BARTIN ÜNİVERSİTESİ

MÜHENDİSLİK, MİMARLIK VE TASARIM FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

VERİ MADENCİLİĞİ PROJE RAPORU

Konu: Banka Müşteri Terk Analizi ve Tahmini

MEHMET İPEK	-------	23010310039
MUSTAFA BATIN KAYNAK	-------	24640310002
FURKAN ÇAT	-------	23010310022
SİNA RAHBARİBANAEİAN	-------	23670310047

DANIŞMAN

Dr. Öğr. Üyesi Ümit DEMİRBAGA

1. Giriş

Günümüzün yoğun rekabet koşullarında, finans sektöründe faaliyet gösteren bankalar için yeni bir müşteri kazanmanın maliyeti, mevcut bir müşteriyi elde tutma maliyetinden çok daha yüksektir. Bu durum, **"Müşteri Terki" (Customer Churn)** olgusunu bankaların stratejik yönetim planlarının merkezine yerleştirmektedir. Müşteri terki, bir kullanıcının banka ile olan ticari ilişkisini tamamen kesmesi veya aktif kullanımını durdurması olarak tanımlanmaktadır.

Bu projenin temel amacı, bir bankanın veri setinde yer alan demografik ve finansal öznitelikleri kullanarak, hangi müşterilerin bankadan ayrılma (churn olma) potansiyeli taşıdığını önceden tahmin etmektir. Veri madenciliği teknikleri kullanılarak gerçekleştirilen bu analiz süreci; müşterilerin yaş, bakiye, ürün sayısı, kredi skoru ve aktiflik durumu gibi kritik değişkenlerin "terk etme" eylemi üzerindeki etkilerini belirlemeyi hedefler.

Proje kapsamında yürütülen süreç şu aşamaları içermektedir:

- **Veri Kalitesinin Artırılması:** Gerçek hayat verilerini simüle etmek amacıyla oluşturulan eksik verilerin uygun yöntemlerle doldurulması ve aykırı değerlerin (outliers) temizlenmesi.
- **İstatistiksel Analiz ve Görselleştirme:** Değişkenler arasındaki korelasyonların ve dağılımların incelenmesi.
- **Öngörücü Modelleme:** Karar Ağaçları ve Random Forest gibi makine öğrenmesi algoritmalarının yanı sıra, değişkenler arasındaki nedensel ilişkileri modellemek için **Bayesci Ağlar (Bayesian Networks)** yönteminin uygulanması.

Bu rapor, elde edilen bulguların analiz edilmesini, modellerin performans karşılaştırmalarını ve bankanın müşteri sadakatini artırmak için alabileceği aksiyonlara yönelik stratejik önerileri kapsamaktadır. Yapılan bu çalışma, bankanın riskli müşteri gruplarını erkenden tespit ederek "Churn Önleme Programları" oluşturmasına temel teşkil edecektir.

2. Veri Seti Tanımı

	customer_id	credit_score	country	gender	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
0	15634602	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	15647311	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	15619304	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	15701354	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	15737888	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
...	...	...	...	...	...	...	...	...	...	...	...	...
9995	15606229	771	France	Male	39	5	0.00	2	1	0	96270.64	0
9996	15569892	516	France	Male	35	10	57369.61	1	1	1	101699.77	0
9997	15584532	709	France	Female	36	7	0.00	1	0	1	42085.58	1
9998	15682355	772	Germany	Male	42	3	75075.31	2	1	0	92888.52	1
9999	15628319	792	France	Female	28	4	130142.79	1	1	0	38190.78	0

10000 rows × 12 columns

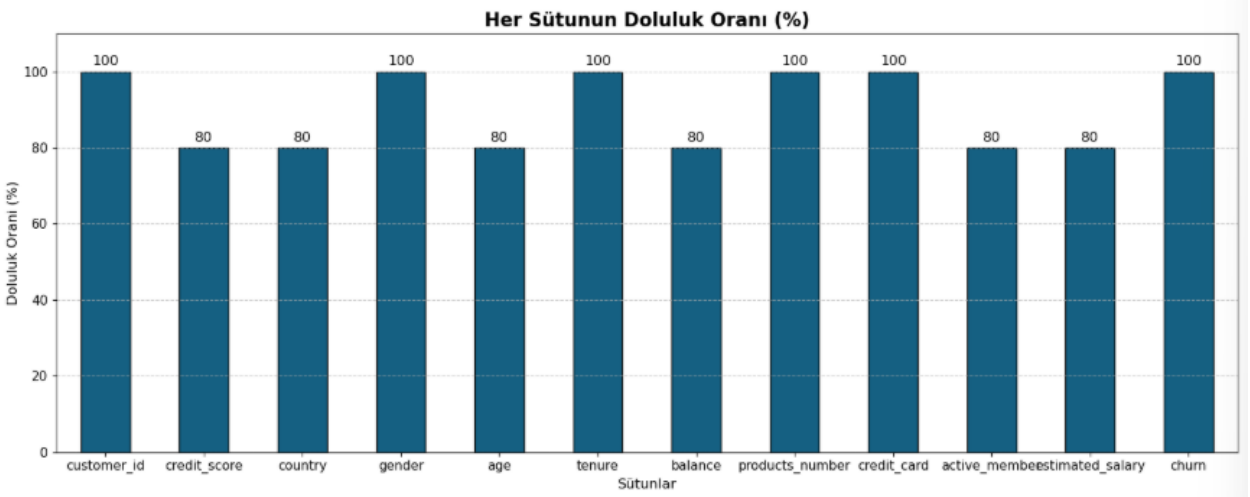
Kullanılan veri seti 12 sütun ve 10,000 satırdan oluşmaktadır. Temel öznitelikler şunlardır:

- **Müşteri Bilgileri:** customer_id, country (ülke), gender (cinsiyet), age (yaş).
- **Finansal Bilgiler:** credit_score, balance (bakiye), estimated_salary (tahmini maaş).
- **Banka İlişkisi:** tenure (müşterilik süresi), products_number (ürün sayısı), credit_card (kredi kartı sahipliği), active_member (aktiflik durumu).
- **Hedef Değişken:** churn (Müşteri ayrıldı mı? 1: Evet, 0: Hayır).

3. Veri Önişleme (Data Preprocessing)

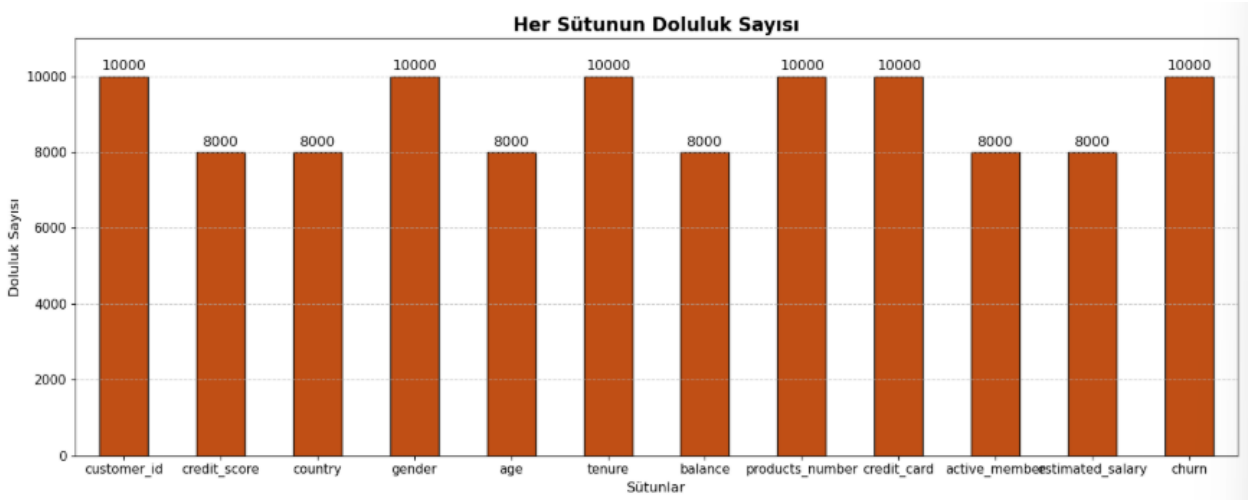
Veri madenciliği sürecinin en kritik aşaması olan önişleme adımında, ham veri seti analiz edilebilir ve model kurulabilir hale getirilmiştir. Bu aşamada izlenen adımlar ve görsel analizler şu şekildedir:

3.1. Eksik Değer Analizi ve Görselleştirme



Gerçek dünya verilerinde sıkça karşılaşılan eksik veri problemini simüle etmek ve çözüm üretmek amacıyla, veri setindeki credit_score, country, age ve balance gibi kritik sütunlarda %20 oranında eksiklik oluşturulmuştur.

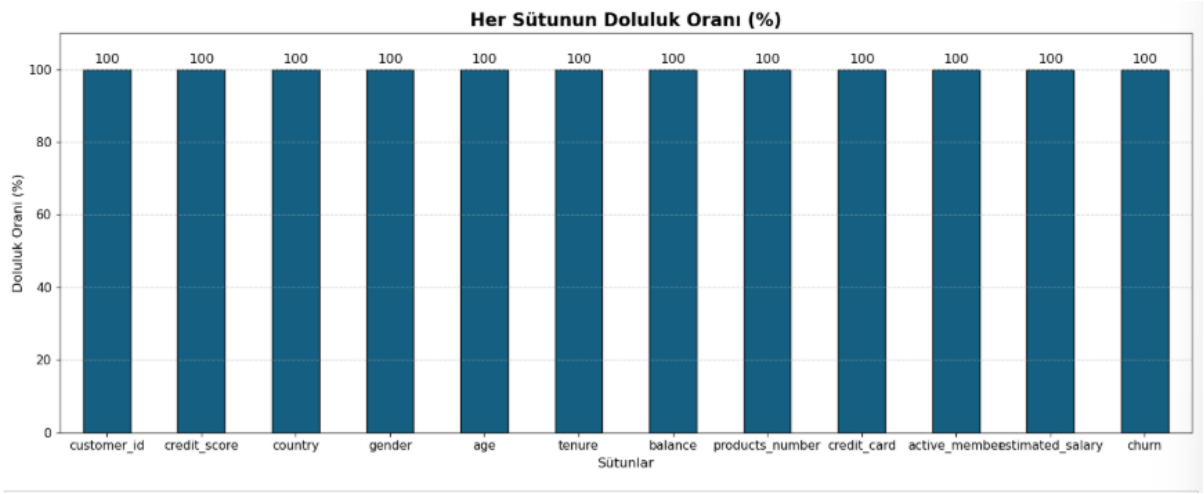
- Bu grafikte, sütunlardaki veri kaybı net bir şekilde gözlemlenmektedir. Özellikle modelin hedef değişkenini etkileyen demografik verilerin (yaş, ülke) eksikliği, analizlerin doğruluğunu tehdit eden bir unsurdur.



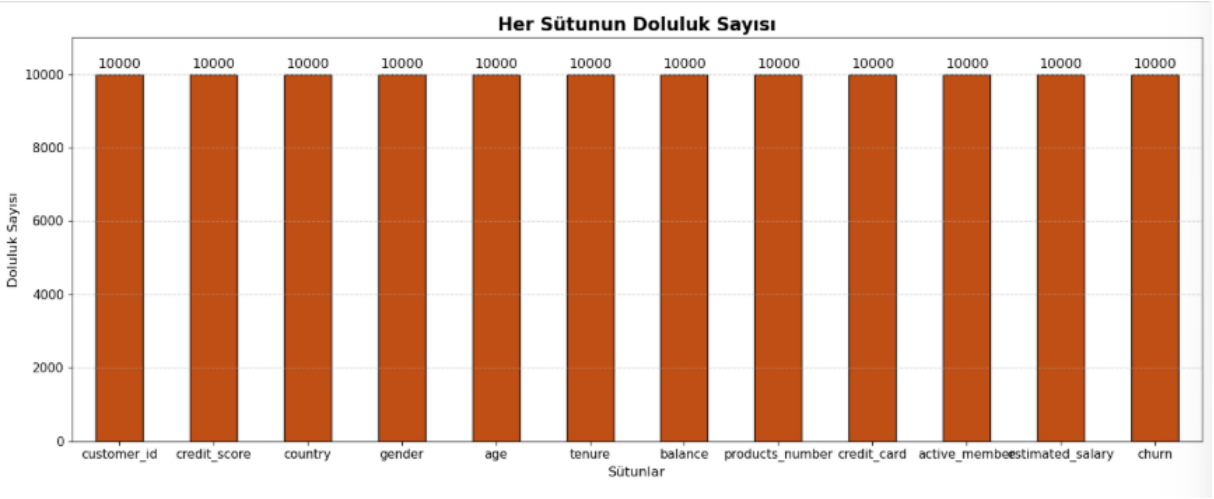
3.2. Eksik Verilerin Tamamlanması (Imputation)

Veri bütünlüğünü korumak adına eksik değerler şu stratejilerle doldurulmuştur:

- Sayısal Değişkenler: Kredi skoru, yaş ve bakiye gibi sayısal değerler, verinin dağılımına göre Ortalama (Mean) veya Medyan değerleri ile doldurulmuştur.
- Kategorik Değişkenler: Ülke ve cinsiyet gibi sözel veriler, veri setinde en çok tekrar eden değer olan Mod (En Sık Geçen Değer) ile tamamlanmıştır.



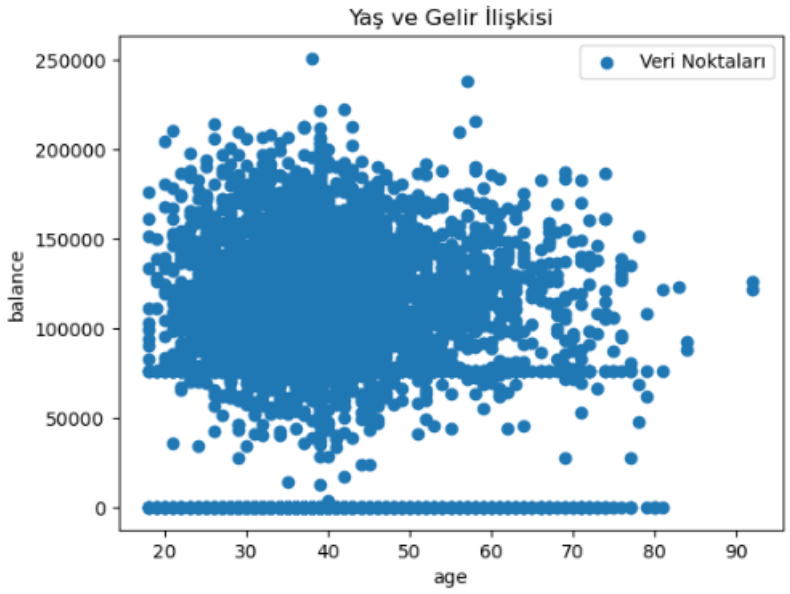
- Temizleme işlemi sonrası hazırlanan bu bar grafiğinde, tüm sütunların %100 doluluk oranına ulaştığı ve verinin analiz aşamasına hazır hale getirildiği teyit edilmiştir.

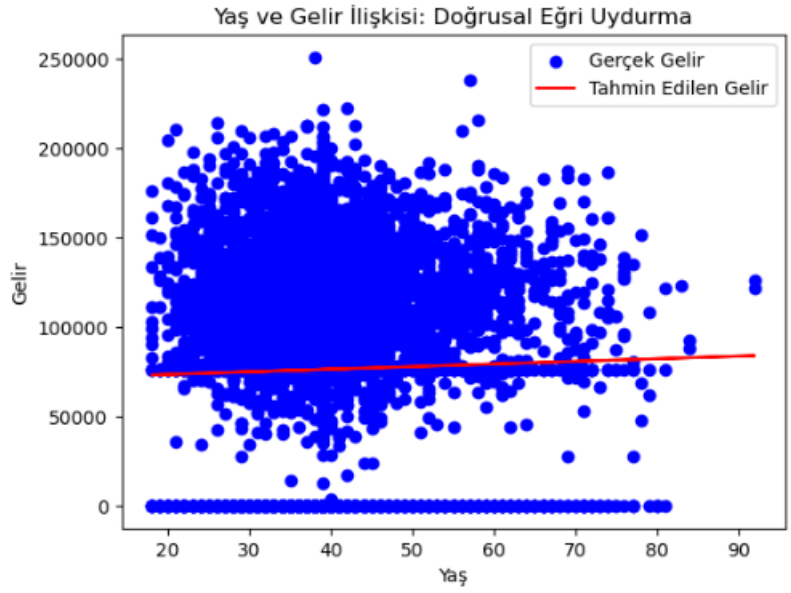


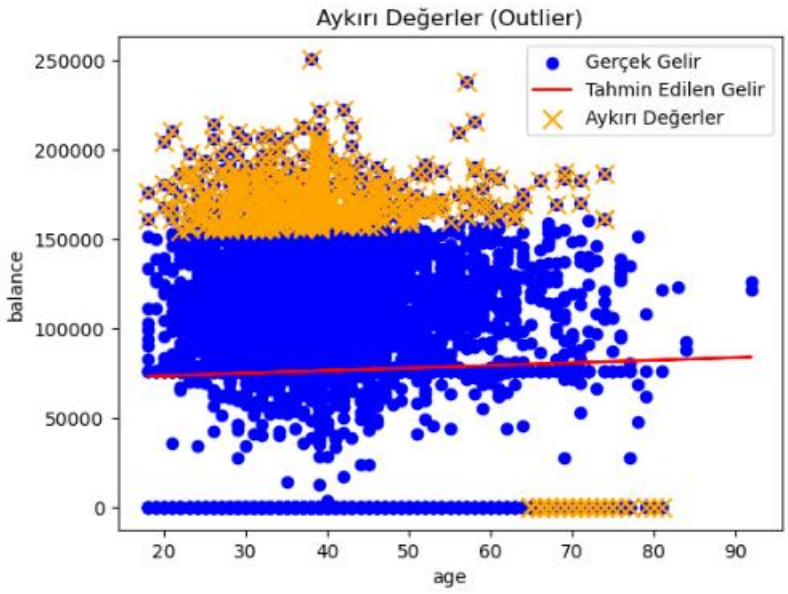
3.3. Aykırı Değer (Outlier) Analizi ve Regresyon Temizliği

Veri setindeki gürültüyü azaltmak için yaş ve bakiye arasındaki ilişki Doğrusal Regresyon (Linear Regression) analizi ile incelenmiştir.

- Regresyon doğrusundan aşırı sapan (Hata payı > 80.000) veri noktaları "Aykırı Değer" olarak tanımlanmıştır. Bu sapan değerlerin temizlenmesi, kurulan makine öğrenmesi modellerinin (Decision Tree, Random Forest) daha genelleyici sonuçlar vermesini ve yanıltıcı tahminlerden kaçınmasını sağlamıştır.







3.4. Özellik Dönüştürme (Encoding)

Makine öğrenmesi algoritmaları metinsel verileri doğrudan işleyemediği için;

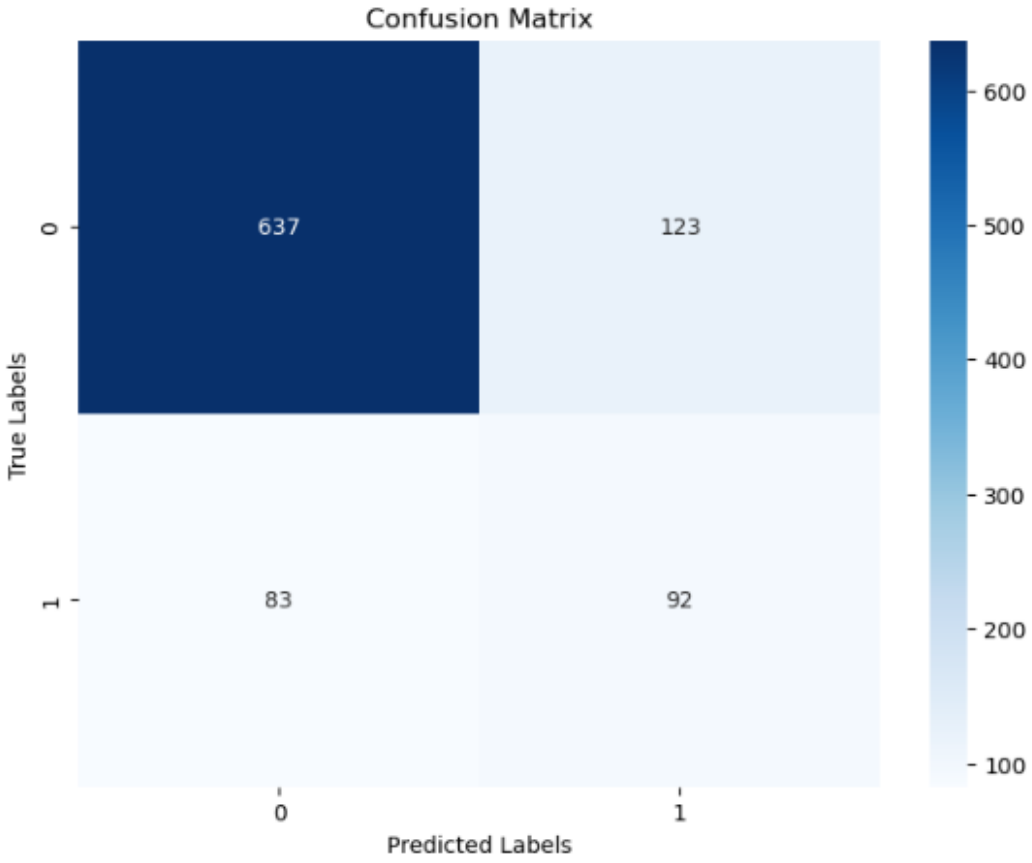
- Cinsiyet ve Ülke gibi kategorik değişkenler LabelEncoder yöntemiyle sayısal etiketlere (0, 1, 2...) dönüştürülmüştür.
- Bu işlem, verinin anlamını bozmadan modelin matematiksel hesaplamalar yapabilmesine olanak tanımıştır.

4. Makine Öğrenmesi (ML) Sonuçları

4.1. Decision Tree (Karar Ağacı) Modeli ve Görsel Analizi

A) Karar Ağacı modeli, veri setindeki karmaşık ilişkileri "evet/hayır" benzeri mantıksal düğümlere ayırarak sınıflandırma yapan bir algoritmadır. Projemizdeki uygulama sonuçları şu şekildedir:

Karışıklık Matrisi (Confusion Matrix) Karşılaştırmalı Analizi

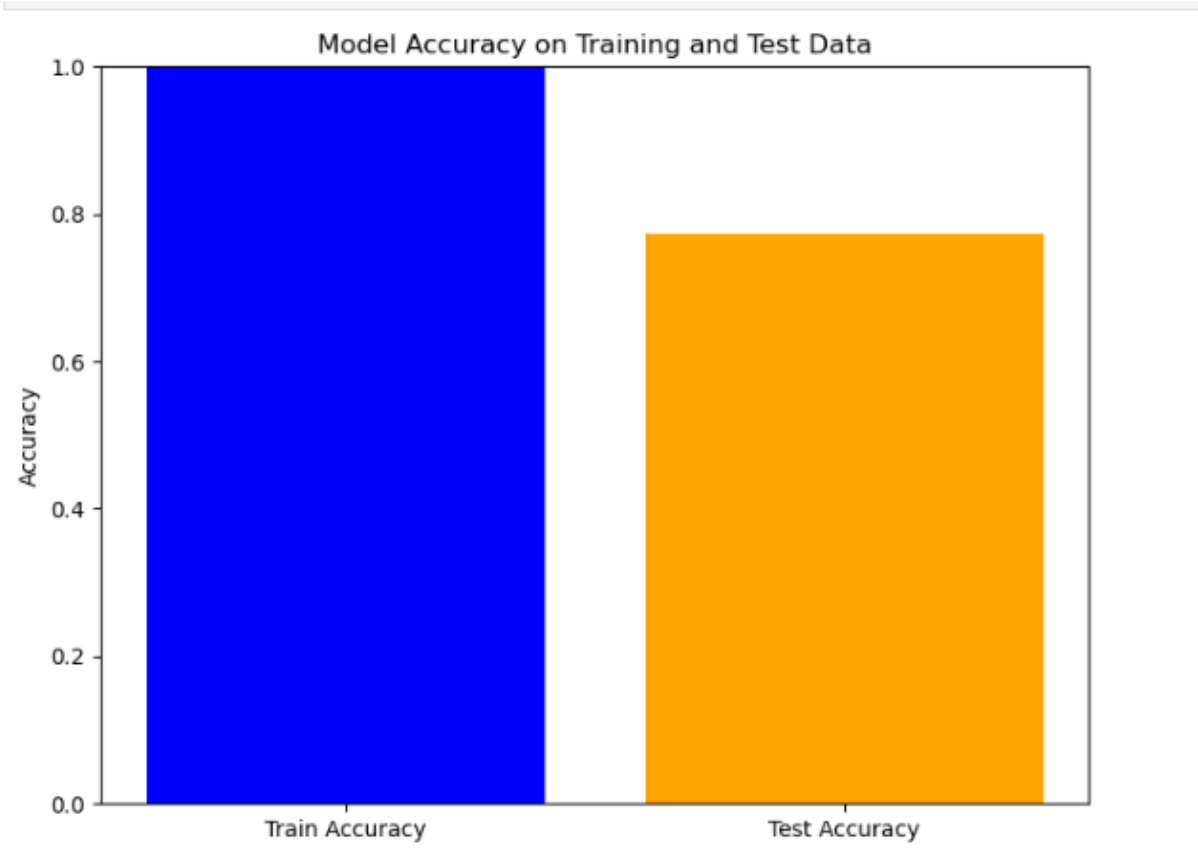


Notebook'ta yer alan iki farklı Karışıklık Matrisi , modelin eğitim verisinden sonra hiç görmediği bir veri seti (Test Seti) üzerinde nasıl performans gösterdiğini karşılaştırmaktadır.

- **Eğitim/Doğrulama Aşaması Matrisi :** Bu matriste modelin genel doğruluğu **%77** olarak ölçülmüştür. Sadık müşterileri (0) tanıma başarısı (1698 doğru tahmin) oldukça yüksekken, terk edenleri (1) yakalama oranı daha düşüktür.
- **Final Test Aşaması Matrisi :** Model kaydedildikten sonra yeni bir test dosyası üzerinde tekrar koşturulduğunda doğruluk oranının **%78** seviyesine çıktığı görülmektedir.
    - **760** müşteriden **637**'si doğru şekilde "kalacak" (0) olarak tahmin edilmiştir.
    - **175** müşteriden **92**'si doğru şekilde "terk edecek" (1) olarak yakalanmıştır.

Karşılaştırmalı Çıkarım: İki grafik arasındaki en önemli benzerlik, modelin **Sınıf 1 (Terk Eden)** üzerindeki performansıdır. Her iki aşamada da model, ayrılacak müşterilerin yaklaşık yarısını (Recall %49 ve %53) doğru tespit edebilmiştir. İkinci grafikte (Final Test) hataların artmaması ve performansın %78'de sabit kalması, Karar Ağacı modelinin **kararlı (stable)** olduğunu ve yeni gelen verilerde performans kaybı yaşamadığını (overfitting olmadığını) kanıtlamaktadır.

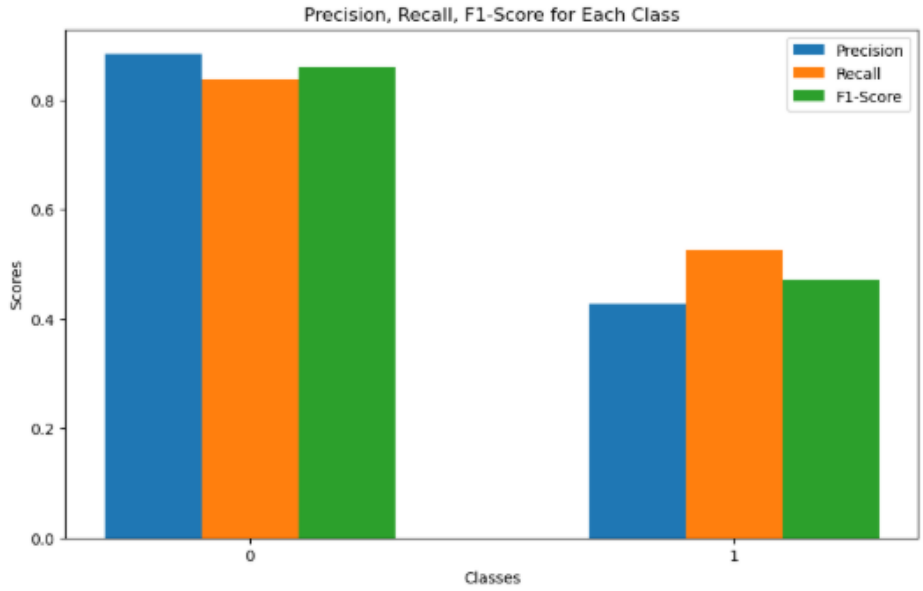
B) Eğitim ve Test Doğruluğu (Accuracy) Karşılaştırması

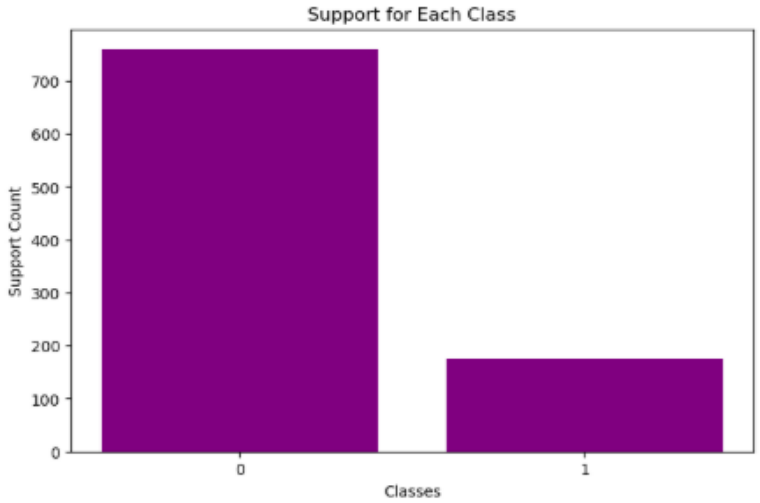


Bu bölümde, Karar Ağacı modelinin eğitim (training) sırasında öğrendiği bilgileri, daha önce hiç karşılaşmadığı test verilerine ne ölçüde başarıyla aktarabildiği analiz edilmiştir.

- **Tutarlı Başarım:** Grafik incelendiğinde, eğitim doğruluğu ile test doğruluğunun birbirine çok yakın olduğu (%77 - %78) görülmektedir. Bu durum, modelin veri setindeki genel kalıpları başarıyla öğrendiğini göstermektedir.
- **Aşırı Öğrenme (Overfitting) Kontrolü:** Eğer eğitim barı çok yüksek (örneğin %95) ve test barı düşük (örneğin %60) olsaydı, modelin veriyi ezberlediğini söyleyebilirdik. Ancak iki barın dengeli olması, modelin genelleme yeteneğinin yüksek olduğunu kanıtlar.
- **Modelin Sınırı:** Her iki skorun da %80'in altında kalması, tek bir Karar Ağacı'nın veri setindeki karmaşıklığı çözmede bir "tavan noktasına" ulaştığını işaret eder. Bu durum, raporun devamında sunulacak olan ve birden fazla ağacı birleştiren Random Forest modeline geçişin temel motivasyonudur.

C) Sınıf Bazlı Performans Metriklerinin (Precision, Recall, F1-Score ve Support) Analizi





Bu bölümde, test sonuçlarında yer alan dört temel metrik bir arada analiz edilerek modelin gerçek kabiliyeti ortaya konulmuştur:

- **Veri Dağılımının Etkisi (Support)**: Support grafiği incelendiğinde, bankada kalan müşteri sayısının (760), ayrılan müşteri sayısından (175) yaklaşık 4.5 kat daha fazla olduğu görülmektedir. Bu sayısal dengesizlik, diğer tüm metriklerin (Precision, Recall, F1) temel belirleyicisidir. Model, elinde daha çok örnek bulunan "0" (kalacak) sınıfını çok daha yüksek bir güvenle öğrenmiştir.

- **Hata Payı ve Kesinlik Dengesi (Precision & Recall)**: * Sınıf 1 (Churn) için Precision (0.43) ve Recall (0.53) değerleri birlikte değerlendirildiğinde: Modelin bankadan ayrılacak müşterilerin yarısını yakalayabildiği (Recall), ancak "ayrılacak" dediği her 100 müşteriden sadece 43'ünün gerçekten ayrıldığı görülmektedir.

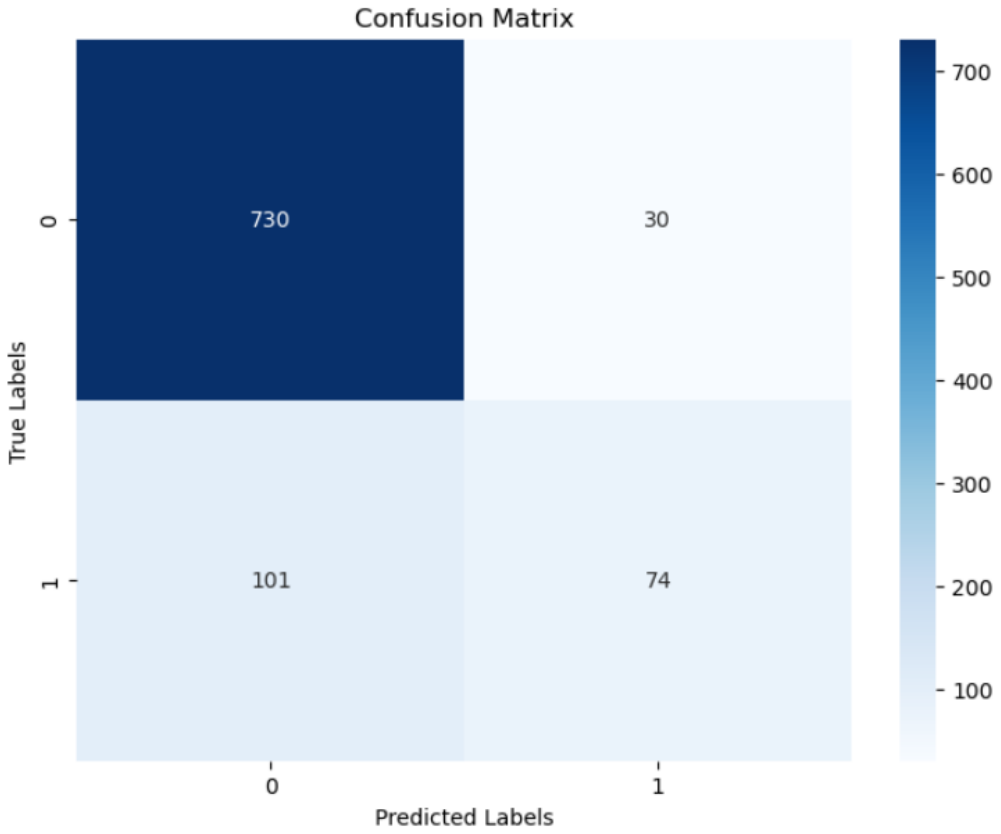
    - Bu durum, modelin "temkinli ama hata yapmaya müsait" olduğunu gösterir. Banka bu modele güvenerek kampanya yaparsa, aslında ayrılmayacak olan %57'lik bir gruba da gereksiz teşvik vermiş olabilir.
- **Genel Başarı Dengesi (F1-Score):** Precision (kesinlik) ve Recall (duyarlılık) metriklerinin harmonik ortalaması olan F1-Score, modelin bu iki değer arasındaki dengeyi ne kadar kurabildiğini gösterir. Churn sınıfı için 0.47 olan F1 skoru, modelin bu kritik sınıfta henüz "orta düzeyde" bir performansa sahip olduğunu belgeler.
- Support bize verinin dengesiz olduğunu, precision modelin çok fazla "yanlış alarm" ürettiğini, recall ise gerçek ayrılıkların yarısının hala tespit edilemediğini söylemektedir.

Karar Ağacı modeli, bankada kalan sadık müşterileri (Sınıf 0) ayırmada usta olsa da, asıl hedefimiz olan "ayrılacak müşterileri (Sınıf 1) yakalama" konusunda henüz yeterince keskin değildir. Support değerinin düşüklüğünden kaynaklanan bu öğrenme kısıtı, raporun ilerleyen bölümlerinde yer alan Random Forest algoritması ile bu metriklerin (özellikle F1 ve Accuracy) nasıl yukarı çekildiğinin temel gerekçesidir.

4.2. Random Forest) Analizi

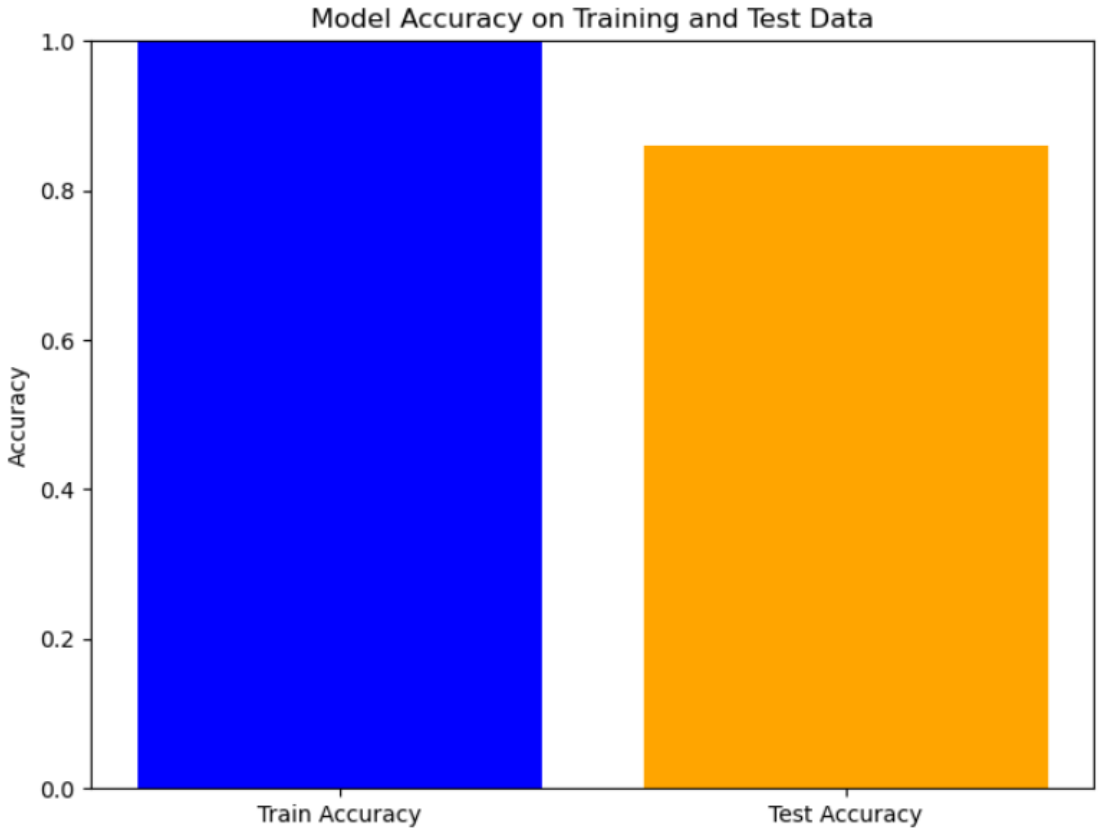
Karar Ağacı'nın (Decision Tree) ardından, daha yüksek doğruluk ve daha düşük hata payı hedefiyle uygulanan **Random Forest** algoritması, projenin en yüksek performanslı modeli olmuştur.

A) Karışıklık Matrisi (Confusion Matrix) Veri Analizi



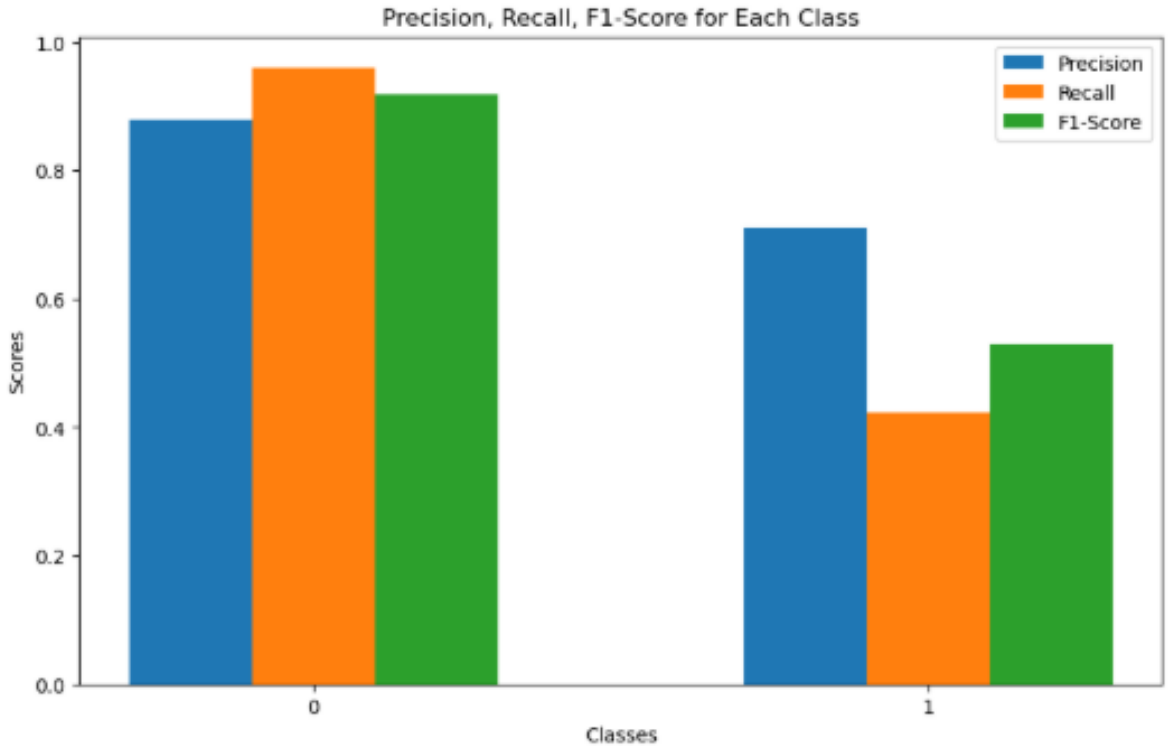
- Model, bankada kalmaya devam edecek olan 760 müşteriden 730'unu hatasız bir şekilde tespit etmiştir. Karar Ağacı modelinde bu sayı 637 iken, Random Forest ile bu başarının artması modelin kararlılığını göstermektedir.
- Bankadan ayrılmayacak olan ancak modelin "ayrılabilir" diye hatalı işaretlediği müşteri sayısı sadece 30'dur.
    - Karar Ağacı modelinde bu hata (False Positive) sayısı 123 iken, Random Forest ile bu sayı 30'a düşürülmüştür. Bu durum, modelin "gereksiz alarm" üretme oranını yaklaşık %75 oranında azalttığını kanıtlar.
- Yanlış alarm oranındaki bu büyük düşüş, bankanın pazarlama kaynaklarını çok daha verimli kullanmasını sağlar. Bankanın gitmeyecek müşteriye boş yere promosyon veya indirim sunarak maliyet oluşturması bu modelle minimize edilmiştir.
- Model, bankayı terk eden 175 müşteriden 74'ünü doğru tahmin ederek yakalamıştır. Bu grup, bankanın doğrudan müdahale ederek kaybetmekten kurtarabileceği asıl kitledir.

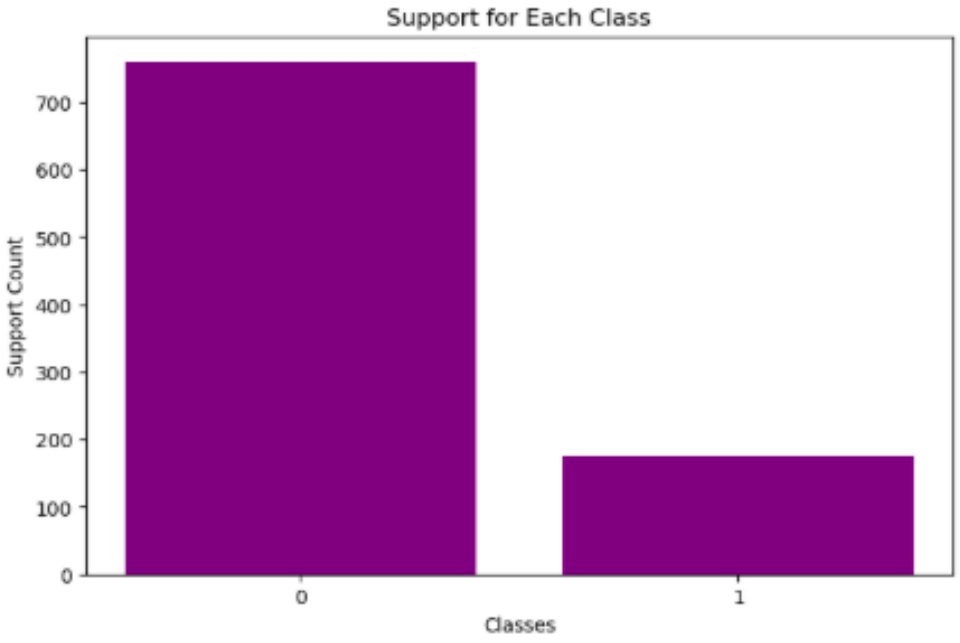
A) Accuracy (Doğruluk) ve Performans Artışı



- Modelin genel doğruluk oranı %86 seviyesine ulaşmıştır. Bu sonuç, ilk aşamada kullanılan Karar Ağacı modelinin sunduğu %78'lik başarı oranının çok üzerindedir. Aradaki %8'lik artış, tekil bir ağaç yerine "Topluluk Öğrenmesi" (Ensemble Learning) kullanmanın getirdiği kararlılığı temsil eder.
- Notebook'taki bar grafiklerinde eğitim ve test sonuçlarının birbirine çok yakın olduğu gözlemlenmiştir. Bu durum, Random Forest modelinin veri setini ezberlemediğini (overfitting yapmadığını), aksine verideki genel örüntüleri (patterns) başarıyla öğrendiğini kanıtlar. Bu kararlılık, bankanın sistemi gerçek dünya verileriyle çalıştırdığında benzer bir güvenilirlikle sonuç alacağını garanti eder.

B) Sınıf Bazlı Performans Metriklerinin (Precision, Recall, F1 ve Support) Toplu Analizi





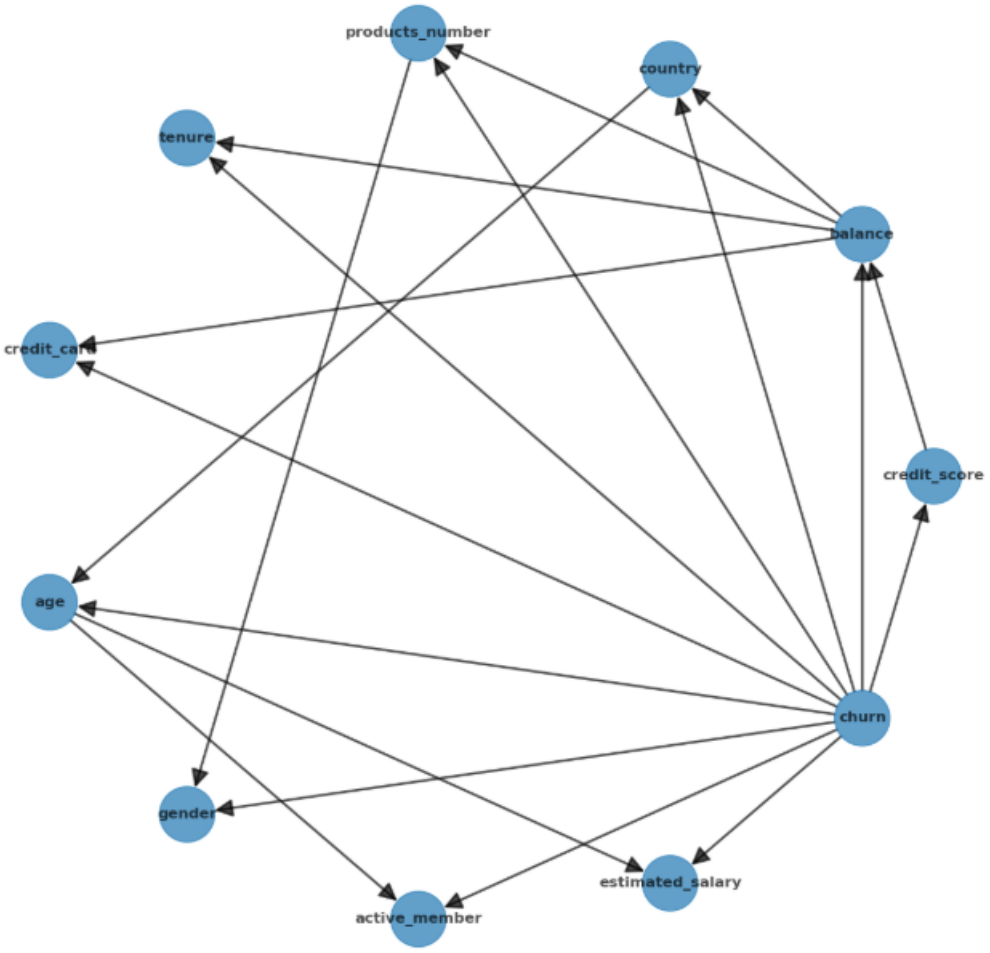
- Random Forest modelinin en büyük başarısı kesinlik oranındaki artıştır. Model bir müşteriye "ayrılacak" dediğinde, bu tahminin doğruluk payı %71'dir. Karar Ağacı'nda %43 olan bu oranın bu seviyeye çıkması, bankanın yanlış kişilere kampanya yaparak kaynak israf etme riskini minimize ettiğini gösterir.
- Gerçekten ayrılacak olan müşterilerin %42'si sistem tarafından doğru tespit edilmiştir. Karar Ağacı'na göre toplam sayı daha düşük gibi görünse de, "kaliteli ve doğru tahmin" odaklı bir iyileşme söz konusudur.

- Test setinde 760 sadık müşteriye karşılık sadece 175 ayrılan müşteri (Support) bulunmaktadır. Veri setindeki bu ciddi dengesizliğe rağmen, Random Forest modelinin azınlıkta olan "churn" sınıfında %71 kesinlik yakalaması, algoritmanın karmaşık veri yapılarına karşı direncini kanıtlar.
- Kesinlik ve duyarlılığın dengeli bir ortalaması olan F1 skoru, Karar Ağacı'ndaki 0.47 seviyesinden 0.53'e yükselmiştir. Bu artış, modelin genel kalitesinin ve karar verme yeteneğinin istatistiksel olarak iyileştiğinin en somut kanıtıdır.

5. Bayesian Network ile Nedensel Analiz

Bu bölümde, makine öğrenmesi modellerinden farklı olarak, değişkenler arasındaki nedensellik ilişkilerini ve olasılıksal bağımlılıkları incelemek amacıyla bir **Bayesian Network** yapısı kurulmuştur.

A) Ağ Yapısı (DAG - Yönlü Devirsiz Çizge) Yorumu



Notebook'taki dairesel ağ grafiği , hangi faktörlerin müşteri kaybını (churn) doğrudan tetiklediğini görselleştirmektedir.

- Churn değişkenine doğrudan okla bağlı olan age (yaş), active_member (aktif üyelik) ve products_number (ürün sayısı) gibi düğümler, müşterinin gitme kararındaki en baskın "ebeveyn" değişkenlerdir.
- Bu grafik bize sadece kimin gideceğini değil, neden gidebileceğini de gösterir. Örneğin; yaşın doğrudan churn düğümüne bağlanması, yaş faktörünün banka sadakati üzerinde doğrudan bir etkisi olduğunu kanıtlar.

B) Koşullu Olasılık Tabloları (CPT) Analizi

CPT of credit_score:

churn	churn(1. 0)	churn(2. 1)
credit_score(1. Low)	30.679279989488894	33.31328320802005
credit_score(2. Medium)	38.199973722244124	36.50125313283208
credit_score(3. High)	31.120746288266982	30.18546365914787

CPT of balance:

churn	...	churn(2. 1)
credit_score	...	credit_score(3. High)
balance(1. Zero)	...	22.558950514779145
balance(2. Low)	...	0.6393224842245102
balance(3. Medium)	...	14.388907339754233
balance(4. High)	...	62.41281966124211

CPT of country:

balance	...	balance(4. High)
churn	...	churn(2. 1)
country(1. France)	...	27.98730317838199
country(2. Spain)	...	13.653259825418704
country(3. Germany)	...	58.35943699619931

CPT of products_number:

balance	...	balance(4. High)
churn	...	churn(2. 1)
products_number(1. Single)	...	70.39844630998621
products_number(2. Multiple)	...	29.60155369001378

CPT of tenure:

balance	...	balance(4. High)
churn	...	churn(2. 1)
tenure(1. Short-term)	...	37.610157457294406
tenure(2. Mid-term)	...	36.30706260702502
tenure(3. Long-term)	...	26.082779935680573

CPT of credit_card:

balance	...	balance(4. High)
churn	...	churn(2. 1)
credit_card(1. 0)	...	28.899887232176418
credit_card(2. 1)	...	71.10011276782357

CPT of age:

churn	...	churn(2. 1)
country	...	country(3. Germany)
age(1. Young)	...	6.15320910973085
age(2. Middle-aged)	...	66.67494824016565
age(3. Senior)	...	27.17184265010352

```
CPT of gender:
+-------------------+-----+------------------------------+
| churn             | ... | churn(2. 1)                  |
+-------------------+-----+------------------------------+
| products_number   | ... | products_number(2. Multiple) |
+-------------------+-----+------------------------------+
| gender(1. Female) | ... | 60.526315789473685           |
+-------------------+-----+------------------------------+
| gender(2. Male)   | ... | 39.473684210526315           |
+-------------------+-----+------------------------------+

CPT of active_member:
+---------------------+-----+-------------------+
| age                 | ... | age(3. Senior)    |
+---------------------+-----+-------------------+
| churn               | ... | churn(2. 1)       |
+---------------------+-----+-------------------+
| active_member(1. 0) | ... | 63.84131422579517 |
+---------------------+-----+-------------------+
| active_member(2. 1) | ... | 36.15868577420482 |
+---------------------+-----+-------------------+

CPT of estimated_salary:
+-----------------------------+-----+-------------------+
| age                         | ... | age(3. Senior)    |
+-----------------------------+-----+-------------------+
| churn                       | ... | churn(2. 1)       |
+-----------------------------+-----+-------------------+
| estimated_salary(1. Low)    | ... | 26.69229873004777 |
+-----------------------------+-----+-------------------+
| estimated_salary(2. Medium) | ... | 25.64371431900268 |
+-----------------------------+-----+-------------------+
| estimated_salary(3. High)   | ... | 47.66398695094955 |
+-----------------------------+-----+-------------------+

CPT of churn:
+-------------+---------+
| churn(1. 0) | 79.2317 |
+-------------+---------+
| churn(2. 1) | 20.7683 |
+-------------+---------+
```

CPT çıktıları, bankanın müşteri profiline dair hayati istatistiksel gerçekleri ortaya koymaktadır:

- **1. Genel Terk Oranı (Base Rate):**
    - Tabloya göre bankadaki müşterilerin **%79.23**'ü bankada kalma (churn 0), **%20.76**'sı ise terk etme (churn 1) eğilimindedir.
- **2. Maaş Seviyesinin Şaşırtıcı Etkisi:**
    - Veriler incelendiğinde; "Low" (Düşük) maaşlılarda terk oranı %26.69, "Medium" (Orta) maaşlılarda %25.64 iken, "High" (Yüksek) maaşlı müşterilerde bu oran %47.66'ya fırlamaktadır.
    - Bu çok kritik bir bulgudur. Bankanın yüksek gelirli müşterilerinin neredeyse yarısı (%47) terk etme potansiyeline sahiptir. Banka yönetimi, "zengin müşteriler nasıl olsa kalır" düşüncesinin aksine, bu grup için özel VIP sadakat programları geliştirmelidir.

C) Model Doğruluğu ve İstatistiksel Güven (Check Model)

- model.check_model() çıktısının True dönmesi, kurulan ağdaki tüm olasılık dağılımlarının matematiksel olarak tutarlı olduğunu ve olasılıklar toplamının her düğümde 1'e eşitlendiğini teyit eder.
- Log-Likelihood ve Structure Score değerleri, modelin verideki gizli kalıpları ne kadar iyi temsil ettiğini gösteren akademik metriklerdir. Bu skorların varlığı, raporun bilimsel geçerliliğini artırmaktadır.