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

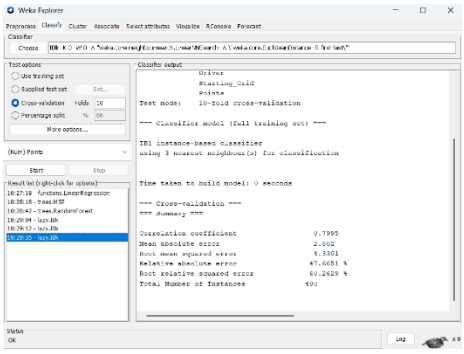
Kullandığım veri seti **"F1_Season_2025_GPs_Final_Results.arff"** adlı 2025 F1 sezonundaki 24 farklı Grand Prix yarışını ve her yarıştaki 20 sürücüyü kapsayan toplam 419 satır ve 5 sütundan oluşan güncel bir veri setidir.

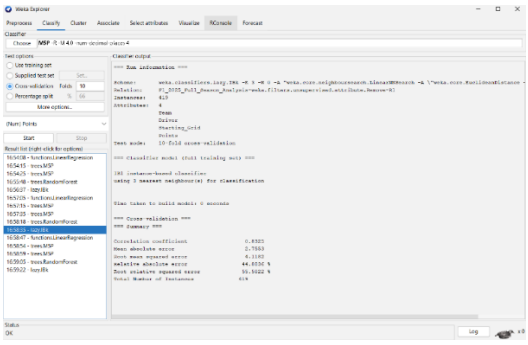
* Sezonda 10 takım ve her takımda güncel olarak 2 sürücü vardır. Sezon boyunca toplam 21 farklı sürücü olmasının nedeni sürücü(driver) "Jack Doohan" İlk 4 yarışa çıktıktan sonra yerine "Franco Colapinto" getirilmesidir. Yine güncel olarak her yarışa en fazla 20 sürücü çıkmıştır.

* 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





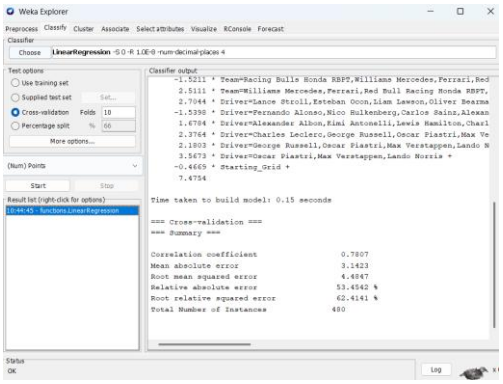
Korelasyon katsayısı: 0.7995 , 0.8325

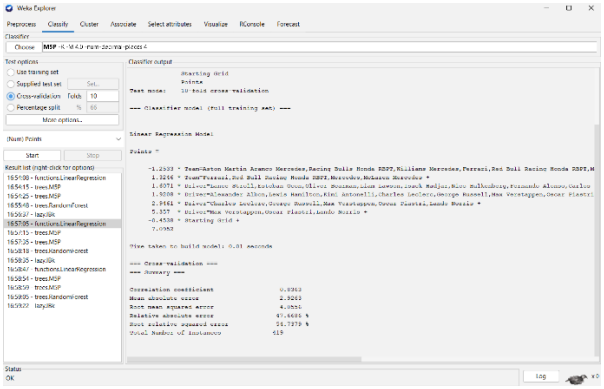
MAE : 2.802 , 2.7553

480 satır , 419 satır

Linear Regression Cross-validation-folds10

DNF-DSQ olanların silinmediği ve silindiği modellerin verimliliği





Korelasyon katsayısı: 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)

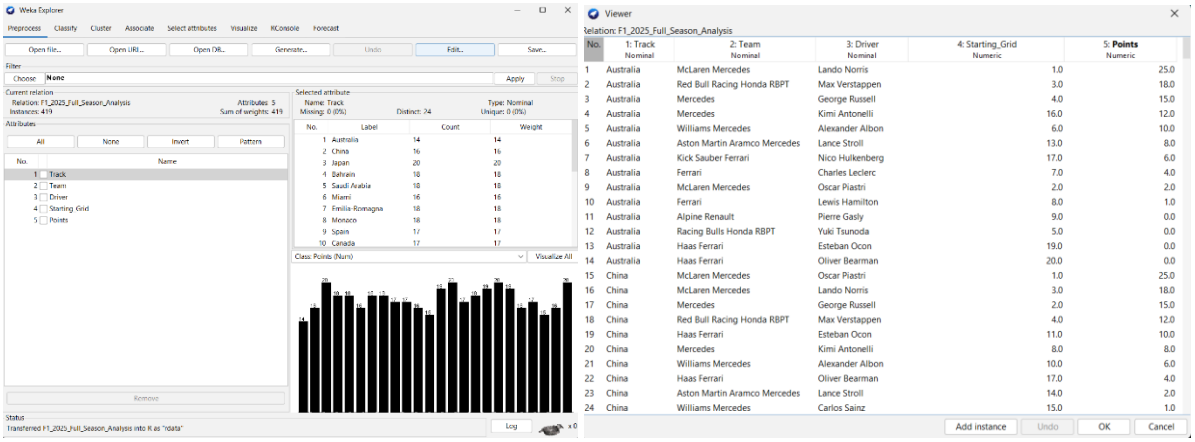
  Starting_Grid ("Yarış başlangıcı için pist üstündeki dizilim" , Numeric)

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

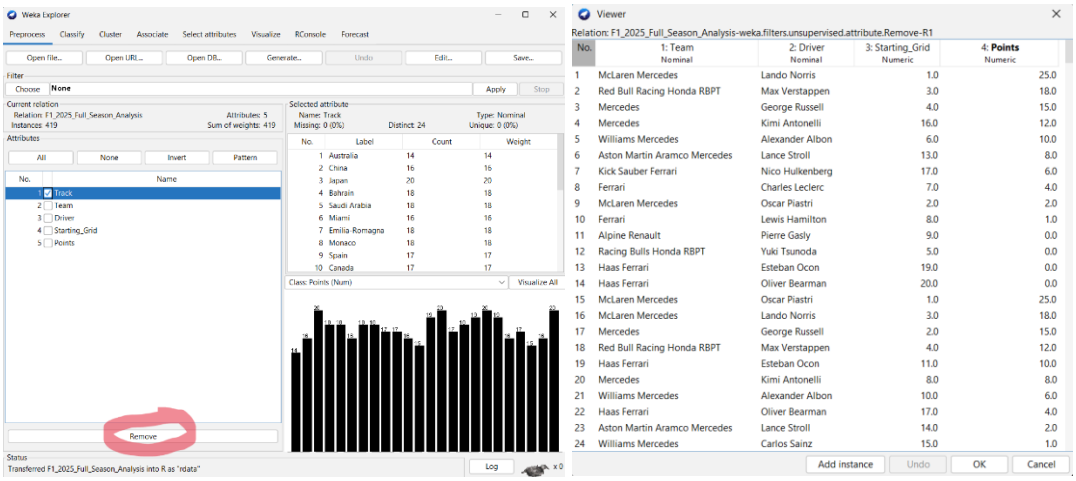
F1 de yarış sonu, bitiriş pozisyonuna bağlı verilen puan tablosu

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

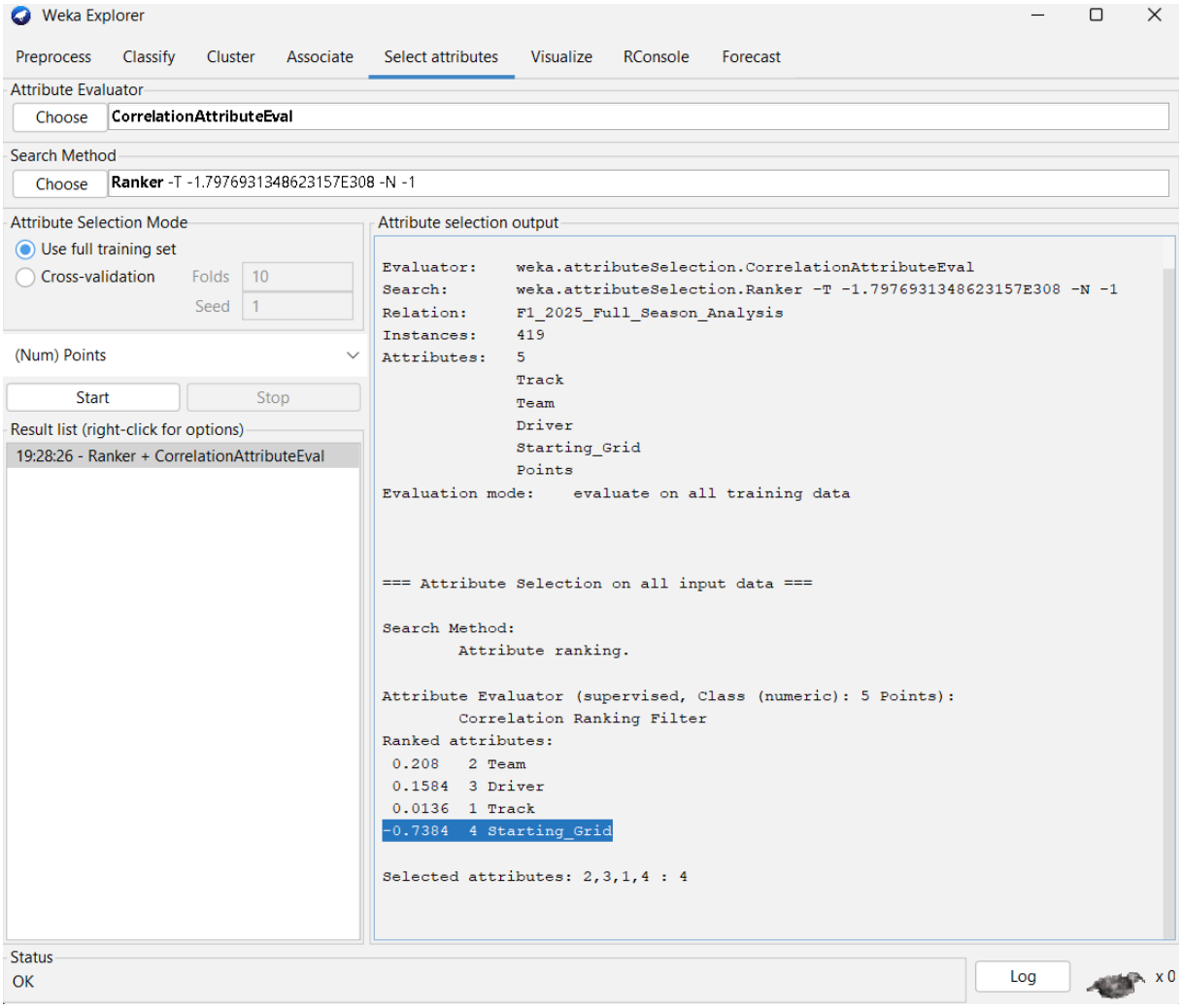


Pist(Track) kaldırıldıktan sonra WEKA’da veri setindeki örnek satırların(instances) görünümü



3. Öznitelik Seçimi (Attribute Selection)

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

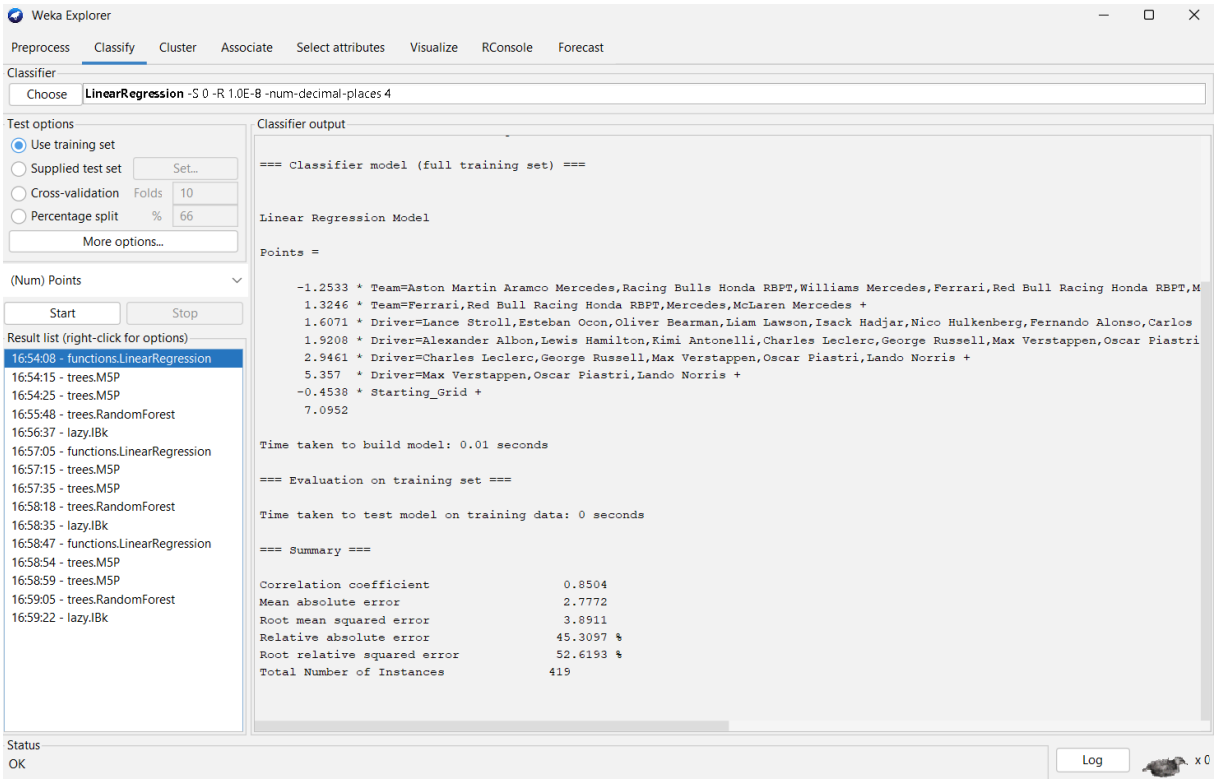
Algoritma	Test Modu	Korelasyon Katsayısı	MAE (Ort. Mutlak Hata)	RMSE (Karekök Hata)
Linear Reg.	Eğitim kümesi	**0.8504**	**2.7772**	**3.8911**
	çapraz doğrulama	**0.8363**	**2.9263**	**4.0556**
	%66 bölme	**0.8515**	**2.9947**	**4.0736**
M5P	Eğitim kümesi	**0.8771**	**2.5195**	**3.5529**
	çapraz doğrulama	**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'lık bölmede en yüksek korelasyon değeri 0.9239 , 0.8634 ile **RandomForest**'tadır.

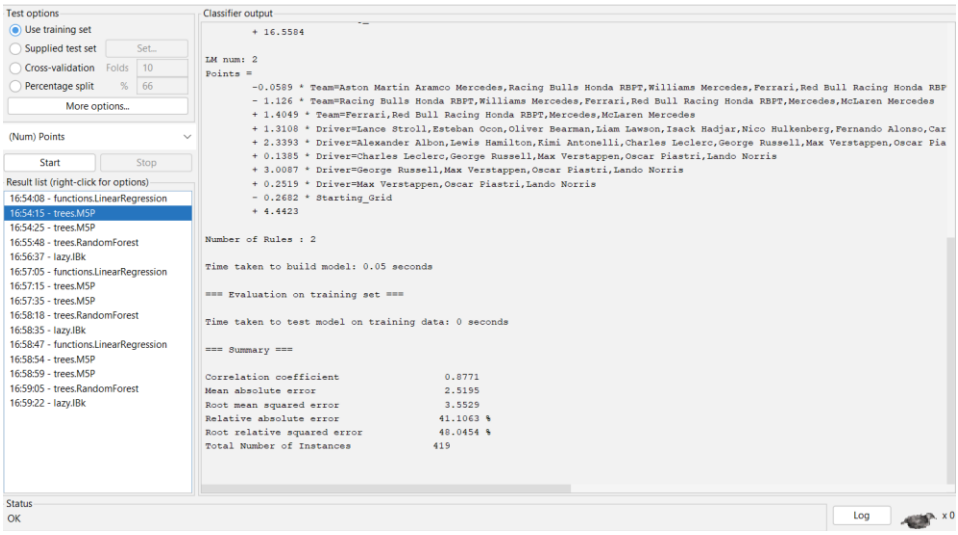
Alt tarafta tabloya konu olan sonuçların ekran görüntüleri koyulmuştur. Ve olan modeller için ağaçları görselleştirilmiştir

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

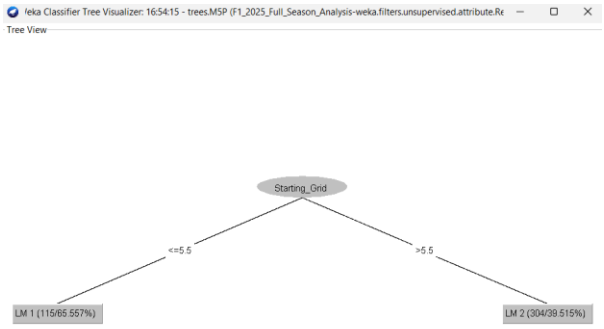
1. Linear Regresyon



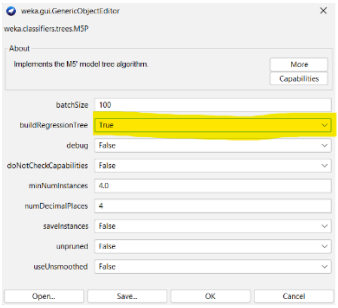
2. M5P

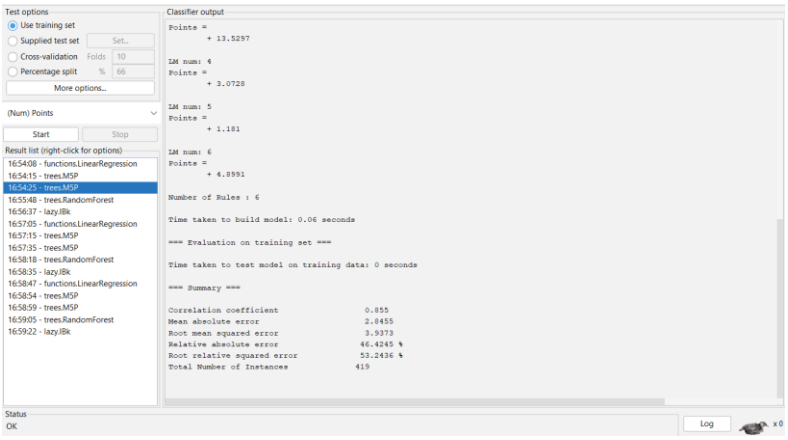


M5P ağacının görselleştirmesi

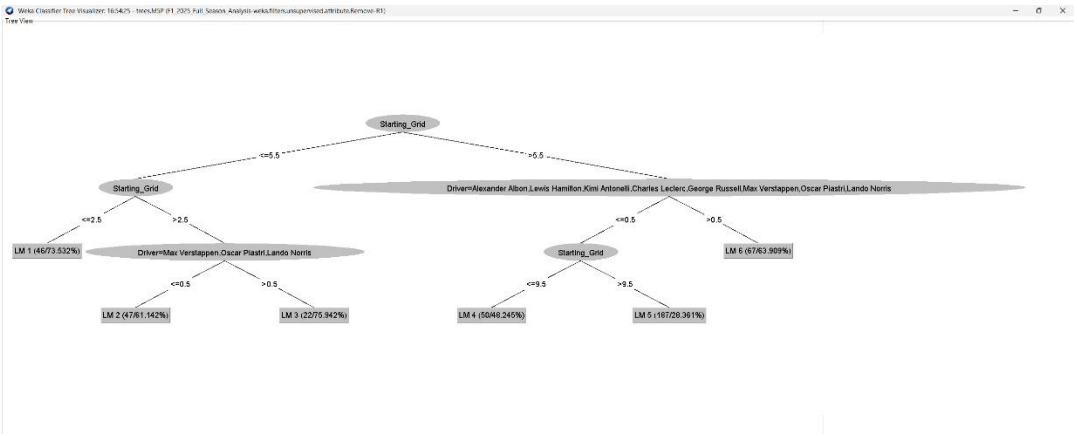


3. M5P (buildRegression True)





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) eşiğine göre iki ana gruba ayırmıştır.

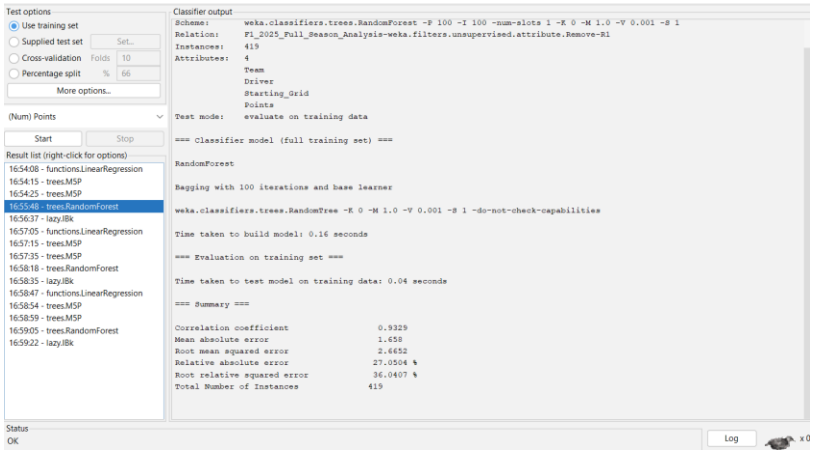
Sol Taraf: Ön Sırada Başlayanlar (Starting_Grid <= 5.5)

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

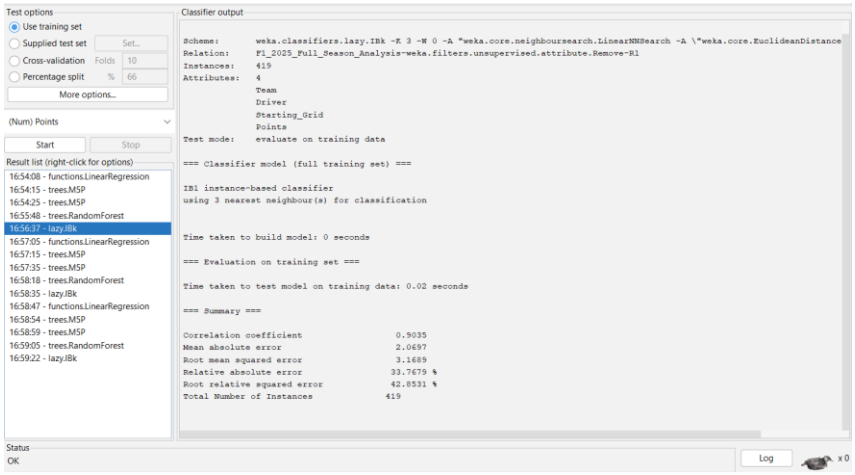
3. Sağ Taraf: Orta ve Arka Sıralar (Starting_Grid > 5.5)

- **Yetenekli Pilotların Telafisi (LM6):** 6. sıradan daha geride başlayan ancak "rekabetçi sürücü 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) öngörülürken, 10. sıradan daha geride (Starting_Grid > 9.5) başlayanlar için model en düşük puan beklentisini (LM5) sunmaktadır.

4. RandomForest

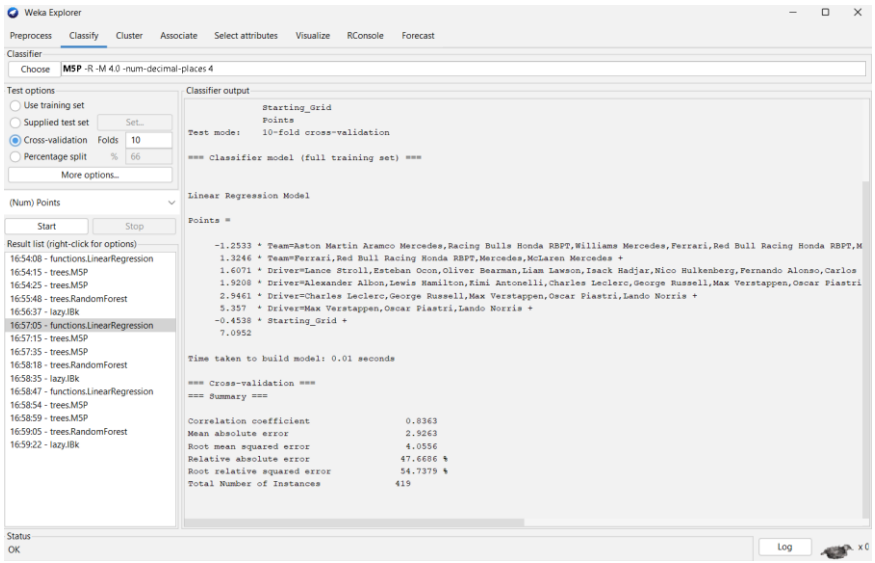


5. IBk K=3

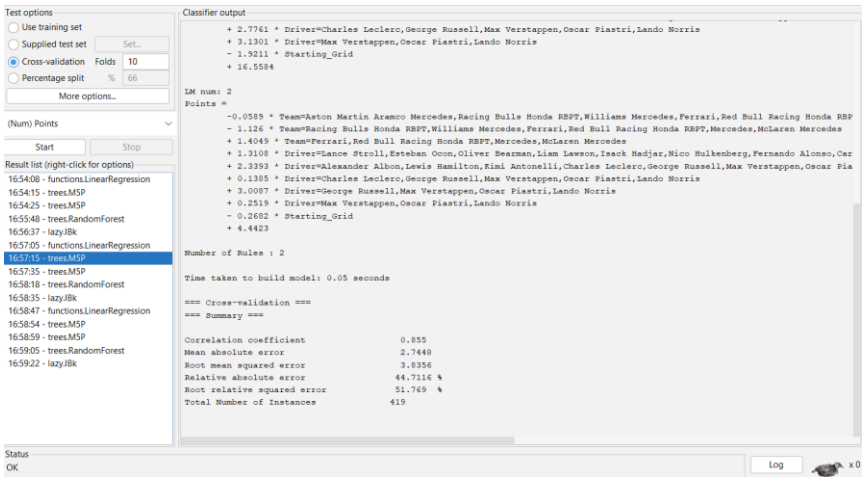


- **10 Katlamalı Çapraz Doğrulama (cross-val-fold-10)**

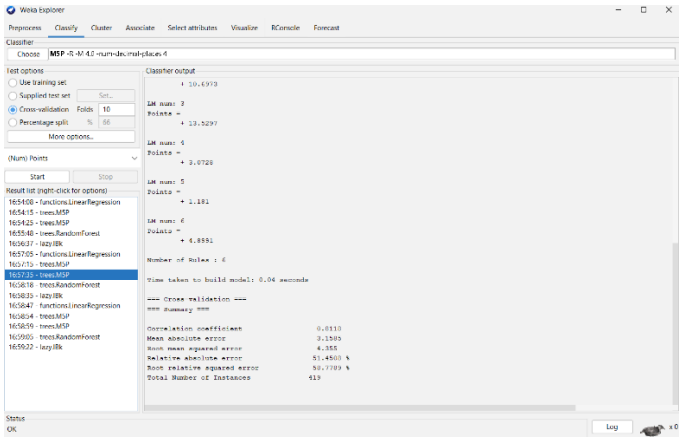
1. Linear Regresyon



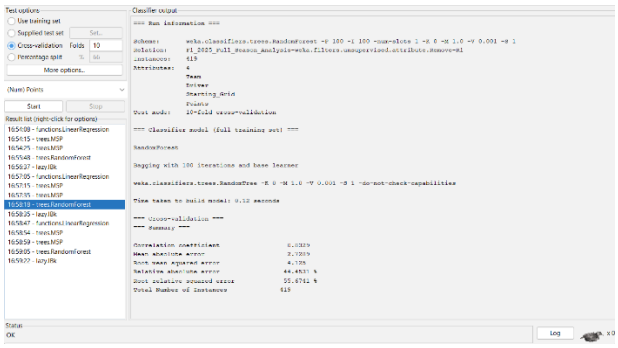
2. M5P



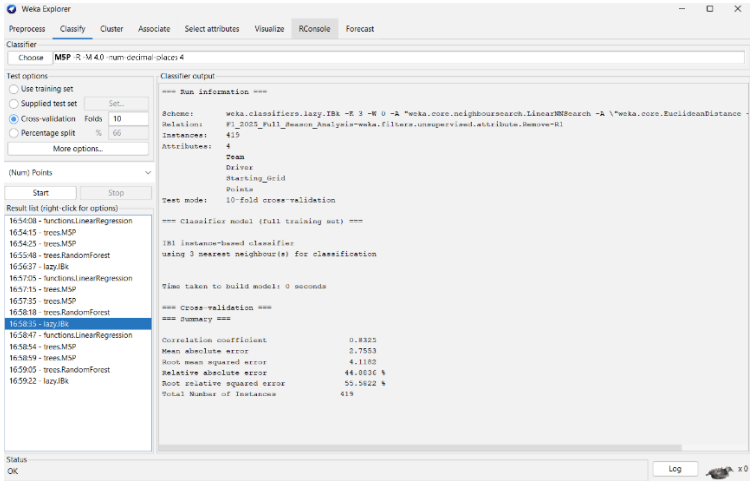
3. M5P (buildRegression True)



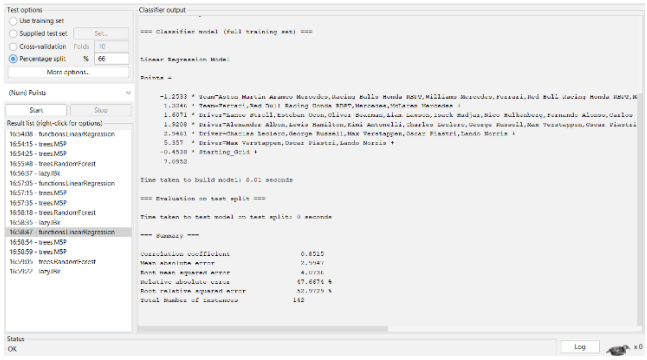
4. RandomForest



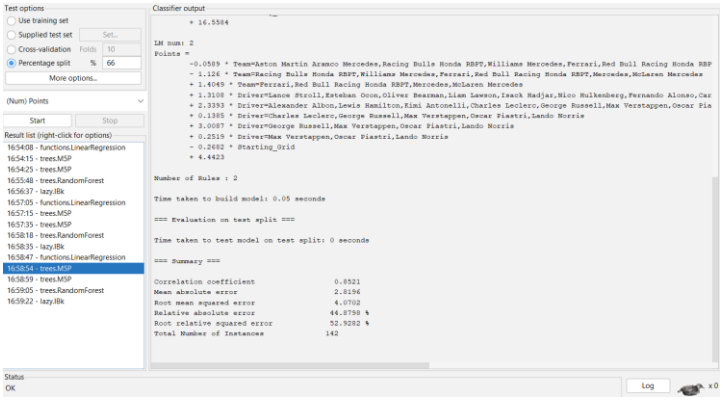
5. IBk K=3



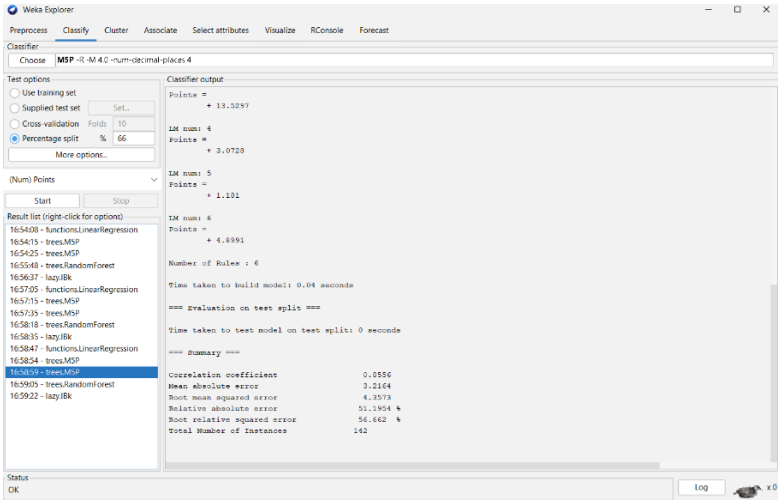
- **%66'lık bölme metoduyla percentage split 66%)**
    1. Linear Regresyon



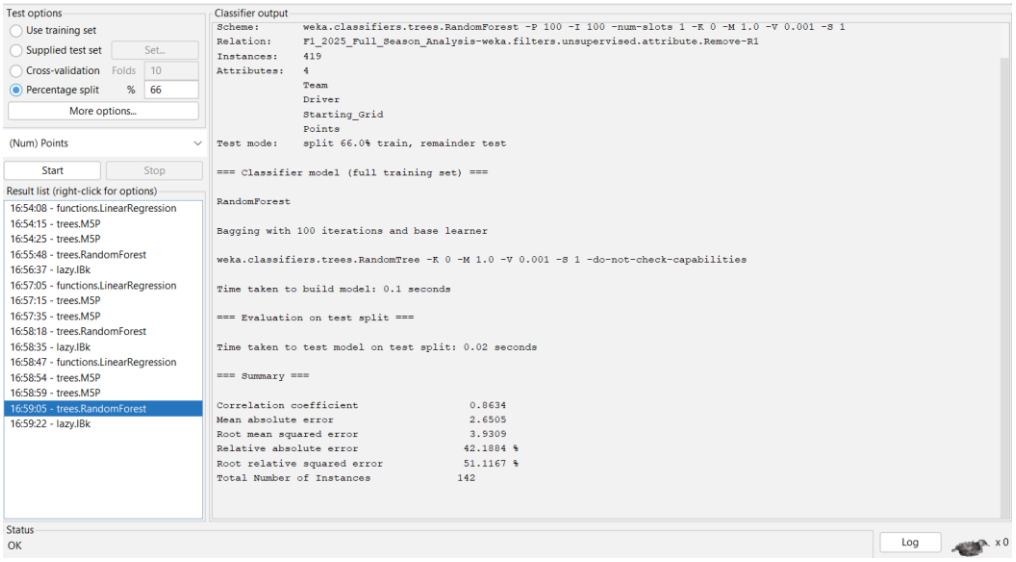
    2. M5P



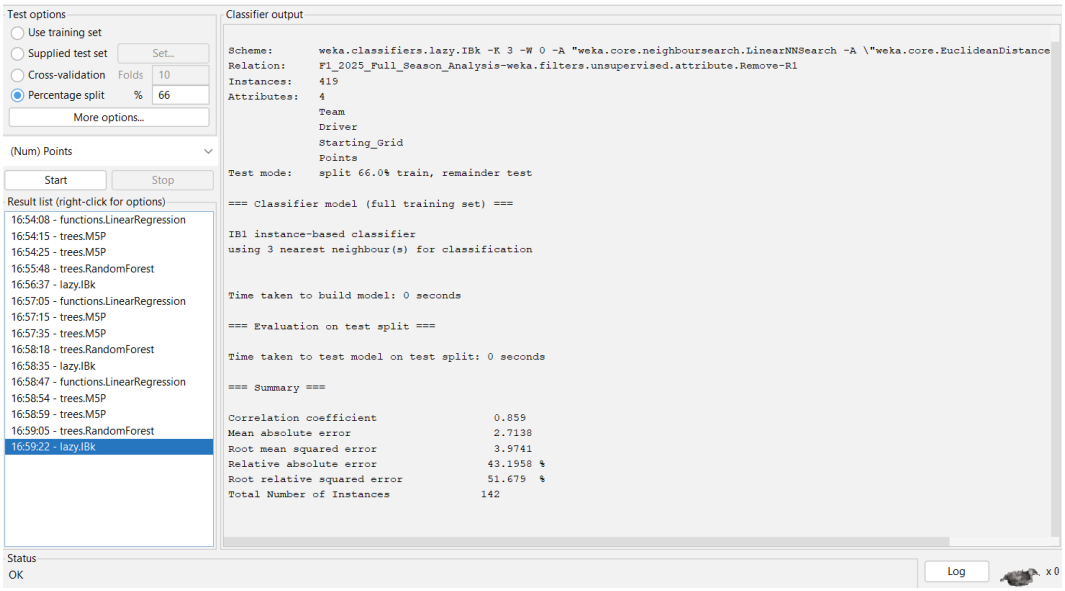
3. M5P buildRegression_true



4. RandomForest



5. IBk K=3

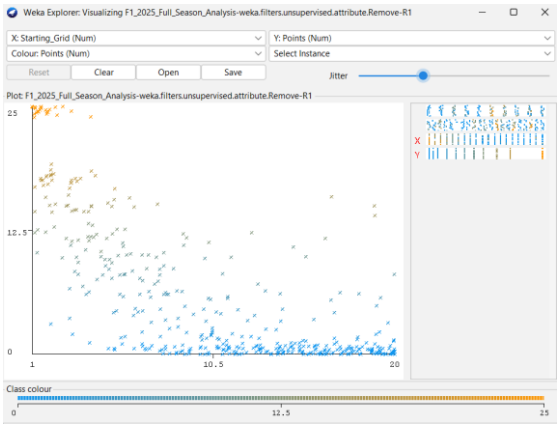

Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Num) Points
Start
Stop
Result list (right-click for options)
16:54:08 - functions.LinearRegression
16:54:15 - trees.M5P
16:54:25 - trees.M5P
16:55:48 - trees.RandomForest
16:56:37 - lazy.IBk
16:57:05 - functions.LinearRegression
16:57:15 - trees.M5P
16:57:35 - trees.M5P
16:58:18 - trees.RandomForest
16:58:35 - lazy.IBk
16:58:47 - functions.LinearRegression
16:58:54 - trees.M5P
16:58:59 - trees.M5P
16:59:05 - trees.RandomForest
16:59:22 - 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 51.679 %
Total Number of Instances 142
Status
OK
Log
x 0

6. Verilerin değerlendirilmesi ve Analiz

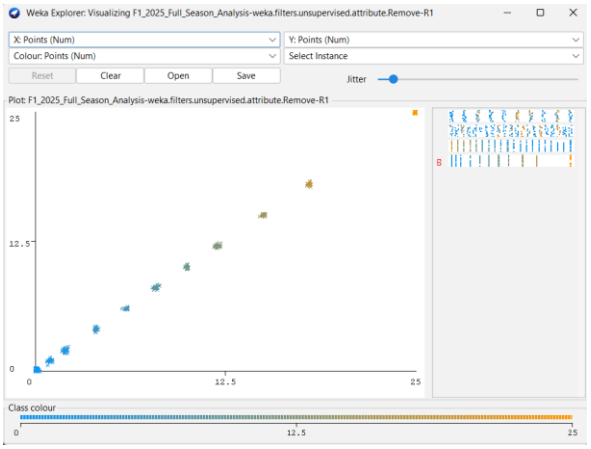
Burda en güvenilir iki modelin (RandomForest ve IBk) orijinal veri setiyle beraber grafiksel karşılaştırmaları yapılmıştır. (%66 Bölme metodu)

- Orijinal Veri setinin görselleştirmesi

(X:Starting_Grid , Y: Points)

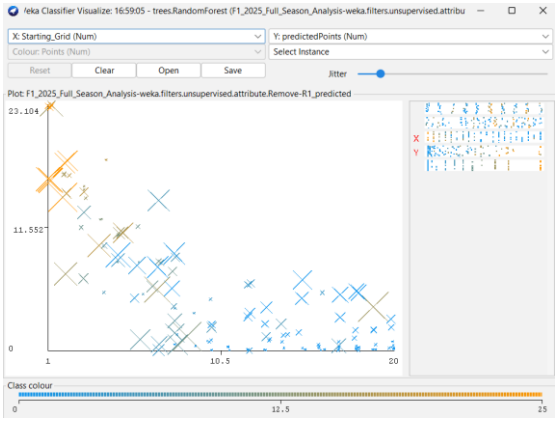


(X:Points , Y: Points)

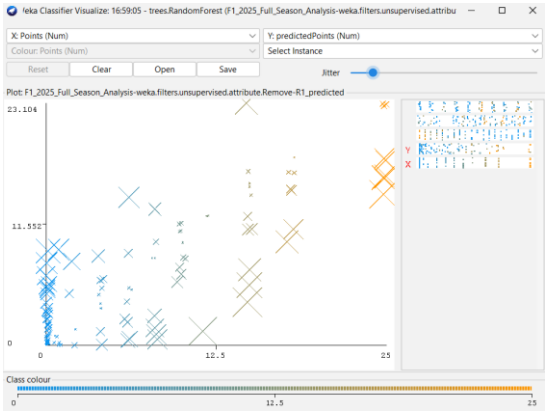


- RandomForest

(X:Starting_Grid , Y: Points)

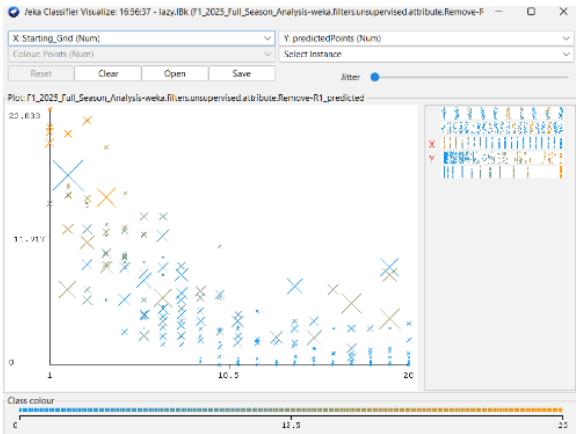


(X:Points , Y: predictedPoints “tahmin edilen puanlar”)

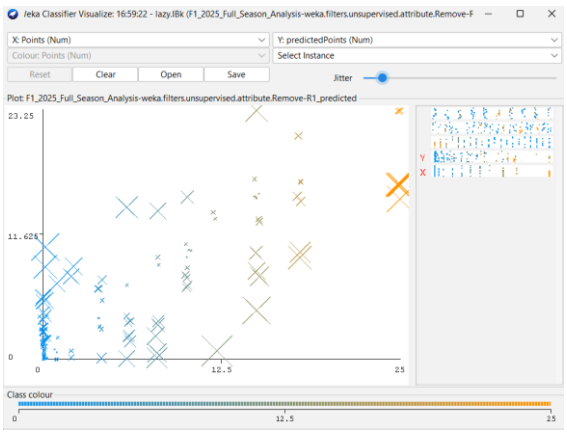


- IBk K=3

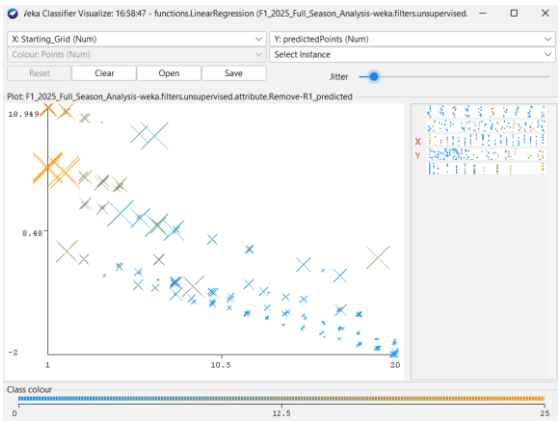
(X:Starting_Grid , Y: Points)

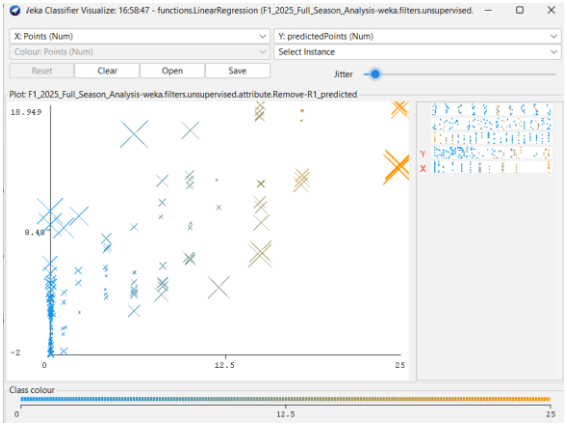


(X:Points , Y: predictedPoints “tahmin edilen puanlar”)

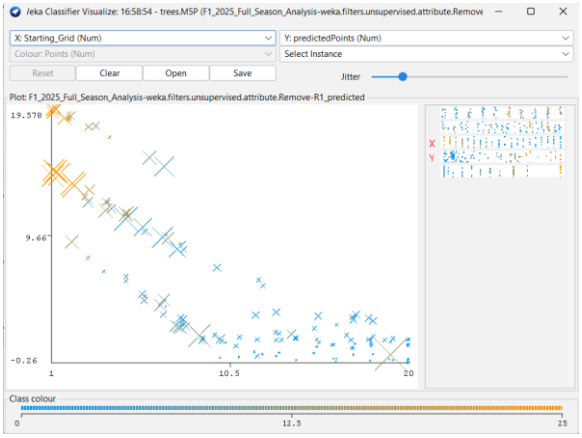


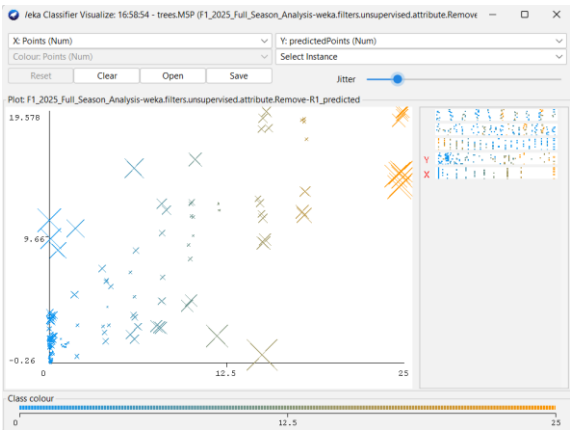
- Linear Regression



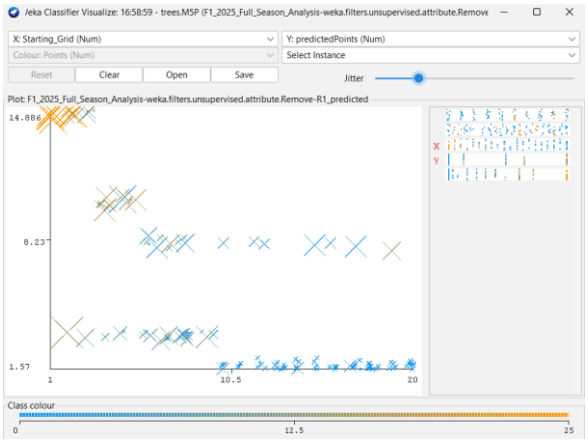


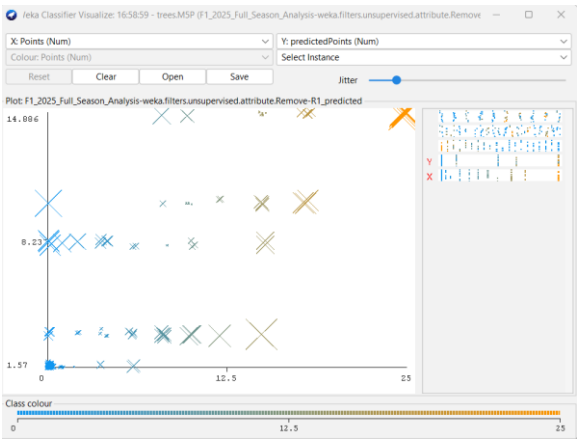
- M5P





- 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 skalası 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