

Veri Madenciliği Proje

Ad Soyad : Selçuk Arda Özcan

**Proje İsmi : 2025 F1 SEZONU VERİLERİYLE REGRESYON VE
KARAR AĞACI MODELLEMESİ**

1. Projenin Tanımı ve Amacı

Bu projede 2025 Formula 1 sezonu verilerini kullanarak yarış sonunda kimin ne kadar puan alması gerektiğini tahmin etmeyi hedeflenmiştir. Temelde, **yarış (Grand Prix, GP)** başlamadan önce pilotun ve takımın **aracının pist üzerindeki diziliminin (starting grid)**, yarış bittikten sonra pozisyonunuza bağlı verilen final puanı(**points**) üzerindeki etkisini matematiksel olarak modellemek ve farklı algoritmaların veriler üzerindeki performansın karşılaştırılması amaçlanmıştır.

2. Veri Setinin Özellikleri

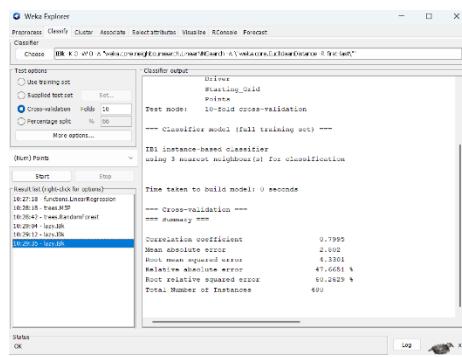
Kullandığım veri seti “**F1_Season_2025_GPs_Final_Results.arff**” adlı 2025 F1 sezonundaki 24 farklı Grand Prix yarışını ve her yarıştaki 20 sürücüyü kapsayan toplam 419 satır ve 5 sütünden oluşan güncel bir veri setidir.

- * Sezonda 10 takım ve her takımda güncel olarak 2 sürücü vardır. Sezon boyunca toplam 21 farklı sürücü olmasının nedeni sürücü(driver) “Jack Doohan” İlk 4 yarışa çıktıktan sonra yerine “Franco Colapinto” getirilmesidir. Yine güncel olarak her yarışa en fazla 20 sürücü çıkmıştır.
- * Yarışı tamamlayamayan (DNF) veya diskalifiye edilen (DSQ) sürücüler, modelin verimliliğini düşürmemek için dosyadan silinmiştir.

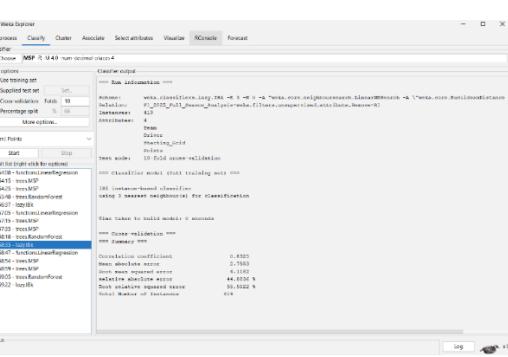
IBK K=3 Cross-validation-folds10

DNF-DSQ olanların silinmediği

ve silindiği modellerin verimliliği



Korelasyon katsayısı: 0.7995



, 0.8325

MAE : 2.802

, 2.7553

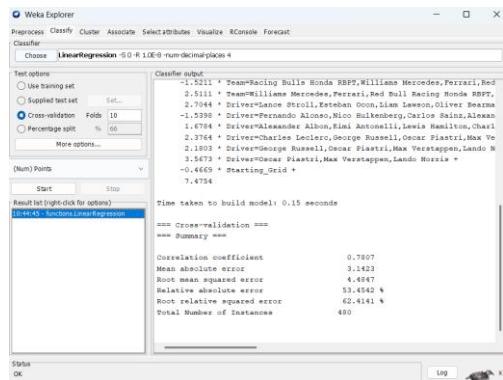
480 satır

, 419 satır

Linear Regression Cross-validation-folds10

DNF-DSQ olanların silinmediği

ve silindiği modellerin verimliliği



Korelasyon katsayı: 0.7807

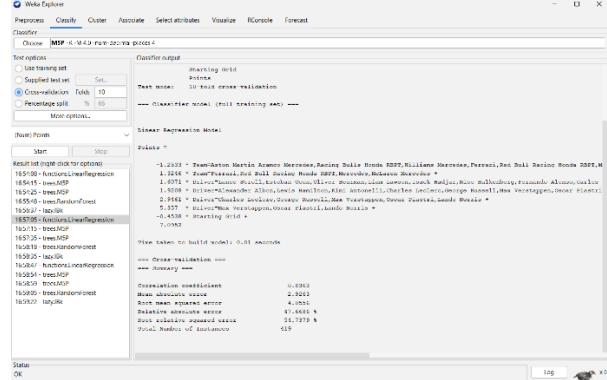
, 0.8363

MAE : 3.1423

, 2.9263

480 satır

, 419 satır



* Pist isimleri(Track) , sadece tek bir sezonu hesapladığımız için gereksiz (noise) bir özniteliktir.

* Baz alınan [kaggle](#)'daki veri seti sezonun İspanya'dan sonraki yarışlarını kapsamayan eksik ve güncel olmayan bir veri setidir. Öznitelik eksiltme ve düzenlemelerle “F1_Season_2025_GPs_Final_Results.arff” oluşturulmuştur.

- **Öznitelikler:** Track (“Pist”, Nominal Değer)

Team (“Takım”, Nominal Değer)

Driver (“Sürücü”, Nominal Değer)

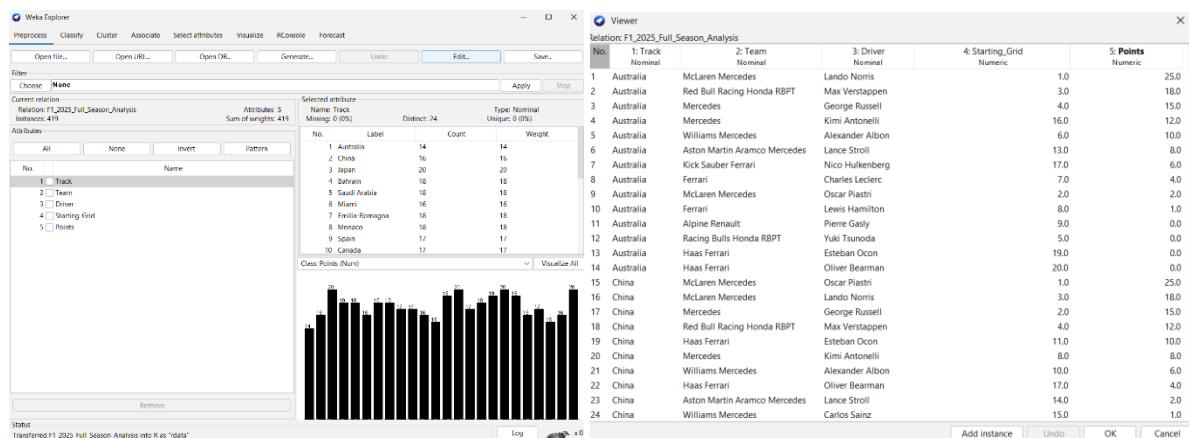
Starting_Grid (“Yarış başlangıcı için pist üzerindeki dizilim”, Numeric)

Hedef Değişken: Points (Numeric) (“Yarış sonu pozisyonunuza bağlı verilen 0-25 arası sürekli değerler”).

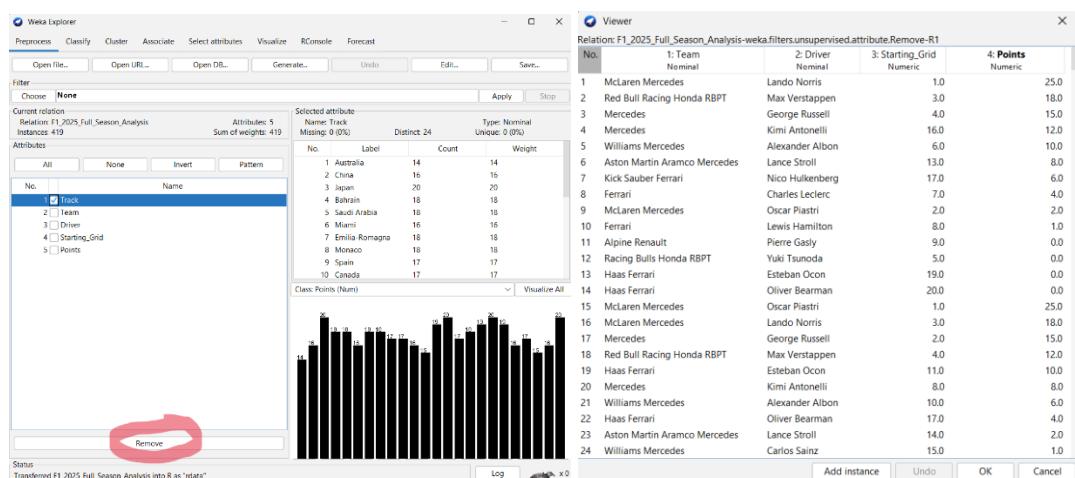
F1 de yarış sonu, bitiriş pozisyonuna bağlı verilen puan tablosu

Positions	Points
1st Position	25 Points
2nd Position	18 Points
3rd Position	15 Points
4th Position	12 Points
5th Position	10 Points
6th Position	8 Points
7th Position	6 Points
8th Position	4 Points
9th Position	2 Points
10th Position	1 Point

WEKA'da veri setindeki örnek satırların(instances) görünümü

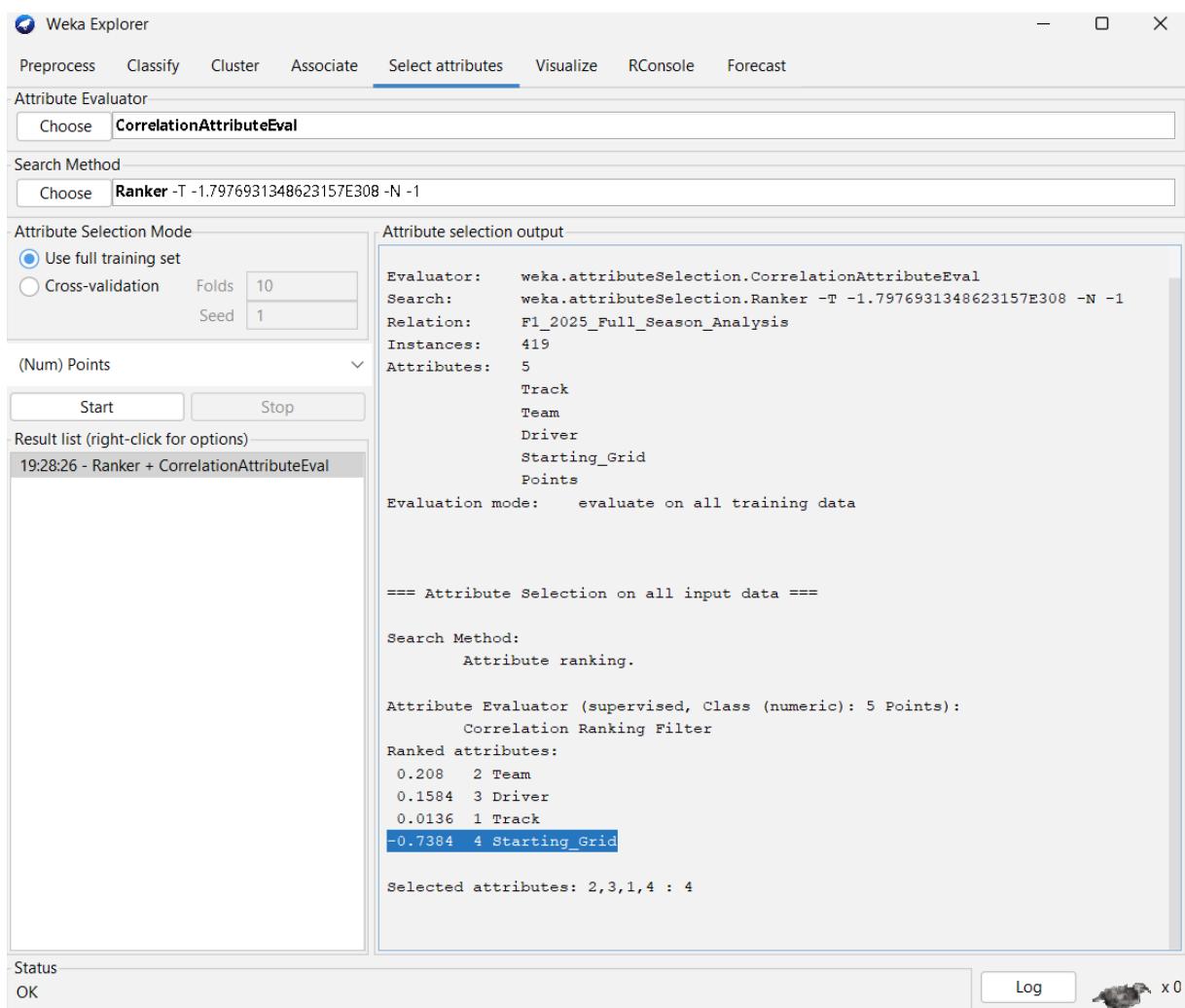


Pist(Track) kaldırıldıktan sonra WEKA'da veri setindeki örnek satırların(instances) görünümü



3. Öznitelik Seçimi (Attribute Selection)

Select Attributes kısmından CorrelationAttributeEval yap sonucunda, puanı(points) belirleyen en güçlü faktörün - 0.7384 skorla Starting_Grid olduğu ispatlanmıştır. Negatif korelasyon, grid sırası küçüldükçe puanın arttığını (ters orantı) bilimsel olarak doğrulamaktadır.
(skor 1.00'e kadar yaklaşrsa o kadar doğrudur.)



4. Kullanılan Algoritmalar

Projede farklı versiyondan 5 tane algoritma kullanılmıştır .Modellerin başarısı **Correlation coefficient (korelasyon katsayısı)** ve **Mean Absolute Error (MAE)** metrikleri ile ölçülmüştür.

- **Linear Regression :** Değişkenler arasındaki doğrusal ilişkiye modelleyen fonksiyondur.
- **M5P (trees) ve M5P buildRegression_true** versiyonları kullanıldı.
- **RandomForest :** 100 farklı ağaç aynı anda çalışır . Karakutu(blackbox) .
- **IBk K=3 (KNN) :** 3-en yakın komşu algoritmasını kullanan örnek tabanlı sınıflama yapar.

ALGORİTMA KARŞILAŞTIRMA TABLOSU

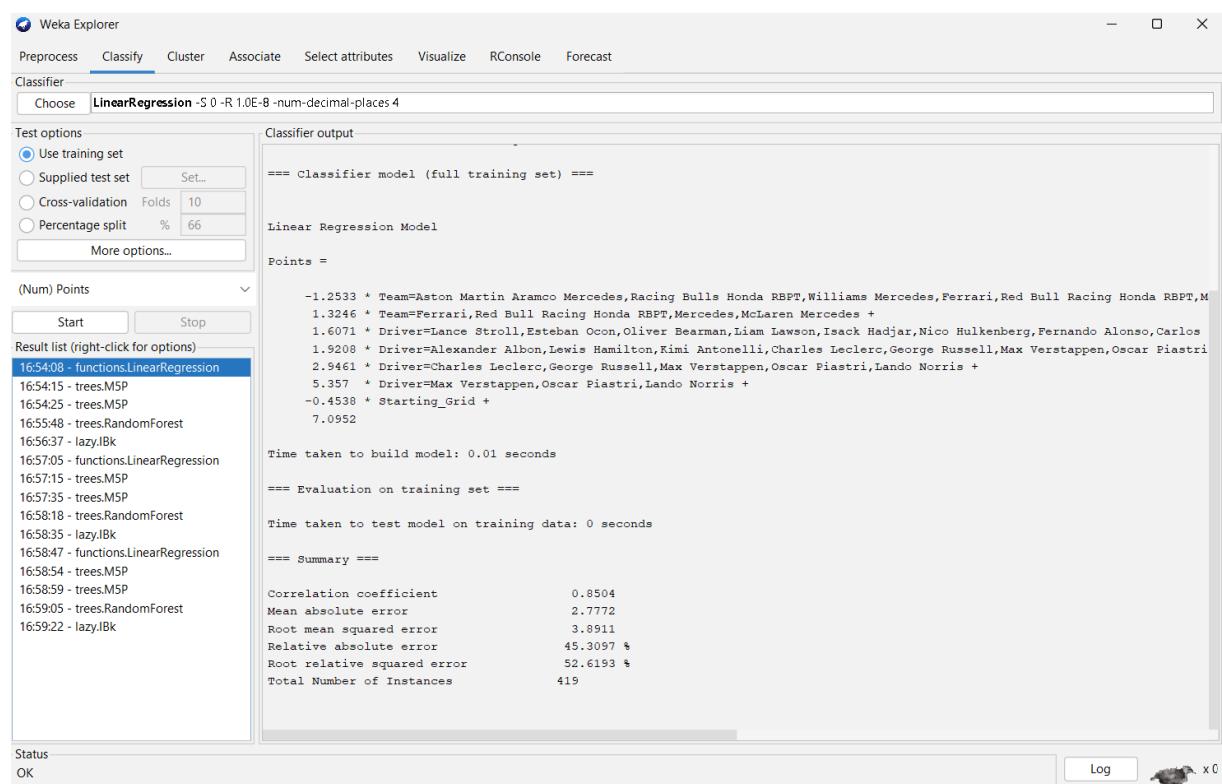
Algoritma	Test Modu	Korelasyon Katsayısı	MAE (Ort. Mutlak Hata)	RMSE (Karekök Hata)
Linear Reg.	Eğitim kümesi	0.8504	2.7772	3.8911
	çapraz doğrulama	0.8363	2.9263	4.0556
	%66 bölme	0.8515	2.9947	4.0736
M5P	Eğitim kümesi	0.8771	2.5195	3.5529
	çapraz doğrulama	0.855	2.7448	3.8356
	%66 bölme	0.8521	2.8196	4.0702
M5PbuildR.	Eğitim kümesi	0.855	2.8455	3.9373
	çapraz doğrulama	0.8118	3.1585	4.355
	%66 bölme	0.8556	3.2164	4.3573
RandomForest	Eğitim kümesi	0.9329	1.658	2.6652
	çapraz doğrulama	0.8329	2.7289	4.125
	%66 bölme	0.8634	2.6505	3.9309
IBk K=3	Eğitim kümesi	0.9035	2.0697	3.1689
	çapraz doğrulama	0.8325	2.7553	4.1182
	%66 bölme	0.859	2.7138	3.9741

Bu tabloda eğitim seti ve %66'luk bölümde en yüksek korelasyon değeri 0.9239 , 0.8634 ile **RandomForest**'tadır.

Alt tarafta tabloya konu olan sonuçların ekran görüntüleri koyulmuştur. Ve olan modeller için ağaçları görselleştirilmiştir

- **Eğitim setiyle çalıştırıldığında (use training set)**

1. Linear Regresyon



2. M5P

The screenshot shows the Weka Classifier Tree Visualizer interface. On the left, there's a 'Test options' panel with 'Use training set' selected. Below it is a 'Result list (right-click for options)' list containing several entries, with '165408 - functions.LinearRegression' highlighted. On the right, the 'Classifier output' pane displays the generated M5P model rules:

```

IM num: 2
Points =
- 0.0589 * Team=Aston Martin Aramco Mercedes,Racing Bulls Honda RBPT,Williams Mercedes,Ferrari,Red Bull Racing Honda RBPT
- 1.105 * Team=Racing Bulls Honda RBPT,Williams Mercedes,Ferrari,Red Bull Racing Honda RBPT,Mercedes,McLaren Mercedes
+ 1.4019 * Team=Ferrari,Red Bull Racing Honda RBPT,Mercedes,McLaren Mercedes
+ 1.3104 * Driver=Lance Stroll,Ezecan Ocan,Oliver Bearman,Liam Lawson,Issack Hadjar,Nico Hulkenberg,Fernando Alonso,Car
+ 2.3393 * Driver=Alexander Albon,Lewis Hamilton,Kimi Anttonelli,Charles Leclerc,George Russell,Max Verstappen,Oscar Piastri,Lando Norris
+ 0.1385 * Driver=Charles Leclerc,George Russell,Max Verstappen,Oscar Piastri,Lando Norris
+ 3.0087 * Driver=George Russell,Max Verstappen,Oscar Piastri,Lando Norris
+ 0.2519 * Driver=Max Verstappen,Oscar Piastri,Lando Norris
- 0.2662 * Driver=Starting_Grid
+ 4.4423

Number of Rules : 2

Time taken to build model: 0.05 seconds

==== Evaluation on training set ====

Time taken to test model on training data: 0 seconds

==== Summary ====

```

Summary statistics:

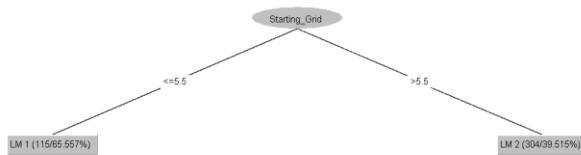
- Correlation coefficient: 0.8771
- Mean absolute error: 2.5195
- Root mean squared error: 3.5529
- Relative absolute error: 41.1063 %
- Root relative squared error: 48.0454 %
- Total Number of Instances: 419

Status: OK

M5P ağacının görselleştirmesi

Weka Classifier Tree Visualizer: 165415 - trees.MSP (F1_2025_Full_Season_Analysis-weka.filters.unsupervised.attribute.Re

Tree View



3. M5P (buildRegression True)

The screenshot shows the Weka GenericObjectEditor for the 'weka.classifiers.trees.MSP' class. The 'buildRegressionTree' parameter is set to 'true' (highlighted in yellow). Other parameters include 'batchSize: 100', 'doNotCheckCapabilities: false', 'minNumInstances: 40', 'numDecimalPlaces: 4', 'saveInstances: false', 'unpruned: false', and 'useUnsmoothed: false'. The 'OK' button is highlighted.

On the right, the 'Classifier output' pane shows the M5P model output for this configuration:

```

Test options
  • Use training set
  • Supplied test set Set...
  • Cross-validation Folds 10
  • Percentage split % 66
  More options...

Classifier output
  Points =
    + 13.5297
  IM num: 4
  Points =
    + 3.0728
  IM num: 5
  Points =
    + 1.181

  Time taken to build model: 0.06 seconds
  Time taken to test model on training data: 0 seconds

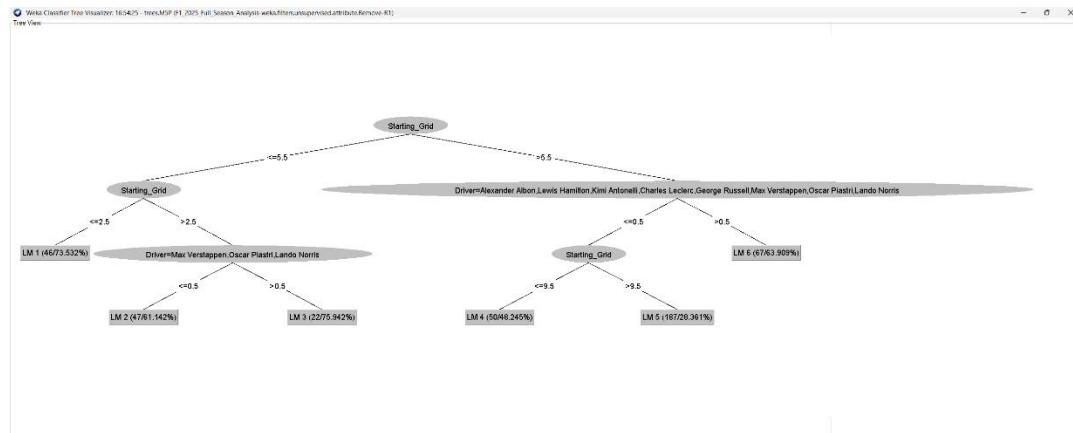
  === Summary ===

  Correlation coefficient: 0.855
  Mean absolute error: 2.0455
  Root mean squared error: 3.9373
  Relative absolute error: 46.4045 %
  Root relative squared error: 53.2436 %
  Total Number of Instances: 419

```

Status: OK

M5P (buildRegression True) görselleştirmesi



Ağacın en tepesinde (kök düğüm) Starting_Grid yer almıştır. Bu durum, bir pilotun yarış sonunda kaç puan alacağını belirleyen en güçlü faktörün başlangıç dizilimi olduğunu gösterir.

Model, veriyi ilk olarak **5.5** (yani ilk 5-6 sıra) eşigine göre iki ana gruba ayırmıştır.

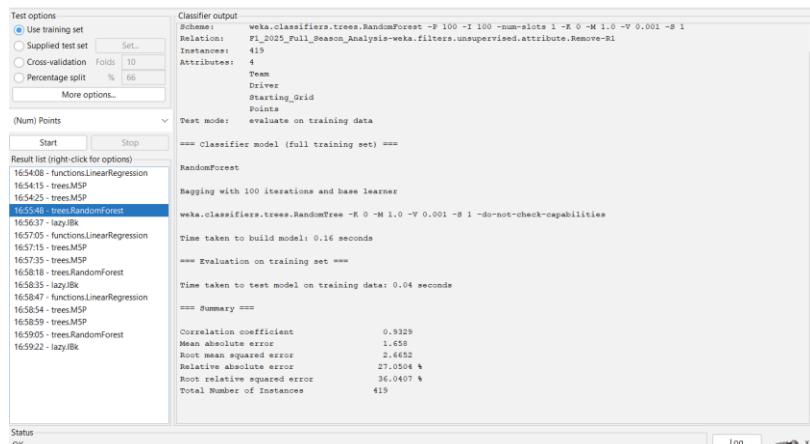
Sol Taraf: Ön Sırada Başlayanlar (Starting_Grid <= 5.5)

- Podyum Grubu (LM1):** İlk 2-3 sırada başlayan sürücüler (Starting_Grid <= 2.5), takımdan bağımsız olarak en yüksek puan potansiyeline sahip grubuktur.
- Sürüci Etkisi (LM2 ve LM3):** 3. ile 5. sıra arasında başlayanlar için model, sürüci kalitesine bakar. Eğer sürücü **Max Verstappen, Oscar Piastri** veya **Lando Norris** üçlüsünden biriyse (LM3), puan tahmini diğer sürücülere (LM2) göre daha yüksektir.

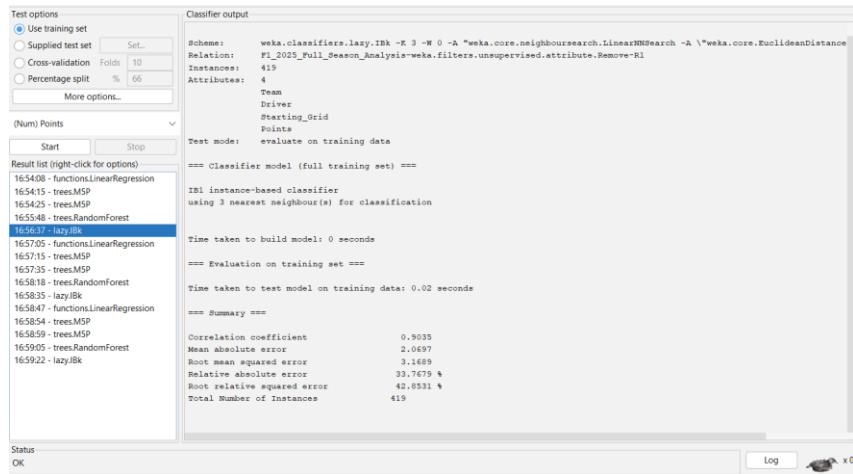
3. Sağ Taraf: Orta ve Arka Sıralar (Starting_Grid > 5.5)

- Yetenekli Pilotların Telafisi (LM6):** 6. sıradan daha geride başlayan ancak "rekabetçi sürüci grubu" (Albon, Hamilton, Leclerc, Russell, Verstappen vb.) içinde yer alan pilotlar, gerilerden başlasalar bile LM6 kuralı ile hala puan alma potansiyeline sahiptir.
- Puan Barajı (LM4 ve LM5):** Rekabetçi grupta olmayan ve 6. sıradan geride başlayan sürücüler için yeni bir baraj oluşur:
 - 9.5 Eşiği:** Eğer bu sürücüler ilk 10'un içinde (Starting_Grid <= 9.5) başlarsa düşük de olsa bir puan (LM4) öngörlürken, 10. sıradan daha geride (Starting_Grid > 9.5) başlayanlar için model en düşük puan beklentisini (LM5) sunmaktadır.

4. RandomForest

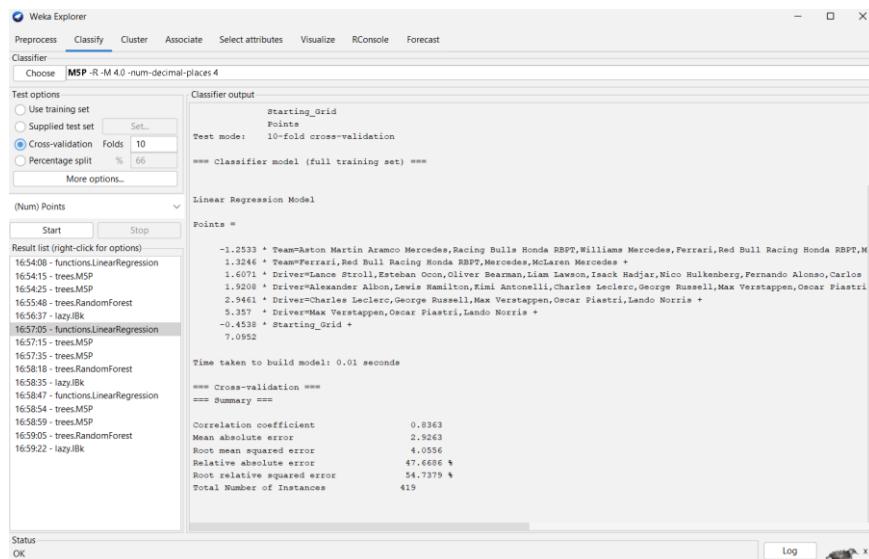


5. IBk K=3



• 10 Katlamalı Çapraz Doğrulama (cross-val-fold-10)

1. Linear Regresyon



2. M5P

The screenshot shows the Weka interface with the 'Classify' tab selected. Under 'Classifier', 'M5P' is chosen. In the 'Test options' section, 'Cross-validation' is selected with 'Folds' set to 10. The 'Classifier output' pane displays the generated regression equation and various performance metrics.

```

Classifier output
+ 2.7761 * Driver=Charles Leclerc,George Russell,Max Verstappen,Oscar Piastri,Lando Norris
+ 3.1301 * Driver=Max Verstappen,Oscar Piastri,Lando Norris
- 1.9211 * Starting_Grid
+ 16.5584

LM num: 2
Points =
- 0.0589 * Team=Aston Martin Aramco Mercedes,Racing Bulla Honda RBPT,Williams Mercedes,Ferrari,Bed Bull Racing Honda RBPT
- 1.126 * Team=Racing Bulla Honda RBPT,Williams Mercedes,Ferrari,Bed Bull Racing Honda RBPT,Mercedes,McLaren Mercedes
+ 1.4049 * Team=Ferrari,Red Bull Racing Honda RBPT,Mercedes,McLaren Mercedes
+ 3.1308 * Driver=Lance Stroll,Esteve Ocon,Oliver Bearman,Liam Lawson,Isack Hadjar,Nico Hulkenberg,Fernando Alonso,Car
+ 2.3393 * Driver=Alexander Albon,Lewis Hamilton,Kimi Anttonelli,Charles Leclerc,George Russell,Max Verstappen,Oscar Pi
- 0.1385 * Driver=Charles Leclerc,George Russell,Max Verstappen,Oscar Piastri,Lando Norris
+ 0.2519 * Driver=George Russell,Max Verstappen,Oscar Piastri,Lando Norris
- 0.2682 * Starting_Grid
+ 4.4423

Number of Rules : 2
Time taken to build model: 0.05 seconds

*** Cross-validation ***
*** Summary ***

Correlation coefficient 0.855
Mean absolute error 2.7448
Root mean squared error 3.8356
Relative absolute error 44.7116 %
Root relative squared error 51.769 %
Total Number of Instances 419

```

3. M5P (buildRegression True)

The screenshot shows the Weka interface with the 'Classify' tab selected. Under 'Classifier', 'M5P' is chosen. In the 'Test options' section, 'Cross-validation' is selected with 'Folds' set to 10. The 'Classifier output' pane displays the generated regression equation and various performance metrics.

```

Classifier output
+ 10.4979

LM num: 2
Points =
+ 13.5297

LM num: 4
Points =
+ 3.0728

LM num: 5
Points =
+ 1.181

LM num: 6
Points =
+ 4.8551

Number of Rules : 4
Time taken to build model: 0.04 seconds

*** Cross-validation ***
*** Summary ***

Correlation coefficient 0.8110
Mean absolute error 3.1590
Root mean squared error 4.1355
Relative absolute error 51.4500 %
Root relative squared error 59.7709 %
Total Number of Instances 419

```

4. RandomForest

The screenshot shows the Weka interface with the 'Classify' tab selected. Under 'Classifier', 'RandomForest' is chosen. In the 'Test options' section, 'Cross-validation' is selected with 'Folds' set to 10. The 'Classifier output' pane displays the generated regression equation and various performance metrics.

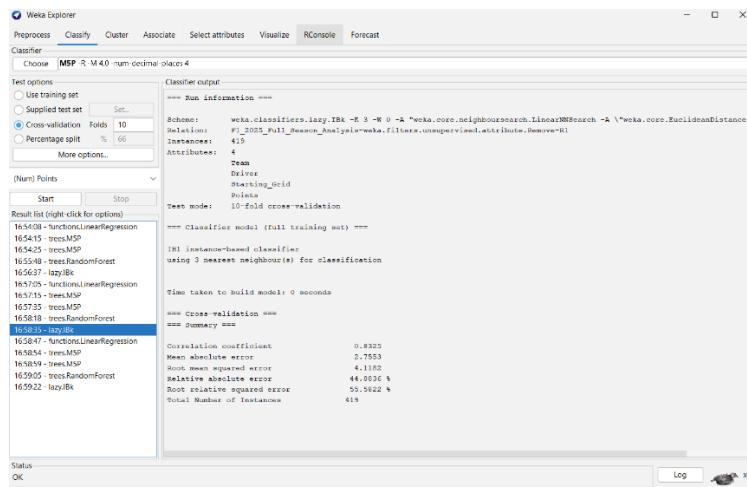
```

Classifier output
*** Run Information ***
Scheme: weka.classifiers.trees.RandomForest -> 100 -num-slots 2 -w -I 0 -r 0.001 -s 1
Selection: v1.2057_vul1_season_AnalyticWeka.filters.unsupervised.attribute.remove@1
Instances: 419
Attributes: 4
Tree:
  Root
    Driver
      Points
    Points
  Decision
    10-fold cross-validation
    **** Classifies model (full training set) ****
    RandomForest
    Debugging with 100 iterations and base learner
    weka.classifiers.trees.RandomTree -E 0 -W 1.0 -P 0.001 -S 1 -do-not-check-capabilities
    Time taken to build model: 0.12 seconds

    *** Cross-validation ***
    *** Summary ***
    Correlation coefficient 0.0319
    Mean absolute error 2.1209
    Root mean squared error 4.1279
    Relative absolute error 44.4511 %
    Root relative squared error 51.4712 %
    Total Number of Instances 419

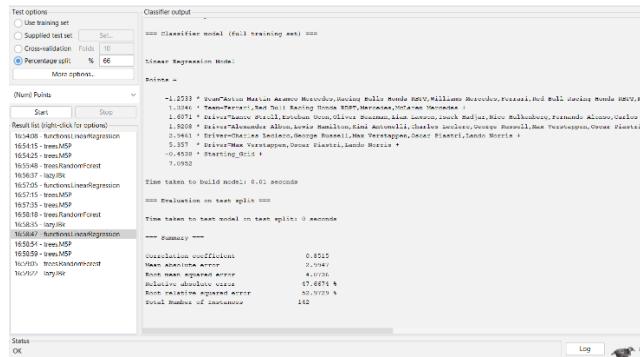
```

5. IBk K=3



- %66'luk bölmeye metoduyla percentage split 66%)

1. Linear Regresyon



2. M5P



3. M5P buildRegression_true

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'M5P -R 4.0 -num-decimal-places 4'. The 'Test options' section shows 'Percentage split % 66' selected. The 'Classifier output' pane displays the following results:

```

Classifier output
Points = + 13.5287
IM num: 4
Points = + 3.0728
IM num: 5
Points = + 1.181
IM num: 6
Points = + 4.8891
Number of Rules : 6
Time taken to build model: 0.04 seconds
*** Evaluation on test split ***
Time taken to test model on test split: 0 seconds
*** Summary ***
Correlation coefficient          0.0556
Mean absolute error            3.2164
Root mean squared error        4.3573
Relative absolute error         51.1504
Root relative squared error    56.4624
Total Number of Instances       142

```

4. RandomForest

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1'. The 'Test options' section shows 'Percentage split % 66' selected. The 'Classifier output' pane displays the following results:

```

Classifier output
Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Relation: F1_2025_Full_Season_Analysis-weka.filters.unsupervised.attribute.Remove-R1
Instances: 419
Attributes: 4
Team
Driver
Starting_Grid
Points
Test mode: split 66.0% train, remainder test
*** Classifier model (full training set) ***
RandomForest
Bagging with 100 iterations and base learner
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities
Time taken to build model: 0.1 seconds
*** Evaluation on test split ***
Time taken to test model on test split: 0.02 seconds
*** Summary ***
Correlation coefficient          0.8634
Mean absolute error            2.6505
Root mean squared error        3.9309
Relative absolute error         42.1884 %
Root relative squared error    51.1167 %
Total Number of Instances       142

```

5. IBk K=3

Test options

Use training set
 Supplied test set Set...
 Cross-validation Folds 10
 Percentage split % 66
More options...

(Num) Points ▾

Start Stop

Result list (right-click for options)

- 165408 - functions.LinearRegression
- 165415 - trees.MSP
- 165425 - trees.MSP
- 165548 - trees.RandomForest
- 165637 - lazy.IBk
- 165705 - functions.LinearRegression
- 165715 - trees.MSP
- 165735 - trees.MSP
- 165818 - trees.RandomForest
- 165835 - lazy.IBk
- 165847 - functions.LinearRegression
- 165854 - trees.MSP
- 165859 - trees.MSP
- 165905 - trees.RandomForest
- 165922 - lazy.IBk

Classifier output

```
Scheme: weka.classifiers.lazy.IBk -K 3 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance"
Relation: F1_2025_Full_Season_Analysis-weka.filters.unsupervised.attribute.Remove-R1
Instances: 419
Attributes: 4
Team
Driver
Starting_Grid
Points
Test mode: split 66.0% train, remainder test

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correlation coefficient          0.859
Mean absolute error              2.7138
Root mean squared error          3.9741
Relative absolute error          43.1958 %
Root relative squared error     51.675 %
Total Number of Instances       142
```

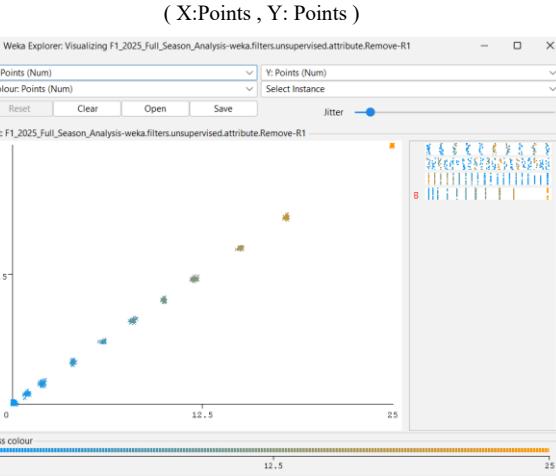
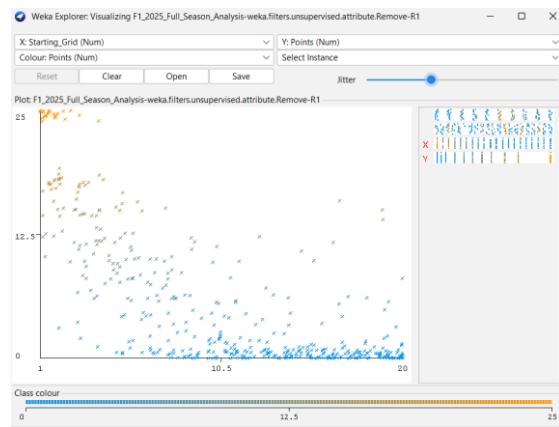
Status OK Log x0

6. Verilerin değerlendirilmesi ve Analiz

Burda en güvenilir iki modelin (RandomForest ve IBk) orijinal veri setiyle beraber grafiksel karşılaştırmaları yapılmıştır. (%66 Bölme metodu)

- Orijinal Veri setinin görselleştirmesi

(X:Starting_Grid , Y: Points)

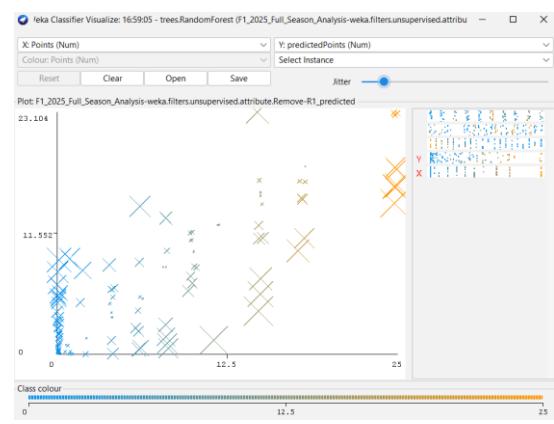


- RandomForest

(X:Starting_Grid , Y: Points)

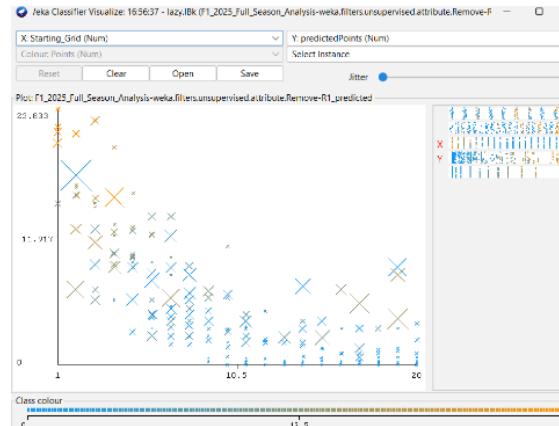


(X:Points , Y: predictedPoints “tahmin edilen puanlar”)

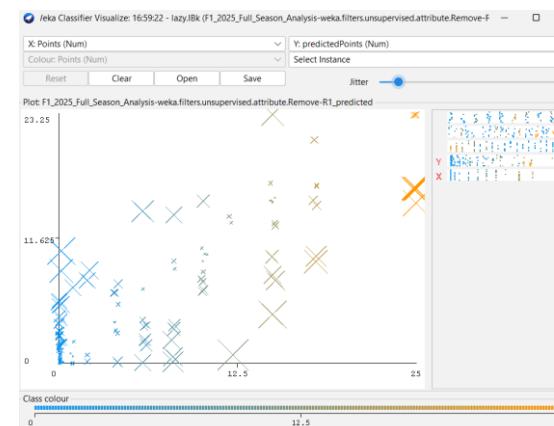


- IBk K=3

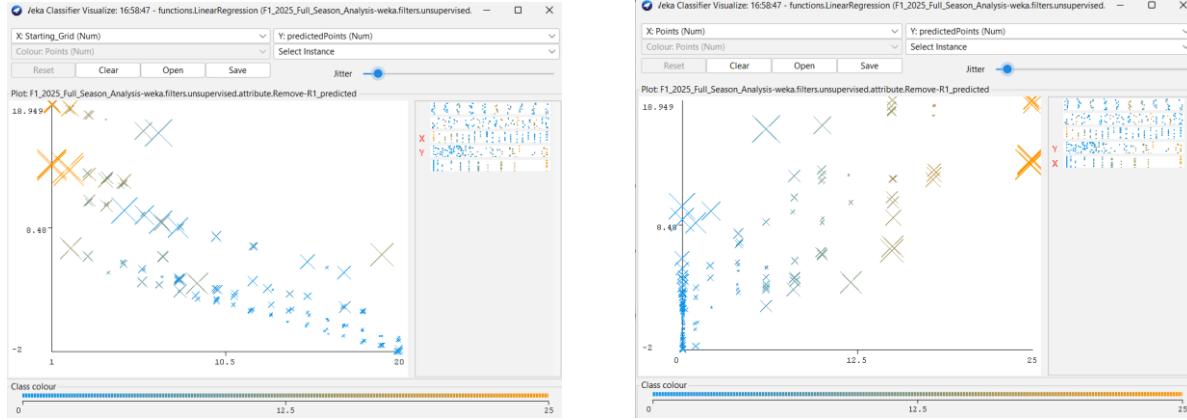
(X:Starting_Grid , Y: Points)



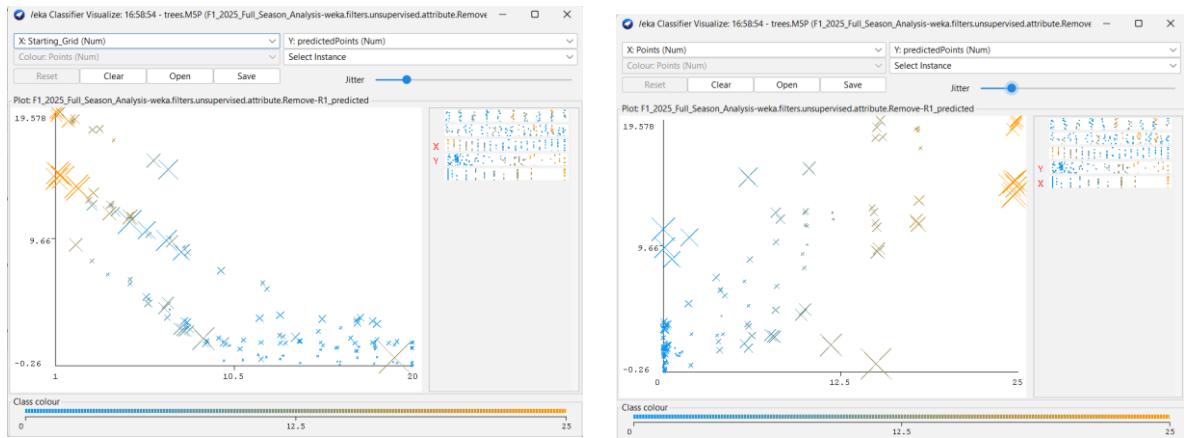
(X:Points , Y: predictedPoints “tahmin edilen puanlar”)



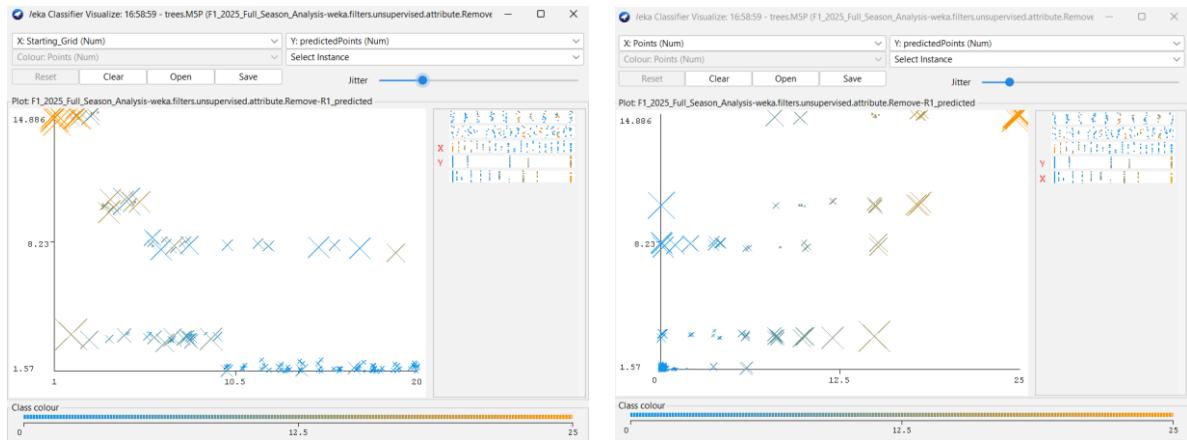
- Linear Regression



- M5P



- M5P buildRegression_true



Grafiklerden çıkarılacak gerçekler

- Örneğin, M5P model ağacında puan(points) **-0.26**'dan başlar ve bu da F1'de imkansızdır.
- Formula 1'de puan skalarası 0 ile 25 arasındadır ve süreklilik arz etmez (kesiklidir); 10. sıradan sonra herkes 0 alır. Lineer modeller bu "0'da sabitlenme" durumunu anlayamaz ve çizgiyi aşağı doğru çekmeye devam eder.

IBk ve RandomForest neden daha gerçekçi?

- IBk (KNN), tahmini yaparken geçmişteki en yakın 3 yarış sonucunun ortalamasını alır. Orijinal veride negatif puan olmadığı için, IBk'nın üreteceği sonuç her zaman **0-25 aralığında** kalmak zorundadır.
- F1'deki puan sistemi (25-18-15...) basamaklı bir yapıdadır. RandomForest, 100 farklı ağaç kullanarak bu basamaklı ve doğrusal olmayan geçişleri, düz bir çizgi çizen regresyondan çok daha iyi yakalar.
- M5P'nin tahmin tavanı yaklaşık **17 puanda** kalırken IBk ve RandomForest gerçek veri noktalarına odaklandığı için **25 puanlık** galibiyetleri çok daha isabetli yansıtır.

Referanslar:

projede kullanılan veri seti için:

<https://www.kaggle.com/datasets/selcukardaozcan/f1-2025-season-grand-prix-results-withs-points/data>

DNF ve DSQ'lerin silinmediği ham veri seti için:

<https://www.kaggle.com/datasets/selcukardaozcan/f1-2025-season-gp-point-results-dnf-dsq-got-0>

ilk baz alınan ve düzenlenen eksik veri seti:

<https://www.kaggle.com/datasets/makslypko/f1-race-result-2025>

Formula 1'de kullanılan puanlama sistemi:

https://f1insiders.com/wp-content/uploads/2022/10/f1_points.jpeg

verisetinde kullanılan yarış sonuçlarının alındığı website:

<https://www.formula1.com/en/results/2025/races>