İskenderun Teknik Üniversitesi

Mühendislik ve Doğa Bilimleri Fakültesi

Bilgisayar Mühendisliği Bölümü

Parkinson Hastalığının Makine Öğrenmesi ile Tespiti

Bitirme Projesi-I

192503043 İrem Cing

Dr. Öğr. Üyesi Ahmet GÖKÇEN

Parkinson Hastalığı Nedir

Parkinson hastalığı, beynin alt kısımlarındaki gri cevher çekirdeklerinin bozukluğuna bağlı bir sinir sistemi hastalığıdır. Parkinson hastalığı, parkinsonizm sendromunun en sık görülen varyantıdır. Adını hastalığı ilk defa 1817'de titremeli felç olarak tarifleyen James Parkinson'dan almıştır. Bu hastalığın görülme sıklığı 1000'de 1'dir. Parkinson, 60 yaşın üzerindeki bireylerin %1'inde görülürken, 85 ve üzeri yaşlarda bu oran %5'lere çıkmaktadır. Parkinson, harekette yavaşlık (bradykinesia), titreme ve kasılma olarak karakterize edilmektedir. Bunlara ek olarak uyku bozukluğu, depresyon belirtileri ve konuşma bozukluğu görülmektedir. Konuşma bozukluğu kısık sesle konuşma, donuk konuşma, konuşmaya başlayamama, telaffuz hataları, konuşurken ses yüksekliğini ayarlayamama gibi sosyal hayatı etkileyebilen zorlukları içermektedir.

Kişinin Parkinson olup olmadığı basit bir test ile anlaşılamamaktadır. Nöroloji uzmanı bir doktor hastalığa teşhis koyabilmek ve rahatsızlığın başka bir hastalık durumundan kaynaklanıp kaynaklanmadığını anlamak amacıyla hastalardan biyokimyasal testler ve beyin tomografisi ister. Ek olarak, bacak ve kolların işlevsel yeterliliği, kas durumu, serbest yürüyüş ve dengeyi sağlayabilme durumlarını değerlendirmek için bazı fiziksel testler istemektedir. Hastalar genellikle 60 ve üzeri yaşlarda olduğu için istenen testler, bu yaşlardaki insanlara zor gelmektedir. Tüm bu zorluklar sebebiyle Parkinson'un tanısında daha basit ve güvenilir yöntemlere ihtiyaç duyulmaktadır.

Parkinson hastalığının teşhisinin hızlı ve güvenli yapılabilmesi için son yıllarda pek çok bilimsel çalışma yapılmıştır. Bu çalışmaların en önemli amacı hastalar için fiziksel zorlukları ortadan kaldırmak ve klinik çalışanlarının üzerindeki iş yükünü azaltmaktır.

Parkinson'un erken evrelerinde ve en sık görülen rahatsızlıklarından biri vokal (konuşma ve ses) problemlerdir. Yakın geçmişte bireylerin konuşma kayıtları (sesleri) kaydedilerek Parkinson teşhisi üzerinde çalışmalar gerçekleştirilmiştir. Vokal çalışmalar literatürde belirgin bir yere sahiptir.

Bu çalışmanın amacı Parkinson teşhisi için makine öğrenmesi tabanlı yeni yaklaşımlar sunmaktır. Günümüzde Parkinson teşhis etmek için kullanılan yöntemler hastalar için fazladan efor sarf etmelerine ve klinik çalışanları için zaman kaybına sebep olabilmektedir. Bu projede amaç bireylerden alınan ses verileri kullanılarak Parkinson'un çok daha hızlı ve daha kolay bir şekilde teşhis edilmesidir. Bireylerden alınan ses verileri, makine öğrenme algoritmalarının kullanımıyla Parkinson teşhisinde önemli rol oynamaktadır. Bu yöntemin uygulamasında hasta ve sağlıklı bireylerden alınan ses verileri kullanılarak makine öğrenmesi işlemi yapılıyor. Öğretilen ağ sayesinde yeni bireyin ses verilerinden hastalık teşhisine (varsa) gidilebiliyor.

MATERYAL VE YÖNTEMLER

Veri setinde dengeleme işlemi yapıldı. Dengeleme işleminde aşağıdaki yöntemler kullanıldı.

1. **Undersampling**
2. **Oversampling**
3. **SMOTE**

1.**Undersampling** (eksik örnekleme), sınıf dağılımları eşit olana kadar çoğunluk olan sınıfının örneklerini ortadan kaldırarak veri kümesini yeniden dengelemeyi amaçlar. Bu yöntemin en büyük dezavantajı, yetersiz gözlem sayısı olan projelerde katkı sağlayabilecek gözlemleri de veri setinden kaldırmasıdır. Ayrıca, az sayıda gözlemin olduğu durumunda örnek uzayın rastgeleliği zarar görebilir.

2. **Oversampling** (aşırı örnekleme), eşit sınıf dağılımları elde edilene kadar azınlık sınıfının örneklerini çoğaltır. Bu konudaki yöntemlerinin çoğu, azınlık sınıfının örneklerini kopyaladığından, aşırı öğrenme(overfiting) olma olasılığı artar. Ayrıca, yüksek düzeyde dengesiz dağılıma sahip büyük bir veri kümesi olması durumunda, aşırı örnekleme hesaplama açısından çok maliyetli olabilir.

3. **SMOTE** (Synthetic Minority Over-Sampling Technique), sentetik veri üretilmesini sağlayan bir aşırı örnekleme sürecidir. Veri bilimi projelerinden en sık kullanılan yöntemlerden biridir.

Yöntemin ana fikri, azınlık sınıfının örnekleri arasında belirli işlemler yaparak yeni azınlık sınıfı örnekleri yaratmaktır.

Sentetik örnekler şu şekilde üretilir:

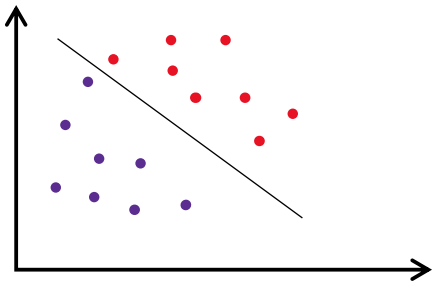
1. İncelenen öznitelik vektörü(Ei) ile en yakın komşusu arasındaki farkı alınır,
2. Bu farkı 0 ile 1 arasında rastgele bir sayı(δ) ile çarpılır,
3. Çıkan sonuç incelenen özellik vektörüne eklenir ve yeni örnek oluşur.

$$E_{yeni} = E_i + (E_i - E_j)\,\delta$$

Gereken aşırı örnekleme miktarına bağlı olarak, en yakın k komşudan komşular rastgele seçilir. Bu işlem, aşırı öğrenme sorununun önüne geçer ve iyi bir sınıflandırma performansı ile sunar.

Veri seti dengelendikten sonra öznitelik seçme algoritması ile en ilgiliden en ilgisize doğru sıralandı. Öznitelik sıralama algoritması olarak Support Vector Machine (Destek Vektör Makineleri) seçildi.

Support Vector Machine, genellikle sınıflandırma problemlerinde kullanılan gözetimli öğrenme yöntemlerinden biridir. Bir düzlem üzerine yerleştirilmiş noktaları ayırmak için bir doğru çizer. Bu doğrunun, iki sınıfının noktaları için de maksimum uzaklıkta olmasını amaçlar. Karmaşık ve küçük, orta veri setleri için uygun olduğundan bu algoritma seçilmiştir.



*En yakın iki tanesi support vector hiper düzleme en yakın olanlardır.

Bu veri setinin düzenlenmesi, dengelenmesi, hesaplanması da dahil tüm işlemler Google Colab ortamında yapıldı.

PARKİNSON HASTALIĞI VERİ SETİ

Çalışmada kullanılan veri seti Extremadura Üniversitesi Matematik Bölümü kaynaklı- Machine Learning Repository- (UCI)'den alınmıştır. Veri seti 240 satır, 48 sütun içermektedir. Veri setinde 40 Parkinson hastası birey ve 40 sağlıklı bireyin 'a' sesini çıkardıkları 3 ses kaydı kopyasından çıkarılan akustik özellikleri içerir.

Veri setinde 0 sağlıklı 1 PH kişileri temsil eder.

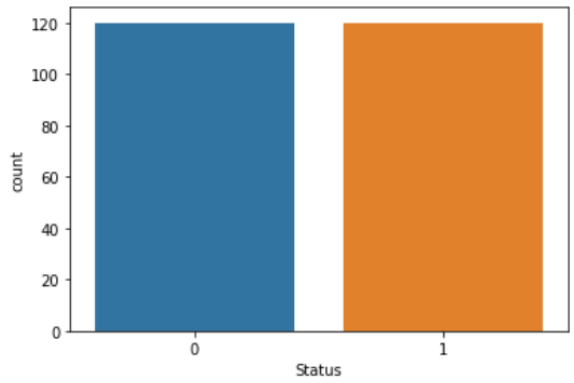
Öznitelik Bilgileri:

1. Kimlik: Deneklerin tanımlayıcısı.
2. Kayıt: Kayıt numarası.
3. Durum: 0=Sağlıklı; 1=PD
4. Cinsiyet: 0=Erkek; 1=Kadın
5. Perde yerel pertürbasyon ölçümleri: göreceli jitter (Jitter_rel), mutlak jitter (Jitter_abs), göreceli ortalama pertürbasyon (Jitter_RAP) ve perde pertürbasyon bölümü (Jitter_PPQ).
6. Genlik pertürbasyon ölçümleri: yerel shimmer (Shim_loc), dB cinsinden shimmer (Shim_dB), 3 noktalı amplitüd pertürbasyon bölümü (Shim_APQ3), 5 noktalı amplitüd pertürbasyon bölümü (Shim_APQ5) ve 11 noktalı amplitüd pertürbasyon bölümü

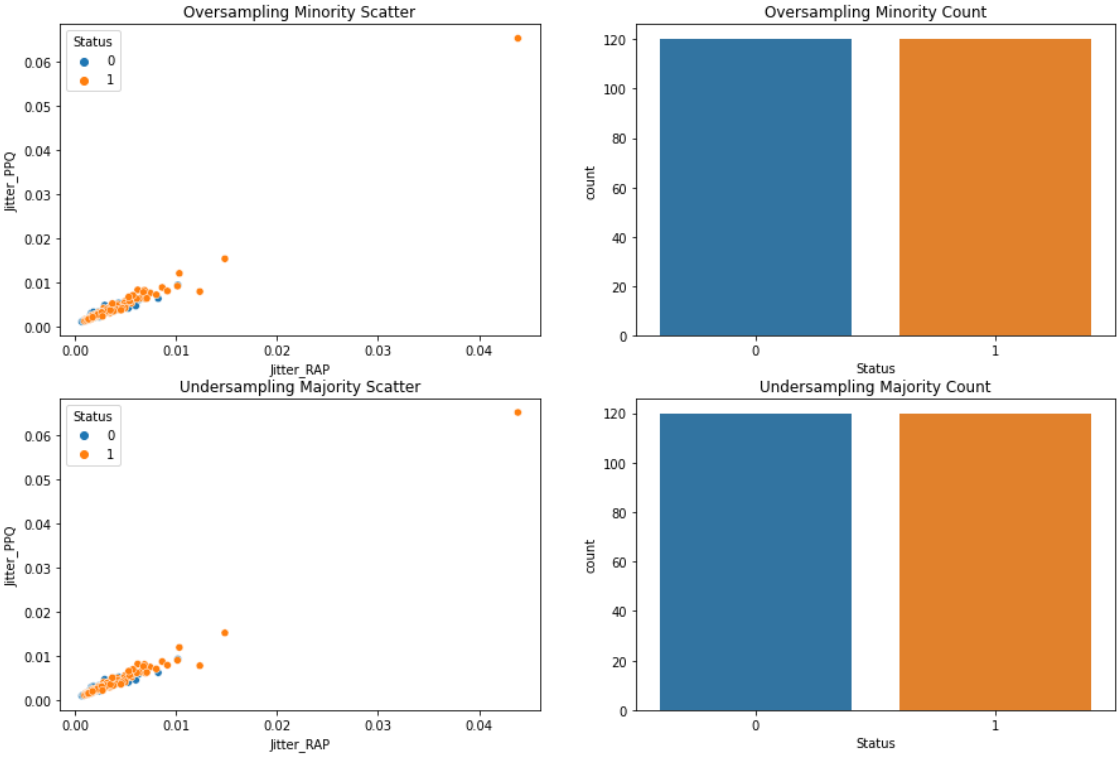
(Shim_APQ11).
7. Harmonik-gürültü oranı ölçümleri: 0-500 Hz (HNR05), 0-1500 Hz (HNR15), 0-2500 Hz (HNR25), 0-3500 Hz frekans bandında harmonik-gürültü oranı Hz (HNR35) ve 0-3800 Hz'de (HNR38).
8. 0'dan 12'ye kadar (MFCC0, MFCC1,..., MFCC12) ve bunların türevleri (Delta0, Delta1,..., Delta12) Mel frekansı cepstral katsayısına dayalı spektral ölçümler.
9. Yineleme periyodu yoğunluk entropisi (RPDE).
10. Trendsiz dalgalanma analizi (DFA).
11. Perde periyodu entropisi (PPE).
12. Glottal-gürültü uyarım oranı (GNE).

BULGULAR

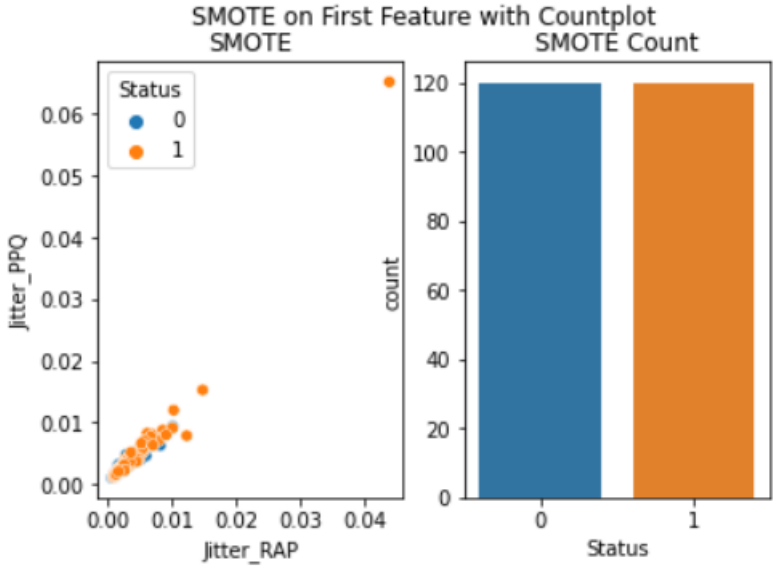
Kaç adet hasta ve sağlıklı birey olduğu kod ile grafik şeklinde gösterilmiştir.



Veri seti dengeleme seçeneklerine göre verinin nasıl dengelenebileceği grafik yardımıyla aktarılmıştır. Oversampling ve undersampling; azınlık dağılımı ve sayısı.



SMOTE ilk özellik grafiği



KODLAR

İmport edilen kütüphaneler:

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn import svm
from sklearn.metrics import accuracy_score
from imblearn.over_sampling import RandomOverSampler
from imblearn.under_sampling import TomekLinks
from imblearn.under_sampling import RandomUnderSampler
from collections import Counter
from imblearn.over_sampling import SMOTE
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn.linear_model import LogisticRegression
import time as time
```

Veri seti bağlantısı:

```
parkinsons_data = pd.read_csv('/content/ReplicatedAcousticFeatures-
ParkinsonDatabase.csv')

parkinsons_data.head()
parkinsons_data.shape
parkinsons_data.info()
parkinsons_data.isnull().sum()
parkinsons_data.ID.nunique()
parkinsons_data.describe()
parkinsons_data.duplicated().sum()
X=parkinsons_data
y=parkinsons_data['Status']
X=X.drop(columns=["Status",'ID'])
print(X.shape,y.shape)
parkinsons_data.Status.value_counts()
sns.countplot(x="Status", data=parkinsons_data)
parkinsons_data.Status.value_counts()
print(str((120/(120+120))*100) + '% Parkinson Hastası Birey')
parkinsons_data.groupby('Status').mean()
ran_under=TomekLinks(sampling_strategy='azinlik olmayan')
X_under, y_under= ran_under.fit_resample(X, y)
print(Counter(y_under))
ran_over = RandomOverSampler(sampling_strategy='minority',random_state=
1)
X_over, y_over = ran_over.fit_resample(X, y)
print(Counter(y_over))
ran_under = RandomUnderSampler(sampling_strategy='majority',random_stat
e=1)
X_under, y_under= ran_under.fit_resample(X, y)
print(Counter(y_under)
park_over=X_over.copy()
park_over['Status']=y_over
park_under=X_under.copy()
park_under['Status']=y_under
figure, axis = plt.subplots(2, 2,figsize=(15,10))

sns.scatterplot(ax=axis[0,0],data=park_over,x='Jitter_RAP',y='Jitter_PP
Q',hue='Status')
axis[0, 0].set_title("Oversampling Minority Scatter")

sns.countplot(ax=axis[0, 1],x="Status", data=park_over)
axis[0, 1].set_title("Oversampling Minority Count")

sns.scatterplot(ax=axis[1, 0],data=park_under,x='Jitter_RAP',y='Jitter_
PPQ',hue='Status')
```

```
axis[1, 0].set_title("Undersampling Majority Scatter")

sns.countplot(ax=axis[1, 1],x="Status", data=park_under)
axis[1, 1].set_title("Undersampling Majority Count")
plt.show()
park_over.duplicated().sum()
sm = SMOTE(sampling_strategy='minority', random_state=1)
X_smote, y_smote= sm.fit_resample(X, y)
print(Counter(y_smote))
X_smote.columns
park_smote=X_smote.copy()
park_smote['Status']=y_smote
fig, (ax1, ax2) = plt.subplots(1,2)
fig.suptitle('SMOTE on First Feature with Countplot')
sns.scatterplot(ax=ax1,data=park_smote,x='Jitter_RAP',y='Jitter_PPQ',hu
e='Status').set(title='SMOTE')
sns.countplot(ax=ax2,x="Status", data=park_smote).set(title='SMOTE Coun
t')
cols=X_smote.columns
norm_smote=pd.DataFrame(preprocessing.normalize(X_smote),columns=cols)
norm_over=pd.DataFrame(preprocessing.normalize(X_over),columns=cols)
norm_under=pd.DataFrame(preprocessing.normalize(X_under),columns=cols)
scale_under=pd.DataFrame(preprocessing.scale(X_under),columns=cols)
scale_over=pd.DataFrame(preprocessing.scale(X_over),columns=cols)
scale_smote=pd.DataFrame(preprocessing.scale(X_smote),columns=cols)
fig, axs = plt.subplots(1,2,figsize=(12,8))
axs[0].hist(norm_smote['Jitter_RAP'], bins = 25, color = '#00A0A0')
axs[0].title.set_text('Normallestirme')
axs[1].hist(scale_smote['Jitter_PPQ'], bins = 25, color = '#00A0A0')
axs[1].title.set_text('Olcekleme')
X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size=0.2
, random_state=2)
print(X.shape, X_train.shape, X_test.shape)
scaler = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)

X_test = scaler.transform(X_test)
print(X_train)
model = svm.SVC(kernel='linear')
model.fit(X_train, Y_train)
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(Y_train, X_train_prediction)
print('Egitim verilerinin dogruluk derecesi : ', training_data_accuracy
)
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(Y_test, X_test_prediction)
print('Test verilerinin dogruluk derecesi : ', test_data_accuracy)
input_data = (#BURAYA VERİ SETİ GİRİLİR)
```

```
input_data_as_numpy_array = np.asarray(input_data)

input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

the data
std_data = scaler.transform(input_data_reshaped)

prediction = model.predict(std_data)
print(prediction)


if (prediction[0] == 0):
  print("Kisi Saglıklı")

else:
  print("Kisi Parkinson")

}
```

Açıklamalarıyla birlikte kodların bulunduğu drive linki
https://drive.google.com/drive/folders/1dTIt82prdwWCcQRmY02tmNrrVOaMlwNj?usp=sharing

KAYNAKÇA

Destek vektör makineleri nedir: https://medium.com/deep-learning-turkiye/nedir-bu-destek-vekt%C3%B6r-makineleri-makine-%C3%B6%C4%9Frenmesi-serisi-2-94e576e4223e

Parkinson hastalığı nedir:

1. https://tr.wikipedia.org/wiki/Parkinson_hastal%C4%B1%C4%9F%C4%B1
2. Parkinson Hastalığında Semptomlar, Konuşma Zorlukları veya Değişiklikleri | DİLGEM (dilgem.com.tr)
3. https://www.turkiyeparkinsonhastaligidernegi.com/

Veri seti dengeleme nedir: https://www.veribilimiokulu.com/dengesiz-veri-setlerinde-modelleme/

Veri seti: https://archive.ics.uci.edu/ml/datasets/Parkinson+Dataset+with+replicated+acoustic+features+#

Öznitelik seçimi nedir: https://medium.com/@gulcanogundur/%C3%B6znitelik-se%C3%A7imi-feature-selection-teknikleri-5cd8cbab7706

https://www.kaggle.com/datasets/debasisdotcom/parkinson-disease-detection/code

https://www.youtube.com/watch?v=HbyN_ey-JVc

https://www.youtube.com/watch?v=kZNkaNATmd8&t=195s

https://www.youtube.com/watch?v=sjH9DHj3Ugw&t=184s

https://www.youtube.com/watch?v=CjFj-nuCGX4&t=53s

https://www.youtube.com/watch?v=ls47CPFU1vE&t=62s

https://www.youtube.com/watch?v=W-5nFcEuauY