Regresyon Analiziyle CPU Özelliklerinden CPU Performans Skorunu Tahmin Etme

Predicting CPU Performance Score from CPU Features with Regression Analysis

İzzet Emre Şen[1], Güney Kaya[2] and Osman Altay[3]

[1,2,3] *Manisa Celal Bayar Üniversitesi, HFTTF Yazılım Mühendisliği, Manisa, Turkey*

Özet— Bu araştırmanın amacı, regresyon analizi kullanarak bir CPU'nun performans skorunu CPU'nun özelliklerinden yola çıkarak tahmin etmektir. CPU performansı, sistem yapılandırması ve tasarımının yanı sıra bilgisayarınızı seçerken değerlendirmek için inanılmaz derecede önemlidir. Saat hızı ve çekirdek sayısı gibi çeşitli CPU özellikleri hakkında toplanan verileri bir makine öğrenmesi modeli eğitmek için kullanacağız. Daha sonra, CPU özelliklerine dayalı olarak belirli bir CPU'nun performansı hakkında tahminler yapmak için bu modelden yararlanacağız. Modelimizin performans skorlarını doğru bir şekilde tahmin edebilmesini ve CPU özellikleri ile performans arasındaki ilişki hakkında faydalı bilgiler sağlamasını bekliyoruz. Bu araştırma, gelecekteki CPU'ların tasarımına ve optimizasyonuna yardımcı olma potansiyeline sahiptir.

Anahtar Kelimeler—Regresyon Analizi, Makine Öğrenmesi, Veri Madenciliği, CPU, CPU Performansı

I. Giriş

Modern dünyamızda bilgisayarlar o kadar yaygın bir şekilde kullanılmaktadır ki, bir bilgisayar sisteminin veya daha spesifik olarak bir bilgisayarın merkezi işlem biriminin (CPU) - performansının değerlendirilmesi, çeşitli kararlarda inanılmaz derecede önemlidir. Bir bilgisayarın CPU'su çeşitli işlevleri yerine getirir: hesaplamalar, mantıksal kararlar, bilgisayarın belleğindeki bir yerden başka bir yere veri taşıma ve masaüstünüzde çalışan uygulamalar arasında geçiş yapma gibi çoklu görevler. Bu nedenle, CPU performansı yalnızca bilgisayar seçimi için değil, aynı zamanda bilgisayar sistemi yapılandırması ve sistem tasarımı için de inanılmaz derecede önemlidir. CPU performans skorunun, CPU özelliklerine göre nasıl değiştiğini anlamak için regresyon analizi kullanılabilir. Regresyon analizi, bir değişkenin başka bir değişkene göre nasıl değiştiğini anlamaya yardımcı olan bir istatistiksel yöntemdir. Bu araştırma makalesi, regresyon analizi kullanarak CPU özelliklerinden CPU performans skorunu tahmin etmeyi amaçlamaktadır. Bu amaç doğrultusunda, çeşitli CPU özelliklerini içeren bir veri kümesi kullanılacak ve regresyon modeli oluşturulacaktır. Modelin doğruluğu ve güvenilirliği ise test verileri kullanılarak değerlendirilecektir. Bu araştırma, CPU performans skorunu daha doğru bir şekilde tahmin etmeye yardımcı olacak ve böylece bilgisayar kullanıcıları için daha iyi bir alışveriş seçeneği sağlayacaktır.Kullanıcılar istedikleri performans seviyesine uygun bir CPU alarak, para ve zaman kaybını önleyebileceklerdir. Ayrıca, bu araştırma CPU üretim sürecine de fayda sağlayacak ve daha optimize CPU üretimine yardımcı olacaktır. Üretici firmalar için optimum tasarımı yapma imkanı sağlayacak, daha az hatalı ürün üretebilecek ve müşteri memnuniyetini artırabilecektir. Bu da firmaların rekabet gücünü artıracak böylece son kullanıcı ve firmalar için olumlu bir sonuç doğuracaktır.

II. Veri Toplama

Bir CPU'nun preformansını etkileyen çeşitli özellikler vardır. Bunların en başında çekirdek hızı, Saat hızı gibi özellikler gelir. Biz de bir CPU'nun performansını tahmin etmek için bu tarz verilere yer verdik.

CPU Name: İşlemcinin adı.

Date: Piyasaya sürüldüğü tarih.

Platform: İşlemcinin hangi platforma çıktığı.

Series: İşlemcinin serisi.

# of Cores: İşlemcinin çekirdek sayısı.Çekirdekler, tek bir bilgi işlem bileşenindeki (yonga veya çip) bağımsız merkezi işlemci birimi sayısını belirten bir donanım terimidir.

# of Threads: İşlemcinin iş parçacığı sayısı. İşlem Parçacığı ve işlem parçacığı işleme, tek bir CPU çekirdeğinden geçen veya bu çekirdekte işlenen basit düzenli talimatlar sırası için kullanılan yazılım terimidir.

Base Clock: İşlemci Taban Frekansı işlemci transistörlerinin açılıp kapandığı hızı tanımlar. İşlemci taban frekansı, TDP'nin tanımlandığı çalışma noktasıdır. Frekans Gigahertz (GHz) türünde veya saniye başına devir türünden hesaplanır.

Max. Boost Clock: Maksimum turbo frekansı, işlemcinin Intel® Turbo Boost Teknolojisi'ni ve varsa Intel® Thermal Velocity Boost'u kullanarak çalışabileceği maksimum tek çekirdek frekansıdır. Frekans Gigahertz (GHz) türünde veya saniye başına milyar devir türünden hesaplanır.

Litography:[1] İşlemcinin silikon üstündeki baskısının kalınlığını ifade eder.

L3 Cache: CPU Önbelleği, işlemcide hızlı belleğin bulunduğu bölgedir.

TDP: Termal Tasarım Gücü (TDP), Intel tarafından belirlenen bir yüksek karmaşıklıklı iş yükü altında işlemcinin tüm aktif çekirdeklerle Taban Frekansında dağıttığı ortalama gücü watt türünden yansıtır.

T_{jmax} : Bağlantı Sıcaklığı, işlemci çipinde izin verilen maksimum sıcaklıktır.

CPU Mark[2]: İşlemcinin benchmark testleri yapıldıktan sonra aldığı puan.

Topladığımız veri setindeki CPU'lar günümüzde en çok kullanılan iki marka olan Intel[3] ve AMD[4] markalı CPU'lardır. Bu CPU'lara ait verileri toplama süreci markaların resmi internet sitelerinden gerçekleşti. Veri setinde kullanmak için seçtiğimiz CPU'lar benchmark puanı 2749 ve 67509 arasındadır. Veri setiyle ilgili istatistiksel veriler aşağıdaki Tablo 1'deki gibidir.

III. YÖNTEMLER

a. Veri Toplama ve Ön İşleme

Bilgisayar biliminde veri madenciliği, büyük hacimli verilerdeki ilginç ve faydalı kalıpları ve ilişkileri keşfetme sürecidir. Veri madenciliği alanı, veri kümeleri olarak bilinen büyük dijital koleksiyonları analiz etmek için istatistik ve yapay zeka araçlarını (sinir ağları ve makine öğrenimi gibi) veritabanı yönetimiyle birleştirir.[5] Veri madenciliği sürecinin adımları aşağıdaki gibi sıralanabilir:

1. Veri Temizlenmesi
2. Veri Entegrasyonu
3. Veri Seçimi
4. Veri Transferi
5. Veri Madenciliği
6. Veri Değerlendirmesi
7. Veri Sunumu

b. Kullanılan Algoritma ve Veri Ölçekleme

1. Lineer Regresyon

İstatistikte lineer regresyon, skaler bir yanıt ile bir veya daha fazla açıklayıcı değişken (bağımlı ve bağımsız değişkenler olarak da bilinir) arasındaki ilişkiyi modellemek için kullanılan doğrusal bir yaklaşımdır. Bir bağımsız değişkenin durumu, basit lineer regresyon olarak adlandırılır; eğer bağımsız değişkenler birden fazlaysa çoklu lineer regresyon olarak adlandırılır.[6][7](1)

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i, i = 1, ..., n \qquad (1)$$

Burada y bağımlı değişken, x ise bağımsız değişkendir. β_0 değeri sabiti ve β_n değerleri ise katsayıları ifade eder. Oluşturulan modelin tahmin ettiği değerler denklemin gösterdiği doğru üzerinde sıralanır.

2. Min Max Scaler

Veri setindeki CPU özelliklerinin ölçekleri arasında farklar olduğu için veri ölçekleme kullanılmasına karar verildi. Bunun için Min Max Scaler yöntemi uygulandı.

Min Max Scaler, bir veri kümesindeki değerleri bir aralıkta (normalleştirme) ölçeklendirme işlemini gerçekleştirir. Bu, verilerin standartlaştırılmasına benzer, ancak standartlaştırma işleminde veriler ortalama değerine göre ölçeklendirilirken, Min Max Scaler kullanılarak veriler belirlenen minimum ve maksimum değerler arasına sıkıştırılır. Bu, özellikle veri kümesinde çok sayıda uç değerler (outliers) varsa yararlı olabilir.

Min Max Scaler, veri kümesindeki her bir özelliği (özellikler arasında öğrenme algoritması tarafından kullanılan nitelikler vardır) aşağıdaki formül kullanılarak ölçeklendirir:

(x - min(x)) / (max(x) - min(x))

Burada, x veri kümesindeki bir özelliğin değeri ve min(x) ve max(x), bu özelliğin minimum ve maksimum değerleridir. Bu formül, veri kümesindeki her bir özelliğin değerlerini 0 ile 1 arasında ölçeklendirir.

IV. DENEYSEL SONUÇLAR VE TARTIŞMA

Tahmin edilen değer ve gerçek değer arasındaki farkı belirlemenin çeşitli yolları vardır. Bizim bu modelde performansı ölçmek için kullandığımız metrikler Mean Absolute Error (MAE)(2), Mean Squared Error (MSE)(3), Root Mean Squared Error (RMSE)(4) ve son olarak R-Squared (R^2) olarak sıralanabilir.

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (2)$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (3)$$

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (4)$$

Burada n frekansı, y_i i'nci gerçek değeri ve $\hat{y}_i$ ise i'nci tahmin edilen değeri ifade eder.

R-kare, bir regresyon modelinin uyum iyiliğini temsil eden istatistiksel bir ölçütdür. R-kare için ideal değer 1'dir. R-kare değeri 1'e ne kadar yakınsa, model o kadar iyi uyum sağlar.

R-kare, artıkların kareler toplamının (SSres) ortalamaya uzaklığın kareler toplamının (SStot) bir karşılaştırmasıdır.

$$SS_{res} = \sum(y_i - \hat{y}_i)^2 \qquad (5)$$

$$SS_{tot} = \sum(y_i - \hat{y}_{ort})^2 \qquad (6)$$

R^2 değeri Denklem (5) ve (6) yardımıyla aşağıdaki gibi bulunabilir.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad (7)$$

Toplanılan veri setindeki verilerin 206 adet verinin %70'i eğitim, %30'u ise sınama olarak ayırılmıştır.

TABLE 1: İSTATISTIKSEL VERILER

Date	Series	# of Cores	# of Threads	Base Clock	Max. Boost Clock	Lito-graphy	L3 Cache	TDP	Tjmax	CPU Mark	
2018,544	6,203	6,805	12,174	3,10	4,00	12,927	13,373	66,063	97,513	14743,984	Ortalama
3,110	1,891	3,757	6,928	0,58	0,75	5,692	10,649	45,928	6,813	10648,6	STD
2019	7	6	12	3,10	4,00	14	12	56,5	100	12714	Medyan
2022	7	4	8	3,00	4,00	14	8	65	100	16979	Mod
12	8	22	32	3,20	2,00	27	61	247	41	63591,08	Range
9,673	3,577	14,11821	48,008	0,33	0,56	32,399	113,410	2109,445	46,417	1,13E+08	Varyans

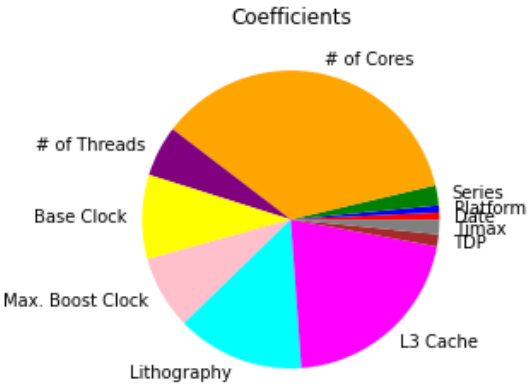

Fig. 1: Coefficients Pie Chart

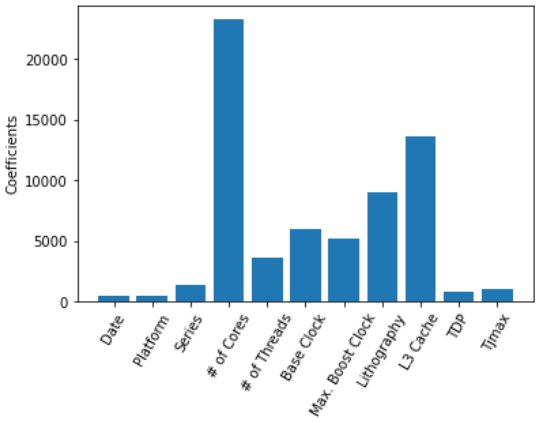

Fig. 2: Coefficients Bar Graph

Coefficient grafiklerini her bir özniteliğin sonucu ne kadar etkilediğini kavramada kullanabiliriz. Bu pasta ve bar grafiklerinde, 11 özniteliğin hedef değişkeni etkileme katsayılarını görebiliyoruz. Buna göre en çok etkileyen özellikler çekirdek sayısı ve L3 belleği, en az etkileyenler ise çıkış tarihi ve platformları olarak belirlenmiştir. Çekirdek sayısı (Number of Cores), CPU saat hızı (Base Clock),bellek boyutu (L3 Cache) ve CPU'nun inceliğini (Lithography) belirten özniteliklerin ayrıca yeni çıkan CPU'ların tanıtım ve reklamlarında en çok tekrarlanan ve öne çıkarılan özellikleri olduğunu varsayarsak, beklenildiği gibi sonuca etkilerinin de oldukça yüksek olduğunu doğrulayabiliyoruz.

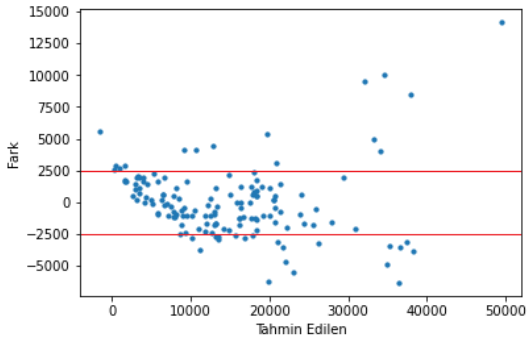

Fig. 3: Residuals / Predictions Scatter Graph

Bu dağılım grafiğinde ise Tahmin edilen değerler ile gerçek değerler arasındaki farkı görebiliyoruz. Kendi çizdiğimiz +2500 ile -2500 fark çizgilerinin arasında kalan alana baktığımızda, tahmin edilen çoğu değerin gerçek değerlerden farkının mutlak değer 2500'den fazla olmadığını anlayabiliyoruz. Bu bize modelin başarımını gösterir.

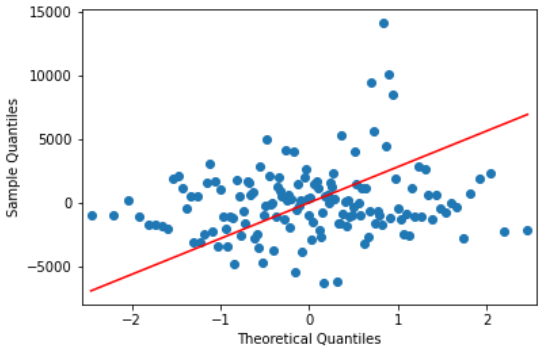

Fig. 4: Q-Q Plot

Bu Q-Q plot grafiği, verilerimizin normal dağılımın durumunu teorik niceliğiyle karşılaştırmamızı sağlar. Veriler merkez hatta (line) ne kadar yakınsa, verilerin yayılımı normal dağılıma o kadar yakın demektir. Verilen çizgi üzerindeki yakınlıklar diğer grafiklerle orantılı bir sonuç göstermektedir.

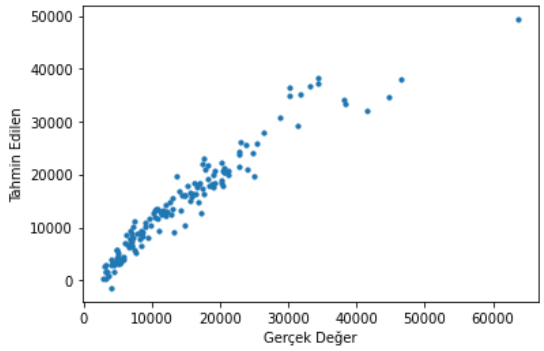

Fig. 5: Predictions / Actual Values Scatter Graph

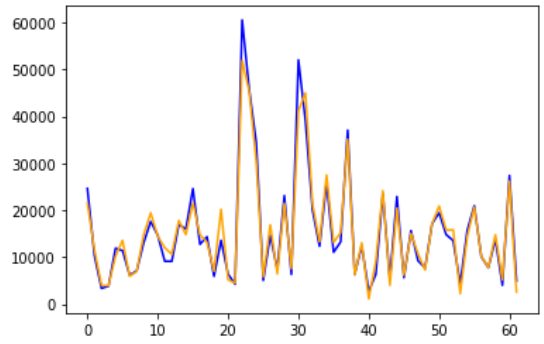

Fig. 6: Predictions / Actual Values Line Graph

Bu dağılım ve çizgi grafikleri bize tahmin edilen değerler ile gerçek değerleri grafiğini veriyor. Çizgi grafiğinde mavi çizgi gerçek değerleri gösterirken turuncu çizgi tahmin edilen değerleri gösteriyor. Burada modelin ortalama olarak 40000 üstündeki CPU Skoru tahminlerinde daha fazla yanıldığını görebiliyoruz. Bunun sebebi olarak bu işlemcilerin oldukça yeni ve güçlü -aynı zamanda da nadir- olması nedeniyle veri setinde eğitim için yeterince bulunmaması gösterilebilir.

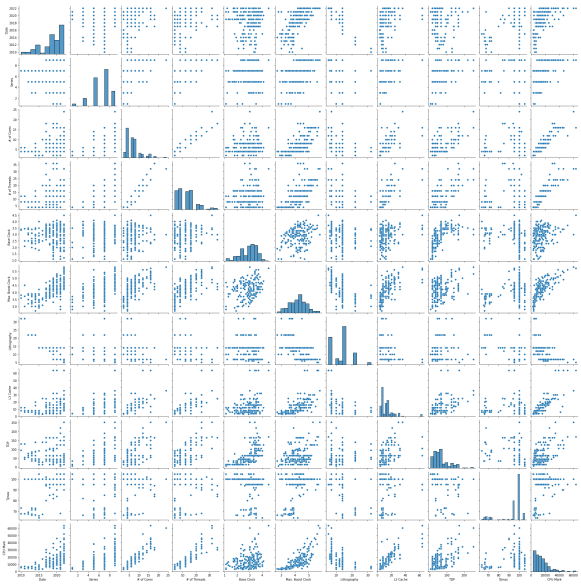

Fig. 7: Pair Plot

Bu grafik, veri kümesi içindeki öznitelikler arasındaki ikili ilişkileri gösterir.

V. Sonuç

Bu çalışmada performans puanları 2700 ile 64000 arasındaki CPU'ların puanları sayısal tahmin yöntemiyle tahmin edilmeye çalışılmıştır. Çalışmada kullanılan veri seti 12 öznitelikten oluşmaktadır. Çoklu Lineer Regresyon algoritması sayısal tahmin için kullanılmıştır. Veri seti üzerinde nominal veriyi sayısal veriye çevirmek ve 0 ile 1 arasında ölçeklendirmek dışında başka bir işlem yapılmamıştır. MAE, MSE, RMSE ve R-kare metrik değerleri hesaplanmıştır.

Çoklu Lineer Regresyon algoritması uygulanan modelimiz sonuç olarak:

- MAE: 1749.9316
- MSE: 6743509.5098
- RMSE: 2596.8268
- R^2: 0.952203

olarak hesaplamıştır.

References

[1] Compare processors and graphics cards in details. https://technical.city/en. Accessed: 2022-11-25.

[2] PassMark Software. https://www.cpubenchmark.net/CPU_mega_page.html. Accessed: 2022-11-25.

[3] Intel Proceccors for PC, Laptop, Server and AI. https://www.intel.com/content/www/us/en/products/details/processors.html. Accessed: 2022-11-25.

[4] AMD Processors | AMD. https://www.amd.com/en/processors/. Accessed: 2022-11-25.

[5] C. Clifton. "encyclopædia britannica: Definition of data mining". Tomado de https://www.britannica.com/technology/data-mining (23/10/2022). Accessed: 2022-12-17.

[6] D. A. Freedman, *Statistical Models: Theory and Practice.* Boston: Cambridge University Press, 2009, p. 26, "A simple regression equation has on the right hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression e right hand side, each with its own slope coefficient".

[7] Wikipedia. Linear regression. Tomado de https://en.wikipedia.org/wiki/Linear_regression (09/12/2022).