1- Veri Kümesi Hazırlanması

Projede kullanılan veri kümesi İngiltere Futbol 1. Ligi olan Premier Lig'in 2021-2022 sezonu maç sonuçlarının istatistiksel bilgilerini içermektedir.
Veri kümesinin hazırlanma aşamasında 'web scraping' yöntemleri kullanılmıştır. Python dili kullanılarak gerçekleştirilmiş olup requests ve BeautifulSoup kütüphaneleri veri çekme işlemlerini sağlarken, pandas kütüphanesi çekilen verilerin düzenlenmesinde kullanılmıştır. Veriler 'https://fbref.com/en/comps/9/Premier-League-Stats' sitesinden elde edilmiştir. Proje 2 aşamadan oluşmaktadır. Maç sonucu tahmininde sınıflandırma, atılan gol veya topa sahip olma gibi özelliklerin tahmininde regresyon modelleri kullanılmıştır.

Veri setinde kullanılan özellikler şunlardır; maç sonucu, atılan gol, yenilen gol, gol beklentisi, gol yeme beklentisi, topa sahip olma, atılan şut, kaleyi bulan şut, şutların kaleye ortalama uzaklığı, atılan serbest vuruş, atılan penaltı ve gole çevrilen penaltı.

```
1 print(matches.shape)
2 matches
```

```
(514, 12)
```

	result	gf	ga	xg	xga	poss	sh	sot	dist	fk	pk	pkatt
1	W	2.0	0.0	2.2	0.5	75.0	13.0	1.0	18.7	1.0	1	1
2	W	4.0	0.0	1.7	0.1	67.0	19.0	7.0	17.5	0.0	0	0
3	D	3.0	3.0	2.1	1.8	69.0	21.0	10.0	16.2	1.0	0	0
4	W	4.0	2.0	2.2	0.1	74.0	18.0	5.0	14.1	0.0	0	0
5	W	6.0	0.0	3.3	0.7	74.0	17.0	9.0	14.8	0.0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
48	L	3.0	4.0	1.7	2.1	41.0	8.0	4.0	16.1	0.0	0	0
49	L	0.0	3.0	0.4	2.3	31.0	6.0	2.0	24.2	0.0	0	0
50	W	1.0	0.0	1.3	1.2	35.0	15.0	4.0	15.5	0.0	0	0
52	L	0.0	2.0	0.2	3.4	47.0	4.0	4.0	19.0	0.0	0	0
54	W	3.0	1.0	2.1	0.9	59.0	19.0	9.0	14.6	0.0	0	0

514 rows × 12 columns

Veri seti 2 aşamada meydana gelmiştir. Maç sonuçlarının bulunduğu sayfa detaylı bilgiler içermediğinden, şut verilerinin

bulunduğu sayfadan veriler çekilip ‘Date’ kolonu üzerinden birleştirilmiştir. Veri seti 514 kayıt, 12 özellikten oluşur.

```
team_data = matches.merge(shooting[["Date", "Sh", "SoT", "Dist", "FK", "PK", "PKatt"]], on="Date")
# matches tablosu 19 sütun, shooting tablosundan 7 sütun ekleniyor (1'i ortak Date o yüzden 6 sütun)
# yeni tablo 19 + 6 = 25 sütun olacaktır
```

2- Bulgular

Sınıflandırma modelleri olarak Logistic Regression, Decision Tree Classifier, Random Forest Classifier, SVC ve Gaussian NB seçilmiştir. Regresyon modelleri olarak Bayesian Ridge, Lasso, Elastic Net, Decision Tree Regressor ve Linear Regression seçilmiştir. Belirtilen modeller ‘sklearn’ kütüphanesinin sunduğu imkanlar ile kullanılmıştır.

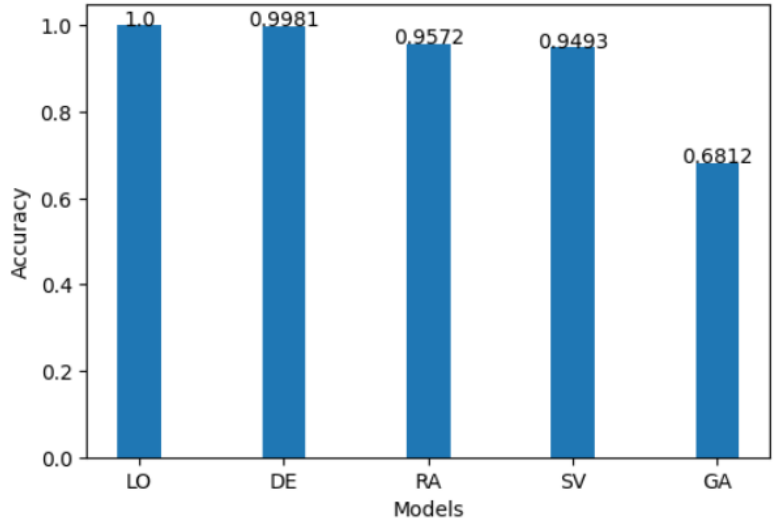
Sınıflandırma işlemlerinde metrik olarak ‘mean accuracy’ kullanılmıştır.

1. Sınıflandırma Örneği: Maç sonucu kolonu sınıflandırılması.

```
X = matches.drop(columns=['result'], axis=1)
Y = matches['result']
```

Bu örnekte atılan ve yenilen gol özellikleri çıkarılmadığından 1 ve 1’e yakın çok yüksek başarı oranları yakalanmıştır. Bu çıkarım göstermek istenilmiştir.

```
LO: Mean Accuracy: 1.0, Standard Deviation: 0.0
DE: Mean Accuracy: 0.998076923076923, Standard Deviation: 0.0057692307692307826
RA: Mean Accuracy: 0.9571644042232277, Standard Deviation: 0.021108191781902316
SV: Mean Accuracy: 0.9493212669683257, Standard Deviation: 0.039465859982189946
GA: Mean Accuracy: 0.6812217194570136, Standard Deviation: 0.07512239701168336
```

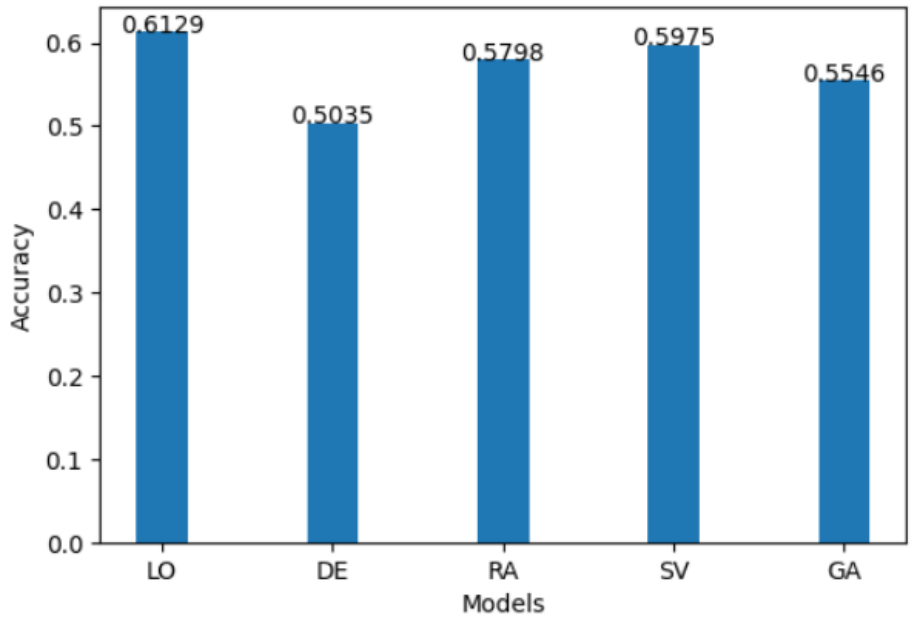


2. Sınıflandırma Örneği: Maç sonucu kolonunun bazı özellikler çıkarılarak sınıflandırılması.

```
X = matches.drop(columns=['result', 'ga', 'gf'], axis=1)
Y = matches['result']
```

Bu örnekte atılan ve yenilen gol özellikleri çıkarılınca başarı oranının ne kadar düştüğü gözlemlenmiştir.

```
LO: Mean Accuracy: 0.6129336349924583, Standard Deviation: 0.04219632236416618
DE: Mean Accuracy: 0.5035067873303167, Standard Deviation: 0.07165635561407993
RA: Mean Accuracy: 0.5798265460030165, Standard Deviation: 0.03626386629174735
SV: Mean Accuracy: 0.5974736048265459, Standard Deviation: 0.04737112418994452
GA: Mean Accuracy: 0.5546380090497737, Standard Deviation: 0.05513316979706163
```



Ayrıca t-test ve p değerleri aşağıda şekilde incelenmiştir.

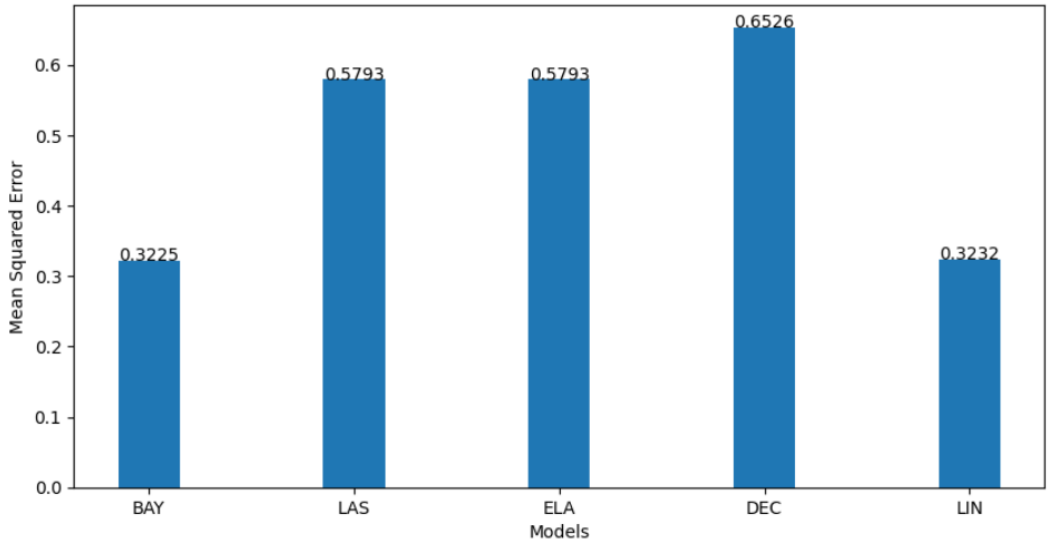
```
LO vs DE: t-statistic = 3.8586104342713092, p-value = 0.003855242968730131
LO vs RA: t-statistic = 2.9340363690556277, p-value = 0.016647373572585857
LO vs SV: t-statistic = 1.1393969637775179, p-value = 0.28395405107285865
LO vs GA: t-statistic = 3.208736716466786, p-value = 0.010680543015916933
DE vs RA: t-statistic = -3.0363021270932156, p-value = 0.014102062532796143
DE vs SV: t-statistic = -3.393390983628674, p-value = 0.007956219857677067
DE vs GA: t-statistic = -1.7784986966830116, p-value = 0.1090343679432196
RA vs SV: t-statistic = -1.1850586881647849, p-value = 0.26634150085059444
RA vs GA: t-statistic = 1.2369531435020304, p-value = 0.24740711176070496
SV vs GA: t-statistic = 2.150868887925689, p-value = 0.05995182890365449
```

t-statistic değeri ortalamaları karşılaştırmak için, p değeri ise gözlenen farkın istatistiksel olarak anlamlı olup olmadığını belirlemek için kullanılır.

Regresyon işlemleri sonuç özelliği çıkarılarak yapılmıştır. Metrik olarak ‘mean squared error’ ve ‘root mean squared error’ kullanılmıştır.

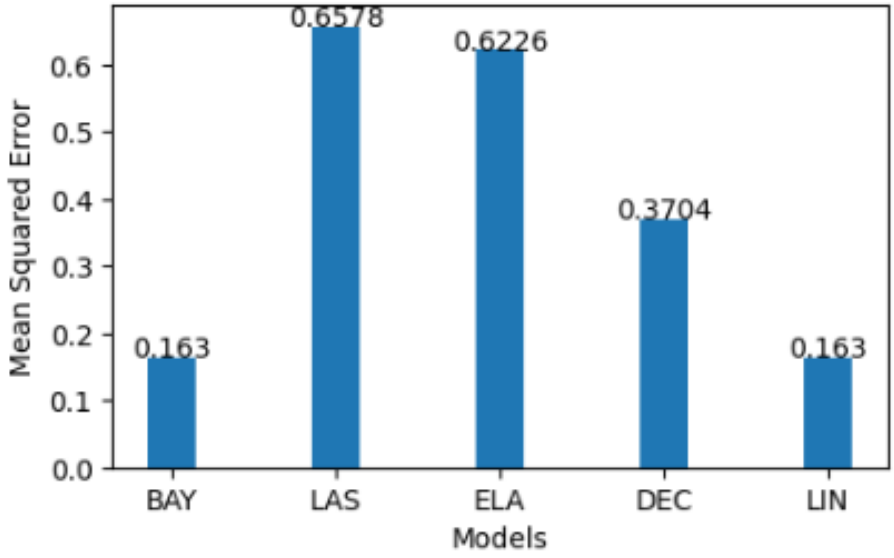
1. Regresyon Örneği: Gol yeme beklentisi tahmini.

```
BAY: Mean Squared Error: 0.3225296467806019,  Standard Deviation: 0.07673331166333051
LAS: Mean Squared Error: 0.5793124427104359,  Standard Deviation: 0.1333125542211606
ELA: Mean Squared Error: 0.5793124427104359,  Standard Deviation: 0.1333125542211606
DEC: Mean Squared Error: 0.6526398944193061,  Standard Deviation: 0.0721609380952775
LIN: Mean Squared Error: 0.32320160843700985,  Standard Deviation: 0.07735565713407526
```



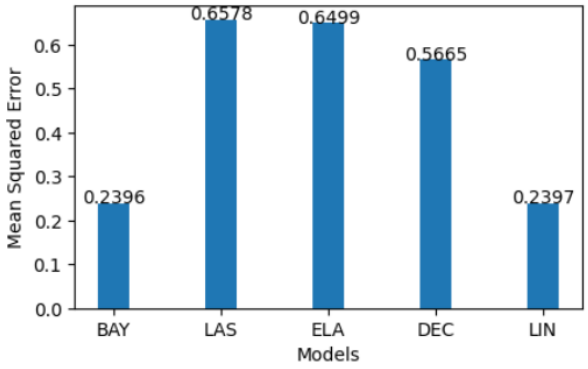
2. Regresyon Örneği: Gol atma beklentisi tahmini.

```
BAY: Mean Squared Error: 0.163,  Standard Deviation: 0.0521
LAS: Mean Squared Error: 0.6578,  Standard Deviation: 0.1995
ELA: Mean Squared Error: 0.6226,  Standard Deviation: 0.1988
DEC: Mean Squared Error: 0.3704,  Standard Deviation: 0.1049
LIN: Mean Squared Error: 0.163,  Standard Deviation: 0.052
```



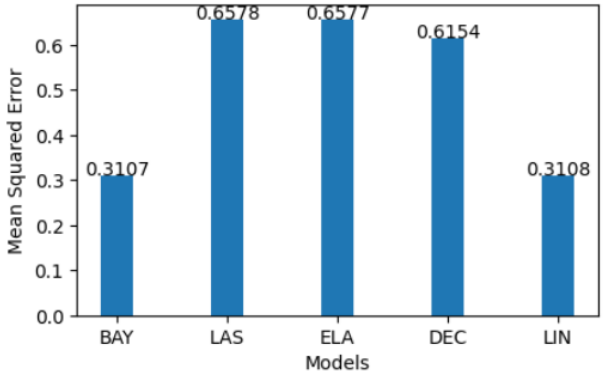
3. Regresyon Örneği: Gol atma beklentisinin atılan şut kayıtlarının çıkarılarak tahmini.

```
BAY: Mean Squared Error: 0.2396,  Standard Deviation: 0.0721
LAS: Mean Squared Error: 0.6578,  Standard Deviation: 0.1995
ELA: Mean Squared Error: 0.6499,  Standard Deviation: 0.2019
DEC: Mean Squared Error: 0.5665,  Standard Deviation: 0.1462
LIN: Mean Squared Error: 0.2397,  Standard Deviation: 0.0716
```



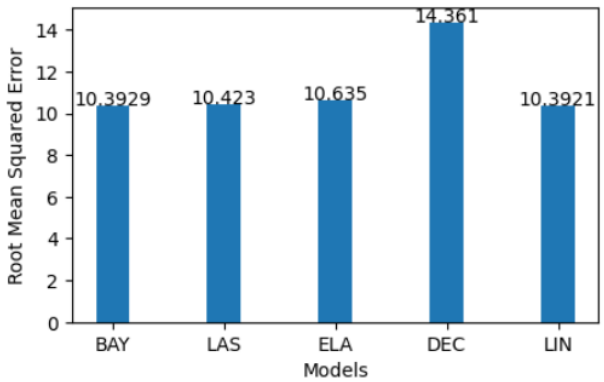
4. Regresyon Örneği: Gol atma beklentisinin atılan şut ve isabetli şut kayıtlarının çıkarılarak tahmini.

```
BAY: Mean Squared Error: 0.3107,  Standard Deviation: 0.089
LAS: Mean Squared Error: 0.6578,  Standard Deviation: 0.1995
ELA: Mean Squared Error: 0.6577,  Standard Deviation: 0.1997
DEC: Mean Squared Error: 0.6154,  Standard Deviation: 0.1467
LIN: Mean Squared Error: 0.3108,  Standard Deviation: 0.0887
```



5. Regresyon Örneği: Topa sahip olma oranının tahmini. (Root Mean Squared Error metriği ile ölçülmüştür.)

```
BAY: Mean Squared Error: 10.3929,  Standard Deviation: 21.4614
LAS: Mean Squared Error: 10.423,  Standard Deviation: 21.2559
ELA: Mean Squared Error: 10.635,  Standard Deviation: 22.6099
DEC: Mean Squared Error: 14.361,  Standard Deviation: 38.0758
LIN: Mean Squared Error: 10.3921,  Standard Deviation: 20.5864
```



3- Yorumlar

Yukarıda gerçekleştirilen örnekler sonucu aşağıdaki çıkarımlar yapılmıştır:

- Sınıflandırma ve regresyon işlemlerinde kullanılacak özellikler oldukça önemlidir. Korelasyon değeri 1 veya -1 özellikler çıkartılmalıdır. Aynı şekilde korelasyon değerinin çok düşük olduğu özellikler sınıflandırma ve tahmin için veri setinde bulunmamalıdır.
- 2. Sınıflandırma örneğinde p değerleri gözlemlendiğinde sayısal olarak yüksek değerler bulunması bu iki modelin kullanılmasının anlamlı olmadığını gösterir. İncelenen örnekte LO ve SV, RA ve SV, RA ve GA modellerinin kullanılması çeşitliliği kısıtlar.
- 1. Regresyon örneğinde modeller genel olarak farklı başarı oranlarına sahiptir, iyi bir örnek olarak sunulabilir.
- 2, 3 ve 4. Regresyon örneklerinde özellik çıkarımının etkisi net bir şekilde gözlemlenir. Gol atma beklentisi, atılan şut ve isabetli şutlarla pozitif şekilde ilişkili olduğundan bu özelliklerin çıkarımı hata payını oldukça arttırır.
- 5. Regresyon örneğinde topa sahiplik oranı incelenmiştir, hata payları oldukça yüksek çıkabildiğinden metrik olarak ‘root mean squared error’ model farklarını başarılı bir şekilde açıklar.

Mustafa Kemal Ekim

18011072