name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ
phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007	0.00370	0.00554
phon_R01_S01_2	122.400	148.650	113.819	0.00968	0.00008	0.00465	0.00696
phon_R01_S01_3	116.682	131.111	111.555	0.01050	0.00009	0.00544	0.00781
phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698
phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908

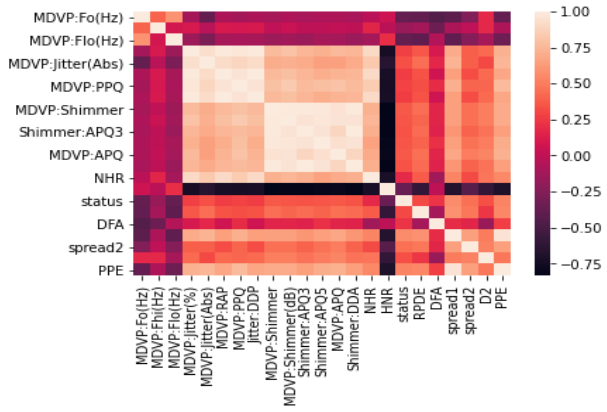
```
count    195.000000
mean       0.753846
std        0.431878
min        0.000000
25%        1.000000
50%        1.000000
75%        1.000000
max        1.000000
Name: status, dtype: float64
```

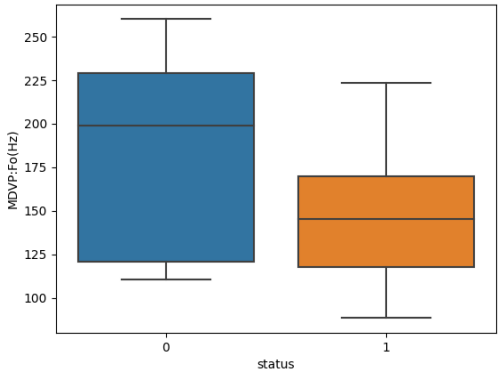
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 24 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   name              195 non-null    object
 1   MDVP:Fo(Hz)       195 non-null    float64
 2   MDVP:Fhi(Hz)      195 non-null    float64
 3   MDVP:Flo(Hz)      195 non-null    float64
 4   MDVP:Jitter(%)    195 non-null    float64
 5   MDVP:Jitter(Abs)  195 non-null    float64
 6   MDVP:RAP          195 non-null    float64
 7   MDVP:PPQ          195 non-null    float64
 8   Jitter:DDP        195 non-null    float64
 9   MDVP:Shimmer      195 non-null    float64
 10  MDVP:Shimmer(dB)  195 non-null    float64
 11  Shimmer:APQ3      195 non-null    float64
 12  Shimmer:APQ5      195 non-null    float64
 13  MDVP:APQ          195 non-null    float64
 14  Shimmer:DDA       195 non-null    float64
 15  NHR               195 non-null    float64
 16  HNR               195 non-null    float64
 17  status            195 non-null    int64
 18  RPDE              195 non-null    float64
 19  DFA               195 non-null    float64
 20  spread1           195 non-null    float64
 21  spread2           195 non-null    float64
 22  D2                195 non-null    float64
 23  PPE               195 non-null    float64
dtypes: float64(22), int64(1), object(1)
memory usage: 36.7+ KB
```

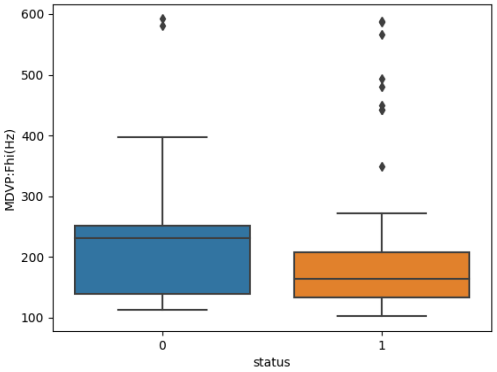
Bu noktada verilerimize genel olarak baktığımızda oldukça sağlıklı bir veri seti elde etmiş olduğumuzu görüyoruz. Herhangi bir outlier durumu görünmemekle birlikte genel olarak hastaların yoğun olduğu bir veri seti karşımızda bulunmakta. Bu aslında algoritmayı sağlıklı ve hastalıklı olarak ayırmada zorlayacak bir durum. Eşit dağılım olsaydı, algoritmalarımız daha kesin çalışırdı fakat yine de çok iyi sonuçlar aldığımızı belirtmem gerek. 195 adet veriden fena olmayan sonuçlar aldık fakat verilerimizin sayısı daha çok olsaydı tahminimce daha iyi eğitilmiş bir algoritmalarımız olabilirdi.

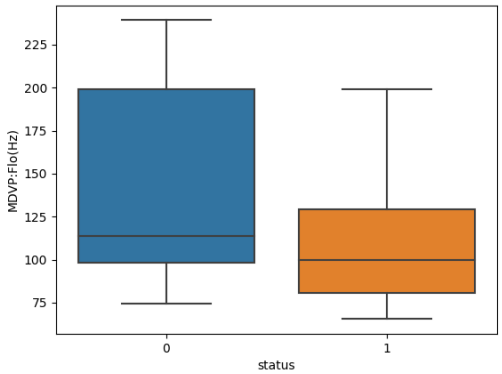
Verilerin görsel ilişkilerine bir göz gezdirelim

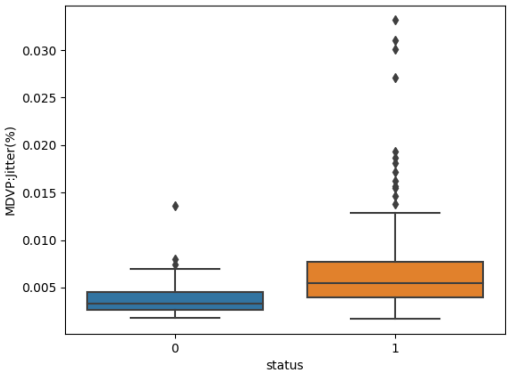
```
sns.heatmap(df.corr());
# data üzerinde korrelasyon spread1,spread2 ve ppe featureslerinde yoğunlukta headmap üzerinden görebiliriz.
```

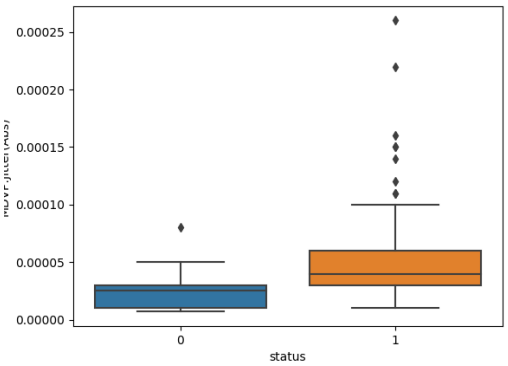


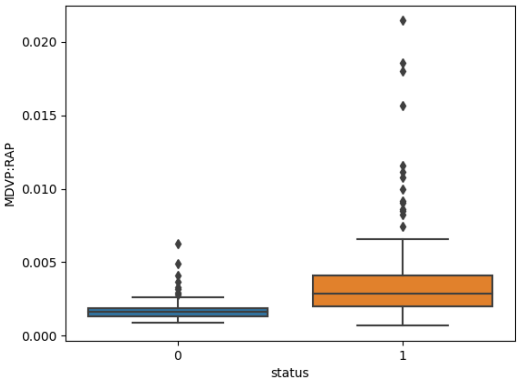

MDVP:Fo(Hz)
250
225
200
175
150
125
100
0
1
status

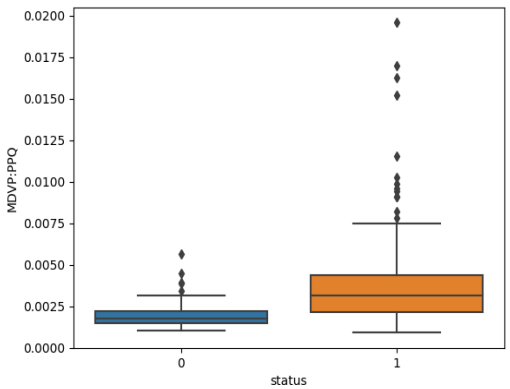

MDVP:Fhi(Hz)
600
500
400
300
200
100
0
1
status

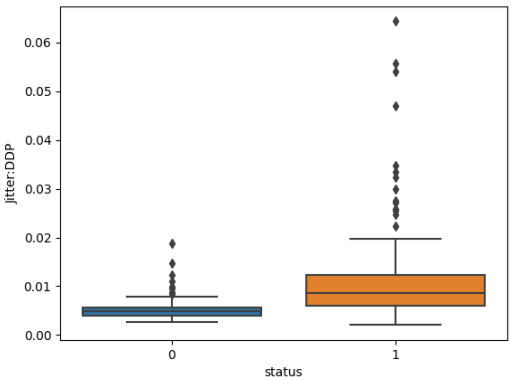

MDVP:Flo(Hz)
225
200
175
150
125
100
75
0
1
status

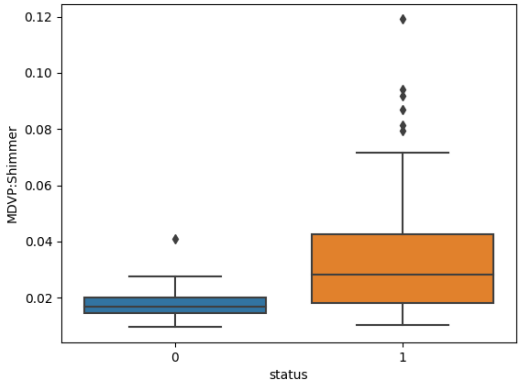

MDVP:Jitter(%)
0.030
0.025
0.020
0.015
0.010
0.005
0
1
status

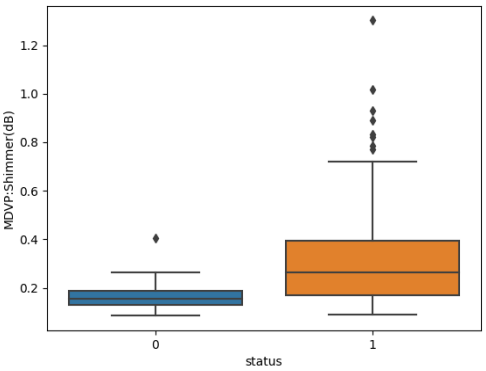

0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
1
status

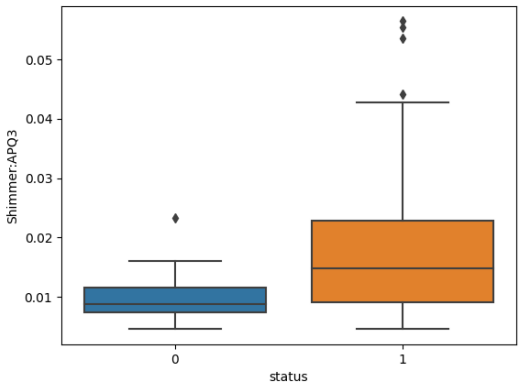

MDVP:RAP
0.020
0.015
0.010
0.005
0.000
0
1
status

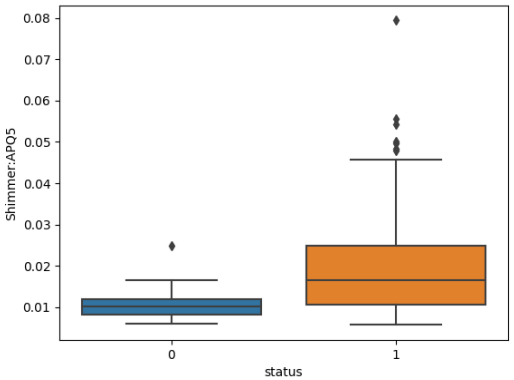

MDVP:PPQ
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
1
status

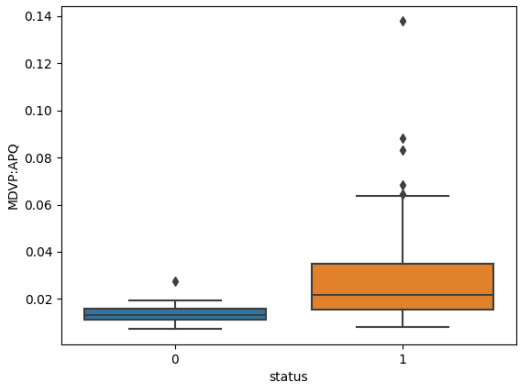

Jitter:DDP
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
1
status

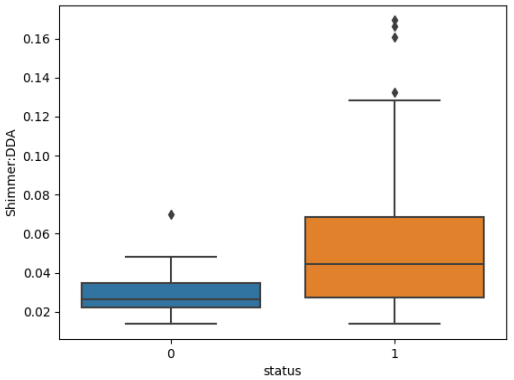

MDVP:Shimmer
0.12
0.10
0.08
0.06
0.04
0.02
0
1
status

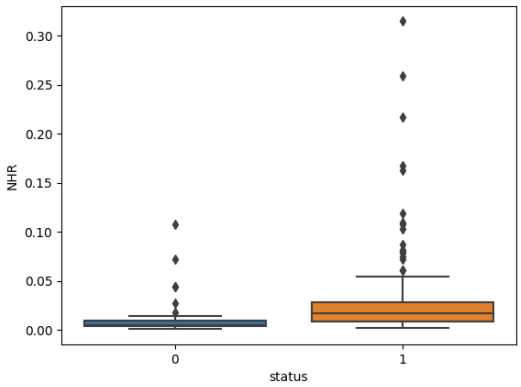

MDVP:Shimmer(dB)
1.2
1.0
0.8
0.6
0.4
0.2
0
1
status

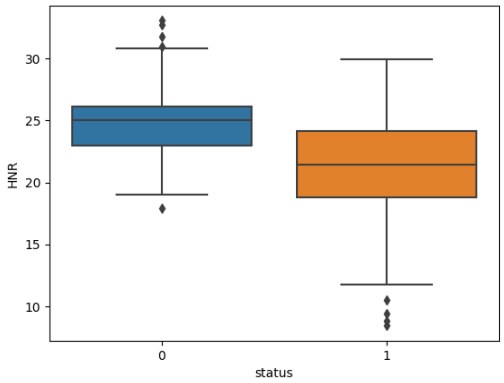

Shimmer:APQ3
0.05
0.04
0.03
0.02
0.01
0
1
status

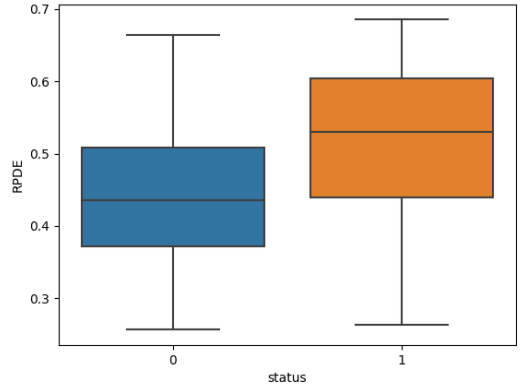

Shimmer:APQ5
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1
status

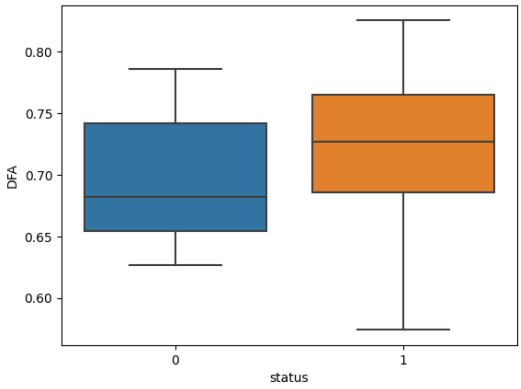

MDVP:APQ
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0
1
status

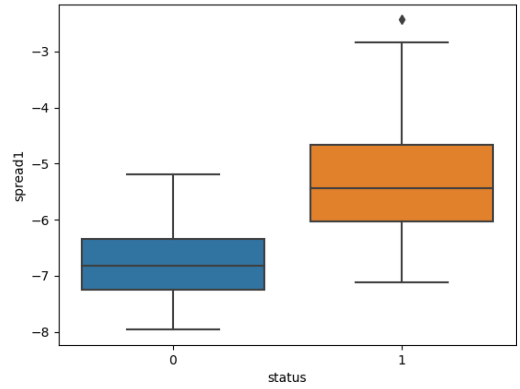

Shimmer:DDA
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0
1
status

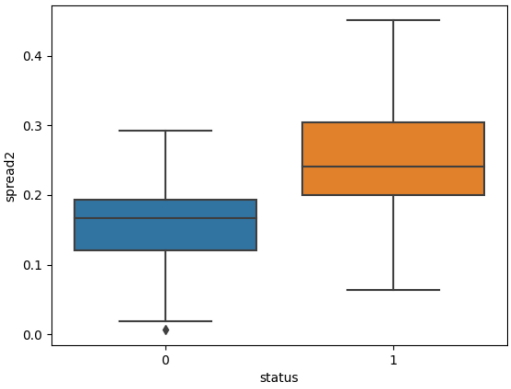

NHR
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
status

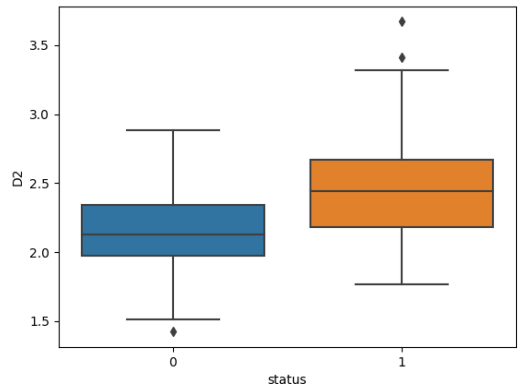

HNR
30
25
20
15
10
0
1
status

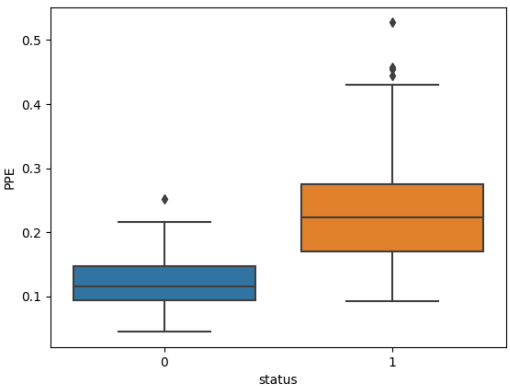




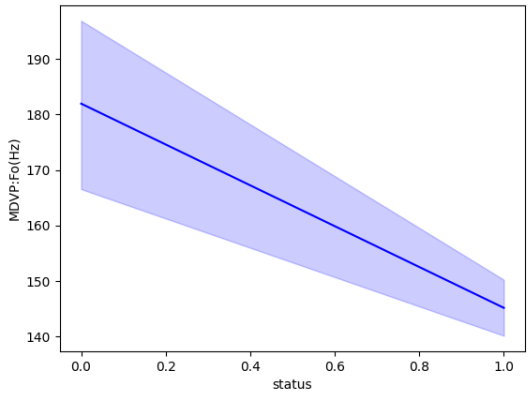


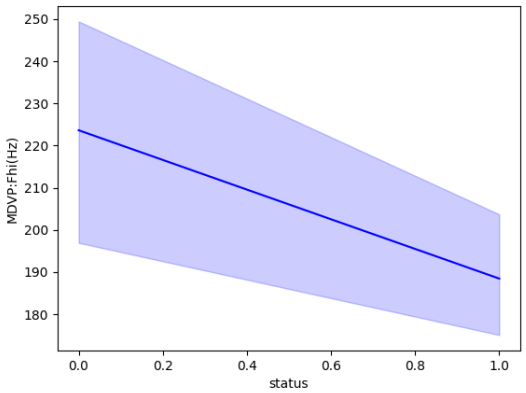


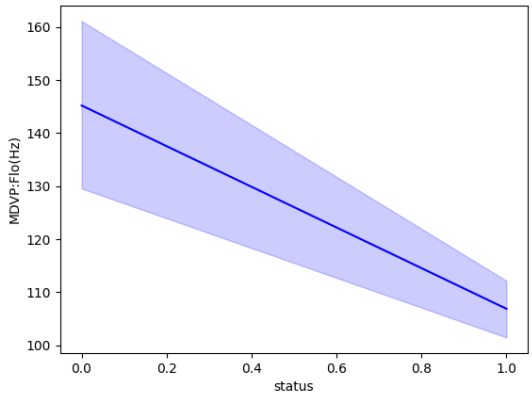


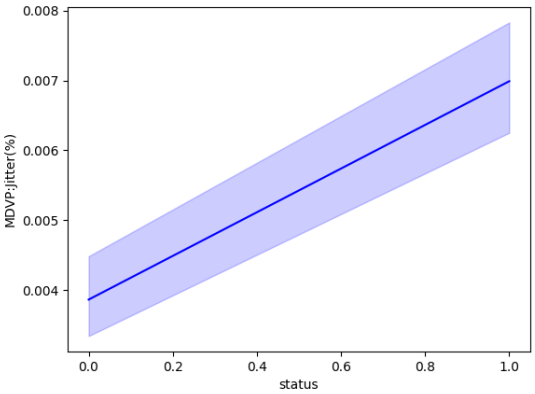


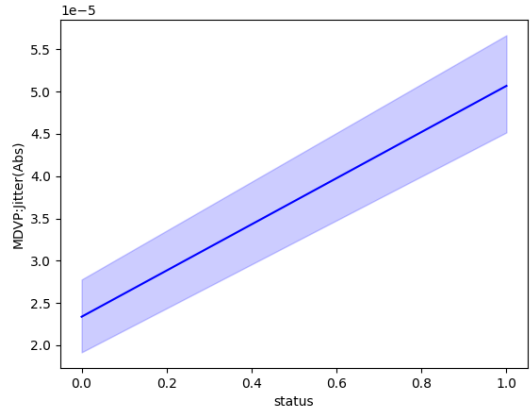
spread1,spread2 ve ppe featureları güzel ayrım yapmış. Değerleri yüksek olanlar genelde hasta, düşük olanlar ise genelde hasta değil çıkarımını yapabiliriz bu noktada.

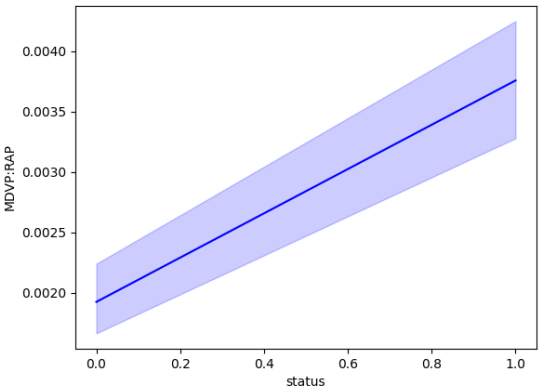

MDVP:Fo(Hz)
status

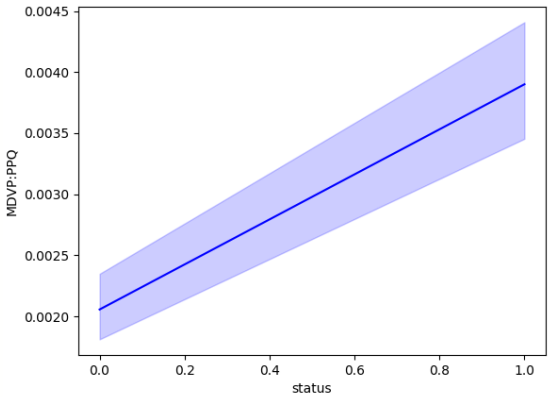

MDVP:Fhi(Hz)
status

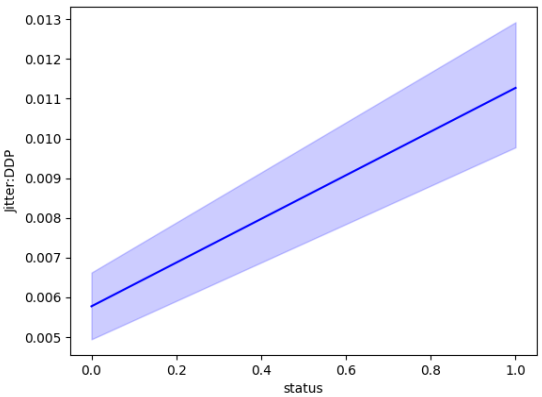

MDVP:Flo(Hz)
status

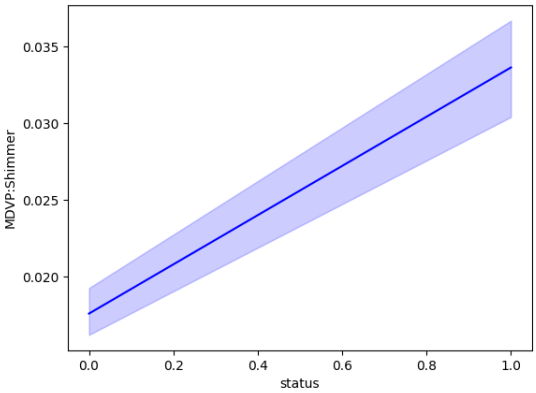

MDVP:Jitter(%)
status

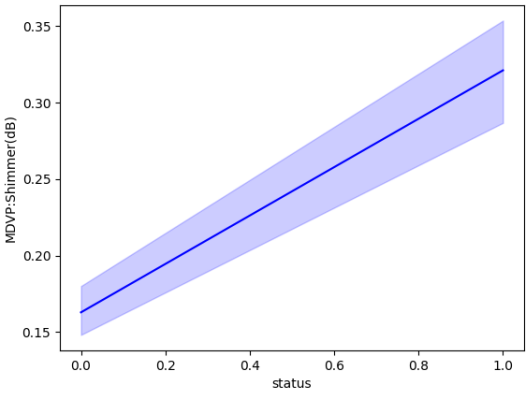

1e−5
MDVP:Jitter(Abs)
status

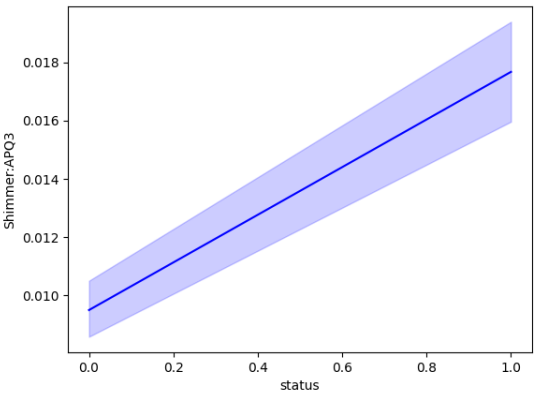

MDVP:RAP
status

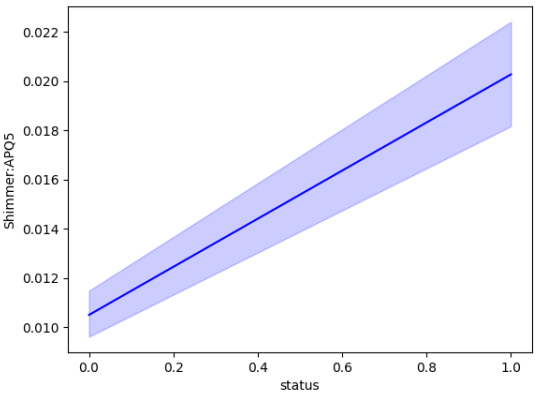

MDVP:PPQ
status

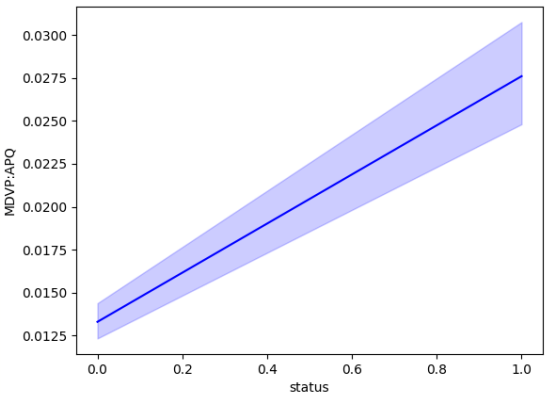

Jitter:DDP
status

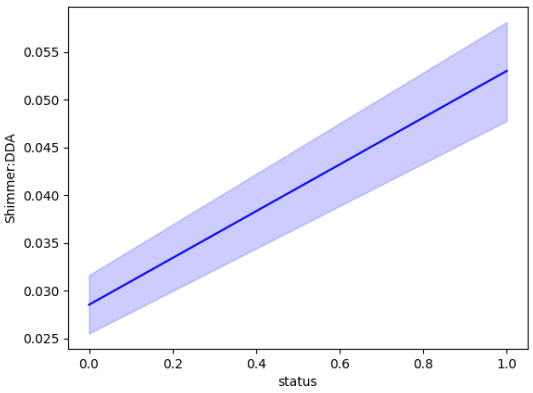

MDVP:Shimmer
0.035
0.030
0.025
0.020
0.0
0.2
0.4
0.6
0.8
1.0
status

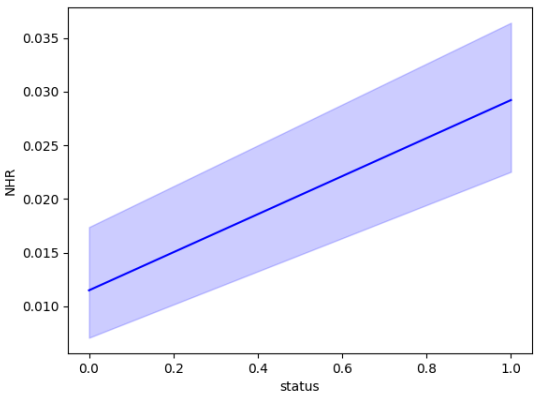

MDVP:Shimmer(dB)
0.35
0.30
0.25
0.20
0.15
0.0
0.2
0.4
0.6
0.8
1.0
status

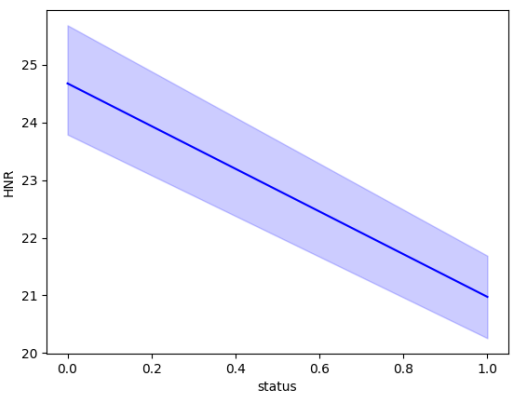

Shimmer:APQ3
0.018
0.016
0.014
0.012
0.010
0.0
0.2
0.4
0.6
0.8
1.0
status

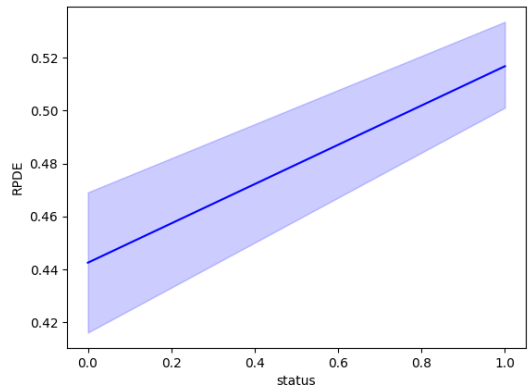

Shimmer:APQ5
0.022
0.020
0.018
0.016
0.014
0.012
0.010
0.0
0.2
0.4
0.6
0.8
1.0
status

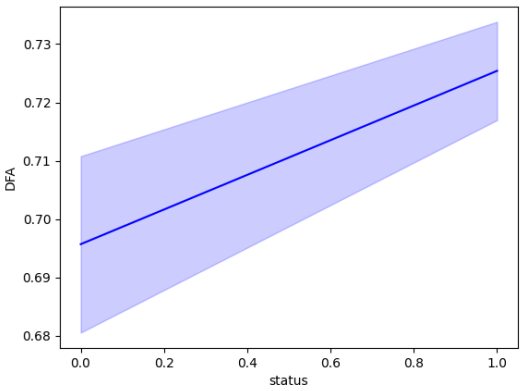

MDVP:APQ
0.0300
0.0275
0.0250
0.0225
0.0200
0.0175
0.0150
0.0125
0.0
0.2
0.4
0.6
0.8
1.0
status

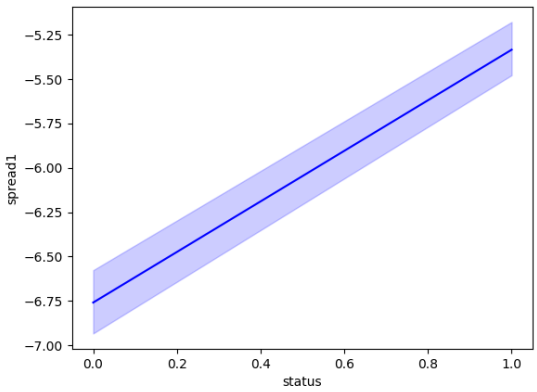

Shimmer:DDA
0.055
0.050
0.045
0.040
0.035
0.030
0.025
0.0
0.2
0.4
0.6
0.8
1.0
status

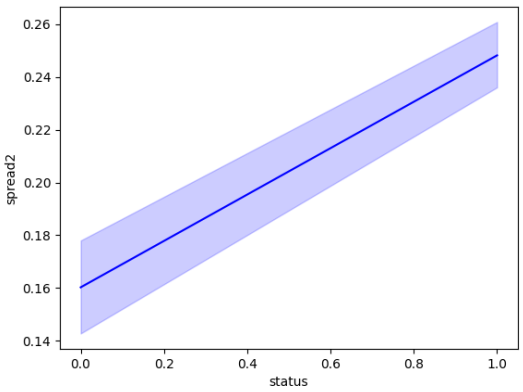

NHR
0.035
0.030
0.025
0.020
0.015
0.010
0.0
0.2
0.4
0.6
0.8
1.0
status

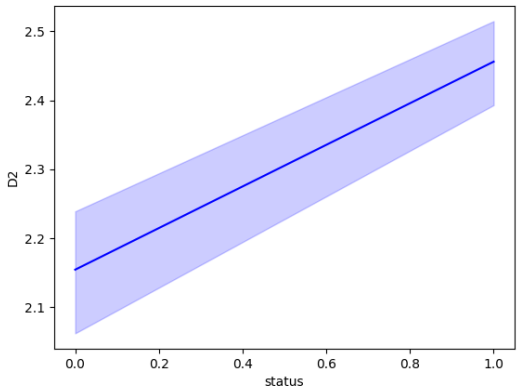

HNR
25
24
23
22
21
20
0.0
0.2
0.4
0.6
0.8
1.0
status

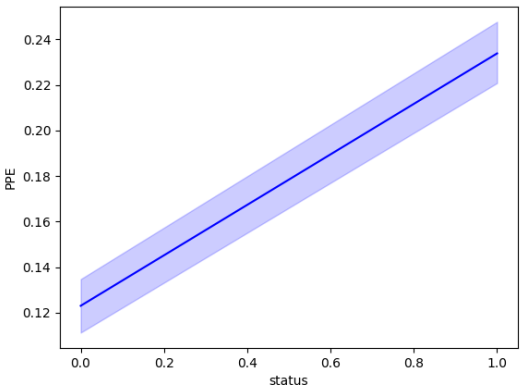












MDVP-FO(Hz),MDVP-Fhi(Hz), MDVP-Flo(Hz) ve HNR değerlerinin düşmesi demek hastalıklı olmanız demektir. Bu değerlerin y üksek olması hastalığı ayıran en net etkenlerden biri olarak görülmektedir.

Verilerimizi Standart şekilde scale ettikten sonra preprocess işlemlerimizi bitirmiş olacağız. Bundan sonra gidişata göre train,test ayrımı yapacağımızı söylemiştim. Deneme yanılma yoluyla uygun ayrımı 75 train, 25 test şeklinde karar verdim. Seçtiğimiz 4 modeli sözlük veri yapısında saklayıp ilgili eğitimini gerçekleştirdik. Verilerimiz hazır olduğuna göre karşılaştırma kısmına geçebiliriz.

```
XGBoost accuraty is: % 95.91836734693877

Logistic_Regression accuraty is: % 81.63265306122449

SVM_Classifier accuraty is: % 89.79591836734694

Desicion_Tree accuraty is: % 83.6734693877551
```

- XGBoost 100 deneme sonucu: 93-97 arasında bir sonuç verdi. Genel olarak kendisi mükemmel bir algoritma olduğu için az veriye rağmen overfitting yapmamaya gayret ediyor. Kendisi maksimum shuffle edilen test verilerinde 3 kere hata yaptı.

- Logistic Regression 100 deneme sonucu: Bazı noktalarda accuraty, %90'ların üstüne çıktı fakat genel olarak 80-84 arasında geziyor. Bununla birlikte yüksek accuraty değerleri ile düşük accuraty değerleri arasındaki fark, düşük veriden kaynaklı overfitting göstergesi olarak yorumluyoruz.

- SVM Classifier 100 deneme sonucu: Bu noktada SVM makinesi, maksimum %91 minimum %84 değerlerinde dolaştı. Burada lojistik regresyondan daha iyi bir sonuç ama genel olarak overfitting durumu görülmekte, eğer C değerini overfitting için yükseltirsek accuraty seviyesi %70'lere düşüp az verinin cezası olarak tahminde fazla hata yapma durumunda kalıyor. Bu noktada varyans arttırılması, daha genel bir öğrenme yaparken az veri için sınıflandırma hatasına neden oluyor. Yani az veri konusunda hala XGBoost 1 numara.

- Desicion Tree 100 deneme sonucu: Karar ağaçları, bazen XGBoost accuraty değerleri yakalasa da bazen de lojistik regresyon kadar düşük değerler verebiliyor. Bu noktada maximum 93, minimum 81 accuarty aralığı seçebiliriz. Karar ağaçlarının değiken yapısı, random-forest tekniğinden dolayı bazen iyi bir çizgi çekerken bazen overfitting durumunda kalıyor. Bu noktada dengesiz bir öğrenim söz konusu diyebilirim. Bunun sebebi ise bariz, verilerimiz az sayıda.

XGBoost Model Classification predict -- Real Data is

```
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
Healthy : ------- : Healthy
infected : ------- : Healthy    Mistake!!
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : infected    Mistake!!
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
```

Total Mistake is : 2

Logistic_Regression Model Classification predict -- Real Data is

```
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : infected    Mistake!!
Healthy : ------- : infected    Mistake!!
infected : ------- : Healthy    Mistake!!
Healthy : ------- : infected    Mistake!!
infected : ------- : infected
infected : ------- : infected
infected : ------- : Healthy    Mistake!!
Healthy : ------- : Healthy
infected : ------- : Healthy    Mistake!!
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : Healthy    Mistake!!
Healthy : ------- : infected    Mistake!!
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : Healthy    Mistake!!
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
```

Total Mistake is : 9

SVM_Classifier Model Classification predict -- Real Data is

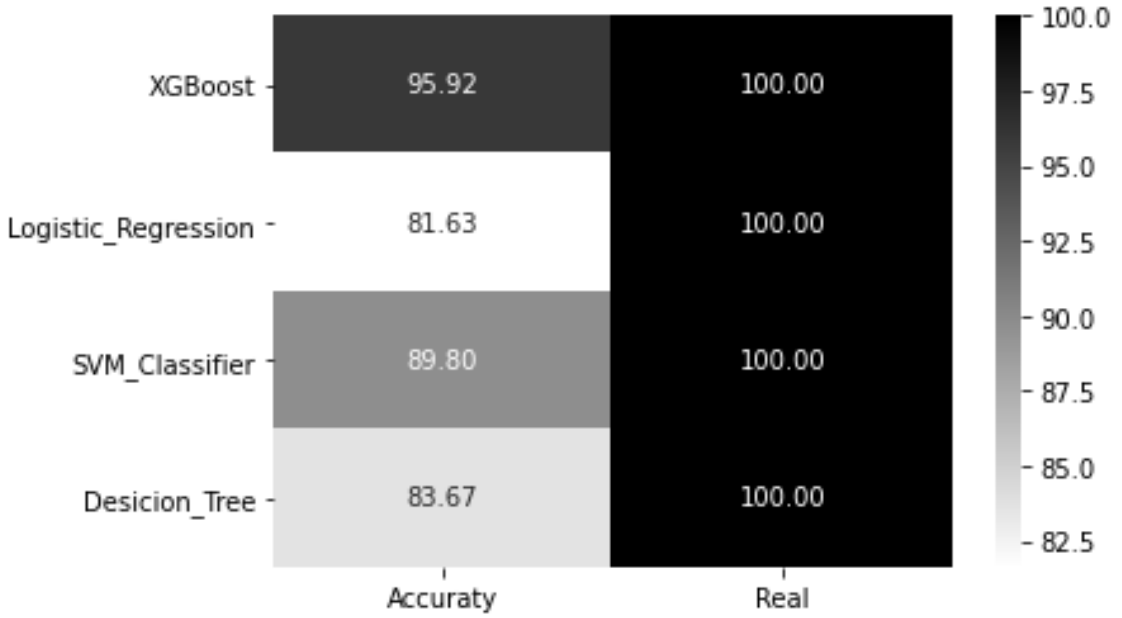
```
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : Healthy    Mistake!!
infected : ------- : infected
infected : ------- : infected
infected : ------- : Healthy    Mistake!!
Healthy : ------- : Healthy
infected : ------- : Healthy    Mistake!!
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : Healthy    Mistake!!
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : Healthy    Mistake!!
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
```

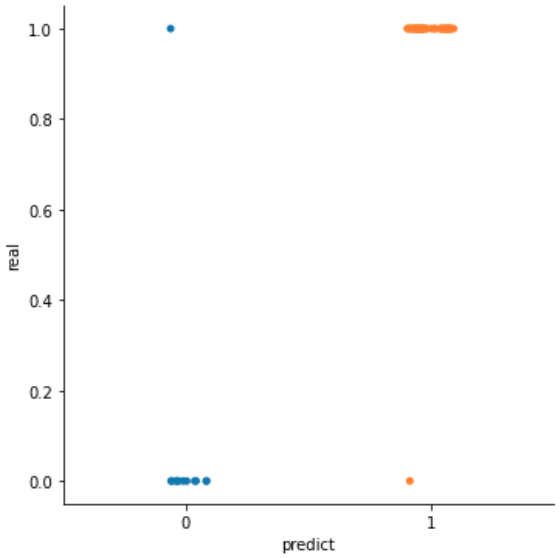
Total Mistake is : 5

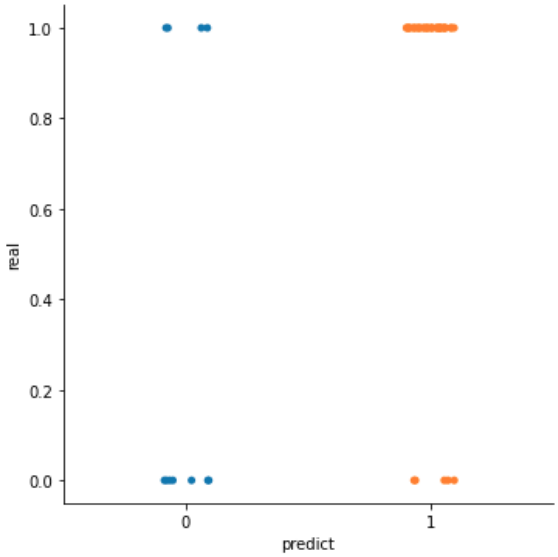
Desicion_Tree Model Classification predict -- Real Data is

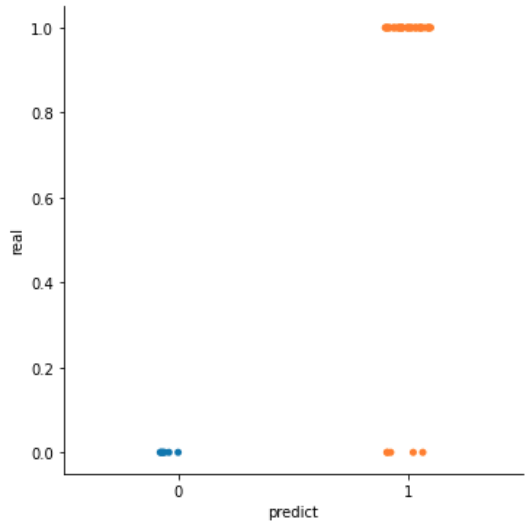
```
Healthy : ------- : infected    Mistake!!
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : infected    Mistake!!
infected : ------- : infected
Healthy : ------- : Healthy
Healthy : ------- : infected    Mistake!!
infected : ------- : infected
infected : ------- : infected
infected : ------- : Healthy    Mistake!!
Healthy : ------- : Healthy
infected : ------- : Healthy    Mistake!!
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : infected    Mistake!!
Healthy : ------- : Healthy
Healthy : ------- : infected    Mistake!!
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : infected    Mistake!!
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
Healthy : ------- : Healthy
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
infected : ------- : infected
Healthy : ------- : Healthy
```

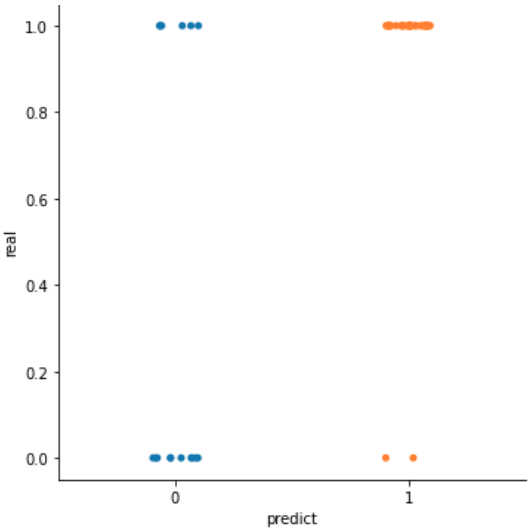
Total Mistake is : 8


XGBoost
Logistic_Regression
SVM_Classifier
Desicion_Tree
95.92
81.63
89.80
83.67
100.00
100.00
100.00
100.00
Accuraty
Real
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5


real
predict
0.0
0.2
0.4
0.6
0.8
1.0
0
1


real
predict
0.0
0.2
0.4
0.6
0.8
1.0
0
1

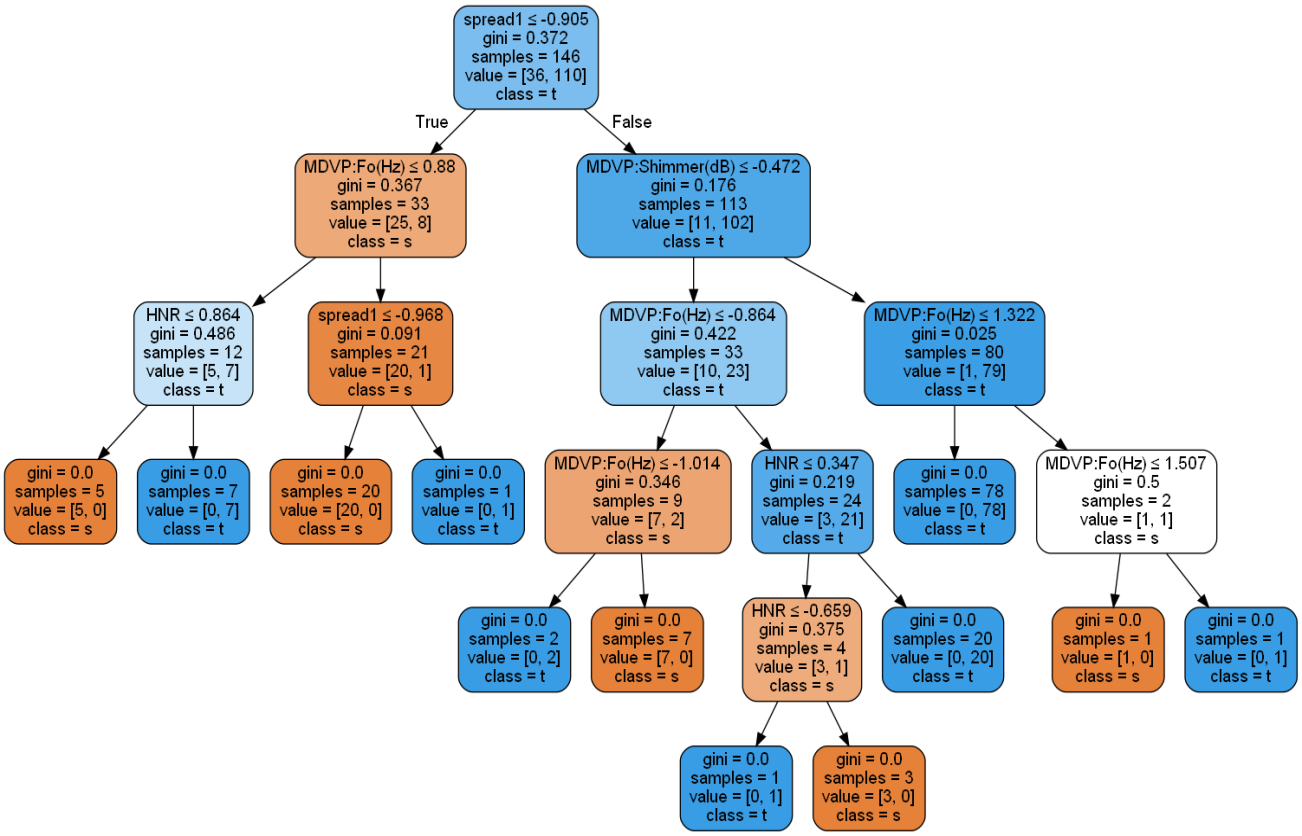

real
predict
0.0
0.2
0.4
0.6
0.8
1.0
0
1


real
predict
0.0
0.2
0.4
0.6
0.8
1.0
0
1

- Sonuç: Düşük veri arasında modeller için karşılaştırma yaptığımızda bir sıralama yaparsak;

1-)XGBoost
2-)SVM Machine
3-)Desicion Tree
4-)Logistic Regression

Sıralamasını elde etmiş bulunmaktayız.



Desicion Tree için örnek bir tree oluşumunu da bu şekilde görmekteyiz. Desicion Tree için gini yerine cross-entropy seçtiğimiz zaman genel accuraty düşüklüğü elde etmiş bulunmaktayız. Az verilerimizden dolayı cross-entropy, çok keskin kararlar veriyor ve birbirlerine yakın hasta-sağlıklı dataları için sınıflandırma hatası yapıyor. Bu yüzden gini, cost fonksiyonu olarak uygun görülmektedir. Bu projede, hyperparameter optimization konusu dâhil değildir. Bu yüzden parametreler, deneme yanılma yoluyla elde edilmiştir.