TBF495 Python Programlama ve Veri Bilimi Dersi

2nci Proje Raporu

(Makine Öğrenmesi)

Hazırlayanlar (Takım Nu.; Öğc. Nu.; Adı ve Soyadı)

* 22196555 Barış Eren Paçcı
* 22194155 Bahadır Acuner
* 22195247 Doruk Dölek

Mayıs 2022

# İÇINDEKİLER

[İÇINDEKİLER 2](#_Toc66019504)

[1. Projenin Tanımı ve Başlatılma Aşaması 3](#_Toc66019505)

[1.1. Proje Çerçevesi 3](#_Toc66019506)

[1.2. Projenin Amaçları 4](#_Toc66019507)

[1.3. Proje Başlatma Belgesi 4](#_Toc66019508)

[1.3.1 Proje Yönetim Modeli 5](#_Toc66019509)

[1.3.2 Projenin Başarı Ölçütleri 5](#_Toc66019510)

[1.3.3 Proje Paydaşları ve Ekibi 6](#_Toc66019511)

[2. Veri Bilimi Aşaması 6](#_Toc66019512)

[2.1. Veri Kaynağı ve Elde Edilmesi 6](#_Toc66019513)

[2.2. Verinin Hazırlanması 6](#_Toc66019514)

[2.2.1 Verinin Temizlenmesi 7](#_Toc66019515)

[2.2.2 Verinin Birleştirilmesi 7](#_Toc66019516)

[2.2.3 Verinin Dönüştürülmesi 7](#_Toc66019517)

[2.2.4 Keşfe Yönelik Veri Analizi (Exploratory Data Analysis) 7](#_Toc66019518)

[3. Makine Öğrenmesi Aşaması 7](#_Toc66019519)

[3.1. Problemin Çözümüne Yönelik Modeller (Algoritma + Veri) 7](#_Toc66019520)

[3.2. Modellerin Eğitimi ve Test Edilmesi 8](#_Toc66019521)

[3.3. Modellerin Değerlendirilmesi 8](#_Toc66019522)

[4. Proje Sonlandırma Aşaması 8](#_Toc66019523)

[4.1. Proje Sonlandırma İşlemleri 8](#_Toc66019524)

[4.2. Proje Görevlerinin Takım Üyelerine Dağılımı 8](#_Toc66019525)

# Projenin Tanımı ve Başlatılma Aşaması

* Proje sonuç raporunda aşağıdaki proje çerçevesinde belirtilen, Ürün-2’e ait veri bilimi ve makine öğrenmesin ait süreçlerle ilgili bilgiler bu raporda yer alacaktır.

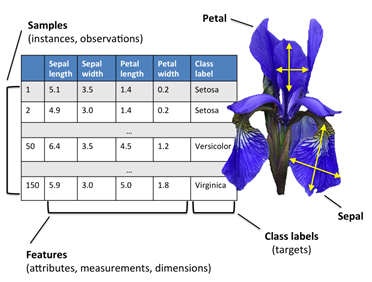
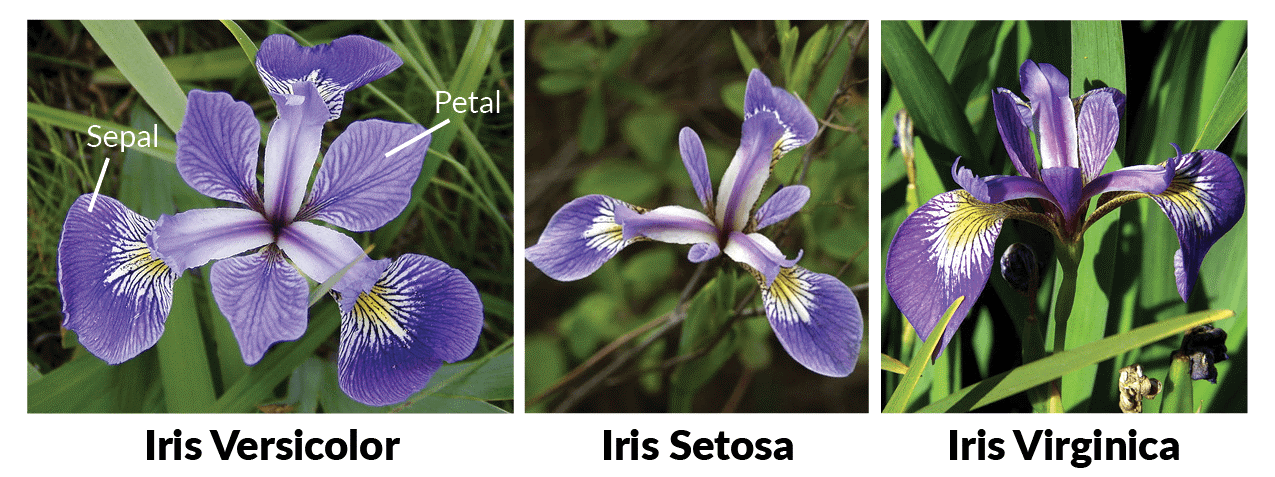
## Proje Çerçevesi

Türkiye’nin çeşitli coğrafi bölgelerinde bulunan Süsen (İris) çiçekleriyle ilgili ülke çapında bir bilimsel araştırma yürütülmektedir. Söz konusu çalışmayı desteklemek amacıyla proje ekibinizden veri bilimi ve makine öğrenmesi destekli bütünleşik bir yazılım geliştirmeniz beklenmektedir. Dolayısıyla dönem projesinde, aşağıda belirtilen ihtiyaçları karşılamak üzere makine öğrenmesi yazılımı ürün olarak geliştirilecektir:

* **Ürün-2:** **İris çiçeklerinin türünü tahmin etmede kullanılacak, Python Programlama Diliyle geliştirilmiş bütünleşik makine öğrenmesi yazılımıdır.**

İris çiçek türleri ve özellikleri Şekil 1.1’de gösterilmiştir.:

**Şekil-1.1**.İris Çiçek Türleri ve Özellikleri (Feature)



Makine öğrenmesi projesi yazılımın ihtiyaçları veri bilimi ve makine öğrenmesi başlıkları altında gruplanarak aşağıdaki tabloda gösterilmiştir:

|  |  |
| --- | --- |
| **S. Nu** | **İhtiyaç Listesi** |
| **Veri Bilimi** | |
| 1 | Verinin Hazırlanması (temizlenmesi, birleştirilmesi, dönüştürülmesi ve görselleştirilmesi) Etkinlikleri |
| 2 | Keşfe Yönelik Veri Analizi Etkinlikleri |
| **Makine Öğrenmesi** | |
| 1 | Modellerinin (Algoritma + Veri) Belirlenmesi ve Uygun Modelin Seçimi |
| 2 | Modellerin Eğitimi ve Test Edilmesi |
| 3 | Modellerin Değerlendirilmesi ve İyileştirilmesi |
| **Sonuç Raporu** | |
| 4 | Proje Sonuçların Raporlanması ve Sunulması |

## Projenin Amaçları

Bu proje, Türkiye’nin çeşitli coğrafi bölgelerinde bulunan İris çiçeklerinin türlerini doğru bir şekilde sınıflandırmak ve İris çiçeklerinin türünü tahmin etmede kullanılacak amacıyla tasarlanmıştır. Ana hedeflerimiz şunlardır:

**Veri Toplama ve Hazırlık:** Türkiye genelindeki İris çiçeklerine ait verileri toplamak, temizlemek ve analiz için uygun hale getirmek.

**Model Geliştirme:** Çeşitli makine öğrenmesi modellerini kullanarak veriler üzerinde eğitim yapmak ve en iyi performans gösteren modeli belirlemek.

**Doğruluk ve Performans:** Seçilen modelin doğruluk oranını maksimize etmek ve çiçek türlerini tahmin etmede yüksek performans sağlamak.

**Uygulamalı Araştırma:** Elde edilen bulguları bilimsel bir çerçevede değerlendirerek Türkiye’nin biyoçeşitliliği üzerine katkıda bulunmak.

**Eğitim ve Bilinçlendirme:** Proje sonuçlarını eğitim materyalleri olarak kullanarak biyoloji ve veri bilimi alanında farkındalık yaratmak.

Bu amaçlara ulaşarak, biyolojik araştırmalarda veri biliminin ve makine öğrenmesinin potansiyelini ortaya koymak ve bu alanlardaki ilerlemeleri desteklemek ve kullanıcılara bu verileri daha doğru ve özet şeklinde tanıtmak.

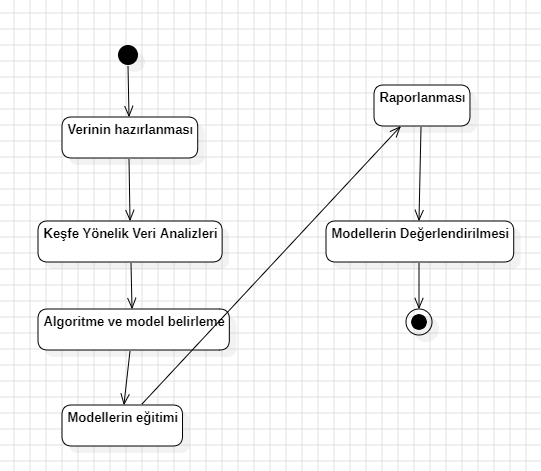
## Proje Başlatma Belgesi

Bu bölümde bir proje başlatma belgesinde bulunması gereken maddelerden önemli olduğu değerlendirilenler belirtilecektir.

### Proje Yönetim Modeli

Projede kullanacağınız proje yönetim modeli bu proje sonuç raporunun içendekiler tablosundaki ilgili başlıklar (veri bilimi ve makine öğrenmesi) seçilerek belirlenecektir. Bu adımlar UML etkinlik çizeneğiyle gösterimi yapılarak aşağıda gösterilecektir.

**Şekil-1.3.1.** Proje Yönetim Modeli Etkinlik Çizeneği (Activity Diagram)



### Projenin Başarı Ölçütleri

**Model Doğruluğu:** Geliştirilen makine öğrenmesi modelinin, test veri seti üzerindeki sınıflandırma doğruluğunun %95 veya üzeri olması hedeflenmektedir. Bu, modelimizin İris çiçek türlerini yüksek doğrulukla tahmin edebileceğini gösterir.

**Kullanıcı Memnuniyeti:** Proje paydaşlarından alınacak geri bildirimler ile kullanıcı memnuniyeti değerlendirilecek. Proje sonuçlarının, kullanıcıların ihtiyaçlarını karşılayacak şekilde teslim edilmesi ve en az %90 kullanıcı memnuniyeti oranına ulaşılması amaçlanmaktadır.

**Bütçe Uyumu:** Projenin onaylanmış bütçe çerçevesinde tamamlanması. Bütçe aşımının %5'in altında tutulması hedeflenmektedir.

Bu başarı ölçütleri, projenin gelişimini takip etmek, sonuçlarını ve İris çiçeklerinin türünü tahmin etmede kullanılacak. Değerlendirmek için kullanılacak temel metriklerdir.

Veri Bilimi yönetimi ve verinin işlenmesi kapsamında çeşitli analiz ve uygulamalar yaparak satış ve pazarlama süreçlerinin etkinliği ve verimliliğinin artırılmasıdır.

Makina öğrenmesi λ Veri bilimi uygulamaları, altyapısı, bilgi birikimi, yazılım ve veri bilimi yeteneklerini kullanarak Makine Öğrenmesi kapsamında yapay zekâ uygulamalarını geliştirmek, mevcut sistemlerle bütünleşik hale getirmektir

### Proje Paydaşları ve Ekibi

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S. Nu.** | **Projedeki Görevi** | **Görev Kodu** | **Öğrenci Numarası** | **Adı ve Soyadı** |
| 1 | Takım Lideri | T-1 | 22196555 | Barış Eren Paçcı |
| 2 | Takım Üyesi | Ü-1 | 22194155 | Bahadır Cemil Acuner |
| 3 | Takım Üyesi | Ü-2 | 22195247 | Doruk Dölek |
| 4 | Takım Üyesi | Ü-3 |  |  |
| 5 | Takım Üyesi | Ü-4 |  |  |

# Veri Bilimi Aşaması

## Veri Kaynağı ve Elde Edilmesi

Projemizde kullanılan veri kaynağı, ünlü İris çiçek veri setidir. Bu veri seti, İris çiçeğinin üç farklı türüne (Setosa, Versicolor ve Virginica) ait ölçümleri içerir. Veri setinde her bir çiçek türünden 50 örnek bulunmakta olup, toplamda 150 çiçek örneği mevcuttur. Her bir örneğin dört farklı özelliği ölçülmüştür: çanak yaprak uzunluğu, çanak yaprak genişliği, taç yaprak uzunluğu ve taç yaprak genişliği.

**Veri Setinin Elde Edilmesi:**

**Kaynak:** İris veri seti, ilk olarak 1936 yılında Ronald Fisher tarafından toplanmış ve analiz edilmiştir. Bu veri seti, makine öğrenmesi ve istatistik alanlarında klasik bir veri seti olarak kabul edilmekte ve genellikle veri bilimi eğitimi ve araştırmalarında kullanılmaktadır.

**Erişim:** Veri setine erişim, Python'un sklearn kütüphanesi aracılığıyla sağlanabilir. sklearn.datasets modülü altında yer alan load\_iris fonksiyonu kullanılarak veri seti doğrudan yüklenebilir. Bu yöntem, veri setini kolayca kullanılabilir bir formatla sunar ve ön işleme gerektirmeden doğrudan analize başlanmasına olanak tanır.

**Kullanımı:** Yüklendikten sonra, veri seti bir DataFrame'e dönüştürülerek analiz ve işleme için hazır hale getirilir. Python'un pandas kütüphanesi, veri manipülasyonu ve temizleme işlemleri için tercih edilen bir araçtır. Bu kütüphane ile veri seti üzerinde çeşitli işlemler yapılabilir, örneğin eksik verilerin yönetimi, veri türlerinin dönüştürülmesi ve istatistiksel analizler.

Bu adımlar, projemizin veri bilimi aşamasının temelini oluşturmakta ve veri kaynağımızın güvenilir ve erişilebilir olmasını sağlamaktadır

## Verinin Hazırlanması

Bu bölüme veri bilimi kapsamında aşağıdaki başlıklar altına projede yapmış olduğunuz etkinlikleri belirtiniz, Python program kodlarını ver ekran görüntülerini paylaşınız

### Verinin Temizlenmesi

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

### Verinin Birleştirilmesi

Ders sunum ve notlarından faydalanılacaktır.

Ayırıp birleştirme

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Ayrı Özellikleri birleştirme

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

### Verinin Dönüştürülmesi

Ders sunum ve notlarından faydalanılacaktır.

Kategorik Dönüştürme

metin, ekran görüntüsü, çizgi, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

İsim Değiştirme

metin, ekran görüntüsü, çizgi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

### Keşfe Yönelik Veri Analizi (Exploratory Data Analysis)

Bu bölümde verinin temsili, gösterimi ve görselleştirilmesiyle ilgi yapılan etkinlikler, ilgili *grafik ve tablolar*, Python program kodları ver ekran görüntüleri yer alacaktır.

Grafik

metin, ekran görüntüsü, sayı, numara, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

Keşfe Yönelik 1

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Keşfe Yönelik 2

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

# Makine Öğrenmesi Aşaması

## Problemin Çözümüne Yönelik Modeller (Algoritma + Veri)

* Bir öğrenme modelini veri ile birlikte kullanılan algoritma (*Algoritma + Veri*) oluşturmaktadır. Bu bölümde, projede ele alınan problemin çözümüne yönelik kullanabileceğiniz modelleri listeleyiniz.
* Bu modelleri sırasıyla kullanarak projenin başarı ölçütlerini karşılayan en uygun model hakkında karar vermek için sonraki bölümlerdeki etkinlikleri gerçekleştiriniz ve ekran görüntülerini paylaşınız.

Burada kullanacağımız öğrenme modelini veri ile birlikte kullanılan algoritma KNN , KMeans, Naive Bayes, Logistic Regression ve Random Forest. Amacımız bunlar arasından en doğru sonuçları bulup en iyi modeli seçmek. Bunları seçme amacımız en iyi sonuçları bunların verebileceğini düşünmemizdir.

## Modellerin Eğitimi ve Test Edilmesi

Bu bölümde projede ele alınan problemin çözümüne yönelik belirlediğiniz modellerin eğitimi ve test adımlarını gerçekleştirerek bu adımların ilgili çıktılarını, program kodlarını ve ilgili ekran görüntülerini paylaşınız.

KNN

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

KMeansmetin, ekran görüntüsü, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

Naive Bayes



Logistic Regression



Random Forest

metin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

## Modellerin Değerlendirilmesi

Bu noktada en uygun model/modeller :Naive Bayes, Random Forest, Logistic Regression ve Knn olacaktır çünkü doğruluk oranları %100 ve ana model ile aynı

metin, diyagram, ekran görüntüsü, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, diyagram, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

# Proje Sonlandırma Aşaması

## Proje Sonlandırma İşlemleri

**Yazılımın Performansını İzleme ve İyileştirme:** Yazılımın performansını düzenli olarak izlemek ve gerekirse iyileştirmek için analiz yapmak önemlidir. Performans sorunları tespit edildiğinde, yazılımı optimize etmek için düzeltmeler yapılmalıdır.

**Güvenlik Güncellemeleri:** Yazılımın güvenliği sürekli olarak takip edilmeli ve güncel kalması sağlanmalıdır. Güvenlik açıkları tespit edildiğinde, hızlı bir şekilde düzeltme yapılmalı ve güncelleme yayınlanmalıdır.

**Yeni Özelliklerin Eklenmesi:** Müşteri gereksinimleri değişebilir veya yeni teknolojiler gelişebilir. Bu durumda, yazılıma yeni özellikler eklemek veya mevcut özellikleri güncellemek gerekebilir. Bu süreçte, kullanıcı geri bildirimleri ve pazar araştırmaları da önemlidir.

**Yazılım Dokümantasyonunun Güncellenmesi:** Her güncelleme veya değişiklik sonrasında yazılım dokümantasyonu da güncellenmelidir. Bu, yeni özelliklerin veya değişikliklerin nasıl kullanılacağına dair net bir kılavuz sağlar.

**Yazılımın Uyumluluğunun İzlenmesi:** Platformlar ve bağımlılıklar zamanla değişebilir. Yazılımın uyumlu olduğundan emin olmak için düzenli olarak testler yapılmalı ve uyumluluk sorunları tespit edildiğinde bunlar giderilmelidir.

**Kullanıcı Geri Bildirimlerinin Değerlendirilmesi:** Kullanıcı geri bildirimleri, yazılımın güncellenmesi ve geliştirilmesi için değerli ipuçları sağlar. Bu geri bildirimler düzenli olarak değerlendirilmeli ve gerektiğinde yazılım üzerinde değişiklikler yapılmalıdır.

**Yazılımın Yedeklenmesi ve Kurtarılabilirliği:** Yazılımın yedeklenmesi ve kurtarılabilirliği düzenli olarak test edilmeli ve güvenilir bir şekilde sağlanmalıdır. Veri kaybı durumunda, yazılımın hızlı bir şekilde eski haline getirilmesi önemlidir.

## Proje Görevlerinin Takım Üyelerine Dağılımı

Aşağıdaki tabloda projeyle ilgili görev ve sorumlulukların takım üyelerine (görev kodları kullanılarak) nasıl paylaştırıldığı, dağılım oranları gösterilecektir.

| **S.**  **Nu.** | **Projedeki Görevlerin Takım Üyelerine Dağılımı** | **Yüzdelik Dağılımları (%25, %50, %75, %100)** | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **T-1**  **(%)** | **Ü-1 (%)** | **Ü-2 (%)** | **Ü-3 (%)** | **Ü-4 (%)** | **Ü-n (%)** |
| **1.** | **Projenin Tanımı ve Başlatılma Aşaması** | | | | | | |
| 1.1 | Projenin tanımı ve başlatma belgesi | 25 | 50 | 25 |  |  |  |
| **2** | **Veri Bilimi Aşaması** |  | | | | | |
| 2.1 | Verinin hazırlanması, temizlenmesi, birleştirilmesi, görselleştirilmesi | 25 | 50 | 25 |  |  |  |
| 2.2 | Keşfe yönelik veri analizi (exploratory data analysis) | 25 | 50 | 25 |  |  |  |
| **3** | **Makine Öğrenmesi Aşaması** |  | | | | | |
| 3.1 | Modellerin belirlenmesi | 25 | 50 | 25 |  |  |  |
| 3.2 | Modellerin eğitimi ve test edilmesi | 25 | 50 | 25 |  |  |  |
| 3.3 | Modellerin değerlendirilmesi | 25 | 50 | 25 |  |  |  |
| **4** | **Proje Sonlandırma Aşaması** |  | | | | | |
| 4.1 | Proje raporunun hazırlanması | 25 | 50 | 25 |  |  |  |
|  | **ORTALAMA (%)** | 25 | 50 | 25 |  |  |  |