Homework 1

1) How would you define Machine Learning?

Makine Öğrenmesi (Machine Learning-ML), yapay zekanın bir alt bilim dalıdır. Makine öğrenmesi tanım gereği matematiksel ve istatiksel işlemler ile veriler üzerinden çıkarımlar yaparak tahminlerde bulunan sistemlerin bilgisayarlar ile modellenmesidir. Bir diğer ifade ile bilgisayarların açıkça programlanmadan örneklerden öğrenmesini sağlayan algoritmalar geliştirmeyi sağlayan bir alandır. Makine Öğrenmesi geleneksel programlamanın aksine giriş ve çıkış verilerinin işlenmesi ile programı oluşturur. Makine öğrenmesi insanlar için tehlike arz eden alanlarda, insan becerilerinin sürekli tekrarlandığı yani otomatikleştirmeye ihtiyaç duyulan alanlarda bunlarla birlikte Büyük Veri (Big Data), tekrarlayan görevler ve büyük insan gücüne ihtiyaç duyulan alanlarda kullanılabilir.

2) What are the differences between Supervised and Unsupervised Learning? Specify example 3 algorithms for each of these.

Denetimli Öğrenme (Supervised Learning):

Denetimli Öğrenmede (Supervised Learning) tüm veriler etiketlidir ve algoritmalar giriş verilerinden çıktıyı tahmin etmeyi öğrenir. Tahmin modelleri geliştirmek için sınıflandırma(Classification) ve regresyon(Regression) teknikleri kullanılır.

- Sınıflandırma (Classification) Teknikleri:
  Verileriniz etiketlenebilir, kategorilere ayrılabilir veya belirli gruplara sınıflara ayrılabiliyorsa sınıflandırma tekniği kullanılabilir.
- Regresyon (Regression) Teknikleri:
  Sürekli bir değeri tahmin etmek için kullanılır.

Bazı Denetimli Öğrenme Algoritma Örnekleri:

- Linear Regression
- Logistic Regression
- Support Vector Machines (SVMs)

Denetimsiz Öğrenme (Unsupervised Learning):

Denetimsiz Öğrenmede (Unsupervised Learning) tüm veriler etiketsizdir ve algoritmalar, giriş verilerinden doğal yapıyı öğrenir. Tahmin modelleri geliştirmek için kümeleme (Clustering) tekniği kullanılır.

- Kümeleme (Clustering) Teknikleri:
  Genel olarak, kategorize edilmemiş verilerden oluşan bir koleksiyonda bir yapı veya model bulma ile ilgilenir. Kümeleme algoritmaları verilerinizi işler ve verilerde varsa doğal kümeleri (grupları) bulur.

Bazı Denetimsiz Öğrenme Algoritma Örnekleri:

- k-Means Kümeleme
- Hiyerarşik Kümeleme
- Temel Bileşen Analizi (PCA)

Temel Farklar:

- Denetimli öğrenmede modeli denetlemek gerekirken, denetimsiz öğrenmede modelin bilgileri keşfetmesi için tek başına çalışmasına izin verilmelidir.
- Denetimsiz öğrenme algoritmaları, denetimli öğrenmeye kıyasla daha karmaşık işleme görevleri gerçekleştirmenizi sağlar.
- Denetimli öğrenmede sistem öğretilirken, denetimsiz öğrenmede sistem öğretilmiyor, verilerden öğreniyor.

3) What are the test and validation set, and why would you want to use them?

Doğrulama Seti (Validation Set):
Eğitim aşamasında elde edilen modelin performansını değerlendirmek için kullanılan alt bir veri setidir. Ayrıca, bu veri seti hangi modelin iyi olduğunu belirlemek ve modeller için en uygun parametreleri ayarlamak için bir test platformu sağlar.
Test Set:
Yalnızca tam olarak belirlenmiş bir sınıflandırıcının performansını değerlendirmek için kullanılan bir dizi veri. Test seti, modelin eğitim seti dışındaki verilerle ne kadar iyi çalıştığını değerlendirmek için kullanılır.

4) What are the main preprocessing steps? Explain them in detail. Why we need to prepare our data?

1-Veri Temizleme:
Aykırı verilerin(outlier) temizlenmesi, gürültü verilerinin düzeltilmesi, tutarsızlıkların giderilmesi, eksik değerlerin doldurulması

- Veriyi yok sayma, işleme alınmaması
- Eksik olanları elle doldurmak (maaliyetli ve zor)
- Otomatik olarak doldurmak

2-Veri Entegrasyonu:

Veri tabanı, veri küpü veya dosya entegrasyonu. Farklı kaynaklardan elde edilen verilerin birleştirilmesi ve kullanıcılara dönüştürülmüş verinin sunulması.

3-Veri Dönüşümü (Data Transformation):

- Normalizasyon:

Normalizasyon, input değeri indirgemek anlamına gelir. Veriler arasında farklılığın çok fazla olduğu durumlarda verileri tek bir düzen içerisinde ele almaktır. Burada amaç farklı sistemde bulunan verileri ortak bir sisteme taşıyarak karşılaştırılabilmesini sağlamaktır. Yaygın üç yöntemden bahsedebiliriz.

- min-max normalizasyon:
- Z-Score Normalizasyonu:
- Ondalık Normalizasyonu:

4-Veri Azaltma (Data Reduction):

Hacmi küçültme, veri sıkıştırma. Önemsiz niteliklerin kaldırılması

5-Veri Ayrıklaştırma:

Verinin azaltılması. Ön işleme aşaması data mining(veri madenciliği) in başarılı olabilmesi için önemlidir. Ön işleme ile veri, sonraki aşamalarda kullanılabilmesi için elverişli hale getirilir. Bu aşamasının başarısı, sonuçtaki başarıyı doğrudan etkiler. Başarılı bir ön işleme aşamasıyla kesin ve net sonuçlara ulaşmak mümkün olacaktır.

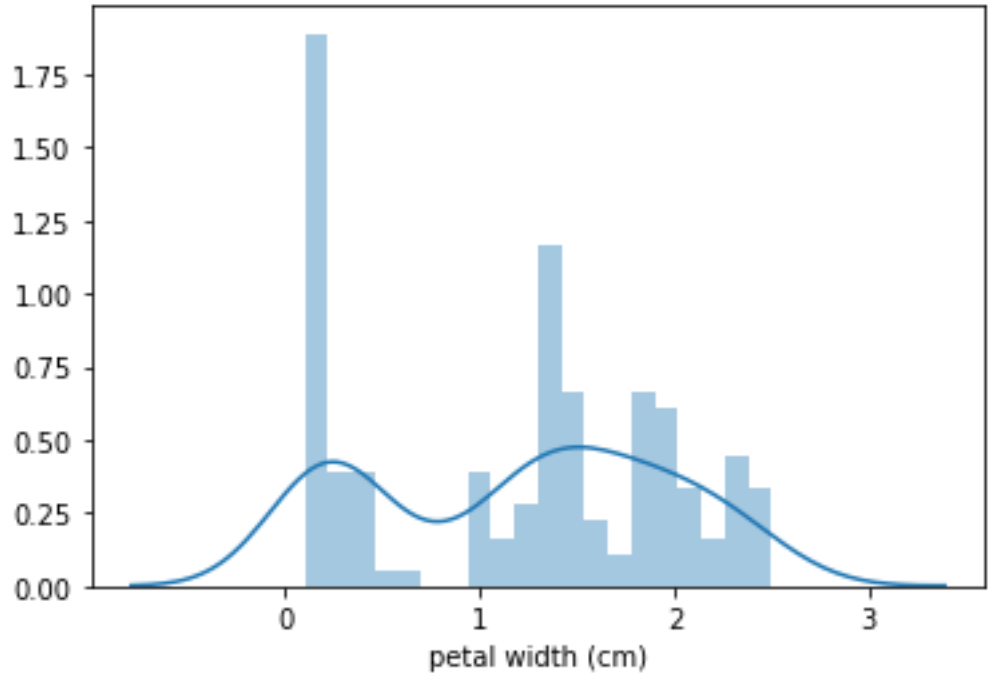
Verilerin Hazırlanması
Verilerimizi uygun bir yere yüklediğimiz ve makine öğrenimi eğitimimizde kullanılmak üzere hazırladığımız aşamadır. Bu aynı zamanda, yararlanabileceğiniz farklı değişkenler arasında ilgili herhangi bir ilişki olup olmadığını görmenize yardımcı olmak ve herhangi bir veri dengesizliği olup olmadığını bize göstermek için verilerinizin uygun görselleştirmelerini yapmak için önemli bir hazırlıktır.

5) How you can explore and analyse countionus and discrete variables?

Ayrık değişken(Discrete Variable), sonsuz veya sayılabilecek şekilde sonsuz sayıda duruma sahip olandır. Bu durumların mutlaka tamsayılar olmadığını unutmayın; ayrıca herhangi bir sayısal değere sahip olmadığı düşünülen durumlar olarak adlandırılabilirler.
Ayrık değişkenler üzerindeki olasılık dağılımı, bir olasılık kütle fonksiyonu (PMF) kullanılarak açıklanabilir.
Sürekli değişken(Countionus Variable), gerçek bir değerle ilişkilendirilir.
Sürekli rastgele değişkenlerle çalışırken, olasılık dağılımlarını bir olasılık kütle işlevi yerine bir olasılık yoğunluk işlevi (PDF) kullanarak tanımlarız.

6) Analyse the plot given below. (What is the plot and variable type, check the distribution and make comment about how you can preproccess it.)

```
Out[1]: <matplotlib.axes._subplots.AxesSubplot at 0x2b8ced01320>
```



Grafik bir histogram grafiğidir. Bu grafik log normal dağılıma benzer. Normal bir dağılım gözl enmediği için outlier bulundurma ihtimali vardır. “zscore” methodu ile outlier detection gerçek leştirilebilir. İmbalanced veriler söz konusu olduğu için median ile doldurma yapılabilir.