MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE GENETİK VARYANTLARIN PATOJENİTE ANALİZİ BRCA1,BRCA2 GENLERİ UYGULAMASI

ONUR KAPLAN – 160202061

MUHAMMED EMRE KARA - 160202094

GİRİŞ

- Bu çalışma, BRCA1 ve BRCA2 genlerine ait varyantların patojenite analizinin yapılması üzerine eğilmiştir.
- Tüm dünyada kadınlarda en sık görülen kanser türü olan meme kanserinin önlenmesinde etkin rol oynayan bu iki gen üzerinde oluşabilen varyantlar uygulama alanı olarak seçilmiştir.
- Bu çalışmada bu sorunu çözmek üzere makine öğrenmesi yöntemlerinden yararlanılması önerilmektedir.

GENEL BİLGİLER

Aşağıda yer verilen maddeler çalışmanın kapsam ve içeriğinin anlaşılabilmesi için bu kısımda ayrıntılı olarak açıklanmıştır.

- Meme Kanseri
- DNA Nedir?
- Gen Nedir?
- BRCA Nedir?
- Gen Dizileme
- İnsan Genom Projesi
- Varyant Nedir?
- BRCA Varyantları
- Patojenite Nedir?
- CADD
- Myvariant.info
- Makine Öğrenmesi Nedir?
- Sınıflandırma nedir?
- Gradient Boosting
- XGBOOST
- LightGBM
- CatBoost
- Lojistik Regresyon
- KNN
- SVM
- Karar Ağacı Sınıflandırma
- Naive Bayes
- Random Forest

MALZEME VE YÖNTEM

- **Veri Seçimi**

Veri Tabanı	BRCA1+BRCA2 Varyantları
dbSNP	55834
CADD	50571
dbNSFP	38781

Veri Tabanı	Özellik Sayısı
dbSNP	58
CADD	135

MALZEME VE YÖNTEM

- **Veri Hazırlama**

cadd.alt	cadd.anc	cadd.annc	cadd.bsta	cadd.chro	cadd.cons	cadd.cons
T	C	Transcript	111	17	intron	INTRONIC
A	G	Transcript	111	17	intron	INTRONIC
A	G	CodingTra	111	17	missense	NON_SYN
A	T	CodingTra	111	17	missense	NON_SYN
G	T	CodingTra	115	17	missense	NON_SYN
A	G	Transcript	112	17	intron	INTRONIC
G	G	Transcript	115	17	intron	INTRONIC
A	T	CodingTra	116	17	synonymc	SYNONYM
C	T	CodingTra	116	17	missense	NON_SYN
T	C	CodingTra	115	17	missense	NON_SYN
G	A	CodingTra	115	17	missense	NON_SYN
G	A	CodingTra	374	13	missense	NON_SYN
C	A	CodingTra	398	13	missense	NON_SYN

MALZEME VE YÖNTEM

- **Veri Önişleme**
- ID içeren sütunların silinmesi(5)
- Doluluk oranını sağlamayan sütunların silinmesi(41)

	Sütun Adı	Boş Değer Sayısı	Boş Değer Yüzdesi
1	motif.ecount	12329	99.9757
2	motif.ehipos	12329	99.9757
3	motif.ename	12329	99.9757
4	motif.escorechng	12329	99.9757
5	motif.dist	12240	99.254
6	motif.toverlap	12240	99.254
7	mirsvr.aln	12214	99.0431
8	mirsvr.e	12214	99.0431
9	mirsvr.score	12214	99.0431
10	cadd.scoresegdup	12186	98.8161
11	1000g.asn	11870	96.2536
12	1000g.eur	11795	95.6455
13	esp.af	11758	95.3454
14	esp.afr	11758	95.3454
15	esp.eur	11758	95.3454
16	tf.bs	11700	94.8751
17	tf.bs_peaks	11700	94.8751
18	tf.bs_peaks_max	11700	94.8751
19	1000g.amr	11657	94.5264
20	1000g.afr	11516	93.3831
21	cadd.intron	11514	93.3669

	Sütun Adı	Boş Değer Sayısı	Boş Değer Yüzdesi
22	1000g.af	11232	91.0801
23	cadd.dst2splice	10706	86.8148
24	cadd.dst2spltype	10706	86.8148
25	encode.occ	10571	85.7201
26	p_val.comb	10571	85.7201
27	p_val.ctcf	10571	85.7201
28	p_val.dnas	10571	85.7201
29	p_val.faire	10571	85.7201
30	p_val.mycp	10571	85.7201
31	p_val.polii	10571	85.7201
32	sig.ctcf	10571	85.7201
33	sig.dnase	10571	85.7201
34	sig.faire	10571	85.7201
35	sig.myc	10571	85.7201
36	sig.polii	10571	85.7201
37	sift.cat	4854	39.361
38	sift.val	4854	39.361
39	cadd.grantham	4853	39.3529
40	polyphen.cat	4853	39.3529
41	polyphen.val	4853	39.3529

MALZEME VE YÖNTEM

- **Veri Hazırlama – Boş Verilerin Doldurulması**
- 15,742/912,568 = 1.73%
- Numerik veriler için : Ortalama(mean)
- Metinsel veriler için : En sık geçen(most_frequent)
- **Veri Hazırlama – Kategorik-Numerik Dönüşümü**
- variants_encoded.csv
- variants_not_encoded.csv

MALZEME VE YÖNTEM

- **Basit Sınıflandırma Metotları**
- Logistic Regression

Ortalama Accuracy :
91.88912876993452
Ortalama Classification Report :

	precision	recall	f1-score	support
Benign	0.87	0.86	0.87	1005
Likely benign	0.94	0.95	0.95	2876
Likely pathogenic	0.12	0.03	0.05	90
Pathogenic	0.91	0.96	0.94	1195
accuracy			0.92	5166
macro avg	0.71	0.70	0.70	5166
weighted avg	0.91	0.92	0.91	5166

MALZEME VE YÖNTEM

- **Basit Sınıflandırma Metotları**
- KNN

Ortalama Accuracy :
71.58358314874144
Ortalama Classification Report :

	precision	recall	f1-score	support
Benign	0.83	0.72	0.77	1005
Likely benign	0.70	0.90	0.79	2876
Likely pathogenic	0.23	0.07	0.10	90
Pathogenic	0.65	0.33	0.44	1195
accuracy			0.72	5166
macro avg	0.60	0.50	0.52	5166
weighted avg	0.71	0.72	0.69	5166

MALZEME VE YÖNTEM

- **Basit Sınıflandırma Metotları**
- SVM

Ortalama Accuracy :
91.52147413872198
Ortalama Classification Report :

	precision	recall	f1-score	support
Benign	0.90	0.80	0.85	1005
Likely benign	0.92	0.96	0.94	2876
Likely pathogenic	0.12	0.02	0.04	90
Pathogenic	0.92	0.96	0.94	1195
accuracy			0.92	5166
macro avg	0.72	0.69	0.69	5166
weighted avg	0.90	0.92	0.91	5166

MALZEME VE YÖNTEM

- **Basit Sınıflandırma Metotları**
- Gaussian Naive Bayes

Ortalama Accuracy :
73.65836486843263
Ortalama Classification Report :

	precision	recall	f1-score	support
Benign	0.78	0.81	0.79	1005
Likely benign	0.95	0.69	0.80	2876
Likely pathogenic	0.06	0.63	0.10	90
Pathogenic	0.95	0.79	0.86	1195
accuracy			0.74	5166
macro avg	0.68	0.73	0.64	5166
weighted avg	0.90	0.74	0.80	5166

MALZEME VE YÖNTEM

- **Basit Sınıflandırma Metotları**
- Decision Tree Classifier

```
Ortalama Accuracy :
90.10830223513793
Ortalama Classification Report :
                    precision    recall  f1-score   support

           Benign       0.83      0.84      0.84      1005
     Likely benign      0.94      0.93      0.93      2876
Likely pathogenic       0.19      0.19      0.19        90
        Pathogenic      0.93      0.93      0.93      1195

          accuracy                           0.90      5166
         macro avg      0.72      0.72      0.72      5166
      weighted avg      0.90      0.90      0.90      5166
```

MALZEME VE YÖNTEM

- **Basit Sınıflandırma Metotları**
- Random Forest Classifier

```
Ortalama Accuracy :
91.83115786398932
Ortalama Classification Report :
                    precision    recall  f1-score   support

            Benign       0.89      0.82      0.86      1005
       Likely benign       0.93      0.96      0.94      2876
Likely pathogenic       0.34      0.12      0.18        90
        Pathogenic       0.92      0.96      0.94      1195

          accuracy                           0.92      5166
         macro avg       0.77      0.72      0.73      5166
      weighted avg       0.91      0.92      0.91      5166
```

MALZEME VE YÖNTEM

- **Boosted Tree Sınıflandırma Metotları**
- XGBoost

Ortalama Accuracy :
91.56023375606907
Ortalama Classification Report :

	precision	recall	f1-score	support
Benign	0.84	0.87	0.85	968
Likely benign	0.95	0.94	0.94	2919
Likely pathogenic	0.12	0.25	0.16	44
Pathogenic	0.96	0.93	0.94	1235
accuracy			0.92	5166
macro avg	0.72	0.75	0.73	5166
weighted avg	0.92	0.92	0.92	5166

MALZEME VE YÖNTEM

- **Boosted Tree Sınıflandırma Metotları**
- LightGBM

```
Ortalama Accuracy :
92.39251695967315
Ortalama Classification Report :
                    precision    recall  f1-score   support

            Benign       0.89      0.85      0.87      1005
     Likely beging       0.94      0.96      0.95      2876
Likely pathogenic        0.33      0.17      0.22        90
        Pathegenic       0.93      0.97      0.95      1195

          accuracy                           0.92      5166
         macro avg       0.77      0.73      0.75      5166
      weighted avg       0.92      0.92      0.92      5166
```

MALZEME VE YÖNTEM

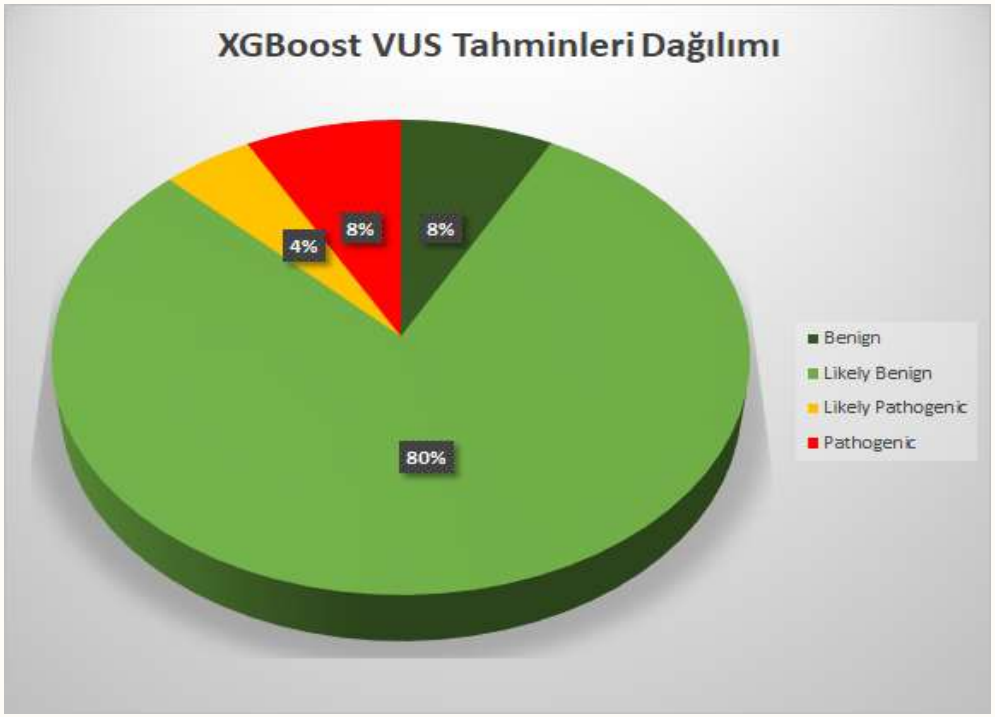
- **Boosted Tree Sınıflandırma Metotları**
- CatBoost

```
average accuracy: 0.9202467508393235
average classification report:
                    precision    recall  f1-score   support

            Benign       0.88      0.85      0.86      1005
     Likely beging       0.94      0.95      0.95      2876
Likely pathogenic        0.25      0.08      0.12        90
        Pathegenic       0.92      0.97      0.95      1195

          accuracy                           0.92      5166
         macro avg       0.75      0.71      0.72      5166
      weighted avg       0.91      0.92      0.92      5166
```

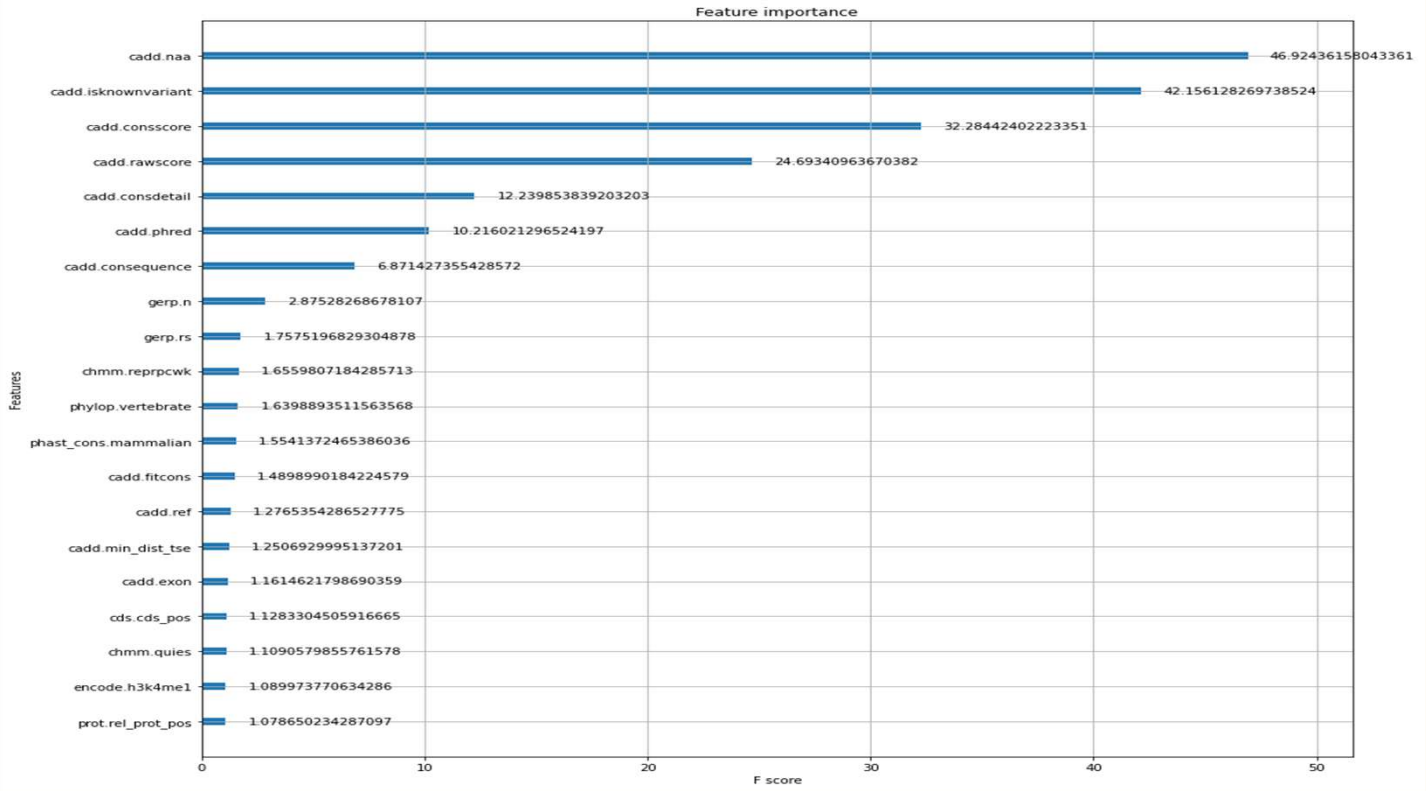
ÇIKTILAR - XGBOOST

Ortalama Accuracy :
91.77307461132717
Ortalama Classification Report :

	precision	recall	f1-score	support
Benign	0.84	0.88	0.86	962
Likely benign	0.95	0.94	0.94	2921
Likely pathogenic	0.18	0.31	0.23	51
Pathogenic	0.96	0.93	0.94	1232
accuracy			0.92	5166
macro avg	0.73	0.76	0.74	5166
weighted avg	0.92	0.92	0.92	5166



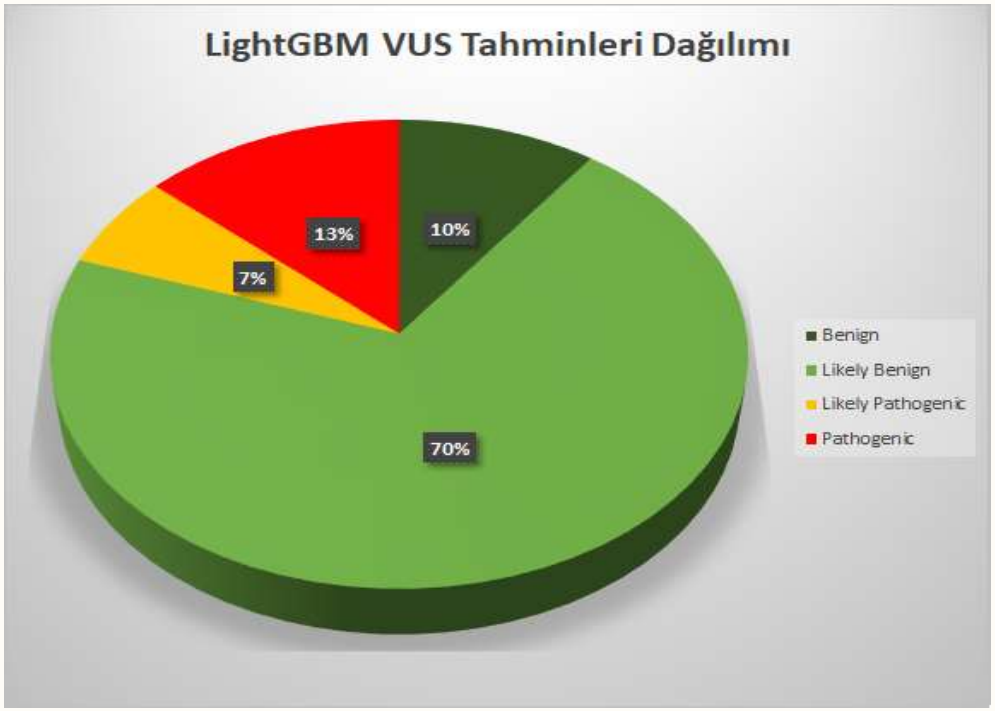
ÇIKTILAR - XGBOOST



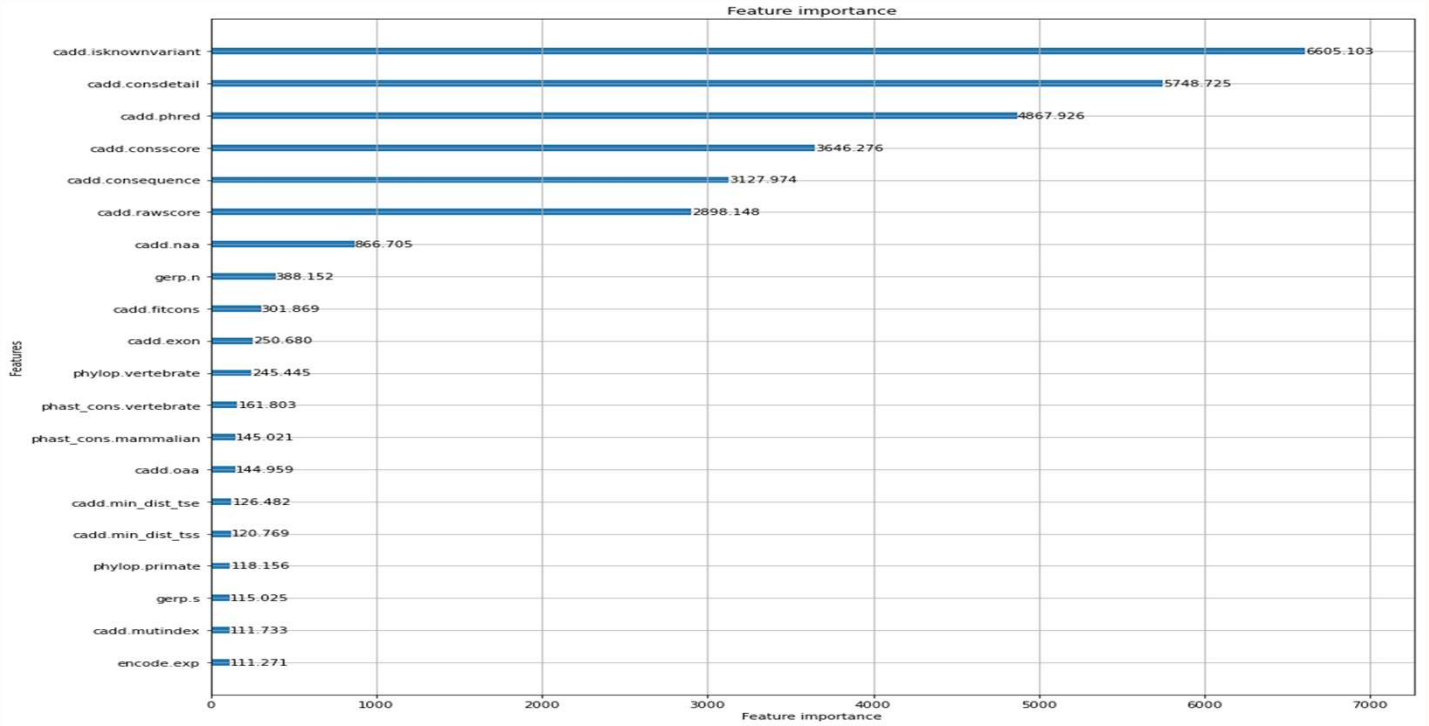
ÇIKTILAR - LİGHTGBM

Ortalama Accuracy :
92.33432136029406
Ortalama Classification Report :

	precision	recall	f1-score	support
Benign	0.88	0.84	0.86	1005
Likely beging	0.94	0.96	0.95	2876
Likely pathogenic	0.35	0.18	0.24	90
Pathegenic	0.94	0.97	0.95	1195
accuracy			0.92	5166
macro avg	0.78	0.74	0.75	5166
weighted avg	0.92	0.92	0.92	5166



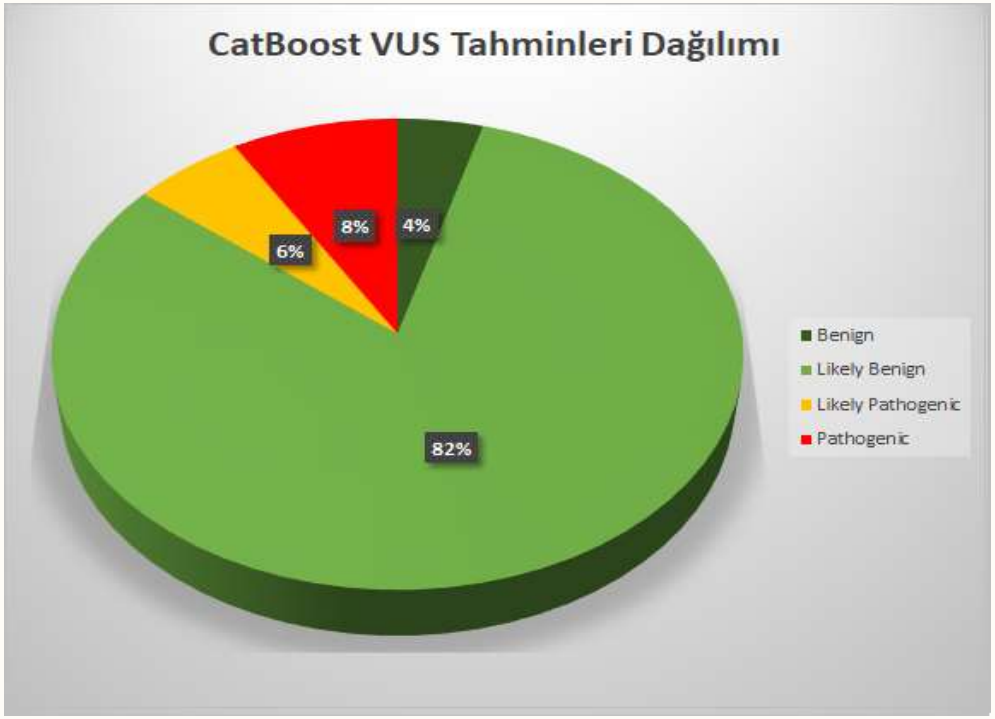
ÇIKTILAR - LİGHTGBM



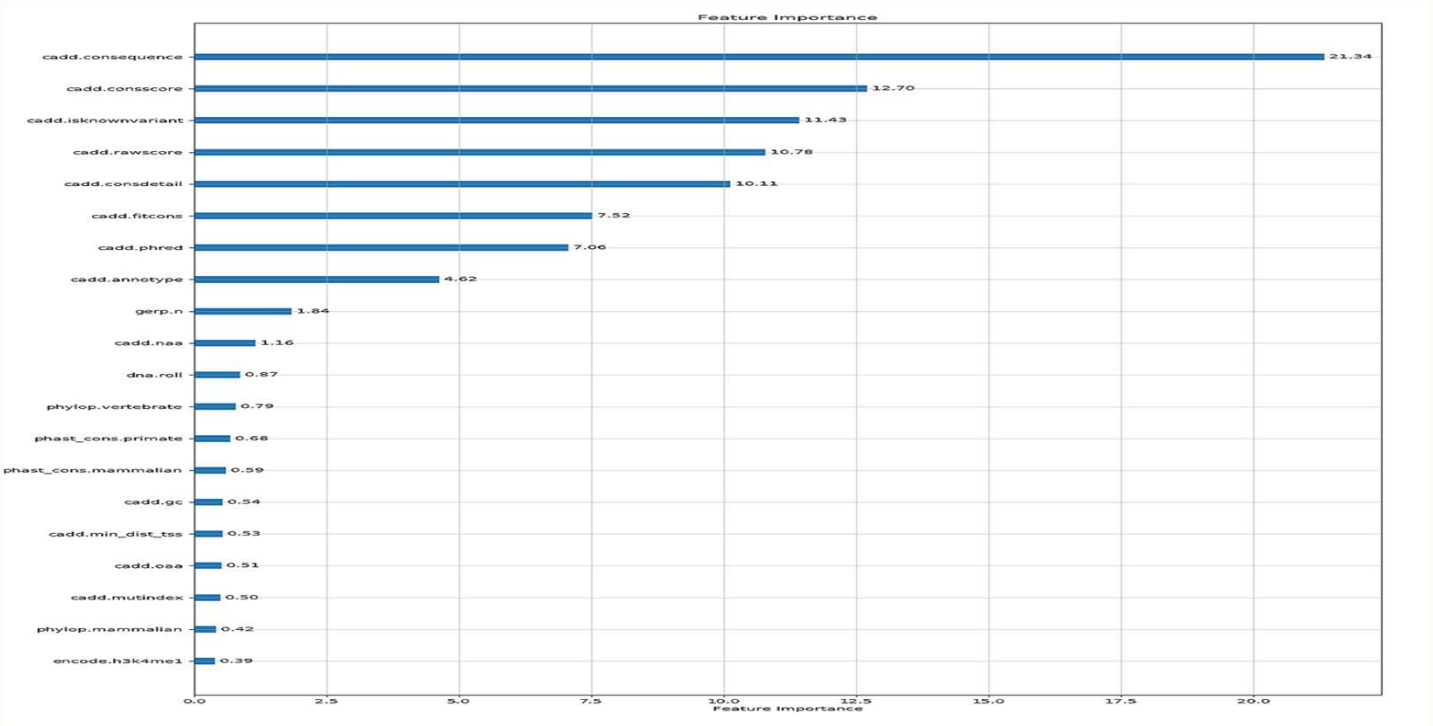
ÇIKTILAR - CATBOOST

average accuracy: 0.9225693319676965
average classification report:

	precision	recall	f1-score	support
Benign	0.89	0.85	0.87	1005
Likely beging	0.94	0.96	0.95	2876
Likely pathogenic	0.29	0.09	0.14	90
Pathegenic	0.92	0.97	0.95	1195
accuracy			0.92	5166
macro avg	0.76	0.72	0.72	5166
weighted avg	0.91	0.92	0.92	5166



ÇIKTILAR - CATBOOST



ÇIKTILAR - RANDOM FOREST

Ortalama Accuracy :
91.40536380675616
Ortalama Classification Report :

	precision	recall	f1-score	support
Benign	0.85	0.84	0.85	1005
Likely benign	0.94	0.94	0.94	2876
Likely pathogenic	0.38	0.23	0.29	90
Pathogenic	0.94	0.95	0.95	1195
accuracy			0.91	5166
macro avg	0.77	0.74	0.75	5166
weighted avg	0.91	0.91	0.91	5166

ÇIKTILAR - LOGİSTİC REGRESSİON

Ortalama Accuracy :
91.9859341910381
Ortalama Classification Report :

	precision	recall	f1-score	support
Benign	0.88	0.86	0.87	1005
Likely benign	0.94	0.95	0.95	2876
Likely pathogenic	0.30	0.12	0.17	90
Pathogenic	0.92	0.96	0.94	1195
accuracy			0.92	5166
macro avg	0.76	0.72	0.73	5166
weighted avg	0.91	0.92	0.92	5166

SONUÇLAR VE ÖNERİLER

- Bu çalışmada BRCA1 ve BRCA2'e ait genetik varyantların patojenitesini tahmin etmek amacıyla Makine Öğrenmesi ile Sınıflandırma yöntemlerini kullandık. Verileri CADD veri tabanından elde ettik. 3 tanesi Boosted Tree algoritmaları olan 9 farklı algoritma üzerinden elde ettiğimiz sonuçlardan Accuracy metriği üzerinde 93%'e kadar başarılı olduk, yine sınıflara ait F-1 metriği üzerinde Likely Pathogenic sınıfı için 34%'e kadar, kalan 3 sınıf için 95%'e kadar başarılı olduk. Bahsedilen 3 Boosted Tree algoritması için Feature Importance grafikleri çizdirerek anlamlı parametrelere vurgu yaptık. Verimizde patojenitesi bilinmeyen 6761 adet VUS olarak adlandırılan varyantlar için modellerimizle tahminlerde bulunduk.

TEŞEKKÜRLER

- ONUR KAPLAN – 160202061
- MUHAMMED EMRE KARA - 160202094