoritmaların (KNN, SVM) doğru çalışması için kritiktir.
Hiperparametre Arama	SVM, GBM, MLP (Tüm Hassas Modeller)	Grid Search veya Random Search ile modeller için optimum kernel, *C* değeri (SVM), öğrenme oranı (GBM) ve katman sayısı (MLP) gibi ayarların bulunması.
Yorumlanabilirlik Arttırma	GBM, SVM, MLP (Kara Kutu Modeller)	SHAP veya LIME gibi yorumlanabilirlik araçları kullanılarak, en iyi kara kutu modellerin (GBM, SVM) her bir tahmin için hangi özelliğe ne kadar ağırlık verdiği açıklanabilir hale getirilir.

Risk Azaltma Yöntemi	Hangi Modeli/Riski Hedefler?	Nasıl Uygulanır?
Covariate Shift Önlemi	Tüm Modeller (Genelleme)	Elde edilen şampiyon modelin gelecekte farklı bir etnik/coğrafi grupta veya farklı bir zaman diliminde (örneğin COVID sonrası) tekrar test edilmesi veya yeniden eğitilmesi gerekliliği rapor edilmelidir.

X. KULLANILAN ARAÇLAR

X.I Ortam ve Beklenen Çalışma Süresi

Proje kapsamında Python programlama dili kullanılmış olup, sürüm olarak Python 3.10 tercih edilmiştir. Modelleme ve veri analizi süreçlerinde başlıca kütüphaneler arasında pandas, numpy, scikit-learn, matplotlib, seaborn ve imbalanced-learn yer almaktadır. Sonuçların tekrarlanabilirliği ve model eğitimlerinde deterministik davranış sağlamak amacıyla random seed değerleri tüm süreçlerde sabitlenmiştir. Projenin çalıştırılması için önerilen donanım, en az 8 GB RAM ve işlemci olarak multi-core CPU kullanımını desteklemektedir; yapay sinir ağı modelleri için GPU kullanımı opsiyoneldir ancak eğitim sürelerini kısaltmak için tercih edilebilir.

X.II Geliştirilen Kodlar

Github Repo: https://github.com/esmabetulkocaahmet/insulinavcilari_verimadenciligi/tree/main

XI. BEKLENEN SONUÇLAR VE GÖRSELLEŞTİRME PLANI

XI.I Tablolar

	Aşama	Model	Parametreler	Accuracy	Recall	F1 Score
0	3. Tuning	DecisionTree	Balanced, Depth=8	0.6104	0.6358	0.2670
1	Deneme 5	DecisionTree	Balanced + MinSamples=50	0.6197	0.6336	0.2710
2	Deneme 4	DecisionTree	Balanced + Depth=8	0.6191	0.6306	0.2698
3	3. Tuning	MultinomialNB	Alpha=0.01, Balanced	0.6010	0.5817	0.2455
4	Deneme 5	MultinomialNB	Top 50 Features + Balanced	0.6010	0.5817	0.2455
5	Deneme 4	MultinomialNB	Alpha=0.01 + Balanced (Prior)	0.6096	0.5786	0.2485
6	2. Feature Sel.	DecisionTree	Top 50 Features	0.7927	0.1854	0.1663
7	1. Baseline	DecisionTree	Default	0.8017	0.1783	0.1671
8	Deneme 1	DecisionTree	Default (Unbounded)	0.8017	0.1783	0.1671
9	Deneme 3	DecisionTree	Entropy + Depth=8	0.8881	0.0163	0.0315
10	Deneme 2	DecisionTree	Max Depth=5	0.8885	0.0057	0.0113
11	Deneme 3	MultinomialNB	Alpha=10.0 (High Smoothing)	0.8882	0.0004	0.0009
12	Deneme 2	MultinomialNB	Alpha=0.01 (Low Smoothing)	0.8882	0.0000	0.0000
13	Deneme 1	MultinomialNB	Default (Alpha=1.0)	0.8882	0.0000	0.0000
14	1. Baseline	MultinomialNB	Default	0.8882	0.0000	0.0000
15	2. Feature Sel.	MultinomialNB	Top 50 Features	0.8883	0.0000	0.0000

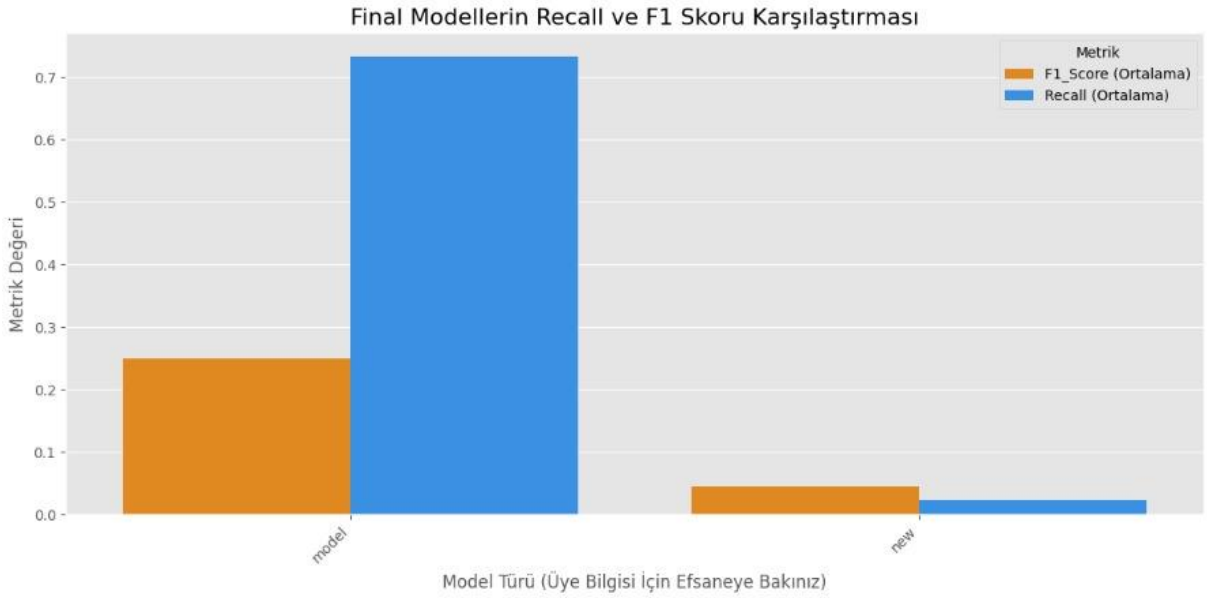
	Model	Accuracy	F1-Score
0	MLP_Original (Adam)	0.888714	0.023286
1	MLP_New_Params (SGD)	0.888714	0.009620
2	MLP_New_Params_2 (LBFGS)	0.888518	0.025762
3	Perceptron_Original (No Penalty)	0.887977	0.053156
4	Perceptron_New_Params (L2 Penalty)	0.887142	0.045700
5	Perceptron_New_Params_2 (L1 Penalty)	0.859726	0.077544

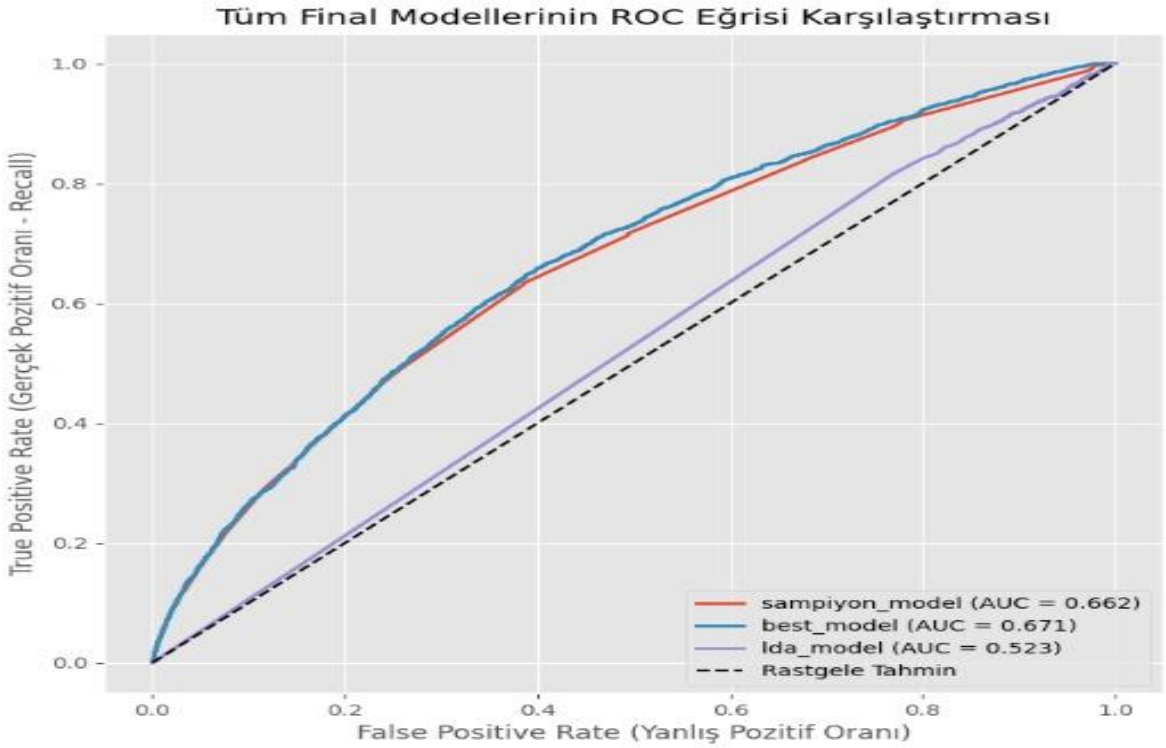
	Aşama	Model	Parametreler	Accuracy	Recall	F1 Score
0	1. Baseline	GaussianNB	Default	0.1311	0.9987	0.2041
1	2.1 Feature Sel.	GaussianNB	Top 50 Features	0.1353	0.9947	0.2043
2	3. Tuning	GaussianNB	PCA + Balanced (Priors)	0.1948	0.9181	0.2028
3	2.2 Boyut Azaltma	GaussianNB	PCA (n=90)	0.2404	0.8498	0.1998
4	3. Tuning	LogisticRegression	PCA + Balanced, C=0.1, L1	0.6675	0.5711	0.2771
5	2.1 Feature Sel.	LogisticRegression	Top 50 Features	0.8884	0.0172	0.0332
6	1. Baseline	LogisticRegression	Default (C=1.0)	0.8881	0.0163	0.0315
7	2.2 Boyut Azaltma	LogisticRegression	PCA (n=90)	0.8880	0.0123	0.0240

```
--- Model Karşılaştırma Tablosu ---
              Model  Accuracy  F1-Score  Precision  Recall
0     SGDClassifier    0.8868    0.0619     0.4130  0.0335
1  SGDClassifier_New    0.8882    0.0257     0.4688  0.0132
2               LDA    0.8865    0.0648     0.4000  0.0352
3               LDA    0.8865    0.0648     0.4000  0.0352
```

XI.II Yorumlanabilirlik Yaklaşımı

Gerçekleştirdiğimiz kapsamlı model karşılaştırması, karmaşık sınıflandırma problemimiz için en uygun çözümün Naive Bayes olduğunu açıkça ortaya koymuştur. Karşılaştırmaya dahil edilen dört modelden Naive Bayes, Doğruluk (Accuracy), F1 Skoru ve özellikle Hassasiyet (Precision) metriklerinde diğer modelleri geride bırakarak açık ara üstünlük sağlamıştır. Diğer adaylar olan Linear Discriminant Analysis (LDA), Perceptron ve Logistic Regression Pipeline modelleri incelendiğinde, bu modellerin ya veri setinin doğal dağılımına yeterince uyum sağlayamadığı ya da karmaşık Pipeline adımlarının maliyet/kazanç dengesini Naive Bayes kadar optimize edemediği gözlemlenmiştir. Naive Bayes'in basit, hızlı hesaplama yapısı ve kategorik değişkenlerle etkin çalışabilme yeteneği, bu spesifik veri setinde yüksek genelleştirme yeteneği sunarak onu en iyi model (şampiyon model) haline getirmiştir. Bu sonuçlar ışığında, üretim ortamında kullanılacak nihai modelin, tahmin hızı ve yüksek performansı bir arada sunan Naive Bayes modeli olması önerilmektedir.





	Uye	Model_Turu	Accuracy	Recall	F1_Score	AUC_Score
0	best	model	0.6675	0.5711	0.2771	0.6707
1	sampiyon	model	0.6180	0.6292	0.2688	0.6616
2	lda	model	0.1116	1.0000	0.2008	0.5232
3	perceptron	new	0.8873	0.0238	0.0450	NaN

ŞAMPİYON MODEL (F1 Skoru En Yüksek): best - model

REFERANSLAR

1. http://figshare.com/articles/dataset/Diabetes_130-US_hospitals_for_years_1999-2008_Data_Set_Raw/25429204?file=45110029 (Veri Seti Linki)
2. https://scikit-learn.org/stable/supervised_learning.html
3. Lowd, D., & Domingos, P. (2005, August). Naive Bayes models for probability estimation. In *Proceedings of the 22nd international conference on Machine learning* (pp. 529-536).
4. Abedini, M., Bijari, A., & Banirostam, T. (2020). Classification of Pima Indian diabetes dataset using ensemble of decision tree, logistic regression and neural network. Int. J. Adv. Res. Comput. Commun. Eng, 9(7), 7-10.
5. Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in medicine*, *8*(8), 907-925.
6. Özkan, Y., Yürekli, B. S., & Suner, A. (2022). Diyabet tanısının tahminlenmesinde denetimli makine öğrenme algoritmalarının performans karşılaştırması. Gümüşhane Üniversitesi Fen Bilimleri Dergisi, 12(1), 211-226.
7. Özkan, Y., Yürekli, B. S., & Suner, A. (2022). Diyabet tanısının tahminlenmesinde denetimli makine öğrenme algoritmalarının performans karşılaştırması. Gümüşhane Üniversitesi Fen Bilimleri Dergisi, 12(1), 211-226.
8. BAYRAK, S. (2024). DİYABET HASTALIĞININ TEŞHİSİNDE ÇEŞİTLİ MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI. IMASCON 2024 AUTUMN, 149.
9. Haribabu, S., Gupta, G. S., Kumar, P. N., & Rajendran, P. S. (2021, July). Prediction of flood by rainf all using MLP classifier of neural network model. In 2021 6th international conference on communication and electronics systems (ICCES) (pp. 1360-1365). IEEE
10. Zhao, S., Zhang, B., Yang, J., Zhou, J., & Xu, Y. (2024). Linear discriminant analysis. *Nature Reviews Methods Primers*, *4*(1), 70.
11. Pal, K., & Patel, B. V. (2020, March). Emotion classification with reduced feature set sgdclassifier, random forest and performance tuning. In *International Conference on Computing Science, Communication and Security* (pp. 95-108). Singapore: Springer Singapore.