1: Veri Seti hikayesi

Titanic veri seti, 15 Nisan 1912 tarihinde ilk yolculuğunda bir buz dağına çarparak batan RMS Titanic adlı geminin yolcularına ait detaylı bilgileri içerir. Bu veri seti, hayatta kalma (survival) durumunu tahmin etmek ve analiz etmek için sıkça kullanılan, popüler bir veri bilim projesidir. Gemi kazası, o dönemin en büyük ve en güvenli yolcu gemisi olarak kabul edilen Titanic'in trajik bir şekilde batmasıyla sonuçlanmış ve 2224 yolcunun yaklaşık %68'i hayatını kaybetmiştir.

2: veri setinin yapısı

* Gözlem Sayısı (Satır Sayısı) : 891
* Değişken Bilgileri (Sütun Sayısı, Değişken Türleri)

Sütun sayısı: 11

Değişken türleri:

Survived category

Pclass int64

Name category

Sex category

Age float64

SibSp int64

Parch int64

Ticket category

Fare float64

Cabin category

Embarked category

dtype: object

* Sayısal Değişkenler için; Betimsel İstatistikler



* Kategorik Değişkenler için; sınıf sayıları ve her bir sınıf frekansları

Kategorik değişkenler sınıf sayıları:

Name sınıf sayısı: 891

Sex sınıf sayısı: 2

Ticket sınıf sayısı: 681

Cabin sınıf sayısı: 147

Embarked sınıf ayısı: 3

Embarked sınıf ayısı: 2

Kategorik değişkenler her bir sınıf frekansı:

Name değişkeni için

Name

Abbing, Mr. Anthony 1

Nysveen, Mr. Johan Hansen 1

Nicholson, Mr. Arthur Ernest 1

Nicola-Yarred, Master. Elias 1

Nicola-Yarred, Miss. Jamila 1

..

Goodwin, Miss. Lillian Amy 1

Goodwin, Mr. Charles Edward 1

Goodwin, Mrs. Frederick (Augusta Tyler) 1

Graham, Miss. Margaret Edith 1

van Melkebeke, Mr. Philemon 1

Name: count, Length: 891, dtype: int64

* Eksik Veri Mevcut mu ? Var ise değişken başına sayıları.

Eksik veri var mı ? True

Survived 0

Pclass 0

Name 0

Sex 0

Age 177

SibSp 0

Parch 0

Ticket 0

Fare 0

Cabin 687

Embarked 2

dtype: int64

3: Veri görselleştirme için yöntemler

**Kullanılan Yöntemler**

**1. Bar Plot (Sütun Grafiği)**

**2. KDE Plot (Yoğunluk Grafiği)**

**3. Box Plot (Kutu Grafiği)**

**Örnek :** **Bilet Sınıfına Göre Yaş Dağılımı (KDE Plot)**



4: Veri setinizdeki aykırı değer ve eksik veri problemler.

**Aykırı Değerin Tespiti:**

* **IQR (Interquartile Range)** yöntemi ile aykırı değerler tespit ettim.

**Aykırı Değerlere Uygulanan Çözümler**

* **Fare (Bilet Ücreti):**
* Aykırı değerler, veri setindeki genel yapıyı bozabileceği için üst sınırdaki aykırı değerler, sınır değerine kırpıldı (capping yöntemi).
* **Age (Yaş):**
* Çok yüksek yaş değerleri (örneğin 80+) outlier olarak değerlendirildi.
* Bu değerler ya yaş dağılımının medyanıyla ya da yaş gruplarına göre belirlenmiş uygun değerlerle değiştirildi.



**2. Eksik Veri Problemlerinin Çözümü**

Eksik veriler Titanic veri setinde önemli bir problem oluşturur. Özellikle **Age**, **Cabin** ve **Embarked** sütunlarında eksik değerler bulunmaktadır. Bu eksik değerler aşağıdaki şekilde çözüldü:

**Eksik Veri Analizi**

* **Age (Yaş):**177 eksik değer (%19.87).
* **Cabin (Kabin):**687 eksik değer (%77.10).
* **Embarked (Biniş Limanı):**2 eksik değer (%0.22).

**Eksik Verilere Uygulanan Çözümler**

* **Age (Yaş):**Eksik yaş değerleri, **bilet sınıfı (Pclass)** ve **cinsiyet (Sex)** gibi kategorilere göre gruplandırılmış **medyan yaş** ile dolduruldu.
* **Cabin (Kabin):Cabin** sütununda çok fazla eksik değer olduğu için silindi
* **Embarked (Biniş Limanı):**Eksik olan 2 değer, en sık görülen değer olan **"S" (Southampton)** ile dolduruldu (mode imputation yöntemi).



5: hangi tür problem sınıfına neden girmektedir

Titanic veri seti ile yapılan bu proje bir **sınıflandırma problemi**dir.

**Neden Sınıflandırma Problemi?**

* Projenin temel amacı, bir yolcunun **hayatta kalma durumu**nu tahmin etmektir.
* Hedef değişken (**Survived**) iki sınıftan oluşur:**0:** Yolcu hayatta kalmadı.,**1:** Yolcu hayatta kaldı.
* Bu nedenle, modelin çıktısı sürekli bir değer yerine bir sınıf etiketi olduğundan bu problem **ikili sınıflandırma (binary classification)** problemidir.

**Bağımlı Değişken (Hedef Değişken):**

* **Survived :** Yolcunun hayatta kalma durumunu belirtir. İkili kategorik bir değişkendir:**0:** Hayatta kalmadı.,**1:** Hayatta kaldı.

**Bağımsız Değişkenler (Girdi Değişkenleri):**

Bağımsız değişkenler, bir yolcunun hayatta kalma ihtimalini etkileyebilecek özelliklerdir. Titanic veri setinde bu değişkenler şunlardır: **Pclass (Bilet Sınıfı), Sex (Cinsiyet),Age (Yaş), SibSp (Kardeş/Eş Sayısı), Parch (Ebeveyn/Çocuk Sayısı), Fare (Bilet Ücreti), Embarked (Biniş Limanı)**, **Cabin (Kabin Numarası), Name (Ad)**

6: makine öğrenmesi modelleri

**Kullanılan Modeller**

* **CART (Decision Tree - Karar Ağacı):**
* **SVM (Support Vector Machine - Destek Vektör Makinesi):**
* **Random Forest (Rassal Orman):**
* **Gradient Boosting (Eğimli Artırma):**

7: kullanmış olduğunuz makine öğrenmesi modellerinin performans değerlendirme ölçütleri

**Performans Değerlendirme Süreci:**

* **Accuracy (Doğruluk):**Tüm tahminlerin ne kadarının doğru olduğunu ölçer.
* **Average Precision (Ortalama Kesinlik):**Modelin tüm sınıflar için kesinliğini ölçer.
* **Average F1-Score (Ortalama F1-Skoru):**Kesinlik ve duyarlılık arasında bir denge sağlar.
* **Recall (Duyarlılık):**Her bir sınıf için duyarlılık ölçülmüştür.

8: makine öğrenmesi sonuçlarının karşılaştırılması



9: boyut indirgeme yöntemi

Projede boyut indirgeme yöntemi olarak Principal Component Analysis (PCA) kullanılmıştır. Bu işlem sonucunda orijinal veri setindeki birçok özellik, daha az sayıda ana bileşenle ifade edilmiştir.

Yeni veri seti daha az boyutlu, ancak varyansın büyük bir kısmını koruyan ve model performansını optimize eden bir yapıdadır.

10: Orijinal veri setini kullanarak en yüksek test tahmin performansını veren makine öğrenmesi modelinin orijinal veri setinde ve seçtiğiniz boyut azaltma yöntemi ile boyutu azaltılmış veri setindeki performans karşılaştırması

