LOTUS AI

YAPAY ZEKA VE BİLİŞİM ÇÖZÜMLERİ A.Ş.

DİAMONDS REGRESYON RAPORU

HAZIRLAYANIN;

ADI: SENANUR
SOYADI: BAYRAM

04/12/2024

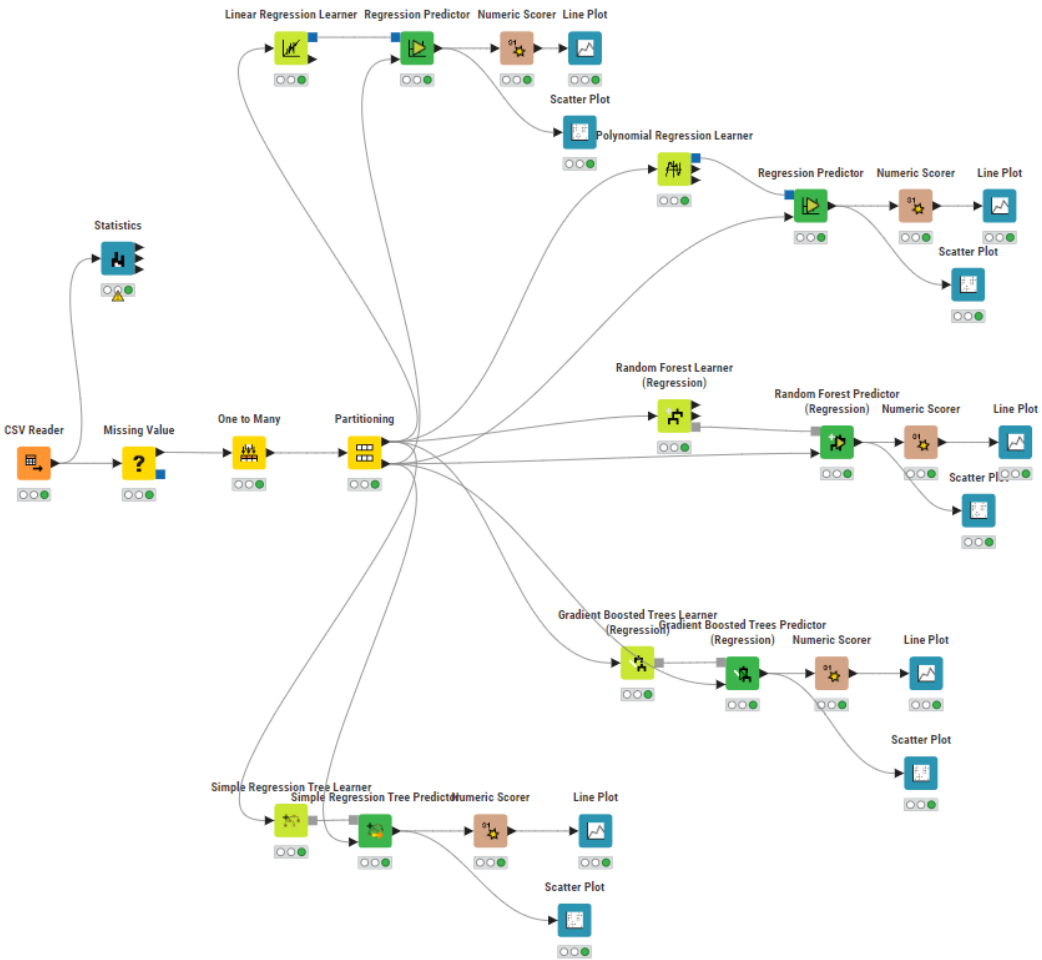
Giriş

Diamonds veri seti, değerli taşların fiyatlarını tahmin etmek amacıyla çeşitli fiziksel ve estetik özellikleri içeren bir koleksiyondur. Bu veri setinde yer alan başlıca özellikler arasında taşın karat ağırlığı, kesim kalitesi, renk, temizlik, derinlik oranı ve masa boyutu gibi faktörler bulunmaktadır. Bu unsurlar, taşların genel değerini belirleyen önemli değişkenler olup, fiyatla olan ilişkilerini anlamak için analiz edilmesi gerekmektedir. Örneğin, taşın karat ağırlığı, boyutu ve kesim kalitesi genellikle fiyat üzerinde doğrudan etkisi olan faktörlerdir. Aynı şekilde, renk ve temizlik gibi kategorik değişkenler de taşın estetik değeri hakkında bilgi verir ve bu da fiyat üzerinde belirleyici bir rol oynar.

Veri setinde her bir satır, bir değerli taşın özelliklerini temsil ederken, "price" sütunu taşın satış fiyatını belirtmektedir. Fiyat tahminleri yapılırken, bu hedef değişkeni, diğer bağımsız değişkenler ile olan ilişkisini inceleyerek tahmin edilebilir. Veri seti üzerinde yapılacak regresyon analizi, taşların özelliklerini göz önünde bulundurarak, fiyat tahminleri yapmak amacıyla güçlü bir model geliştirilmesini sağlar.

Bu proje kapsamında, **KNIME** ve **Python** araçları kullanılarak regresyon analizi yapılabilir. Her iki araç da, veri seti üzerinde regresyon analizi yaparak taşların fiyatlarını tahmin etmeye yönelik güçlü modeller geliştirilmesine olanak tanır. KNIME'ın görsel arayüzü ve Python'un esnekliği, kullanıcıya farklı analiz seçenekleri sunar ve her iki platformda yapılan çalışmaların karşılaştırılmasını sağlar. Bu şekilde, değerli taşların fiyatlarını tahmin etmeye yönelik etkili bir model elde edilmesi amaçlanır.

1. KNIME



Bu akış, veri seti üzerinde farklı regresyon yöntemlerini karşılaştırmak ve analiz etmek amacıyla oluşturulmuş bir iş akışıdır. İlk adımda veri seti okunur, eksik değerler doldurulur ve kategorik değişkenler "One to Many" yöntemiyle sayısal formatta dönüştürülür. Daha sonra veri seti eğitim ve test verisi olarak ikiye ayrılır ve çeşitli regresyon algoritmaları kullanılarak tahmin modelleri oluşturulur. Modellerin performansı ise çeşitli görseller ve metriklerle değerlendirilir.

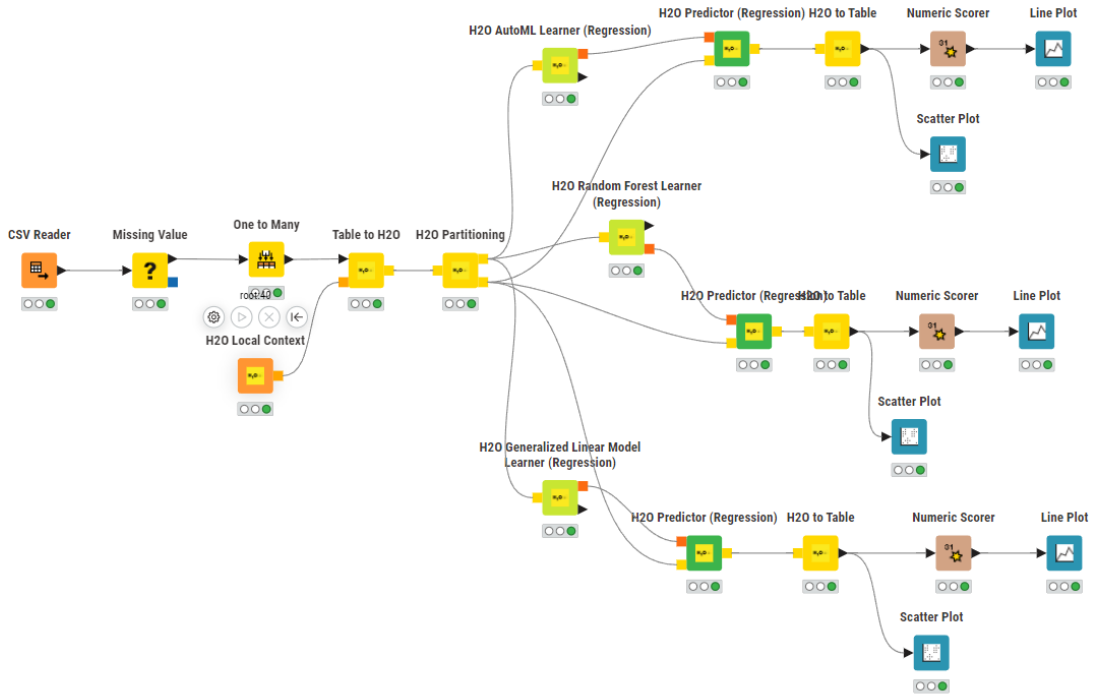
Düğümler ve İşlevleri:

1. **CSV Reader**: Veri setini okur.
2. **Missing Value**: Eksik verileri temizler veya doldurur.
3. **One to Many**: Kategorik değişkenleri birden fazla sütuna ayırarak sayısal formata çevirir.
4. **Partitioning**: Veriyi eğitim ve test olmak üzere ikiye böler.
5. **Linear Regression Learner**: Doğrusal regresyon modeli oluşturur.
6. **Regression Predictor**: Doğrusal regresyon modelini kullanarak test verisi üzerinde tahmin yapar.
7. **Numeric Scorer**: Modelin performansını değerlendirir.
8. **Line Plot ve Scatter Plot**: Tahmin sonuçlarını görselleştirir.
9. **Polynomial Regression Learner**: Polinom regresyon modeli oluşturur.
10. **Random Forest Learner**: Rastgele orman algoritması ile model oluşturur.
11. **Random Forest Predictor**: Test verisi üzerinde rastgele orman modeli ile tahmin yapar.
12. **Gradient Boosted Trees Learner**: Gradyan destekli ağaç yöntemiyle model oluşturur.
13. **Gradient Boosted Trees Predictor**: Test verisi üzerinde gradyan destekli model ile tahmin yapar.
14. **Simple Regression Tree Learner**: Basit regresyon ağaçları yöntemiyle model oluşturur.
15. **Simple Regression Tree Predictor**: Test verisi üzerinde basit regresyon ağaçları modeli ile tahmin yapar.

Sonuçta, bu iş akışı her bir modelin tahmin doğruluğunu karşılaştırarak en iyi performans gösteren algoritmayı belirlemeyi hedefler. Metrik sonuçları ise şu şekildedir:

	Lineer Regresyon	Polinomal Regresyon	Random Forest Regresyon	Gradient Boosted Trees R.	Simple Regresyon Tree
R^2	0.858565	0.868078	0.882583	0.892352	0.776122
Mean Absolute Error (MAE)	874.674960	833.990099	760.673849	726.074912	1028.033180
Mean Squared Error (MSE)	2185026.379786	2038048.513715	1813965.062655	1663049.508932	3458685.232300
Root Mean Squared Error (RMSE)	1478.183472	1427.602365	1346.835202	1289.592768	1859.754078
Mean Signed Difference (MSD)	12.980166	11.028594	22.799118	-57.226156	29.452529
Mean Absolute Percentage Error	0.279838	0.240301	0.197847	0.188768	0.265969
Adjusted R^2	0.858565	0.868078	0.882583	0.892352	0.776122

Optimize edilmiş akış ise şu şekildedir:



Bu iş akışı, H2O platformu kullanılarak regresyon problemlerini çözmek için tasarlanmıştır. Amaç, farklı H2O modellerini (AutoML, Random Forest ve Generalized Linear Model) kullanarak veri seti üzerinde tahmin modelleri oluşturmak, bu modellerin performansını karşılaştırmak ve sonuçları görselleştirmektir. İş akışı, veri hazırlığı adımlarını içermekte ve daha sonra çeşitli H2O regresyon algoritmalarını değerlendirmektedir.

Düğümler ve İşlevleri:

1. **CSV Reader**: Veri setini dosyadan okur.
2. **Missing Value**: Eksik verileri temizler veya doldurur.
3. **One to Many**: Kategorik değişkenleri birden fazla sütuna ayırarak sayısal formata çevirir.
4. **Table to H2O**: Veri setini H2O formatına dönüştürür.
5. **H2O Local Context**: H2O ortamını başlatır ve işlemleri yürütmek için gerekli bağlantıyı kurar.
6. **H2O Partitioning**: Veri setini eğitim ve test olarak ikiye böler.
7. **H2O AutoML Learner (Regression)**: H2O AutoML algoritması kullanarak en iyi regresyon modelini otomatik olarak seçer ve eğitir.
8. **H2O Random Forest Learner (Regression)**: Rastgele orman regresyon modeli oluşturur.
9. **H2O Generalized Linear Model Learner (Regression)**: Genel doğrusal model (GLM) ile regresyon analizi yapar.
10. **H2O Predictor (Regression)**: Eğitilen modelleri kullanarak test veri seti üzerinde tahminler yapar.
11. **H2O to Table**: Tahmin sonuçlarını H2O formatından tabloya dönüştürür.
12. **Numeric Scorer**: Model performansını ölçmek için değerlendirme metrikleri (R^2, MAE vb.) hesaplar.
13. **Line Plot ve Scatter Plot**: Tahmin sonuçlarını görselleştirir.

Metrik sonuçları ise şu şekildedir:

	Auto Ml	H2O Random Forest	H2O Generalized Lineer Model
R^2	0.8867431683312058	0.884285216750212	0.613995260999914
Mean Absolute Error (MAE)	764.0430701218988	772.4000005742769	1730.2613091641751
Mean Squared Error (MSE)	1802694.7782438241	1841817.6851355727	6143988.994922191
Root Mean Squared Error (RMSE)	1342.6446954588635	1357.1358388663873	2478.7071216507593
Mean Signed Difference (MSD)	1.5777738466871307	1.718003999213312	12.405111183405614
Mean Absolute Percentage Error	0.1959089139739133	0.19854614502217144	0.8597942243274321
Adjusted R^2	0.8867431683312058	0.884285216750212	0.613995260999914

2. PYTHON

Bu proje, elmasların fiyatını tahmin etmek amacıyla farklı makine öğrenimi modellerinin performansını değerlendiren bir çalışmadır. Elmaslar, farklı fiziksel ve kimyasal özelliklere sahip değerli taşlar olarak piyasada büyük bir ekonomik öneme sahiptir. Bu özelliklerin (örneğin, kesim kalitesi, renk, berraklık ve karat ağırlığı) fiyat üzerindeki etkisinin doğru bir şekilde analiz edilmesi, fiyatlandırma stratejilerinin iyileştirilmesine katkı sağlayabilir. Bu bağlamda, bu çalışmanın temel amacı, çeşitli regresyon modellerini kullanarak elmas fiyatlarının tahmin edilmesi ve bu modellerin performanslarının karşılaştırılmasıdır.

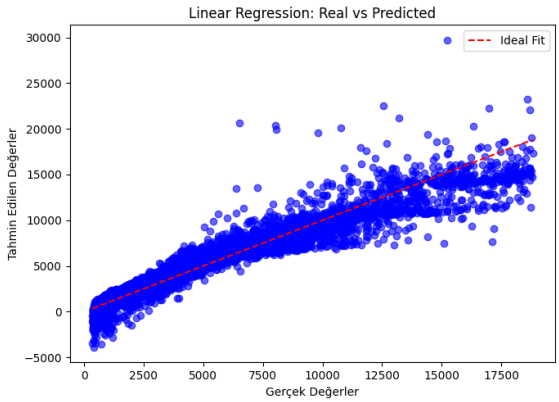
Proje, temel olarak şu aşamalardan oluşmaktadır: İlk aşamada, veri seti hazırlanmış ve fiyat tahmini için gerekli olan özellikler ayrıştırılmıştır. Kategorik değişkenler sayısal verilere dönüştürülerek modellerin analiz yapabilmesi için uygun hale getirilmiştir. Veri, eğitim ve test setlerine ayrılarak modellerin genelleştirme performansı değerlendirilmiştir.

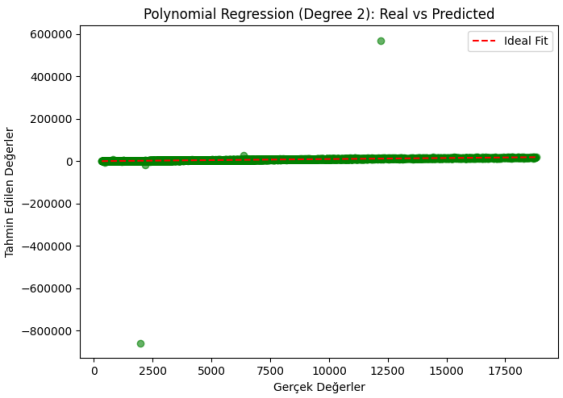
Farklı regresyon teknikleri bu çalışmada test edilmiştir. Lineer regresyon, basit doğrusal ilişkileri modellemek için kullanılmıştır. Polinomsal regresyon, özellikler ile fiyat arasındaki doğrusal olmayan ilişkileri ele almak için tercih edilmiştir. Bunun yanı sıra, random forest regresyonu ve karar ağacı regresyonu gibi daha karmaşık modeller, veri setinin özelliklerinden en iyi şekilde yararlanarak fiyat tahmini yapmak için uygulanmıştır. Random Forest ve Karar Ağacı modellerinin hiperparametre optimizasyonu, tahmin doğruluğunu artırmak için gerçekleştirilmiştir.

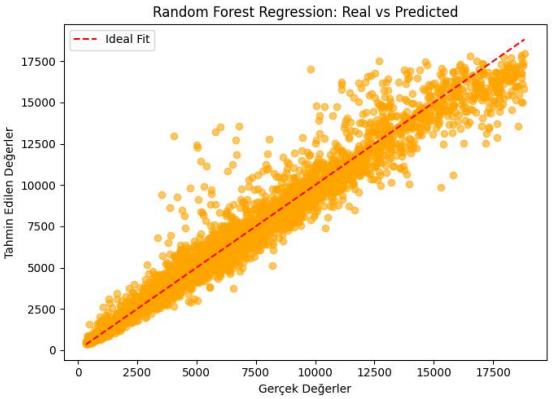
Projenin bir diğer önemli aşaması, modellerin performans değerlendirmesidir. Ortalama Mutlak Hata (MAE) ve R^2 skoru gibi metrikler kullanılarak her modelin doğruluğu ve genel açıklayıcılığı ölçülmüştür. Son olarak, görselleştirme teknikleri aracılığıyla modellerin tahminleri ve gerçek fiyat değerleri arasındaki ilişki analiz edilmiştir. Scatter plotlar kullanılarak, tahminlerin gerçek değerlere ne kadar yaklaştığı görsel olarak incelenmiştir.

Bu çalışmanın sonucunda, elmas fiyat tahmininde en iyi performans gösteren modelin belirlenmesi ve fiyatlandırma süreçlerinin iyileştirilmesi adına faydalı bilgiler elde edilmiştir. Model karşılaştırmaları ve analizler, hem akademik araştırmalar hem de pratik uygulamalar için önemli bir referans oluşturabilir.

Çıktılar şu şekildedir:







Model	Mean Absolute Error (MAE)	R^2 Score
Linear Regression	718.92	0.923
Polynomial Regression (Degree 2)	534.37	-5.81
Random Forest Regression	319.50	0.972
Decision Tree Regression	390.82	0.960

- **En iyi model:** Random Forest regresyonu, **R^2 skoru** ve **MAE** metrikleri açısından en iyi performansı gösteriyor. Yüksek doğruluk ($R^2 = 0.972$) ve düşük hata payı (MAE = 319.50) ile en başarılı model olarak öne çıkmaktadır.
- **En kötü model:** Polinomsal regresyon, R^2 skoru negatif olduğu için en kötü performansı sergileyen modeldir. Bu model, veri setine uyum sağlayamamıştır ve doğrusal regresyon modelinden daha kötü sonuçlar elde edilmiştir.
- **Diğer modeller:** Lineer regresyon ve karar ağacı regresyonu da iyi sonuçlar sunuyor, ancak Random Forest'ın gerisinde kalıyorlar. Lineer regresyon özellikle hızlı bir model olmasına rağmen, karar ağacı daha iyi sonuçlar verirken, polinomsal regresyon fazla karmaşık olmasına rağmen başarısız olmuştur.

Sonuç olarak, **Random Forest regresyonu** bu veriler üzerinde en etkili model olarak seçilebilir.