KSÜ Mühendislik Bilimleri Dergisi, 29(3), 2026
Araştırma Makalesi

KSU J Eng Sci, 29(3), 2026
Research Article

Kahramanmaras Sutcu Imam University
Journal of Engineering Sciences

Geliş Tarihi : 09.02.2026
Kabul Tarihi : 28.07.2026

Received Date : 09.02.2026
Accepted Date : 28.07.2026

SAĞLIK SİGORTASI PRİM TAHMİNİNDE UÇ KIRPMA DUYARLILIĞI VE SHAP TABANLI AÇIKLANABİLİRLİK

WINSORIZATION SENSITIVITY AND SHAP-BASED EXPLAINABILITY IN HEALTH INSURANCE PREMIUM PREDICTION

Tuba IRMAK[1] (ORCID: 0000-0003-4749-5226)*

[1] Atatürk Üniversitesi, Endüstri Mühendisliği Bölümü, Erzurum, Türkiye

* Sorumlu Yazar / Corresponding Author: Tuba IRMAK, tuba.adar@atauni.edu.tr

ÖZET

Sağlık sigortası primlerinin doğru tahmini, bireyler için gereksiz maliyetlerin önlenmesi ve sigorta şirketleri için sürdürülebilir kârlılığın sağlanması açısından kritik öneme sahiptir. Bu çalışmada, Kaggle'da açık erişimli "Insurance Premium Prediction" veri seti kullanılarak bireylerin yıllık sağlık sigortası prim tutarları makine öğrenmesi tabanlı regresyon yöntemleriyle tahmin edilmiştir. Doğrusal, örnek tabanlı, çekirdek tabanlı ve ağaç tabanlı topluluk yaklaşımlarından oluşan sekiz model eğitilmiş; performansları 5-katlı çapraz doğrulama ve ayrılmış test kümesi üzerinde karşılaştırılmıştır. En iyi sonuç 0,9007 R^2 değeriyle Gradient Boosting modelinden elde edilmiş, bu model için RandomizedSearchCV ile hiperparametre optimizasyonu yapılmış ve özellikle MAPE değerinde iyileşme sağlanmıştır. 'bmi' ve 'charges' değişkenlerine %1–%99 aralığında uç kırpma uygulanarak yürütülen duyarlılık analizinde, uç kırpmanın özellikle ağaç tabanlı modellerde R^2 değerlerini artırdığı ve hata metriklerini iyileştirdiği görülmüştür. SHAP analizi, prim tutarları üzerinde en etkili değişkenin sigara kullanım durumu olduğunu ortaya koymuştur. Bulguların genellenebilirliği bağımsız bir veri setinde sınanmış ve ağaç tabanlı topluluk modellerinin üstünlüğü doğrulanmıştır. Sonuçlar, uç değer işleme ile birleştirilen ağaç tabanlı topluluk modellerinin sigorta prim tahmini için güvenilir bir temel sunduğunu göstermektedir.

Anahtar Kelimeler: Sağlık sigortası prim tahmini, gradient boosting, hiperparametre optimizasyonu, uç değer analizi, SHAP analizi.

ABSTRACT

Accurate prediction of health insurance premiums is critical both for preventing unnecessary costs for individuals and for ensuring sustainable profitability for insurers. In this study, annual health insurance premiums are predicted using machine learning-based regression methods on the openly accessible "Insurance Premium Prediction" dataset shared on Kaggle. Eight models comprising linear, instance-based, kernel-based, and tree-based ensemble approaches were trained and compared through 5-fold cross-validation on the training set and on a held-out test set. The best performance was achieved by Gradient Boosting with an R^2 of 0.9007; hyperparameter optimization using RandomizedSearchCV yielded a further improvement, particularly in MAPE. A sensitivity analysis applying winsorization at the 1st–99th percentile range to the 'bmi' and 'charges' variables showed that outlier clipping increased R^2 values and improved error metrics, especially for tree-based models. SHAP analysis identified smoking status as the most influential variable on premium amounts. The generalizability of the findings was tested on an independent dataset, where tree-based ensemble models again outperformed the alternatives. These results suggest that tree-based ensembles combined with outlier treatment provide a reliable basis for insurance premium estimation.

Keywords: Health insurance premium prediction, gradient boosting, hyperparameter optimization, winsorization, SHAP analysis.

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

GİRİŞ

Belirsizlik ve tehlikelerle dolu olan günümüz dünyasında bireyler, oluşabilecek olumsuz durumlara/risklere karşı savunmasızdır. Ölüm, hastalık ve mülk kaybı gibi riskler bireylerin yaşam kalitesini ve ekonomik dengesini önemli derecede etkileyebilen risk kaynaklarıdır (Kaushik et al., 2022). Bu tür risklerin olası mali sonuçlarını azaltmak veya minimum düzeye indirmek amacıyla finansal kaynaklar kullanılarak çeşitli ürün ve hizmetler geliştirilmiştir. Bu ürünlerden biri olan sigorta, oluşabilecek çeşitli risklerin masraflarını azaltmayı veya tazmin etmeyi amaçlayan bir poliçedir (Topal, 2024).

Sağlık sigortası ise kişinin hastalık, kaza, ameliyat, tedavi, hastane yatışı gibi yapacağı sağlık harcamaları karşısında oluşabilecek mali yükü azaltmak amacıyla sigorta şirketi ile yaptığı sözleşmeye dayanan bir güvence sistemidir. Sigortalının ödeyeceği prim karşılığında ve sözleşmede tanımlanan teminatlar kapsamında sağlık giderlerinin tamamı veya bir kısmı sigorta şirketi tarafından karşılanır. Prim tutarı doğru belirlenebilirse birey açısından gereksiz yüksek maliyetler önlenmiş olur. Sigorta firması açısından ise sürdürülebilir karlılığın korunmasını sağlar. Yani prim değerinin doğru belirlenmesi kritik bir öneme sahiptir.

Prim değerlerinin hesabı tek bir değişkene bağlı değildir ve değerde dalgalanmaya sebebiyet verecek birçok faktör mevcuttur (Dutta et al., 2023). Yaş, cinsiyet, beden kitle indeksi, aile yapısı (çocuk sayısı vs.), geçmiş hastalıklar, poliçe kapsamı ve teminat düzeyi, harcama geçmişi, sigara kullanımı, yaşam tarzı seçimi gibi faktörler sağlık sigortasının maliyetini önemli ölçüde etkilemektedir (Kapse et al., 2025). Bu değişkenlerin sayıca fazla olması ve aralarındaki ilişkilerin doğrusal olmayan bir yapıya sahip olması nedeniyle, uygun bir tahmin modelinin geliştirilmesi güçleşmektedir. Bu bağlamda çok değişkenli ve karmaşık ilişkileri öğrenebilme kapasitesine sahip makine öğrenmesi tabanlı regresyon yaklaşımları, sağlık sigortası primlerinin daha doğru ve tutarlı biçimde tahmin edilmesi için etkili bir çözüm alternatifi sunmaktadır (Kulkarni et al., 2022). Prim değerinin otomatik olarak belirlenmesi başvuracak kişinin risk profilinin hızlıca analiz edilmesini ve uygun prim değerinin üretilmesini sağlar. Bu konuda literatürde birden fazla çalışma bulunmaktadır.

Mevcut literatürdeki çalışmaların önemli bir kısmı ağırlıklı olarak model doğruluğuna odaklanmakta olup, model performansının veri dağılımındaki değişimlere karşı kararlılığı ve sonuçların yorumlanabilirliği çoğu zaman sınırlı düzeyde ele alınmaktadır. Özellikle aykırı gözlemlerin model performansı üzerindeki etkisinin sistematik biçimde incelendiği ve model seçimi, hiperparametre optimizasyonu ile sağlamlık analizinin bütüncül bir çerçevede birlikte değerlendirildiği çalışmaların sayısı oldukça sınırlıdır. Bu çalışma, söz konusu boşluğu doldurmayı amaçlayarak yalnızca model performansını değil, aynı zamanda model kararlılığını ve yorumlanabilirliğini de dikkate alan kapsamlı bir değerlendirme yaklaşımı önermektedir.

Bu çalışmada, Kaggle'da açık erişimli olarak paylaşılan bir sağlık sigortası veri seti kullanılarak bireylerin yıllık sağlık sigortası prim tutarlarının makine öğrenmesi tabanlı regresyon yöntemleriyle tahmin edilmesi amaçlanmıştır. Çalışma kapsamında veri seti ön işleme adımlarından geçirilmiş; eksik veri kontrolü yapılmış, yinelenen kayıtlar temizlenmiş ve kategorik değişkenler uygun şekilde sayısallaştırılarak modellemeye hazır hale getirilmiştir. Ardından doğrusal (Linear Regression – LR, Ridge, ElasticNet), örnek tabanlı ve çekirdek tabanlı (KNN – k-Nearest Neighbors, SVR – Support Vector Regression) yöntemler ile ağaç tabanlı topluluk modelleri (Random Forest – RF, Gradient Boosting – GB, XGBoost) olmak üzere toplam sekiz farklı yaklaşım karşılaştırılmıştır. Modellerin performansı eğitim kümesi üzerinde 5-katlı çapraz doğrulama ile değerlendirilmiş, nihai başarı ise hold-out test kümesinde RMSE, MAE, MAPE ve R^2 ölçütleri ile raporlanmıştır. En iyi model için RandomizedSearchCV ile hiperparametre optimizasyonu gerçekleştirilmiş; ayrıca aykırı değerlerin etkisini incelemek amacıyla uç kırpma temelli duyarlılık analizi uygulanarak sonuçların kararlılığı değerlendirilmiştir. Son olarak bağımsız değişkenlerin hedef değişken üzerindeki etki oranları SHAP (SHapley Additive exPlanations) yöntemi ile analiz edilmiştir. Son olarak, elde edilen bulguların tek bir veri setine bağımlı olmadığını göstermek amacıyla, farklı bir öznitelik yapısına sahip ikinci ve bağımsız bir veri seti üzerinde genellenebilirlik analizi gerçekleştirilmiştir.

Çalışmanın geri kalanı şu şekilde organize edilmiştir: ikinci bölümde sağlık sigortası prim/harcama tahmini konusunda makine öğrenmesi modellerinin kullanıldığı çalışmalar özetlenmiştir. Üçüncü bölümde, çalışmanın genel akış süreci, kullanılan veri seti ve analiz grafikleri, makine öğrenmesi modelleri, hiperparametre optimizasyon yöntemi, uç kırpma ve duyarlılık analizi ve sonuçların değerlendirilmesinde kullanılan performans metrikleri anlatılmıştır. Dördüncü bölümde, beş katlı çapraz doğrulama ve hold-out test kümesi üzerinden elde edilen deneysel sonuçlar sunulmuş; en iyi model için elde edilen optimum parametre değerleri ve uç kırpma temelli duyarlılık analizi

bulgularına yer verilmiştir. Ayrıca bu bölümde, SHAP yöntemiyle yapılan açıklanabilirlik analizi sonuçlarına yer verilmiş ve değişkenlerin prim tutarları üzerindeki etkileri yorumlanmıştır. Beşinci bölümde çalışmanın genel sonuçları özetlenmiş ve gelecekte yapılabilecek çalışmalar için öneriler sunulmuştur.

LİTERATÜR ÖZETİ

Bu bölümde, sağlık sigortası prim/harcama tahmini problemine yönelik olarak makine öğrenmesi yöntemlerinin kullanıldığı çalışmalar incelenmiştir. Literatürdeki çalışmaların sayıca fazla olması nedeniyle, özellikle 2024–2025 döneminde yayımlanan güncel araştırmalar; kullanılan veri setleri, öznitelikler, uygulanan makine öğrenmesi modelleri ve elde edilen bulgular açısından ayrıntılı biçimde değerlendirilmiş ve ilgili bilgiler Tablo 1'de özetlenmiştir.

Tablo 1'de özetlenen güncel çalışmalar incelendiğinde, sağlık sigortası prim/harcama tahmininde çoğunlukla doğrusal regresyon yöntemleri ile ağaç tabanlı topluluk modellerinin (özellikle RF, GB ve XGBoost) karşılaştırıldığı ve birçok çalışmada topluluk yaklaşımlarının üstün performans sergilediği görülmektedir (Hu, 2024; Bau & Hanif, 2024; Orji & Ukwandu, 2024; Sijie at al., 2024; Rao et al., 2025). Bununla birlikte literatürdeki çalışmaların bir bölümü yalnızca sınırlı sayıda modeli raporlamakta (Sharma et al., 2024), bazıları ağırlıklı olarak uygulama geliştirme (GUI/Flask) boyutuna odaklanmakta (Mahesh et al., 2025) veya açıklanabilirlik analizlerini (SHAP/ICE/LIME) ön plana çıkarmaktadır (Billa & Nagpal, 2024; Bhongade et al., 2024). Ayrıca uç değerlerin model performansına etkisinin sistematik biçimde incelenmesi ve ön işleme tercihlerinin sonuçlara duyarlılığının raporlanması çoğu çalışmada sınırlı kalmaktadır. Bu çalışma, bu yönüyle, literatürde sıklıkla yalnızca model performansına odaklanan yaklaşımlardan farklı olarak, sigorta primi tahmini problemine bütüncül ve sağlamlık odaklı bir değerlendirme çerçevesi sunmayı amaçlamaktadır. Bu kapsamda çalışmanın özgün katkıları aşağıda özetlenmektedir:

- Çalışmada öncelikle veri seti kapsamlı bir keşifsel veri analizi ile incelenmiş, değişkenler arası ilişkiler ve veri yapısı detaylı biçimde ortaya konmuştur.
- Sekiz farklı regresyon yaklaşımı (doğrusal modeller, örnek/çekirdek tabanlı yöntemler ve ağaç tabanlı topluluk modelleri) tek bir deneysel protokol altında sistematik olarak karşılaştırılmıştır. Bu yönüyle çalışma, farklı model ailelerini aynı koşullar altında değerlendiren bütüncül bir karşılaştırma sunmaktadır.
- Model performansı yalnızca tek bir ölçütle değil, RMSE, MAE, MAPE ve R^2 gibi birden fazla hata metriği üzerinden değerlendirilerek daha dengeli ve güvenilir bir performans analizi gerçekleştirilmiştir.
- En iyi performansı veren model için RandomizedSearchCV kullanılarak hiperparametre optimizasyonu yapılmış ve ham model ile optimize edilmiş model sonuçları birlikte sunularak optimizasyonun etkisi açık biçimde ortaya konmuştur.
- Çalışmanın önemli katkılarından biri, uç gözlemlerin model performansı üzerindeki etkisini incelemek amacıyla uç kırpma tabanlı bir duyarlılık analizi gerçekleştirilmesidir. Bu analiz ile model sonuçlarının veri dağılımındaki değişimlere karşı ne ölçüde kararlı olduğu değerlendirilmiştir.
- Model yorumlanabilirliğini desteklemek amacıyla SHAP analizi uygulanmış ve özellikle değişkenlerin tahmin üzerindeki göreli etkileri incelenmiştir. Bu analiz, model çıktılarının yalnızca doğruluk açısından değil, açıklanabilirlik açısından da değerlendirilmesini sağlamaktadır.
- Önerilen değerlendirme çerçevesinin tek bir veri setine bağımlı olmadığını göstermek amacıyla, farklı ve daha zengin bir öznitelik yapısına sahip ikinci ve bağımsız bir veri seti üzerinde genellenebilirlik analizi gerçekleştirilmiş; ağaç tabanlı topluluk modellerinin üstünlüğü bu bağımsız veri setinde de doğrulanmıştır.

Bu yönleriyle çalışma, yalnızca en yüksek doğruluğa sahip modeli belirlemeyi hedefleyen geleneksel yaklaşımlardan ayrılarak, model seçimi, hiperparametre optimizasyonu ve sağlamlık analizini entegre eden daha kapsamlı ve uygulamaya dönük bir değerlendirme çerçevesi sunmaktadır. Bu çalışma ayrıca, sigorta sektöründe karar vericiler için model seçimi sürecinde yalnızca doğruluk değil, model kararlılığı ve yorumlanabilirliğin de dikkate alınması gerektiğini ortaya koymaktadır.

Tablo 1. 2024–2026 Yıllarındaki Sağlık Sigortası Prim/Harcama Tahmini Üzerine Makine Öğrenmesi Tabanlı Çalışmaların Karşılaştırmalı Özeti

Yazar, Yıl	Çalışmanın Katkısı	Veri Seti	Kullanılan Modeller	Hiperparametre Optimizasyonu	Test Verisindeki En Yüksek Sonuç	Açıklanabilirlik Analizi
Hu, 2024	Sağlık sigorta maliyeti tahmin edilmiştir.	*	LR, Decision Tree – DT, RF	-	0,8564 R^2 - RF	-
Sijie et al., 2024	ANN temelli topluluk modelleri önerilmiştir.	*	ANN + XGBoost, SVM, DT, RF	-	0,8880 R^2 - ANN+RF	-
Billa & Nagpal, 2024	DT ve RF modellerini kullanan bir GUI arayüzü tasarlanmıştır	*	LR, Ridge, DT, KNN, RF, SVM, GB	-	0,8679 R^2 - GB	SHAP
Sharma et al., 2024	Sağlık sigorta maliyeti tahmin edilmiştir.	*	Lineer, Polinomial, Ridge, Lasso	-	0,80 R^2 - Polinomial	-
Bau & Hanif, 2024	Sağlık sigorta maliyeti tahmin edilmiştir.	*	Multiple Linear, Ridge, XGBoost, RF	GridSearchCV	0,86 R^2 - RF	-
Orji & Ukwandu, 2024	Özellik analizinde SHAP ve ICE karşılaştırması yapılmıştır.	**	XGBoost, GBM, RF	GridSearchCV	0,8647 R^2 - XGBoost	SHAP ve ICE
Bhongade et al., 2024	Özniteliklerin sonuca etkisi analiz SHAP ile analiz edilmiştir.	**	Histogram GB, GB, RF	-	0,8528 R^2 - HGB	SHAP
Kapse et al., 2025	(1) Veri toplanarak orijinal veri seti oluşturulmuştur. (2) Değişkenlerin fark yaratıp yaratmadığını tespit etmek için hipotez testleri (ANOVA, T-testi) yapılmıştır.	***	Multiple Regresyon, GB, RF, XGBoost	-	0,61 R^2 - GB	SHAP
Mahesh et al., 2025	Flask web uygulaması sunulmuştur.	*	DT, RF, LR	-	0,963 R^2 - RF	-
Rao et al., 2025	SHAP tabanlı açıklanabilirlik sunulmuştur.	*	XGBoost, KNN, GB, LightGBM, RF, DT, SVR, Lineer Regresyon	RandomizedSearchCV	0,9552 R^2 - XGBoost	SHAP
Liu & Zhang, 2026	Öznitelik önem dereceleri SHAP analizi ile incelenmiştir.	**	DT, KNN, SVM, XGBoost, LightGBM, RF	-	0,939 R^2 - LightGBM	SHAP
Irmak, 2026	**Uç kırpma temelli duyarlılık analizi; SHAP tabanlı açıklanabilirlik analizi yapılmıştır.**	*	**GB, RF, XGBoost, Ridge, ElasticNet, Lineer Regresyon, KNN, SVR**	**RandomizedSearchCV**	**0,9007 R^2 - GB / 0,9107 – GB (uç kırpma uygulanmış)**	**SHAP**

*1338 kayıt – 7 öznitelik (age, sex, bmi, children, smoker, region, charges) - Erişim Bağlantısı: https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction

** 986 kayıt – 11 öznitelik (age, diabetes, blood pressure problems, any transplants, any chronic disesases, height, weight, known allergies, history of cancer in family, number of major surgeries, premium price) – Erişim Bağlantısı: https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction/data

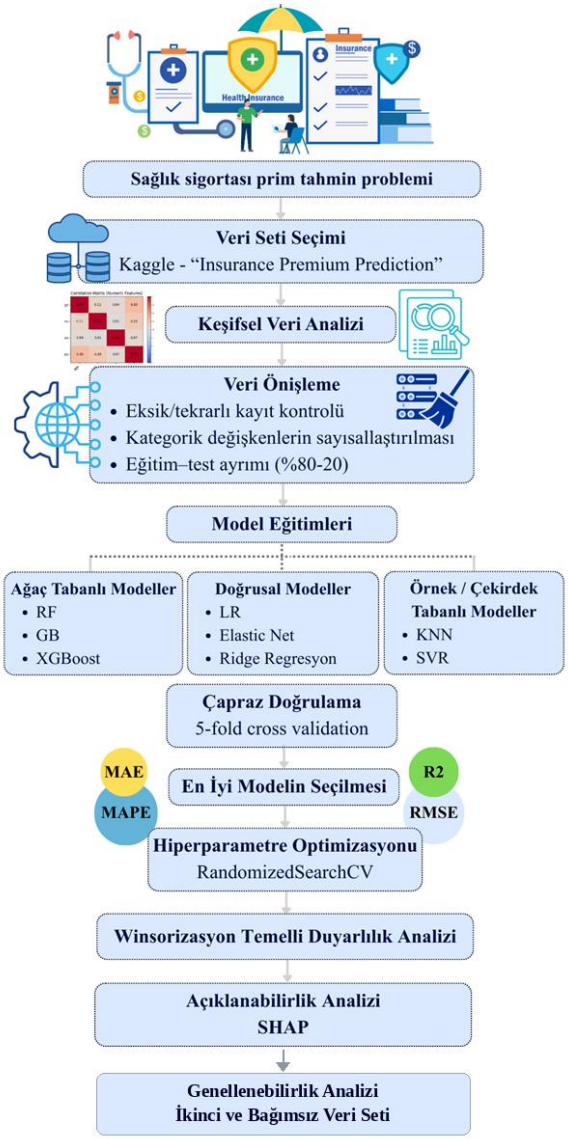
*** 145 kişi – 22 öznitelik (age, gender, bmi, city, smoking, diabetic, high blood pressure, heart disease, health insurance, for only yourself, for only parents, for only spouse, for only children, number of members, type of insurance, insurance provider, premium amount (target), policy holding duration, secondary insurance, number of secondary members, type secondary insurance, secondary insurance premium) – Veri seti açık erişimli olarak paylaşılmamıştır.

MATERYAL ve YÖNTEM

Bu bölümde, öncelikle makine öğrenmesi modellerine girdi olarak kullanılan veri seti ile değişken yapısı ayrıntılı biçimde tanıtılmıştır. Model eğitimi öncesinde, veri setinin yapısını anlamak ve olası aykırı değerleri tespit etmek amacıyla keşifsel veri analizi yapılmıştır. Ardından, deneysel çalışmada değerlendirilen regresyon tabanlı makine öğrenmesi modelleri açıklanmıştır. Hiperparametre optimizasyonunda kullanılan RandomizedSearchCV yönteminin çalışma mantığından bahsedilmiştir. Uç kırpma temelli duyarlılık analizi ile alakalı bilgi verilmiştir. Model çıktılarının yorumlanabilirliğini artırmak amacıyla, özelliklerin modele olan katkılarının analizinde kullanılan SHAP yaklaşımına yer verilmiştir. Son olarak, elde edilen sonuçların nicel olarak değerlendirilmesinde kullanılan performans ölçütleri tanımlanmıştır. Tüm bu aşamaları gösteren akış diyagramı Şekil 1'de sunulmaktadır.

Veri Seti

Bu çalışmada kullanılan veri seti, Kaggle'da açık erişimli olarak paylaşılan 'Insurance Premium Prediction' veri setidir. Veri seti, bireylere ait demografik ve yaşam tarzı özellikleri ile sağlık sigortası prim tutarlarını içermektedir. Veri seti toplam 1338 gözlem ve 7 değişkenden oluşmaktadır. Bağımlı değişken olarak kullanılan charges, bireylerin yıllık sağlık sigortası maliyetlerini temsil etmektedir. Bağımsız değişkenler ise yaş (age), cinsiyet (sex), vücut kitle indeksi (bmi), bakmakla yükümlü olunan çocuk sayısı (children), sigara kullanımı durumu (smoker) ve coğrafi bölge (region) değişkenlerinden oluşmaktadır. Veri setinde hem sayısal (age, bmi, children) hem de kategorik (sex, smoker, region) değişkenler yer almaktadır. Eksik gözlem bulunmayan veri seti, ön işleme adımlarının ardından modelleme aşamasında kullanılmıştır. Tablo 2'de veri önişleme öncesi ham veri setinin ilk beş satırı verilmiştir.



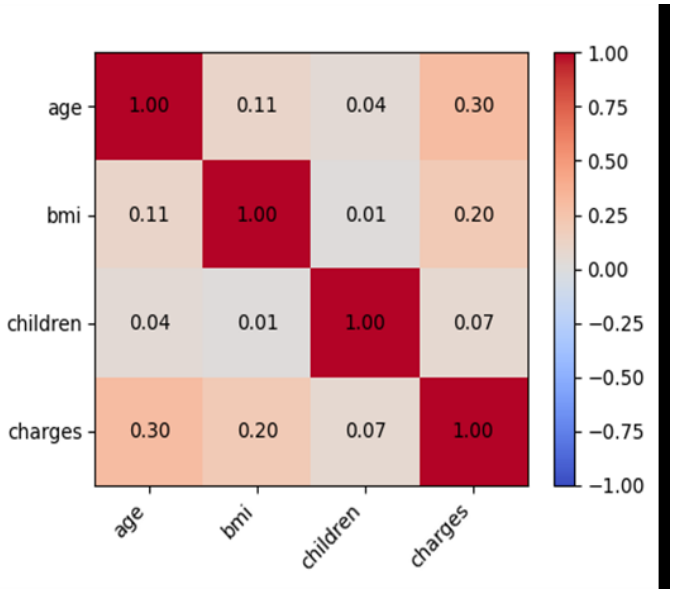
Şekil 1. Çalışma Akış Diyagramı

Tablo 2. Veri Setine Ait İlk Beş Gözlem

age	sex	bmi	children	smoker	region	charges
19	female	27,900	0	yes	southwest	16884,92400
18	male	33,770	1	no	southeast	1725,55230
28	male	33,000	3	no	southeast	4449,46200
33	male	22,705	0	no	northwest	21984,47061
32	male	28,880	0	no	northwest	3866,85520

Keşifsel Veri Analizi

Bu bölümde, makine öğrenmesi model eğitimleri yapılmadan önce veri setinin yapısını daha iyi anlamak ve değişkenlerin hedef değişken üzerindeki potansiyel etkilerini incelemek amacıyla keşifsel veri analizi yapılmıştır. Bu amaçla, öncelikle veri setinde yer alan sayısal değişkenler arasındaki doğrusal ilişkileri incelemek amacıyla oluşturulan korelasyon matrisi Şekil 2'de sunulmuştur.



Şekil 2. Sayısal Değişkenler için Korelasyon Matrisi

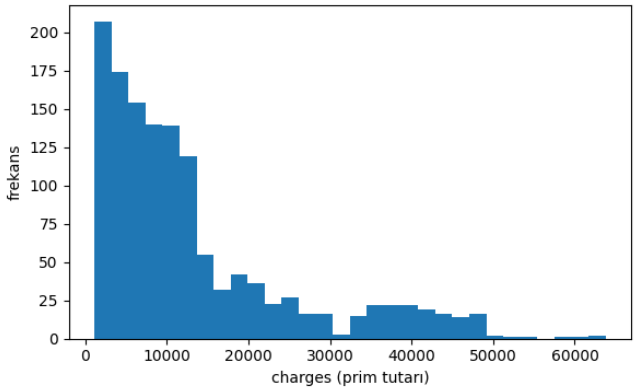
Korelasyon matrisi incelendiğinde, yaş değişkeni ile sağlık sigortası maliyetleri arasında orta düzeyde pozitif bir korelasyon olduğu görülmektedir. Benzer şekilde, vücut kitle indeksi ile sigorta maliyeti değişkeni arasında da pozitif yönlü ancak görece daha zayıf bir ilişki bulunmaktadır. Buna karşın, children değişkeninin hem diğer bağımsız değişkenlerle hem de hedef değişkenle olan korelasyonu oldukça düşük seviyededir. Elde edilen bulgular, özellikle yaş ve vücut kitle indeksinin sigorta primleri üzerinde daha belirgin bir etkiye sahip olabileceğine işaret etmekte; bu değişkenlerin modelleme aşamasında önemli girdiler olacağını göstermektedir.

Aşağıdaki şekilde sağlık sigorta prim tutarlarının dağılımı verilmiştir (Şekil 3). Grafik incelendiğinde, prim tutarlarının belirgin biçimde sağa çarpık olduğu görülmektedir. Dağılımın sağ kuyruğunu yüksek maliyetli ve görece az sayıdaki gözlemler oluştururken, geriye kalan kısmı ise düşük ve orta düzey primlere aittir. Bu durum, sigorta harcamalarının bireyler arasında homojen dağılmadığını ve belirli bir alt grubun oldukça yüksek maliyetlere sahip olduğunu göstermektedir. Elde edilen bulgular, modelleme aşamasında aykırı değerlerin ve dağılım çarpıklığının dikkate alınmasının önemine işaret etmektedir.

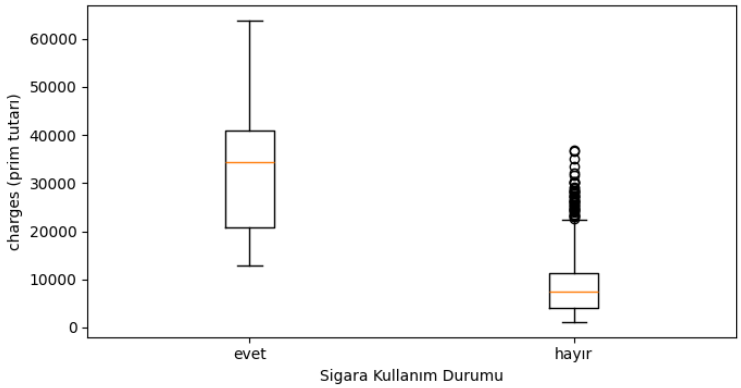
Şekil 4'te sigara kullanma durumunun sağlık sigortası primleri üzerindeki etkisi kutu grafiği aracılığıyla incelenmiştir.

Şekil 4 incelendiğinde, sigara kullanan bireyler ile sigara kullanmayan bireyler arasında sağlık sigortası primleri açısından belirgin bir farklılık olduğu görülmektedir. Sigara kullanan bireylerin medyan prim tutarları, sigara kullanmayanlara kıyasla önemli ölçüde daha yüksektir. Ayrıca, sigara kullanan grupta hem dağılımın genişlediği hem de yüksek maliyetli aykırı gözlemlerin sayısının arttığı dikkat çekmektedir. Bu bulgular, sigara kullanımının sağlık

sigortası maliyetlerini artıran en güçlü belirleyicilerden biri olduğunu göstermekte ve modelleme aşamasında bu değişkenin kritik bir rol oynayacağını işaret etmektedir.

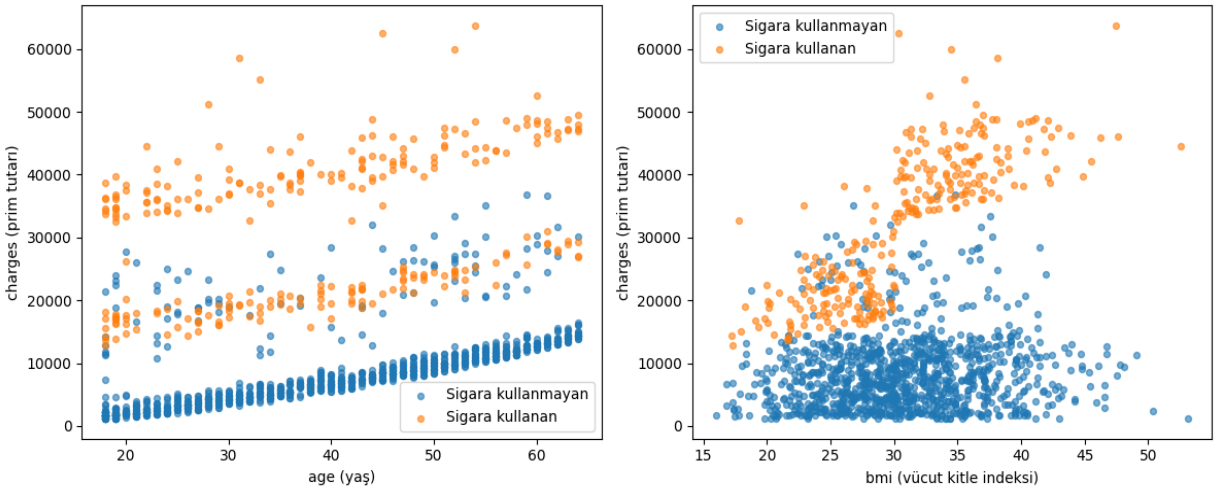


Şekil 3. Sağlık Sigortası Prim Tutarlarının Dağılımı



Şekil 4. Sigara Kullanım Durumuna Göre Sağlık Sigortası Primlerinin Dağılımı

Yaş ve vücut kitle indeksi ile sağlık sigortası primleri arasındaki ilişkiler, sigara kullanım durumu dikkate alınarak Şekil 5'te görselleştirilmiştir.



Şekil 5. Sigara Kullanım Durumuna Göre Yaş ve Vücut Kitle İndeksi ile Sağlık Sigortası Primleri Arasındaki İlişki Grafikleri

Şekil 5'teki grafikler incelendiğinde, yaş – sağlık sigortası primleri arasında pozitif yönlü bir ilişki olduğu ve bu ilişkinin sigara kullanan bireylerde daha belirgin hâle geldiği görülmektedir. Sigara kullanan bireylerin tüm yaş gruplarında sistematik olarak daha yüksek prim tutarlarına sahip olduğu dikkat çekmektedir. Benzer şekilde, vücut kitle indeksi ile prim tutarları arasındaki ilişki sigara kullanım durumu dikkate alındığında daha net bir biçimde ortaya

çıkmaktadır. Özellikle yüksek vücut kitle indeksine sahip ve sigara kullanan bireylerin prim tutarlarının önemli ölçüde arttığı gözlemlenmektedir.

Keşifsel veri analizi sonucunda elde edilen bulgular, değişkenler arasındaki etkileşimlerin doğrusal olmayan yapılar içerebileceğine işaret etmektedir. Bu etkileşimleri yakalayabilen regresyon tabanlı makine öğrenmesi modellerinin kullanımını gerekli kılmaktadır.

Makine Öğrenmesi Modelleri

Bu bölümde, sağlık sigortası prim tutarlarının tahmininde kullanılan makine öğrenmesi modelleri sunulmaktadır. Ayrıntılar aşağıda verilmektedir:

Gradient Boosting Machine (GBM): GBM yönteminin formülasyonu ilk olarak Friedman tarafından 2001 yılında tanıtılmıştır. GBM, tahmin ve sınıflandırma problemlerinde yüksek başarım sağlayan denetimli öğrenme yöntemlerinden biridir. Yöntemin temel amacı, birden çok zayıf öğreniciyi (çoğunlukla karar ağaçlarını) ardışık olarak birleştirerek önceki modellerin hatalarını öğrenen yeni modeller eklemek ve bu sayede kayıp fonksiyon değerini yinelemeli olarak azaltmaktır (Natekin & Knoll, 2013; Saraiva et al., 2026). GBM, artışlı bir model kurarak çalışır ve her iterasyonda mevcut modelin hatalarına küçük bir karar ağacı uydurur ve bu ağacı mevcut modele ekleyerek seçilen kayıp fonksiyonunu gradyan iniş ile azaltır (Friedman, 2002). Bu işleyiş, farklı kayıp fonksiyonlarında (kare hata, LAD, Huber vb.) ve karar ağacı tabanlı zayıf öğrenicilerle hem regresyon hem de sınıflandırma problemlerinde yüksek doğruluk ve esneklik sağlar.

Extreme Gradient Boosting (XGBoost): XGBoost yöntemi, ağaç tabanlı gradient boosting yönteminin ölçeklenebilir, düzenlileştirilmiş ve yüksek performanslı bir uygulamasıdır (Chen & Guestrin, 2016). Temel olarak GBM çerçevesini izler; ancak hesaplama verimliliği, düzenlileştirme ve paralel işlemeye yönelik çeşitli iyileştirmeler içerir. Model, art arda eklenen karar ağaçları ile kayıp fonksiyonunu minimize etmeye çalışırken, aynı zamanda L1 ve L2 düzenlileştirme terimleriyle model karmaşıklığını kontrol eder (Bentéjac et al., 2021). Eksik verilerle başa çıkabilmesi, sütun alt örnekleme, öğrenme oranı, ağaç derinliği gibi zengin hiperparametre yapısı ve paralel hesaplama desteği sayesinde hem doğruluk hem de hesaplama süresi açısından büyük veri setlerinde yaygın olarak tercih edilmektedir (Ali et al., 2023).

Random Forest (RF): RF modeli, karar ağaçlarına dayalı bir topluluk öğrenme yöntemidir (Lin et al., 2026). Breiman'ın 2001 yılında yaptığı çalışmadaki tanımına göre RF, her biri bootstrap örnekleri üzerinde eğitilen çok sayıda karar ağacının çıktısını birleştirerek sınıflandırma veya regresyon tahmin değeri üretir (Breiman, 2001). Her düğümde tüm öznitelikler yerine rastgele seçilen bir alt küme üzerinden en iyi bölme aranması, ağaçlar arası korelasyonu azaltarak genelleme hatasını düşürür (Liaw & Wiener, 2002).

Linear Regression (LR): Klasik lineer regresyon yaklaşımında, tahmin edilen değerler ile gerçek gözlemler arasındaki hata kareler toplamı minimize edilerek model parametreleri tahmin edilmeye çalışılır. Yani, katsayılara hiçbir ceza uygulamadan en küçük kareler tahmini kullanılır (Bayo et al., 2021). Ancak, bağımsız değişkenler arasında güçlü çoklu doğrusal bağlantı olduğunda, öznitelik sayısının fazla olduğu veya veri boyutunun sınırlı olduğu durumlarda OLS katsayılarının varyansı şişer; tahminler kararsız, güven aralıkları geniş ve tahmin performansı zayıf hale gelir (Usman et al., 2021). Buna rağmen LR yöntemi, düzenlileştirilmiş yöntemlerle karşılaştırma yaparken referans model olarak sıkça kullanılmaktadır. Bu olumsuzlukları gidermek amacıyla Ridge, Elastic Net gibi düzenlileştirme tabanlı regresyon yöntemleri geliştirilmiştir.

Ridge Regresyon (RR): Bu yöntemde OLS amaç fonksiyonuna kareler toplamı cezası (L2 normu) eklenerek katsayılar sıfıra doğru küçültülür, fakat genellikle tamamen sıfırlanmaz. Bu sayede varyans belirgin biçimde azaltılır ve çoklu doğrusal bağlantı durumunda daha kararlı tahminler üretilir (Saputro et al., 2025). Bu yöntemin önemli farkı, değişken seçimi yapmamasıdır; yani tüm değişkenleri modelde tutar. Dolayısıyla yorumlanabilirlikten çok tahmin kararlılığının önemli olduğu, tüm değişkenlerin modele girmesinin istendiği durumlar için uygundur (Usman et al., 2021).

Elastic Net Regresyon (ENR): Elastic Net, ceza teriminde hem L1 (Lasso tipi) hem L2 (Ridge tipi) düzenlileştirmeyi birleştirir ve böylece hem değişken seçimi yapabilir hem de çoklu doğrusal bağlantı altında Ridge benzeri grup etkisi sergiler (Farhadi et al., 2019). Yani Ridge ve Lasso regresyon yöntemlerinin avantajlarını bir araya getiren hibrit bir

yöntemdir. Yüksek korelasyonlu ve yüksek boyutle veri setlerinde değişken seçimi yapabilmesi ve model kararlılığını koruması nedeniyle literatürde sıklıkla kullanılan bir yöntemdir.

Support Vector Regression (SVR): SVR yönteminin temel amacı, tahmin edilen değerlerin gerçek gözlemlerden belirli bir hata toleransı (ε) içerisinde kalmasını sağlarken, model karmaşıklığını minimumda tutmaktır (Khajeh et al., 2026). Yöntem, doğrusal bir regresyon doğrusunu yüksek boyutlu bir özellik uzayında kurar; girdiler bu uzaya bir çekirdek fonksiyonu ile dolaylı olarak taşınır (Smola & Schölkopf, 2004; Awad & Khanna, 2015). Gözlenen hata yerine genelleme hatası sınırı minimize edilerek; fonksiyonun etrafında ε-duyarsız bir tüp tanımlanır ve bu tüpün dışına düşen noktalar cezalandırılırken içeride kalan küçük hatalar yok sayılır. Böylece yalnızca tüp dışındaki bazı noktalar destek vektörü olur ve model bu az sayıdaki kritik noktaya dayanarak yüksek boyutlu uzayda düzgün, iyi genelleşen bir fonksiyon öğrenir.

K-Nearest Neighbors (KNN): KNN regresyonu ise örnek temelli ve parametrik olmayan bir yaklaşımdır. Bu yaklaşımda model eğitimi sırasında açık bir fonksiyon öğrenilmez; bunun yerine, tahmin aşamasında eğitim kümesindeki öklidyen, Manhattan, Minkowski vb. bir mesafeye göre en yakın k komşu bulunur ve hedef değer bu komşuların genellikle ortalaması/ ağırlıklı ortalaması alınarak tahmin edilir (Halder et al., 2024). k değerinin seçimi, mesafe metriği ve ölçeklendirme, KNN'in performansını doğrudan belirler; yanlış seçimler aşırı uyum veya yetersiz uyuma yol açabilir (Zhang, 2016).

Hiperparametre Optimizasyonu: RandomizedSearchCV

RandomizedSearchCV, hiperparametre optimizasyonunda rastgele arama ile çapraz doğrulamayı birleştiren basit ama güçlü bir yöntemdir. Temel fikir, her hiperparametre için bir olasılık dağılımı (örneğin uniform, log-uniform, ayrık dağılım) tanımlamak ve bu uzaydan rastgele örneklenen sınırlı sayıda konfigürasyonu, çapraz doğrulama ile değerlendirmektir (Franceschi et al., 2025). Her denemede seçilen hiperparametre setiyle model eğitilir, CV hata/başarı skoru hesaplanır ve en iyi ortalama performansa sahip konfigürasyon seçilir.

Uç Kırpma Yöntemi

Uç kırpma yöntemi, uç değerleri tamamen silmek yerine belirli bir eşikte kırpıp daha az uç bir değere taşımak için kullanılan, sağlam bir istatistiksel ön-işleme tekniğidir. Amaç, uç değerlerin ortalama, varyans ve model tahminleri üzerindeki etkisini azaltırken veri yapısını ve gözlem sayısını korumaktır. İki taraflı uç kırpmada değişkenin alt ve üst uçlarındaki belirli bir yüzde (ör. %5–%5) veya eşik belirlenir. Bu sınırların dışındaki gözlemler, sınır değerle değiştirilir; yani atılmaz (Zubedi et al., 2022; Dash et al., 2023).

Açıklanabilirlik Analizi: SHAP (SHapley Additive exPlanations)

SHAP yöntemi, makine öğrenmesi modellerinin karar verme süreçlerini yorumlamak için oyun teorisine dayalı bir özellik katkı yöntemidir. Temeli, işbirlikçi oyun teorisindeki Shapley değerlerine dayanır: bir oyunda oyuncular toplam kazancı aralarında adil biçimde paylaşırken, her oyuncunun katkısı tüm olası oyuncu sıralamaları üzerinden ortalama marjinal katkı ile tanımlanır; SHAP'ta oyuncular özellikler, kazanç ise model çıktısıdır (Lundberg & Lee, 2017).

Lundberg ve Lee, SHAP'ı "katkısal özellik önem ölçüleri" sınıfını tanımlayarak formüle eder ve bu sınıfta belirli aksiyomları (yerelleştirme, katkısallık, tutarlılık gibi) aynı anda sağlayan tekil bir çözüm olduğunu, bunun da Shapley tabanlı değerler olduğunu gösterir. Bu çerçevede, her bir gözlem için model çıktısı, bir referans değerden başlayıp her özelliğin SHAP değeri kadar katkı yaptığı lineer bir ayrışım olarak yazılır; böylece SHAP, tahmini "özelliklere ayrıştıran" yerel bir açıklama üretir (Lundberg & Lee, 2017; Ponce-Bobadilla et al., 2024).

Shapley değeri, bir özniteliğin modele eklenmesiyle tahmin çıktısında meydana gelen marjinal katkının, tüm olası öznitelik kombinasyonları üzerinden ortalaması alınarak hesaplanmaktadır. SHAP yönteminin en önemli avantajlarından biri, farklı model türlerinden (doğrusal modeller, ağaç tabanlı modeller, kernel tabanlı yöntemler ve derin öğrenme modelleri) bağımsız olarak uygulanabilmesidir. Ayrıca, SHAP değerleri kullanılarak özniteliklerin model çıktısı üzerindeki pozitif veya negatif etkileri hem sayısal hem de görsel olarak analiz edilebilmektedir. Özellikle öznitelik önem sıralaması, bağımlılık grafikleri ve örnek bazlı kuvvet grafikleri, model davranışının daha şeffaf biçimde yorumlanmasına olanak tanımaktadır (Lundberg & Lee, 2017).

Performans Metrikleri

Kullanılan makine öğrenmesi modellerinin tahmin performansını değerlendirmek için R^2, MAE, MAPE ve RMSE metrikleri kullanılmıştır. Tablo 3'te bu metriklerin açıklamaları ve formülleri verilmiştir.

Tablo 3. Performans Metrikleri (Irmak, 2025)

Metrik	Açıklama	Formül
R^2 Determinasyon Katsayısı	Modelin açıklama gücünü gösterir. 1'e ne kadar yakınsa, model veriyi o kadar iyi açıklamaktadır.	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
MAE Ortalama Mutlak Hata	Gerçek değerlerle tahminler arasındaki farkların mutlak ortalamasıdır. Düşük olması istenir.	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - y_i\rvert$
RMSE Kök Ortalama Kare Hata	Hata karelerinin ortalamasının kareköküdür. Büyük hatalara daha duyarlıdır.	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
MAPE Ortalama Mutlak Yüzde Hata	Modelin ortalama hata oranını yüzde cinsinden gösterir.	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$

DENEYSEL BULGULAR

Bu bölümde, öncelikle deneysel çalışmaların yapıldığı ortamdan ve programlama dilinden bahsedilmiştir. Ardından sağlık sigortası verisi üzerinde kurulan regresyon modellerinin performans sonuçları verilmiştir. Modeller 5-katlı çapraz doğrulama ile eğitimi verisi üzerinde karşılaştırılmış ve ayrılan test verisi üzerinde genel performansları doğrulanmıştır. En iyi tahmin performansına sahip yöntem için hiperparametre optimizasyonu yapılmıştır ve elde edilen en iyi parametre kombinasyonu sunulmuştur. Ayrıca aykırı değerlerin performans ölçütlerine etkisi duyarlılık analizi ile test edilmiştir ve sonuçlar verilmiştir. En iyi model için elde edilen tahminlerin yorumlanabilirliğini artırmak amacıyla SHAP analizi uygulanmış ve elde edilen sonuçlar yorumlanmıştır. Son olarak, elde edilen bulguların tek bir veri setine bağımlı olmadığını göstermek amacıyla, farklı ve daha zengin bir öznitelik yapısına sahip ikinci ve bağımsız bir veri seti üzerinde genellenebilirlik analizi gerçekleştirilmiş ve sonuçlar yorumlanmıştır.

Çalışma Ortamı

Bu çalışmada gerçekleştirilen tüm deneysel uygulamalar, Google Colaboratory (Colab) ortamında Python programlama dili kullanılarak yapılmıştır. Colab, herhangi bir program kurulum gereksinimi olmaksızın tarayıcı üzerinden erişilebilen bulut tabanlı yapısı sayesinde yüksek hesaplama kaynaklarına hızlı erişim, otomatik kütüphane yönetimi ve GPU/TPU desteği sunmaktadır. Python'un sade sözdizimi, modüler yapısı ve zengin kütüphane desteği, veri analizi ve makine öğrenmesi uygulamalarının hızlı ve güvenilir biçimde geliştirilmesine olanak sağlamaktadır.

Veri Ön İşleme

Modelleme sürecinden önce veri seti, veri ön işleme aşamasından geçirilmiştir. Yinelenen kayıtların incelenmesi sonucunda 1 adet tekrar eden gözlem veri setinden çıkarılmış ve analizler 1337 gözlem üzerinden gerçekleştirilmiştir. Tüm değişkenler için yapılan kontrollerde eksik veri bulunmadığı tespit edilmiştir. Değişkenler türlerine göre ayrıştırılmış; age, bmi ve children değişkenleri sayısal, sex, smoker ve region değişkenleri ise kategorik olarak tanımlanmıştır. Sayısal değişkenler üzerinde IQR (Interquartile Range) yöntemi ile aykırı değer analizi yapılmış; age ve children değişkenlerinde aykırı gözleme rastlanmazken, bmi değişkeninde 9, charges değişkeninde ise 139 aykırı değer tespit edilmiştir. Ancak özellikle sigorta primlerine ilişkin yüksek değerlerin gerçek yaşamda karşılaşılan maliyet farklılıklarını yansıtabileceği değerlendirilmiş ve bu nedenle aykırı gözlemler ilk etapta veri setinden çıkarılmamıştır. Sayısal değişkenler için ölçekleme ve kategorik değişkenler için uygun sayısallaştırma işlemleri ayrı ön işleme hatları içerisinde yapılandırılmıştır. Tüm dönüşüm parametreleri yalnızca eğitim verisi üzerinde öğrenilmiş, test verisine ise aynı dönüşümler uygulanarak veri sızıntısı riski önlenmiştir.
Keşifsel veri analizi bölümünde bulunan bulgular, veri ön işleme aşamasında gerçekleştirilen IQR tabanlı aykırı değer analizinin motivasyonunu oluşturmuş ve aykırı değerlerin veri ön işleme sürecinde sistematik olarak değerlendirilmesine temel sağlamıştır. Veri ön işleme adımında hedef değişken ve bağımsız değişkenler tanımlandıktan sonra veri seti %80'i eğitim ve %20'si test olmak üzere ayrıştırılmıştır.

Eğitim Kümesinde Çapraz Doğrulama Sonuçları

Veri seti, veri önişleme adımında eğitim-test verisi olarak ayrıştırıldıktan sonra eğitim verisi üzerinde 5-katlı çapraz doğrulama ile makine öğrenmesi modellerinin karşılaştırmaları yapılmıştır. Elde edilen sonuçlar Tablo 4'te verilmiştir. Çapraz doğrulamanın sadece eğitim kümesinde uygulanmasının sebebi test verisine sızıntıyı önlemektir. Yani nihai model performansları modelin daha öncesinde görmediği, hold-out test kümesi üzerinden değerlendirilmiştir.

Tablo 4. Eğitim Kümesi Üzerinde 5-Katlı Çapraz Doğrulama ile Elde Edilen Regresyon Performans Sonuçları (Ortalama ± Standart Sapma)

Model	RMSE_ mean	RMSE-std	MAE-mean	MAE_std	MAPE_ mean	MAPE_ std	R2_mean	R2_std
GB	**4710,98**	**193,78**	**2650,36**	**100,81**	**31,43**	**4,51**	**0,8345**	**0,029**
RF	4884,8	265,91	2725,72	163,48	32,59	4,29	0,8216	0,035
XGBoost	4998,96	206,48	2887,87	117,08	36,56	3,39	0,8137	0,032
Ridge	6123,64	210,65	4226,79	231,42	42,54	5,32	0,7228	0,031
ElasticNet	6123,62	211,21	4224,02	232,42	42,47	5,32	0,7227	0,031
LinearRegres	6123,65	211,64	4221,96	233,16	42,43	5,32	0,7227	0,031
KNN	6709,70	262,54	4043,66	140,31	41,87	5,19	0,6679	0,031
SVR	12128,1	754,71	7860,21	319,39	92,90	10,20	-0,0796	0,035

Tablo 4'e bakıldığında en iyi çapraz doğrulama performansının R^2=0,8346±0,029 değeriyle Gradient Boosting modeliyle elde edildiği görülmektedir. Bu modele ait diğer performans metriklerinin diğer model sonuçlarına kıyasla daha iyi olduğu da görülmüştür. Bu modeli R^2=0,8217±0,035 ile RF ve R^2=0,8137±0,032 değeri ile XGBoost modeli izlemektedir. Doğrusal regresyon modelleri (Linear/Ridge/ElasticNet) Gradient Boosting modeline kıyasla düşük performans göstermektedir. Hata metrikleri hedef değişken olan "charges" değişkeninin orijinal parasal ölçeğinde raporlandığından, değerlerin binler düzeyinde olması beklenen bir durumdur.

Test Kümesindeki Model Performans Sonuçları

Çapraz doğrulama yapılan modeller aynı eğitim-test ayrımında hold-out setinde değerlendirilmiş ve sonuçlar Tablo 5'te verilmiştir. Test sonuçları eğitim verisi üzerinde yapılan 5-katlı çaprazlamayı doğrulamaktadır.

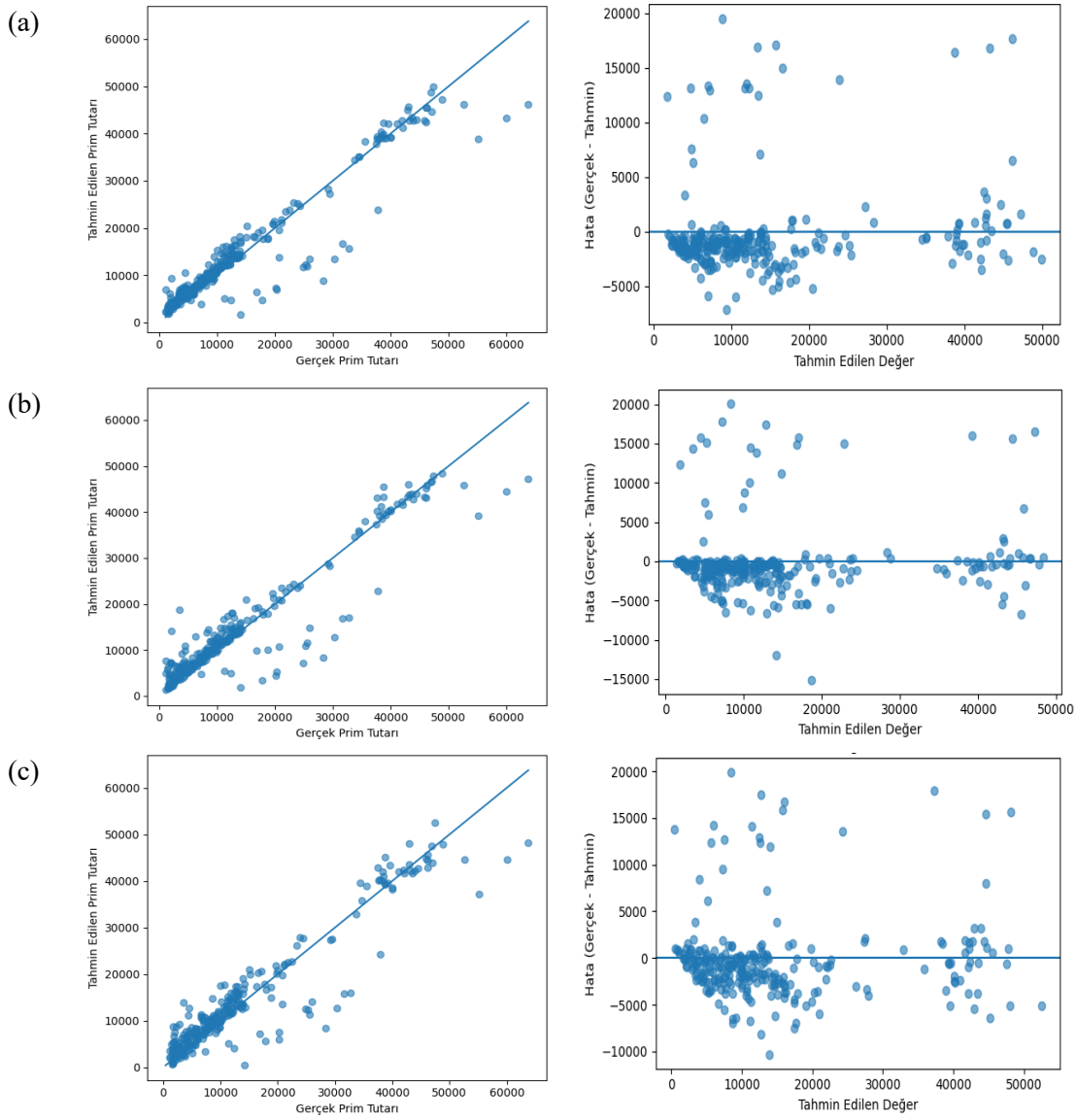
Tablo 5'e bakıldığında en yüksek başarı Gradient Boosting modeli ile elde edilmiştir ve R^2=0,9007 olarak hesaplanmıştır. Bu modeli XGBoost ve RF modelleri takip etmiştir. SVR modeli ise en düşük performansı göstermiştir. Bu sonuçlara bakıldığında ağaç tabanlı modellerinin daha başarılı olduğu görülmektedir. Bunun sebebi ise hedef değişken olan sigorta prim tutarının yaş, BMI ve sigara kullanımı gibi değişkenlerle doğrusal olmayan ilişkiler içermesi ve bu karmaşık ilişkilerin bu modeller tarafından iyi yakalanabilmesidir.

Hold-out test sonuçlarına göre en iyi sonuç veren üç model için gerçek ve tahmin edilen değerlerin karşılaştırması Şekil 6'da sunulmuştur. Sağ taraftaki grafikler bu modeller için artıkların (gerçek-tahmin) tahmin edilen değerlere göre dağılımını gösterirken sol taraftaki grafikler gerçek ve tahmin edilen sağlık sigortası prim değerlerinin saçılımını göstermektedir.

Şekil 6'ya göre üç modelde genel eğilimi yakalamıştır. Ancak sapmalar ve artıkların saçılımının yüksek primli gözlem değerlerinde arttığı görülmektedir.

Tablo 5. Test Kümesi Üzerinde Elde Edilen Regresyon Performans Sonuçları

Model	RMSE_test	MAE_test	MAPE_test	R2_test
GradientBoosting	**4271,88**	**2535,02**	**32,99**	**0,9007**
XGBoost	4566,63	2828,16	36,47	0,8865
RandomForest	4613,54	2550,16	35,12	0,8841
LinearRegression	5956,34	4177,04	41,39	0,8069
ElasticNet	5960,57	4181,52	41,46	0,8066
Ridge	5964,27	4185,40	41,52	0,8064
KNN	7426,94	4395,66	40,04	0,6998
SVR	14262,59	8957,62	94,14	-0,1070



Şekil 6. Test Kümesinde En Yüksek Performans Gösteren Üç Modelin (a – GB; b – RF; c – XGBoost) Gerçek–Tahmin Karşılaştırmaları (Sol) ve Artık Dağılımları (Sağ)

Hiperparametre Optimizasyon Sonuçları

En iyi model olarak belirlenen Gradient Boosting için RandomizedSearchCV yöntemiyle hiperparametre optimizasyonu yapılmıştır. 40 farklı parametre kombinasyonu değerlendirilmiş ve her kombinasyon 5-katlı çapraz doğrulama ile test edilmiştir. Böylece toplam 200 model eğitimi gerçekleştirilmiştir. Optimizasyon sonucunda en iyi parametre kombinasyonu Tablo 6'da verilmiştir.

Tablo 6. Gradientboosting Modelinin En İyi Hiperparametre Değerleri

Parametre	Değer
subsample	0,6
n_estimators	800
min_samples_split	10
min_samples_leaf	4
max_features	None

Bu parametre değerleri ile eğitim verisi için R^2 değeri 0,8444 seviyesine yükselmiştir. Hiperparametre optimizasyonu yapılan modelin test setindeki performansı RMSE=4289,87, MAE=2536,64, MAPE=30,57 ve R^2=0,8999 olarak elde edilmiştir. Bu sonuçlara bakıldığında optimizasyon sonrası test setinde R^2 değeri ihmal edilebilir düzeyde azalmıştır (ΔR^2=−0,0008). RMSE ve MAE değerlerinde çok sınırlı bir artış olmasına karşılık MAPE değerinde belirgin bir şekilde iyileşme sağlanmıştır (32,99 → 30,57, ΔMAPE=−2,42). Özetle, optimizasyonun özellikle göreli hata açısından tahmin kalitesini artırdığı görülmüştür. GB modeli için tüm performans ölçütlerinin test setindeki optimizasyon öncesi ve sonrası değerleri ve fark değerleri Tablo 7'de verilmiştir.

Tablo 7. Test Setinde GB Modelinin Hiperparametre Optimizasyonu Öncesi ve Sonrası Performans Karşılaştırması

Metrik	Optimizasyon Öncesi	Optimizasyon Sonrası	Fark (Δ)
RMSE	4271,88	4289,87	+17,98
MAE	2535,02	2536,64	+1,62
MAPE	**32,99**	**30,57**	**-2,42**
R2	0,9007	0,8999	-0,0008

Optimizasyon sonrası elde edilen sonuçlar detaylı olarak incelendiğinde, farklı hata metriklerinin model davranışını farklı açılardan yansıttığı görülmektedir. RMSE ve MAE değerlerinde gözlenen sınırlı artış, modelin bazı gözlemler için mutlak hata miktarının bir miktar arttığını göstermektedir. Bu durum, hiperparametre optimizasyonu ile modelin eğitim verisine aşırı uyumunun azaltılması ve daha genel bir yapı kazanması ile ilişkilendirilebilir.

Buna karşılık, MAPE değerindeki belirgin iyileşme, modelin özellikle göreli hata açısından daha dengeli tahminler ürettiğini göstermektedir. Bu durum, yüksek değerli gözlemlerin etkisinin dengelenmesi ve modelin veri dağılımına daha uyumlu hale gelmesi ile açıklanabilir. Özellikle sigorta primleri gibi geniş aralıkta değişen hedef değişkenlerde, göreli hata ölçütleri model performansını daha anlamlı şekilde yansıtmaktadır.

R^2 değerinde gözlenen çok sınırlı düşüş ise modelin açıklayıcılık gücünde anlamlı bir kayıp olmadığını, buna karşın daha dengeli ve genellenebilir bir tahmin performansı elde edildiğini göstermektedir. Bu bulgular, hiperparametre optimizasyonunun modelin bias-variance dengesini iyileştirdiğini ve farklı performans metrikleri birlikte değerlendirildiğinde daha sağlıklı sonuçlar elde edildiğini ortaya koymaktadır.

Duyarlılık Analizi Sonuçları

Aykırı değerlerin model performansına etkisini değerlendirmek amacıyla "bmi" ve hedef değişken "charges" sütunlarına %1–%99 aralığında uç kırpma uygulanmış ve tüm modeller aynı eğitim/test setiyle yeniden değerlendirilmiştir. Elde edilen sonuçlar Tablo 8'de verilmiştir.

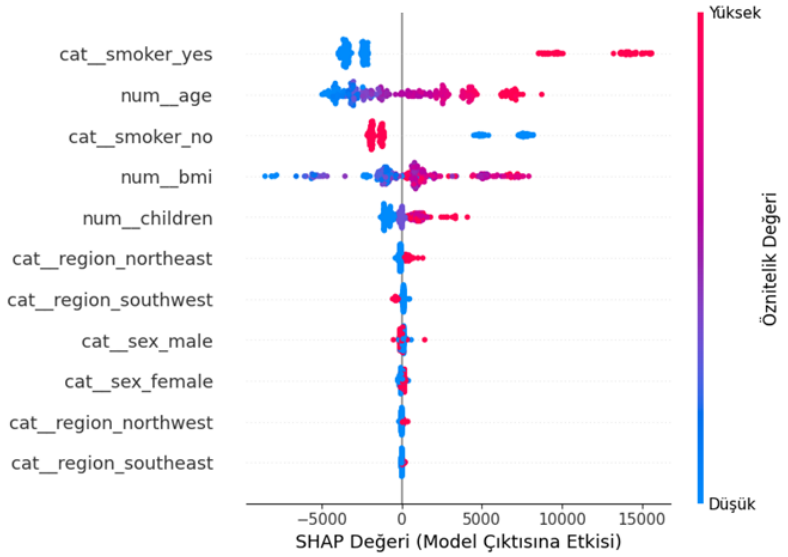
Uç kırpma senaryosunda en iyi performans yine GB ile elde edilmiştir (R^2=0,9107, RMSE=3924,67, MAE=2307,24, MAPE=30,43%). Ham veri senaryosuna kıyasla R^2 değeri 0,01 artmıştır (0,9007 → 0,9107). Benzer şekilde XGBoost'ta R^2 0,0086, RF modelinde R^2 0,0092 artmıştır. Bu sonuçlar, uç gözlem değerlerin kırpılmasının hata metriklerini olumlu yönde etkilediğini göstermektedir.

Tablo 8. Uç Kırpma Uygulanmış Veri Setinde Test Performans Sonuçları

Model	RMSE_test_uc	MAE_test_uc	MAPE_test_uc	R2_test_uc
GradientBoosting	**3924,67**	**2307,24**	**30,43**	**0,9107**
XGBoost	4253,27	2612,91	34,19	0,8951
RandomForest	4289,29	2379,64	34,50	0,8933
LinearRegression	5551,11	4049,82	41,03	0,8213
ElasticNet	5554,94	4054,27	41,10	0,8211
Ridge	5558,31	4058,13	41,16	0,8208
KNN	6859,67	4264,48	39,73	0,7272
SVR	13814,79	8816,71	93,38	-0,1064

Açıklanabilirlik Analiz Sonuçları

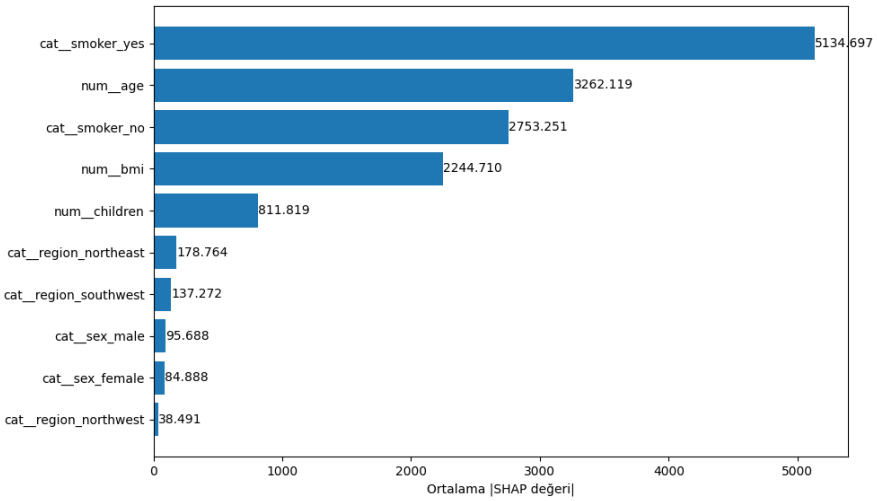
Bu bölümde, test kümesinde en yüksek performansı gösteren Gradient Boosting modeli için elde edilen tahminlerin yorumlanabilirliğini artırmak amacıyla uygulanan SHAP analizi sonuçları verilmiştir. Gradient Boosting modelinin ağaç tabanlı bir topluluk yöntemi olması nedeniyle, SHAP değerleri hesaplanırken hesaplama açısından verimli ve teorik olarak tutarlı sonuçlar sunan TreeSHAP yaklaşımı kullanılmıştır. Elde edilen SHAP değerleri grafiği Şekil 7'de verilmiştir.



Şekil 7. Gradient Boosting Modeli için Treeshap Yöntemi Kullanılarak Elde Edilen SHAP Özet Grafiği

Şekil 7'deki grafiğe bakıldığında, sigara kullanım durumunun sigorta primlerini belirlemede en etkili değişken olduğu görülmektedir. Sigara kullanan bireyler için hesaplanan pozitif SHAP değerlerinin oldukça yüksek olması, bu değişkenin model çıktısını güçlü biçimde artırdığını göstermektedir. Buna karşılık, sigara kullanmayan bireyler model tarafından sistematik olarak daha düşük primlerle ilişkilendirilmiştir. Yaş değişkeni, sigara kullanımından sonra en etkili ikinci faktör olup, yaşın artmasıyla birlikte sigorta primlerinde belirgin bir artış gözlenmektedir. Vücut kitle indeksi de model tahminlerine anlamlı düzeyde katkı sağlamakta; yüksek *bmi* değerleri daha yüksek sigorta maliyetleri ile ilişkilendirilmektedir. Çocuk sayısının etkisi sınırlı olmakla birlikte pozitif yönde olup, primler üzerinde ikincil bir rol oynamaktadır. Buna karşın cinsiyet ve bölge gibi demografik değişkenlerin SHAP değerlerinin sıfır etrafında yoğunlaştığı ve model karar mekanizmasında görece düşük etkiye sahip olduğu görülmektedir.

Şekil 8'de sunulan ortalama mutlak SHAP değerlerine dayalı değişken önem grafiği ise Gradient Boosting modelinin tahminlerini hangi değişkenlerin ne ölçüde etkilediğini nicel olarak ortaya koymaktadır.



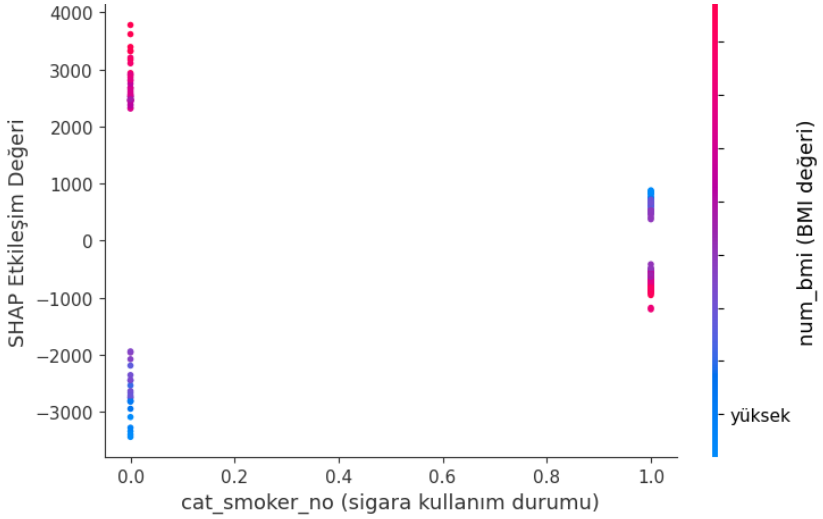
Şekil 8. Değişkenlerin Model Çıktısına Ortalama Mutlak Katkısı

SHAP analizi, bu çalışmada model çıktılarının yorumlanabilirliğini desteklemek amacıyla kullanılmıştır. Elde edilen bulgular, özellikle sigara kullanımının prim tutarı üzerindeki baskın etkisini doğrulamakta olup, bu durum keşifsel veri analizi ve model performans sonuçları ile de tutarlılık göstermektedir. Kullanılan veri setinin görece düşük boyutlu ve değişkenler arası yüksek ayrışabilirlik içeren yapısı, tek bir değişkenin (sigara kullanımı) baskın hâle gelmesine yol açmaktadır. Bu durumun yorumlanabilirlik analizleri üzerindeki etkisini değerlendirmek ve bulguların genellenebilirliğini sınamak amacıyla, farklı bir öznitelik yapısına sahip ikinci bir veri seti analize dahil edilmiş; bu veri setine ilişkin SHAP bulguları 'Genellenebilirlik Analizi' başlığı altında sunulmuştur.

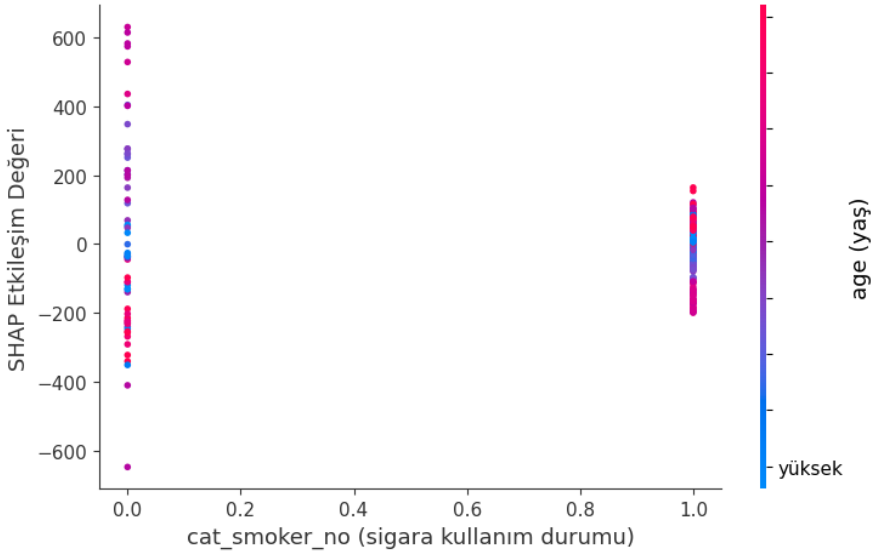
Bunun yanı sıra, modelin yalnızca bireysel değişken etkileri ile değil değişkenler arası etkileşimleri de ne ölçüde yakalayabildiğini incelemek amacıyla SHAP etkileşim analizi gerçekleştirilmiştir. Bu analiz sayesinde, özellikle sigara kullanımı ile diğer değişkenler arasındaki ortak etkilerin model çıktıları üzerindeki yansımaları daha detaylı biçimde değerlendirilmiştir.

Etkileşim Etkileri Analizi

Gerçekleştirilen SHAP etkileşim analizi sonuçları, modelin değişkenler arası ilişkileri ne ölçüde yakalayabildiğini ortaya koymak amacıyla incelenmiştir. Bu kapsamda, özellikle model çıktısı üzerinde belirleyici olduğu düşünülen değişkenler arasındaki etkileşimler detaylı olarak analiz edilmiştir. Elde edilen etkileşim grafikleri, değişkenlerin yalnızca tekil etkilerini değil, aynı zamanda birlikte nasıl etki oluşturduklarını da göstermektedir. Bu doğrultuda, sigara kullanımı ile beden kitle indeksi (BMI) ve yaş değişkenleri arasındaki etkileşimler ön plana çıkmaktadır. İlgili etkileşimlere ait grafikler Şekil 9 ve Şekil 10'da sunulmuştur.



Şekil 9. Sigara Kullanımı ile Beden Kitle İndeksi (BMI) Değişkeni Arasındaki SHAP Etkileşim Grafiği



Şekil 10. Sigara Kullanımı ile Yaş Değişkeni Arasındaki SHAP Etkileşim Grafiği

Şekil 9 ve Şekil 10'daki etkileşim grafikleri incelendiğinde, sigara kullanımı ile diğer değişkenler arasında belirgin etkileşimler olduğu görülmektedir. Özellikle sigara kullanımı ile beden kitle indeksi (BMI) arasındaki etkileşim dikkat çekicidir. Sigara kullanan bireylerde BMI değerinin artmasıyla birlikte prim tutarındaki artışın daha belirgin hale geldiği gözlemlenirken, sigara kullanmayan bireylerde bu etkinin daha sınırlı olduğu görülmektedir. Bu durum, risk faktörlerinin birlikte değerlendirilmesi gerektiğini ortaya koymaktadır.

Benzer şekilde, sigara kullanımı ile yaş değişkeni arasındaki etkileşim incelendiğinde, yaşın prim üzerindeki etkisinin sigara kullanım durumuna bağlı olarak farklılaştığı görülmektedir. Özellikle ileri yaş grubundaki sigara kullanan bireylerde prim değerlerinin daha yüksek seviyelere ulaştığı dikkat çekmektedir. Elde edilen bu bulgular, sigorta primlerinin yalnızca bireysel değişkenlerin etkisiyle değil, değişkenler arası etkileşimler ile birlikte şekillendiğini göstermektedir. Bu durum, ağaç tabanlı topluluk modellerinin karmaşık ve doğrusal olmayan ilişkileri başarılı bir şekilde yakalayabildiğini ortaya koymaktadır.

Genellenebilirlik Analizi

Çalışmada elde edilen bulguların tek bir veri setine özgü olup olmadığını değerlendirmek ve önerilen değerlendirme çerçevesinin genellenebilirliğini sınamak amacıyla, farklı bir öznitelik yapısına sahip ikinci ve bağımsız bir veri seti analize dahil edilmiştir. Bu veri seti, Kaggle üzerinde açık erişimli olarak paylaşılan ve literatürde de kullanılan (Orji & Ukwandu, 2023; Liu & Zhang, 2026) "Medical Insurance Premium Prediction" veri seti olup 986 gözlem ve 11 değişken içermektedir. Birincil veri setinden farklı olarak bu set, demografik bilgilerin (yaş, boy, kilo) yanı sıra ağırlıklı olarak sağlık durumu göstergelerinden (diyabet, tansiyon problemi, organ nakli, kronik hastalık, ailede kanser öyküsü, geçirilen büyük ameliyat sayısı) oluşmaktadır. Hedef değişken, bireyin yıllık sağlık sigortası prim tutarıdır.

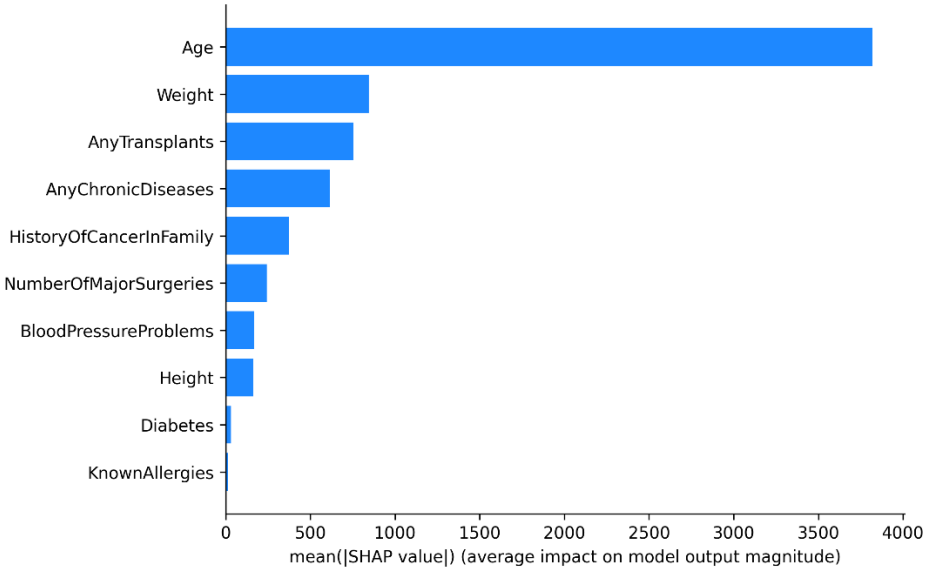
Birincil analizle metodolojik tutarlılığı korumak amacıyla aynı protokol uygulanmıştır: veri seti %80 eğitim ve %20 test olarak ayrılmış, sayısal öznitelikler yalnızca eğitim verisi üzerinde öğrenilen dönüşümlerle ölçeklenmiş (veri sızıntısı önlenmiş), modeller eğitim kümesinde 5-katlı çapraz doğrulama ile karşılaştırılmış ve nihai başarı hold-out test kümesinde RMSE, MAE, MAPE ve R^2 ölçütleriyle raporlanmıştır. Aynı sekiz regresyon modeli (LR, Ridge, ElasticNet, KNN, SVR, RF, GB, XGBoost) değerlendirilmiştir. Test kümesi sonuçları Tablo 9'da sunulmaktadır.

Tablo 9 incelendiğinde, en yüksek başarının R^2=0,8755 değeriyle Random Forest modeliyle elde edildiği; bu modeli Gradient Boosting (R^2=0,8554) ve XGBoost (R^2=0,7987) modellerinin izlediği görülmektedir. Beş katlı çapraz doğrulama da aynı sıralamayı doğrulamıştır (RF için R^2=0,751±0,092). Böylece, birincil veri setinde gözlemlenen temel bulgu olan ağaç tabanlı topluluk modellerinin doğrusal ve örnek/çekirdek tabanlı yöntemlere kıyasla üstün performansı, tamamen farklı bir öznitelik yapısına sahip bu bağımsız veri setinde de korunmuştur. Bu durum, önerilen değerlendirme çerçevesinin tek bir veri setine bağımlı olmadığını ve genellenebilir olduğunu göstermektedir. Ayrıca, prim tutarlarının sıfıra yakın değer içermemesi nedeniyle bu veri setinde MAPE ölçütü daha kararlı ve

yorumlanabilir değerler üretmiştir (en iyi model için %4,52). En iyi performansı gösteren Random Forest modeli için yürütülen SHAP analizi sonuçları Şekil 11'de sunulmaktadır.

Tablo 9. İkinci Veri Seti Üzerinde Elde Edilen Test Performans Sonuçları

Model	RMSE_test	MAE_test	MAPE_test	R2_test
RandomForest	2304,50	1034,75	4,52	0,8755
GradientBoosting	2483,33	1501,73	6,37	0,8554
XGBoost	2929,51	1340,52	6,12	0,7987
LinearRegression	3495,95	2586,23	10,96	0,7134
Ridge	3496,89	2587,22	10,97	0,7132
ElasticNet	4176,47	3176,36	14,21	0,5910
KNN	4396,19	3024,24	13,23	0,5468
SVR	6689,94	5306,18	23,16	-0,0495



Şekil 11. İkinci Veri Seti için Random Forest Modeline ait Ortalama Mutlak SHAP Değerleri

Bu veri setinde öznitelik öneminde en yüksek etkiyi yaş (Age) gösterirken, bunu kilo (Weight), organ nakli (AnyTransplants), kronik hastalık (AnyChronicDiseases) ve ailede kanser öyküsü (HistoryOfCancerInFamily) gibi sağlık durumu göstergeleri izlemektedir. Bu dağılmış önem profili, önerilen yorumlanabilirlik yaklaşımının, etkinin çok sayıda öznitelik arasında paylaşıldığı farklı veri yapılarında da tutarlı ve anlamlı sonuçlar ürettiğini göstermektedir. Yaşın en etkili belirleyici olarak öne çıkması, aynı veri setini kullanan Liu & Zhang (2026) çalışmasının bulgularıyla da tutarlıdır.

SONUÇLAR

Bu çalışmada, Kaggle "Insurance Premium Prediction" veri seti kullanılarak sekiz farklı makine öğrenmesi tabanlı regresyon modeliyle sağlık sigortası prim tutarları tahmin edilmiştir. Modeller, ortak bir deneysel protokol altında 5-katlı çapraz doğrulama ve hold-out test kümesiyle, RMSE, MAE, MAPE ve R^2 metrikleri üzerinden karşılaştırılmıştır.

Elde edilen bulgular, ağaç tabanlı topluluk yaklaşımlarının (özellikle Gradient Boosting, ardından Random Forest ve XGBoost) hem çapraz doğrulama hem de test verisinde klasik doğrusal modeller ve örnek/çekirdek tabanlı yöntemlere kıyasla belirgin biçimde daha yüksek R^2 değerleri ve daha düşük hata metrikleri sunduğunu göstermektedir; örneğin test kümesinde Gradient Boosting ile $R^2 \approx 0{,}90$ seviyesine ulaşılmıştır.

En iyi model için RandomizedSearchCV ile yapılan hiperparametre optimizasyonu, özellikle göreli hata (MAPE) açısından anlamlı iyileşme sağlamış, mutlak hata metriklerinde ise yalnızca sınırlı değişimlere yol açmıştır. Uç kırpma temelli duyarlılık analizi hem ağaç tabanlı hem de doğrusal modellerde R^2 artışı ve hata metriklerinde azalma

sağlayarak uç değerlerin model performansı üzerindeki olumsuz etkisini doğrulamıştır. Böylece çalışma, prim tahmininde yalnızca doğruluk değil, aynı zamanda model seçimi, optimizasyon ve sağlamlık boyutlarını birlikte ele alan bütüncül bir çerçeve sunmaktadır.

Çalışmanın bulgularının tek bir veri setine bağımlı olup olmadığını sınamak amacıyla, farklı bir öznitelik yapısına sahip ikinci ve bağımsız bir veri seti üzerinde genellenebilirlik analizi gerçekleştirilmiştir. Ağaç tabanlı topluluk modellerinin üstünlüğü bu bağımsız veri setinde de korunmuş; ayrıca öznitelik önemlerinin tek bir baskın değişkende toplanmayıp çok sayıda demografik ve sağlık göstergesi arasında dağıldığı gözlemlenmiştir. Bu sonuçlar, önerilen değerlendirme çerçevesinin belirli bir veri yapısına özgü olmadığını ve farklı koşullar altında da tutarlı biçimde çalıştığını göstermektedir.

Çalışmanın bazı sınırlılıkları bulunmaktadır. Birincil analizde kullanılan veri setinin görece küçük ve düşük boyutlu olması, bulguların genellenebilirliği açısından bir kısıt oluşturmaktadır; bu sınırlılık ikinci ve bağımsız bir veri seti ile yürütülen doğrulama analiziyle kısmen giderilmiş olsa da sağlık sigortası gibi kişisel ve ekonomik bilgi içeren veri setlerine erişimin sınırlı olması, daha büyük ve çeşitli veriler üzerinde çalışmayı güçleştirmektedir. Ayrıca, kurumsal kaynaklardan temin edilen verilerde eksik, düzensiz veya tutarsız kayıtlarla karşılaşma olasılığı da dikkate alınması gereken bir kısıt olarak değerlendirilmektedir.

Gelecek çalışmalarda, çok merkezli veya ülke bazlı büyük ölçekli veri setlerinin kullanılması, geliştirilen modellerin genellenebilirliğinin daha geniş ölçekte değerlendirilmesine olanak tanıyacaktır. Bunun yanı sıra, derin öğrenme mimarilerinin, hibrit topluluk öğrenme yapılarının veya otomatik öznitelik seçimi içeren yöntemlerin bu çerçeveye entegre edilmesi başarım ve esneklik açısından ek kazanımlar sağlayabilir. Uç kırpma dışındaki sağlam ön işleme stratejilerinin sistematik biçimde karşılaştırılması, uç değer yönetimi ile model başarımı arasındaki ilişkinin daha ayrıntılı anlaşılmasına katkı sunacaktır. Son olarak, SHAP, ICE, LIME ve permütasyon temelli önem ölçütlerinin birlikte ele alındığı karşılaştırmalı analizler hem karar vericiler hem de düzenleyici kurumlar için şeffaf ve güvenilir prim tahmin sistemlerinin geliştirilmesine zemin hazırlayacaktır.

Yapay Zekâ Katkı Beyanı

Bu makale tamamen herhangi bir yapay zeka aracının yardımı olmadan yazılmış, düzenlenmiş, analiz edilmiş ve hazırlanmıştır. Metin, veri analizi ve şekiller dahil tüm içeriğin yalnızca yazarlar tarafından oluşturulduğunu beyan ederim.

KAYNAKLAR

Ali, Z. A., Abduljabbar, Z. H., Tahir, H. A., Sallow, A. B., & Almufti, S. M. (2023). eXtreme gradient boosting algorithm with machine learning: A review. *Academic Journal of Nawroz University*, *12*(2), 320-334. https://cir.nii.ac.jp/crid/1360021397008256384

Awad, M., & Khanna, R. (2015). *Support vector regression. In Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 67-80). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4302-5990-9_4

Bau, Y. T., & Hanif, S. A. M. (2024). Comparative Analysis of Machine Learning Algorithms for Health Insurance Pricing. *JOIV: International Journal on Informatics Visualization, 8*(1), 481-491. https://doi.org/10.62527/joiv.8.1.2282

Bayo, A. K., Rafiu, A. B., Funmilayo, A. T., & Oluyemi, O. I. (2021). Investigating the impact of multicollinearity on linear regression estimates. *Malaysian Journal of Computing (MJoC), 6*(1), 698-714. https://doi.org/10.24191/mjoc.v6i1.10540

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54*(3), 1937-1967. https://doi.org/10.1007/S10462-020-09896-5

Bhongade, A., Dubey, Y., Palsodkar, P., & Fulzele, P. (2024, December). Explainable AI Model for Medical Insurance Premium Prediction using Ensemble Learning with SHAP Analysis. In *2024 IEEE Pune Section International Conference (PuneCon)* (pp. 1-6). IEEE. https://doi.org/10.1109/PuneCon63413.2024.10895118

Billa, M. M., & Nagpal, T. (2024). Medical insurance price prediction using machine learning. *J. Electr. Syst, 20*(7s), 2270-2279. https://doi.org/10.52783/jes.3962

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. https://doi.org/10.1023/a:1010933404324

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Dash, C. S. K., Behera, A. K., Dehuri, S., & Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal, 6*, 100164. https://doi.org/10.1016/j.dajour.2023.100164

Dutta, S., Bose, P., & Bandyopadhyay, S. K. (2023). Forecasting health insurance premium using machine learning approaches. *Asia-Pacific Journal of Science and Technology, 28*(06), APST-28. https://doi.org/10.14456/apst.2023.96

Farhadi, Z., Belaghi, R. A., & Alma, O. G. (2019). Analysis of penalized regression methods in a simple linear model on the high-dimensional data. *American Journal of Theoretical and Applied Statistics, 8*(5), 185-192. https://doi.org/10.11648/j.ajtas.20190805.14

Franceschi, L., Donini, M., Perrone, V., Klein, A., Archambeau, C., Seeger, M., ... & Frasconi, P. (2025). Hyperparameter optimization in machine learning. *Foundations and Trends® in Machine Learning, 18*(6), 975-1109. https://doi.org/10.1561/2200000088

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis, 38*(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data, 11*(1), 113. https://doi.org/10.1186/s40537-024-00973-y

Hu,Y. (2024). Comparison and Analysis of the Effectiveness of Linear Regression, Decision Tree, and Random Forest Models for Health Insurance Premium Forecasting. *Advances in Economics, Management and Political Sciences, 79*, 347-353. https://doi.org/10.54254/2754-1169/79/20241754

Insurance Premium Prediction Dataset, Erişim Tarihi: 11.09.2025. Erişim Bağlantısı: https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction

Irmak, T. (2025, July). Analysis of Order Cancellation Rates and Feature Weighting with Machine Learning: An E-Commerce Case Study. In *International Conference on Intelligent and Fuzzy Systems* (pp. 763-770). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-98304-7_82

Kapse, M., Sharma, V., Vidhale, R., & Vellanki, V. (2025). Customization of health insurance premiums using machine learning and explainable AI. *International Journal of Information Management Data Insights, 5*(1), 100328. https://doi.org/10.1016/j.jjimei.2025.100328

Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums. *International journal of environmental research and public health, 19*(13), 7898. https://doi.org/10.3390/ijerph19137898

Khajeh, M., Ghaffari-Moghaddam, M., Piri, J., Barkhordar, A., & Şenol, H. (2026). Sustainable Beetroot Bioactives Extraction with NADES-MAE: A Spider Algorithm-Optimized Hybrid SVR-XGBoost Model. *Applied Food Research,* 101940. https://doi.org/10.1016/j.afres.2026.101940

Kulkarni, M., Mesharam, D. D., Patil, B., More, R., Sharma, M., & Patange, P. (2022). Medical insurance cost prediction using machine learning. *International Journal for Research in Applied Science & Engineering Technology, 10*. https://doi.org/10.22214/ijraset.2022.47923

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

Lin, K., Diab, L. S., Zhang, Y., & ELshreif, A. S. (2026). Analyzing the performance of machine learning techniques: Applications in dance education and the radiation sector. *Journal of Radiation Research and Applied Sciences, 19*(3), 102522. https://doi.org/10.1016/j.jrras.2026.102522

Liu, Y., & Zhang, S. (2026). Application of Machine Learning Algorithms for Medical Insurance Cost Prediction and Feature Analysis. *Emerging Markets Finance and Trade, 62*(1), 1-17. https://doi.org/10.1080/1540496X.2025.2522233

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*. https://doi.org/10.48550/arXiv.1705.07874

Mahesh, M., Priyatham, B. S., SaiKumar, G., & Reddy, K. V. V. (2025, May). Predictive Analytics for Medical Insurance Premiums. In *2025 5th International Conference on Pervasive Computing and Social Networking (ICPCSN)* (pp. 303-309). IEEE. https://doi.org/10.1109/ICPCSN65854.2025.11034884

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics, 7*, 21. https://doi.org/10.3389/fnbot.2013.00021

Orji, U., & Ukwandu, E. (2024). Machine learning for an explainable cost prediction of medical insurance. *Machine learning with applications, 15*, 100516. https://doi.org/10.1016/j.mlwa.2023.100516

Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and translational science, 17*(11), e70056. https://doi.org/10.1111/cts.70056

Rao, D. M., Somalaraju, L. K. R., Bodagala, V. S. M., & Edapalapati, V. B. J. S. (2025, May). Predicting Medical Insurance Costs: A Machine Learning Approach for Smarter Risk Assessment. In *2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICCRTEE64519.2025.11052963

Saputro, D. R. S., Wahyu, N. L., & Widyaningsih, Y. (2025). Performance of Ridge Regression, Least Absolute Shrinkage and Selection Operator, and Elastic Net in Overcoming Multicollinearity. *Journal of Multidisciplinary Applied Natural Science, 5*(2), 370-382. https://doi.org/10.47352/jmans.2774-3047.251

Saraiva, F. M. G., Guimarães, A. C. R., da Silva Júnior, O. S., Monteiro, S. N., & Coelho, L. M. (2026). Performance analysis of Gradient Boosting Regressor Using K-fold and Group K-fold for Predicting Permanent Deformation in Tropical Soils. *Journal of Materials Research and Technology*. https://doi.org/10.1016/j.jmrt.2025.12.152

Sharma, A., Malik, A., Sharma, S., Batra, I., & Hussain, M. (2024, November). Empowering Healthcare Accessibility: Machine Learning-Driven Predictive Modeling for Health Insurance Premiums in India. In *2024 IEEE 9th International Conference on Engineering Technologies and Applied Sciences (ICETAS)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICETAS62372.2024.11119999

Sijie, L., Sia, F., Alfred, R., & Moung, E. G. (2024, August). Ensemble Machine Learning Method for Health Insurance Premium Prediction. In *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)* (pp. 642-646). IEEE. https://doi.org/10.1109/IICAIET62352.2024.10730721

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing, 14*(3), 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Topal, B. (2024). Özel Sağlık Sigortası Primleri Üzerinde Makroekonomik Göstergelerin Etkisi. Yüksek Lisans Tezi, Sağlık Yönetimi Anabilim Dalı, Düzce Üniversitesi.

Usman, M., SIS, D., & Alhaji, B. B. (2021). Comparing the Prediction Accuracy of Ridge, Lasso and Elastic Net Regression Models with Linear Regression using Breast Cancer Data. *Bayero Journal of Pure & Applied Sciences, 14*(2), 134-149. https://doi.org/10.4314/bajopas.v14i2.16

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine, 4*(11), 218. https://doi.org/10.21037/atm.2016.03.37

Zubedi, F., Sartono, B., & Notodiputro, K. A. (2022). Implementation of Winsorizing and random oversampling on data containing outliers and unbalanced data with the random forest classification method. *Jurnal Natural, 22*(2), 108-116. https://doi.org/10.24815/jn.v22i2.25499