**YMH418 - Yazılım Mühendisliği Güncel Konular Genel Rapor**

1. **Giriş**

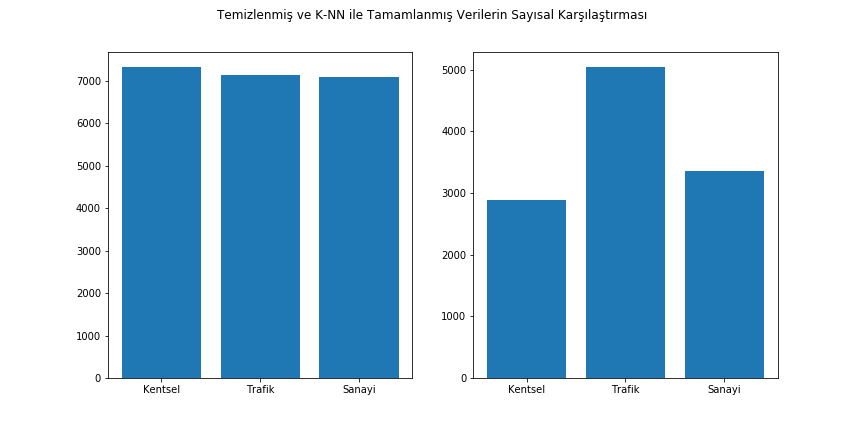
Bu rapor önceki aşamaları, riskleri ve B planını içermektedir.

1. **Rapor**

Birinci aşamada, veri toplamak için tek şehir kullanılmaya karar verildi ve Kayseri’deki 3 istasyondan 2000 yılının başından 2020 Nisana kadar veri toplandı. Toplanan bu veride çok fazla eksik olması ve sayısının az olması nedeniyle, şehir değiştirildi ve Ankara seçildi. Ankara ilinden 3 istasyon tipinde en çok veri bulunduran birer istasyon seçildi. Bu istasyonlar; Bahçelievler, Siteler, Sıhhiye istasyonlarıdır. Veri setinde seçilen alanlar tüm istasyonlardaki ortak olan alanlar olarak belirlendi, daha sonra Gece/Gunduz ve Haftaici/Haftasonu alanları eklendi. Boş satırlar tamamen temizlendi.

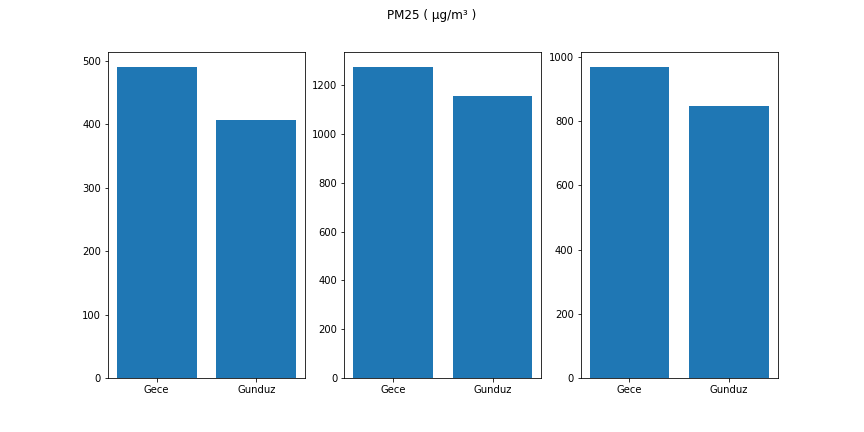
İkinci aşamada, veri görselleştirilmesi yapıldı. Nisan ayında değerlerde bir azalma olduğu görüldü. Bu iş için büyük ölçüde Python dilinde bulunan matplotlib kütüphanesi kullanıldı.

Üçüncü aşamada, eksik değer bulunduran tüm veriler silindi. Aynı zamanda tüm veriler daha sonra lazım olması ihtimaline karşın saklandı. Yapılan bu işlem veri sayısını yarı yarıya düşürdü. Bu temizlenmiş verinin bir özeti çıkarıldı. Bu özetin içerisinde; ortalama, standart sapma gibi bilgiler bulunmaktadır.

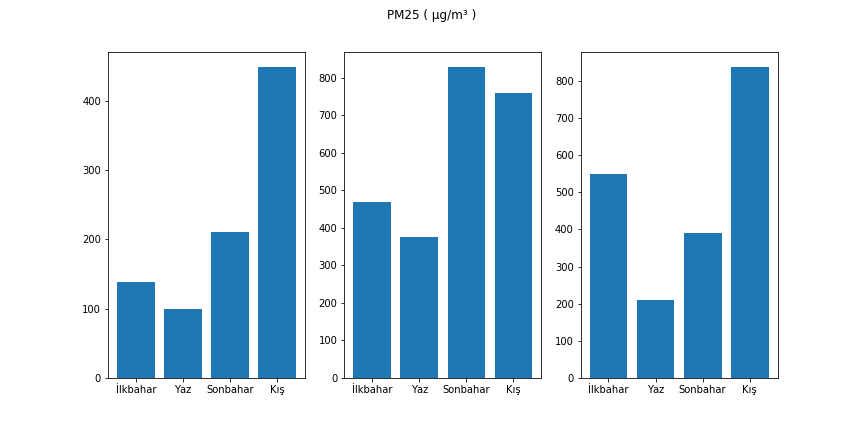


Şekil 1. Tamamlanmış (solda) ve temizlenmiş veri sayılarının grafiği.

Temizlenen veri setinde bulunan tüm gaz ölçümleri 4 mevsim ve gece/gündüz olmak üzere grafikler oluşturularak gösterildi.



Şekil 2.1. PM25 gece/gündüz karşılaştırması grafiği



Şekil 2.2. PM25 mevsime göre karşılaştırması grafiği

Dördüncü aşamada, verinin hangi tip istasyondan geldiğini tahmin eden bir model oluşturuldu fakat bu modelin başarımının çok düşük olduğu gözlendi. Başarımın yükselmesi için bazı araştırmalar yapıldı. Sebebin veri sayısı olabileceği anlaşıldıktan sonra sklearn kütüphanesinin K-NN ile veri doldurma özelliği kullanılarak eski verilerdeki boş değerler tamamlanıp veri seti değişitirildi. Tarih alanı sadece yıl olacak şekilde güncellendi. Yeniden model oluşturulduğunda bu sefer daha yüksek olduğu görüldü daha fazla artırmak için LSTM modeli denendi ve başarımın arttığı gözlendi. Veri setinin bir parçası test verisi olarak kullanıldı ve modelin başarımı ölçüldü. Eğitim Google Colab üzerinde GPU ile yapıldı daha sonra model kaydedildi fakat bazı dosyaların Google Colabten alınırken kaybolması nedeniyle kaydedilen model kullanılamadı bu nedenle model yeniden eğitilecektir.

1. **Risk ve B Planı**

Geliştirilen modelin doğruluk değeri yeterli olsa da kayıp değeri de yüksektir. Yapılan araştırmalar sonucu bu durumun nedeninin overfitting olabileceği anlaşılmıştır bu nedenle eğer model bakanlık sisteminden alınan verilerde başarılı olamazsa projenin kapsamı genişletilerek İç Anadolu’da bulunan başka şehirlerden alınan veriler nüfusa göre düzenlenip eski verilere benzetildikten sonra projeye devam edilecektir.