Berk Eray Yumuşak

eraayp@gmail.com

Data Science Intern Case Study

1. Giriş (Introduction)

Bu çalışmada, ilaç yan etkileriyle ilgili bir veri seti üzerinde kapsamlı bir analiz gerçekleştirilmiştir. Amacımız, veri setini ayrıntılı bir şekilde inceleyip, veri hazırlığı ve ön işleme adımlarıyla modellemeye uygun hale getirmektir. Özellikle, veri bilimi projelerinde verinin doğru bir şekilde işlenmesi ve modele uygun hale getirilmesi, elde edilecek sonuçların doğruluğu açısından kritik önem taşımaktadır.

Projenin başlangıç aşamasında, Keşifçi Veri Analizi (EDA) ile veri setinin genel yapısı, eksik veriler ve anormallikler analiz edilmiştir. Bu aşamada Python kütüphaneleri olan Pandas, Seaborn, ve Matplotlib kullanılarak veri görselleştirmeleri yapılmış ve değişkenler arasındaki ilişkiler araştırılmıştır.

EDA sonucunda veri setindeki eksik verilerin olmadığı tespit edilmiştir. Kategorik değişkenler ve sayısal değişkenler birbirinden ayrılmış, ardından kategorik değişkenler için One-Hot Encoding işlemi, sayısal değişkenler için ise StandardScaler kullanılarak standartlaştırma işlemi gerçekleştirilmiştir.

Bu adımlar, ilerleyen süreçte tahmin modelleri oluşturmak için gerekli olan veriyi hazır hale getirmekte önemli bir rol oynamıştır. Kodlama sırasında modüler bir yapı benimsenmiş, kodun tekrarlanabilirliği ve anlaşılabilirliği sağlanmıştır. Bu süreç boyunca PEP-8 standartlarına uygun yazım kuralları da göz önünde bulundurulmuştur.

Bu çalışma, ilaç yan etkilerini analiz etmeyi ve modelleme süreçlerine sağlam bir temel oluşturmayı hedeflemektedir.

2. Keşifçi Veri Analizi (EDA)

Veri Seti Görünümü: Veri setimizde 2357 gözlem ve 18 değişken bulunmaktadır. Değişkenlerimizin 12 tanesi kategorik, 2 tanesi sayısal ve 4 tanesi tarih değişkenleri olarak girişmiştir.

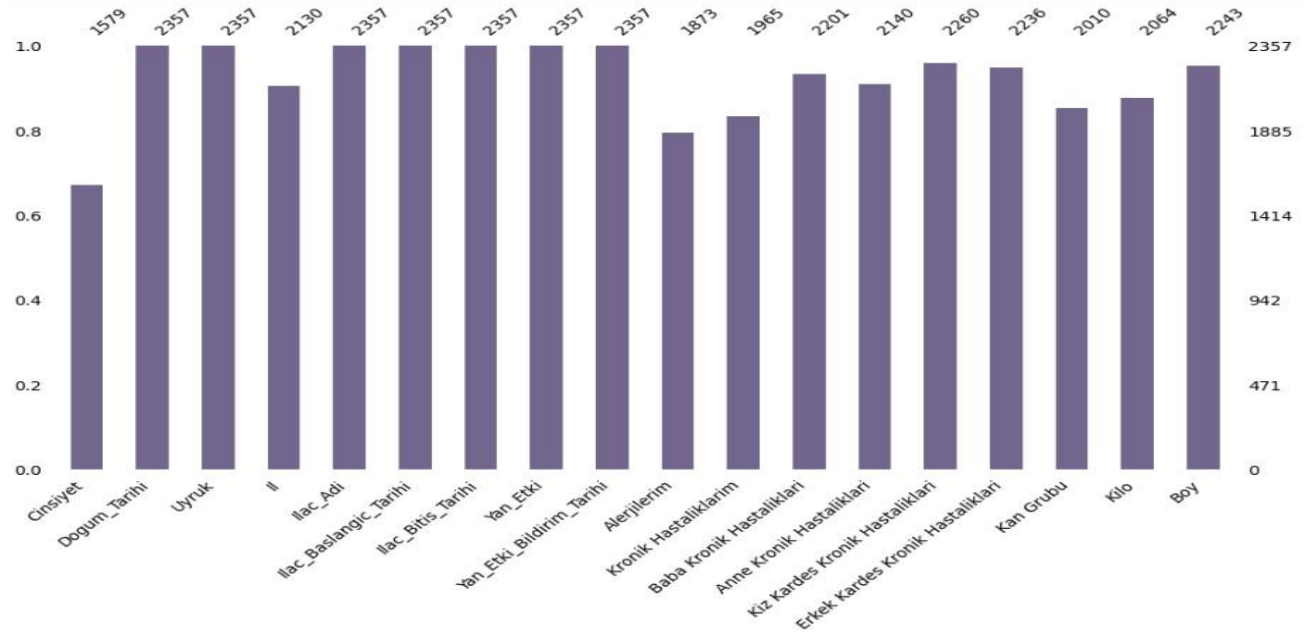
Veri Seti Problemleri

-Eksik Veri Problemi: Değişkenlerin eksik veri frekansları ve yüzdesel olarak oranları aşağıdaki gibidir;

	n_miss	ratio
Cinsiyet	778	33.01
Alerjilerim	484	20.53
Kronik Hastaliklarim	392	16.63
Kan Grubu	347	14.72
Kilo	293	12.43
Il	227	9.63
Anne Kronik Hastaliklari	217	9.21
Baba Kronik Hastaliklari	156	6.62
Erkek Kardes Kronik Hastaliklari	121	5.13
Boy	114	4.84
Kiz Kardes Kronik Hastaliklari	97	4.12

En fazla eksik veri ‘Cinsiyet’ değişkenindedir. Bu değişkendeki eksiklikler veri girişinden kaynaklanabileceği gibi kişinin belirtmek istememesinden de olabilir. Diğer eksik verilere baktığımızda ‘Alerjilerim’, ‘Kronik Hastaliklarim’ ve ‘Kan Grubu’ gibi eksik veriler bilinen bir hastalığı olmadığına işarettir. Veri Ön Hazırlık sırasında eksik veriler problemi ele alınırken bu durumlar göz önüne alınmıştır.

Bkz. Eksik Veri Tablosu



-Tekilleştirme Problemi: ‘Kronik Hastaliklarim’, ‘Anne Kronik Hastaliklari’, ‘Baba Kronik Hastaliklari’ gibi değişkenlerde tek satırda birden fazla veri girişi olduğu için bu bir problemdir.

Kronik Hastaliklarim	Baba Kronik Hastaliklari
Hipertansiyon, Kan Hastaliklari	Guatr, Hipertansiyon
NaN	Guatr, Diger
Kalp Hastaliklari, Diyabet	Diyabet, KOAH
Diyabet, Diger	Kalp Hastaliklari, Diger
Diyabet, Kalp Hastaliklari	Alzheimer, Hipertansiyon
...	...

Her satır tek değişken taşıması gerektiği için ilgili veriler üzerinde bu işlemler yapılmıştır.

-Veri Girişi Problemleri: Veri girişinden kaynaklanabilecek problemler mevcuttur. Örneğin ilaçları kremler olarak filtreleyip yan etkilerine baktığımızda aşağıdaki gibi bir tablo ile karşı karşıya kalıyoruz.

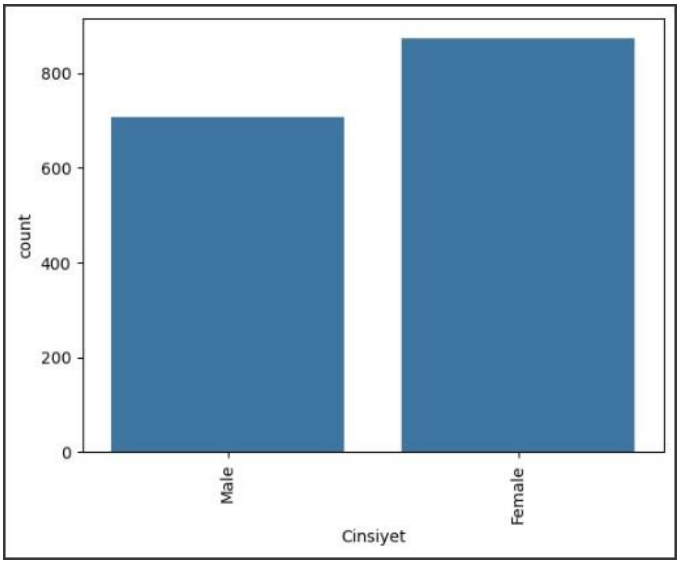
Yan_Etki	
Agizda Farkli Bir Tat	15
Gormede Bulaniklik	12
Tansiyon Yukselme	11
Ishal	11
Kabizlik	9
Karin Agrisi	9
Yorgunluk	8
Mide Bulantisi	8
Az Uyuma	7
Kas Agrisi	6
Istah Artisi	4
Uykululuk Hali	4
Gec Bosalma	4
Deride Morarma	4

Kremlerin ‘Ağızda Farklı Bir Tat’, ‘Görmede Bulanıklık’ veya ‘Tansiyonda Yükselme’ gibi yan etkilerinin görünmesi normal bir durum değildir. Bu durum veri girişinden mi kaynaklanıyor yoksa kremlerin yanlış kullanımından mı kaynaklanıyor bilinmediği için veriler üzerinde herhangi bir düzeltltme işlemi yapılmamıştır.

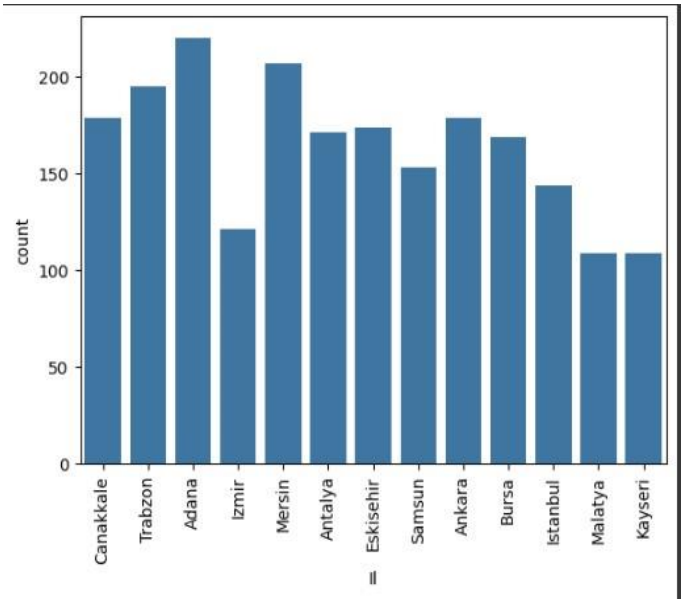
(Lütfen kremleri yemeyiniz!)

BULGULAR

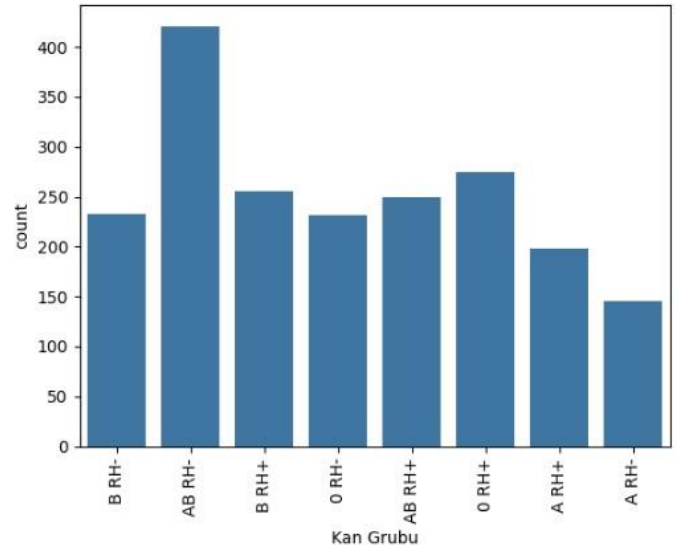
Kategorik verilerin frekans grafikleri



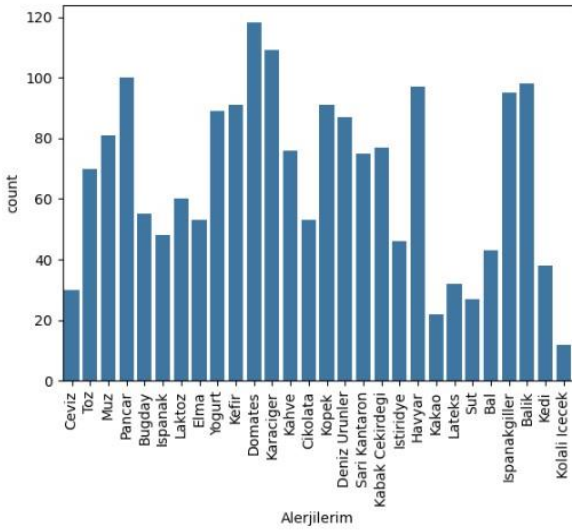
Kadın hasta sayısı, erkek hasta sayısına göre daha fazladır.



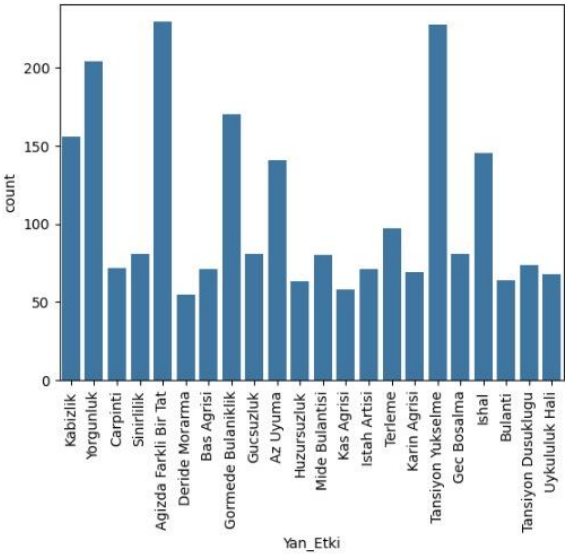
İl bazlı baktığımız zaman verilerimizdeki en yüksek katılım Adana'dan daha sonra Mersin'dendir. En az katılım iste Malatya, Kayseri ve İzmir'dendir.



Kan grubu frekanslarına baktığımız zaman AB RH- hariç diğer kan gruplarının frekansları birbirine benzerken AB RH- en çok görünen kan grubudur.

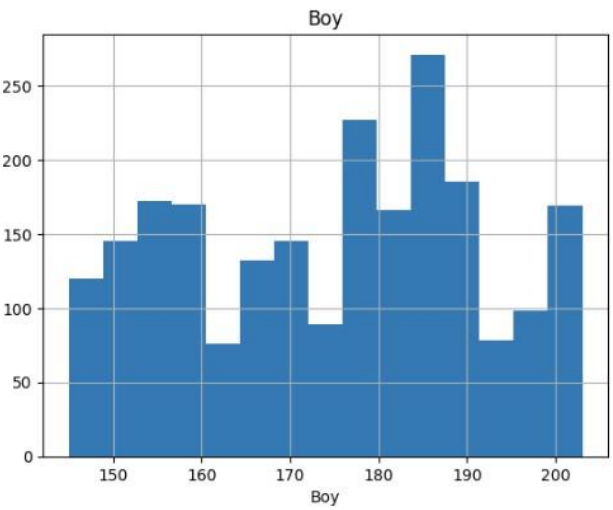


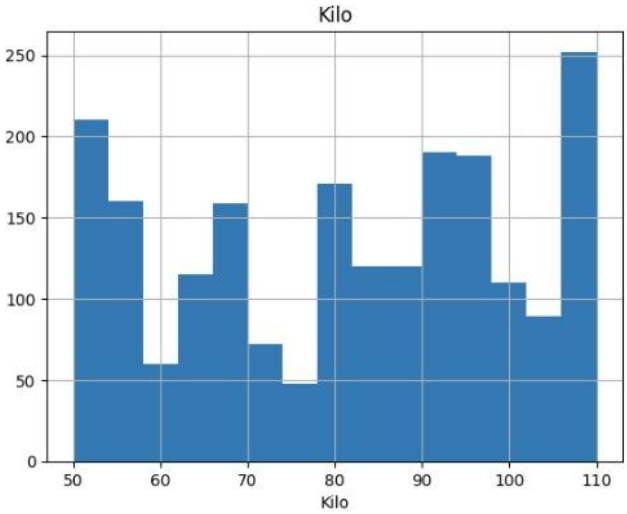
Alerjilerin frekanslarına baktığımız zaman ek sık görülen alerjiler ‘Domates’ ve ‘Karaciğer’dir.



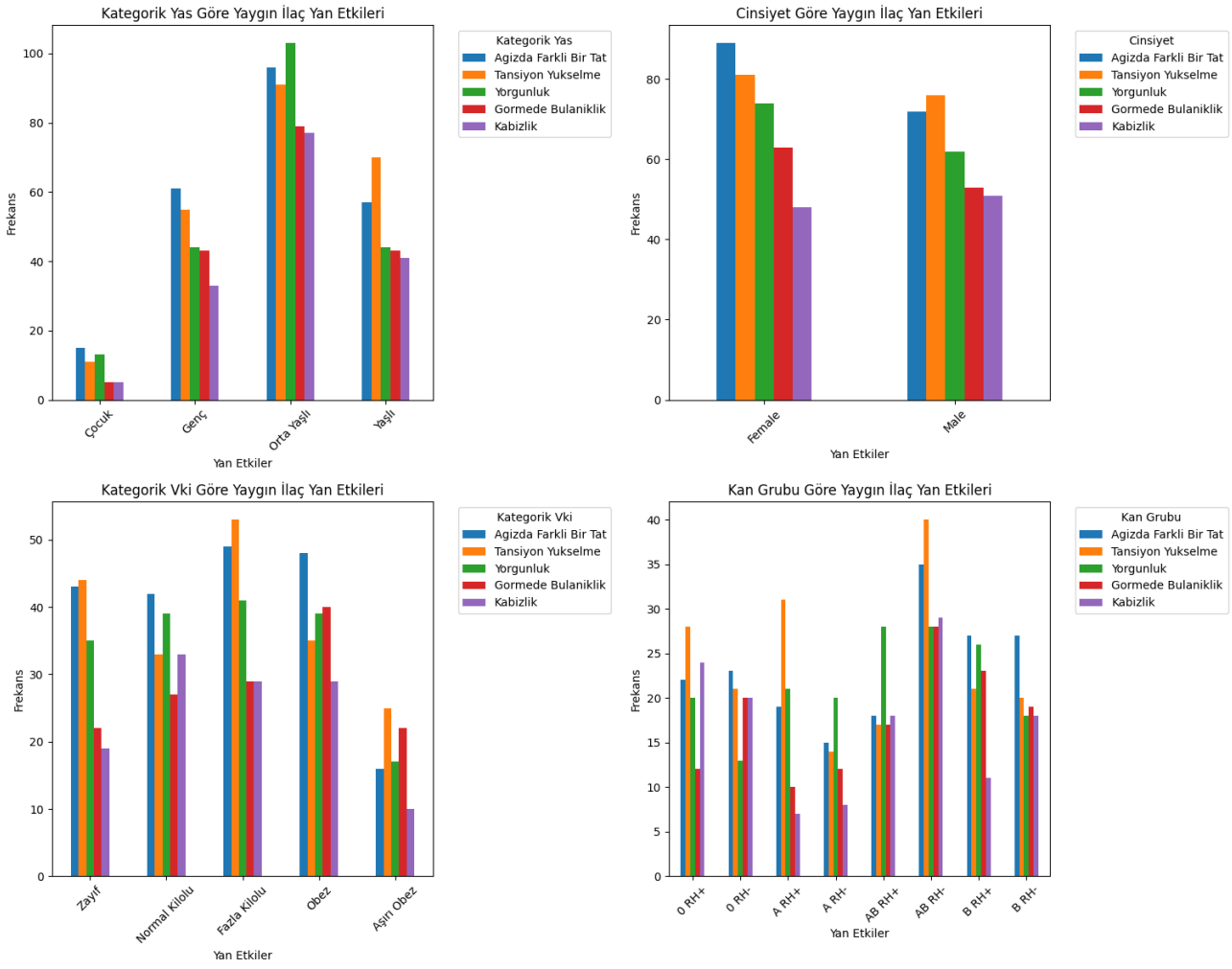
Yan etkiler çalışmamızın ana eksenini oluşturmaktadır. Bu nedenle yan etkilerin frekanslarını verdikten sonra yan etkiler ile yaptığımız diğer analiz sonuçlarını da vereceğiz. Buradaki grafikten de göründüğü gibi ‘Ağızda Farklı Bir Tat’, ‘Tansiyonda Yükselme’ ve ‘Yorgunluk’ en fazla görülen yan etkilerdir.

Sayısal Veriler





Analizimizi derinleştirmek için en fazla görülen 5 yan etkiyi filtreliyoruz.

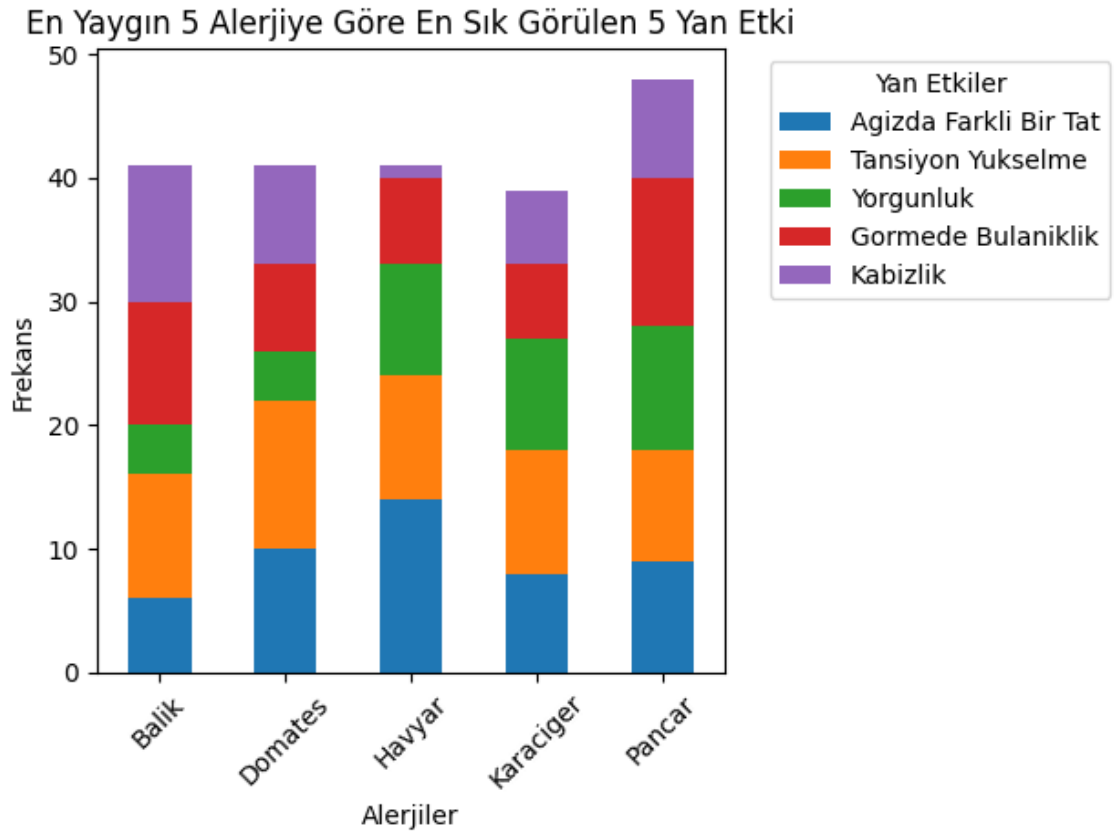


Bu grafiklere baktığımız zaman kadınlarda en fazla görülen yan etki ‘Ağızda Farklı Bir Tat’ iken erkeklerde ‘Tansiyonda Yükselme’dir. Buradan çıkarılacak olan yorum erkeklerin kadınlara oranla tansiyonda yükselme yan etkisiyle daha fazla karşı karşıya olması olasıdır.

Diğer bir yorum yaş kategorilerine göre baktığımız zaman ‘Yorgunluk’ probleminden en fazla etkilenen ‘Orta Yaşlı’lardır. Her ne kadar bu durum frekanslara göre de olsa orta yaşlılarda yorgunluk yan etkisi daha fazla görüşmüştür.

Kilo kategorilerine ayırdığımız grafiğe baktığımız zaman ‘Ağızda Farklı Bir Tat’ ve ‘Yorgunluk’ yan etkisi sadece ‘Obez’ kategorisinde diğer yan etkilerden fazla çıkmıştır. Belirli vücut kitle endeksine sahip insanların bu yan etkileri daha fazla görmesi olasıdır.

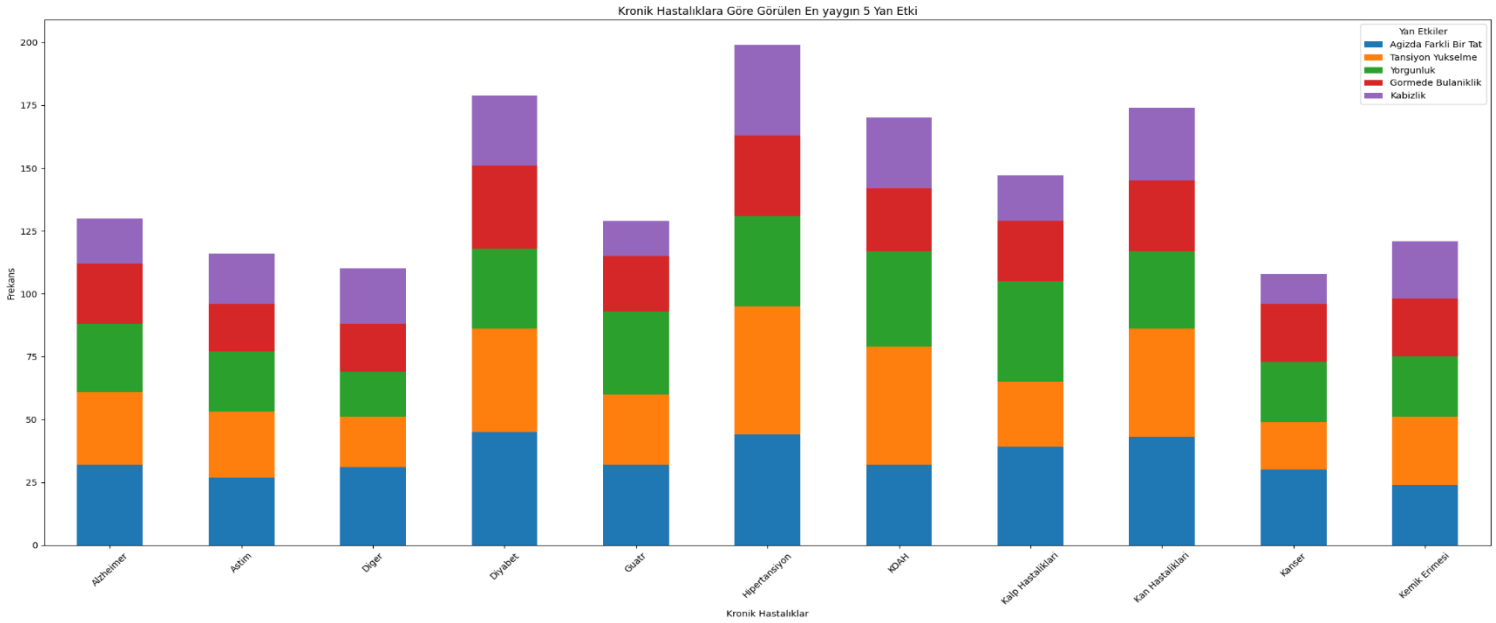
Kan grubu ile ilgili grafiğimize baktığımız zaman ‘B+’ ve ‘B-‘ kan grubunda ‘Ağızda Farklı Bir Tat’ yan etkisi ‘Tansiyonda Yükselme’ye göre gözle görünür derecede fazla çıkmıştır. B Grubuna dahil olanların ‘Tansiyonda Yükselme’ yan etkisini ‘Ağızda Farklı Bir Tat’ yan etkisine göre daha az göreceğini söylemek mümkündür. ‘A RH+’ kan grubunda ise ‘Tansiyonda Yükselme’ yan etkisini diğer yan etkilere göre çok daha fazla görülmüştür.



En fazla görülen 5 alerjiye göre en yaygın 5 yan etkiyi gösteren tabloya baktığımız zaman farklı yorumlar yapmak mümkündür. 'Havyar' alerjisi olanların diğer alerjilere göre 'Kabızlık' yan etkisini görme olasılığı daha düşüktür. Fakat bu alerjiye sahip olanlar 'Ağızda Farklı Bir Tat' yan etkisini diğerlerine göre daha fazla görmüştür.

Son bulgumuz ise tekilleştirdiğimiz kronik rahatsızlıkların frekansı ve bu kronik rahatsızlıklarda görülen yan etkilerdir.

Bu grafiğe baktığımız zaman tekilleştirilmiş kronik hastalıklarda en fazla görülen kronik rahatsızlık Hipertansiyon olmuştur. Bu 'Tansiyonda Yükselme' yan etkisinin fazla görülmesin bir nedeni olabilir.



Tekilleştirilmiş kronik hastalıklara baktığımız zaman tahmin ettiğimiz gibi ‘Tansiyonda Yükselme’ ‘Hipertansiyon’ hastalarında daha fazla görülmüştür. Bu hastaların kullandığı ilaçları daha iyi seçmesi gerekir. Bu konuda doktorunun önerdiği ilaçlar dışında ilaç kullanımı kronik rahatsızlığını tetikleyebilir.

3. Veri Ön İşleme Adımları (Data Preprocessing)

- Eksik Veriler: Eksik veriler doldurulurken farklı yaklaşımlar izlenmiştir. Sayısal değerleri doldururken KNNImputer yöntemi doldurma işlemi yapılmıştır. Kategorik veriler doldurulurken ‘Il’ değişkeni en çok tekrar eden değerle doldurulmuştur. Hastalıklarla ilgili kategorik veriler ‘Bilinen Bir Hastalığı Yok’ ile doldurulurken diğer kategorik veriler ‘Bilinmiyor’ ile doldurulmuştur. Bunun en önemli sebebi hastalıkla ilgili eksik değer farklı anlamlara gelebilir.

- **Tekilleştirme işlemi:** Tekilleştirme işlemi yapılırken virgül ile ayrılan değerler ayrılmış ve her bir değer ayrı bir satıra aktarılmıştır.

-**Kategorik Verilerin Kodlanması:** Kategorik veriler kodlanırken OneHotEncoder kullanılmıştır.

-**Sayısal Değerlerin Standartlaştırılması:** StandartScaler ile sayısal veriler standartlaştırılmış ve veriler modelleme aşamasına hazır hale getirilmiştir.

4. Kod Yapısı (Code Structure)

Kodlar PEP-8 kodlama standartlarına uygun olarak yazılmıştır.

5. Sonuç (Conclusion)

Bu çalışmada ilaç yan etkileriyle ilgili geniş bir veri seti üzerinde kapsamlı bir analiz gerçekleştirilmiştir. Keşifçi Veri Analizi (EDA) sonucunda, veri setindeki eksik veriler tespit edilmiş ve verinin genel yapısı hakkında önemli bilgiler elde edilmiştir. Kategorik ve sayısal veriler uygun şekilde işlenmiş ve modellemeye hazır hale getirilmiştir.

Ana Bulgular

1. Kategorik Veriler:
    - Kadın hasta sayısı erkek hasta sayısından fazladır.
    - İl bazında en yüksek katılım Adana ve Mersin'den, en düşük katılım ise Malatya, Kayseri ve İzmir'dendir.
    - Kan grubu frekansları incelendiğinde AB RH- en sık görülen kan grubudur.
    - En sık görülen alerjiler domates ve karaciğerdir.
    - En yaygın yan etkiler ağızda farklı bir tat, tansiyonda yükselme ve yorgunluktur.
2. Sayısal Veriler:
    - Kadınlarda en fazla görülen yan etki ağızda farklı bir tat iken, erkeklerde tansiyonda yükselmedir.
    - Orta yaşlılar, yorgunluk yan etkisinden en fazla etkilenen gruptur.
    - Obez bireylerde ağızda farklı bir tat ve yorgunluk yan etkileri daha yaygındır.
    - B+ ve B- kan gruplarında ağızda farklı bir tat yan etkisi, tansiyonda yükselmeye göre daha fazladır. A RH+ kan grubunda ise tansiyonda yükselme yan etkisi diğer yan etkilere göre daha fazladır.
3. Kronik Hastalıklar ve Yan Etkiler:
    - Hipertansiyon en fazla görülen kronik rahatsızlıktır ve bu hastalarda tansiyonda yükselme yan etkisi daha fazla görülmüştür.
    - Kronik hastalıkları olan bireyler, özellikle tansiyonda yükselme riski nedeniyle ilaç seçiminde dikkatli olmalıdır.

Veri Ön İşleme Adımları

- Eksik veriler, sayısal değerler için KNNImputer, kategorik veriler için ise en sık tekrar eden veya uygun bir kategorik değerle doldurulmuştur.
- Virgül ile ayrılan değerler tekilleştirilmiş ve her bir değer ayrı bir satıra aktarılmıştır.
- Kategorik veriler One-Hot Encoder ile kodlanmış, sayısal veriler ise StandardScaler ile standartlaştırılmıştır.

Bu çalışma, ilaç yan etkilerini analiz etmeye ve modelleme süreçlerine sağlam bir temel oluşturmaya odaklanmıştır. Elde edilen bulgular ve uygulanan veri işleme adımları, gelecekte yapılacak tahmin modellerinin doğruluğunu artıracaktır. Kodlama sürecinde modüler yapı ve PEP-8 standartlarına uygunluk sağlanarak kodun tekrarlanabilirliği ve anlaşılabilirliği desteklenmiştir. Bu sayede, veri bilimi projelerinde verinin doğru işlenmesi ve modele uygun hale getirilmesi önemle vurgulanmıştır.