

|  |
| --- |
| Betül BİÇER  Mühendislik Fakültesi |
| *Bilgisayar Mühendisliği Bölümü* |
| *TOBB ETÜ* |
| *Ankara, Türkiye* |
| [betul.bicer@etu.edu.tr](mailto:betul.bicer@etu.edu.tr) |

BİL 470

ATP TENİS MAÇLARI DATASETİ İLE YAPAY ÖĞRENME PROJESİ

***Bu çalışmada, maç istatistikleri ve oyuncu özelliklerine dayanarak bir tenis maçının sonucunun tahmini gerçekleştirilmektedir. ATP Tenis Maçları 2020-2024 arası verileri kullanılmıştır. Veri analizi ve temizlemesi yapılıp 3 model ile eğitilmiştir. Makine öğrenme modelleri Destek Vektör Makineleri (SVM), Rastgele Orman (RF) ve Aşırı Gradyan Artırma (XGBoost) karşılaştırmalı olarak sınıflandırma performansları değerlendirilmiştir. SVM en iyi performans gösteren model olmuştur.***

I. Introduction

*A. Problem Tanımı*

Tenis, dünya genelinde popülerliği yüksek olan, iki veya dört oyuncu arasında oynanan bir spordur. Her bir maç, katılan oyuncuların performansı, teknik becerileri ve maç sırasındaki koşullar gibi birçok değişkene bağlı olarak farklı sonuçlar gösterebilir. Bu dinamik ve belirsiz yapı, maç sonuçlarını tahmin etmeyi zor bir hale getirmektedir.

*B. Motivasyon*

Son yıllarda, artan veri toplama teknolojileri ve gelişmiş istatistiksel analiz yöntemleri, bilim insanlarını ve analistleri maç sonuçlarının tahmin edilmesi konusunda daha doğru yöntemler geliştirmeye teşvik etmiştir. Bu tahminler, bahis pazarlarından antrenman programlarının optimize edilmesine kadar geniş bir yelpazede uygulama alanlarına sahiptir.

*C. Sınıflandırma/Regresyon*

Bu çalışmada, maç sonuçlarını tahmin etmek için üç farklı makine öğrenme modeli kullanılmaktadır: Destek Vektör Makineleri (SVM), Rastgele Orman (RF) ve Aşırı Gradyan Artırma (XGBoost). Bu modeller, maçların kazananını veya kaybedenini sınıflandırmak için kullanılmakta olup, her bir modelin maç istatistikleri ve oyuncu özelliklerine dayalı performansı karşılaştırmalı olarak değerlendirilmektedir.

*D. Amaçlar*

Bu çalışmanın amacı, ATP Turu'nda 2000 ile 2024 yılları arasında gerçekleşen maçlara ait verileri kullanarak, makine öğrenme tekniklerinin etkinliğini değerlendirmektir.

Çalışma, önceki literatürde yeterince incelenmemiş olan, farklı makine öğrenme modellerinin tenis maç sonuçları üzerindeki performans farklılıklarını ve özellik önemliliklerini inceleyerek teorik ve pratik katkılarda bulunmayı hedeflemektedir.

*E. Başarım Metrikleri*

* Accuracy
* Precision
* Recall
* F1 Score
* ROC AUC Score

II. Literatür Araştırması

*A. İlgili Çalışmaların Özeti*

**1. A COMPARISON BETWEEN DIFFERENT CLASSIFIERS FOR TENNIS MATCH RESULT PREDICTION** [1]

Bu araştırma, 2019 yılında yapılmış olup tenis maç verileri üzerinden maç sonucu tahmini yapmaktadır. Kullandığı modeller DT, LVQ, SVM’dir. Datayı %70-30 oranında dağıtmıştır. 10 fold-cross validation kullanmıştır. Bu araştırmada, DT 97.87 accuracy ile en iyi tahmini yapan modeldir.

**2. Predicting the Outcome of Tennis Match Based on Match Statistics and Player Characteristics- Carlijn Van Rooij** [2]

Bu araştırma, 2021 yılında yapılmış olup 1991-2016 yılları arasındaki tenis maç verilerini SVM, XGBOOST ve RF modellerini eğiterek maç tahmini yapmıştır. Modeller arasında performans farklarını incelemiştir. SVM modeli %94.41 accuracy ile en iyi sonucu veren modeldir. Break points serve total ve Break points saved öznitelikleri model eğitiminde önemli bir yer kaplamaktadır.

**3. Predicting tennis match using machine learning**[3]

Bu araştırma, 20219-2020 yılları arasında yapılmış olup tenis maçı tahmini için rank baseline, logistic regression ve neural network modelleri kullanmıştır. En iyi sonucu %68.2 accuracy ile neural network modeli vermiştir.

III. Veri Seti ve Özellikler[4]

1. *Veri Kaynağı*

Bu projede kullanılan veri seti, Kaggle’dan alınan Huge Tennis Dataset’tir.

1. *Veri Seti Özeti*

Veri seti, 81 sütun 947720 datadan oluşmaktadır. player\_id, first\_name, last\_name, hand, birth\_date, country\_code, height (cm), age, ace sayısı gibi sütunlar içermektedir.

1. *Öznitelik Açıklamaları*
   * 1. Sayısal Veriler

-rank

-age

- height

- ace

- service point

- break puan sayısı

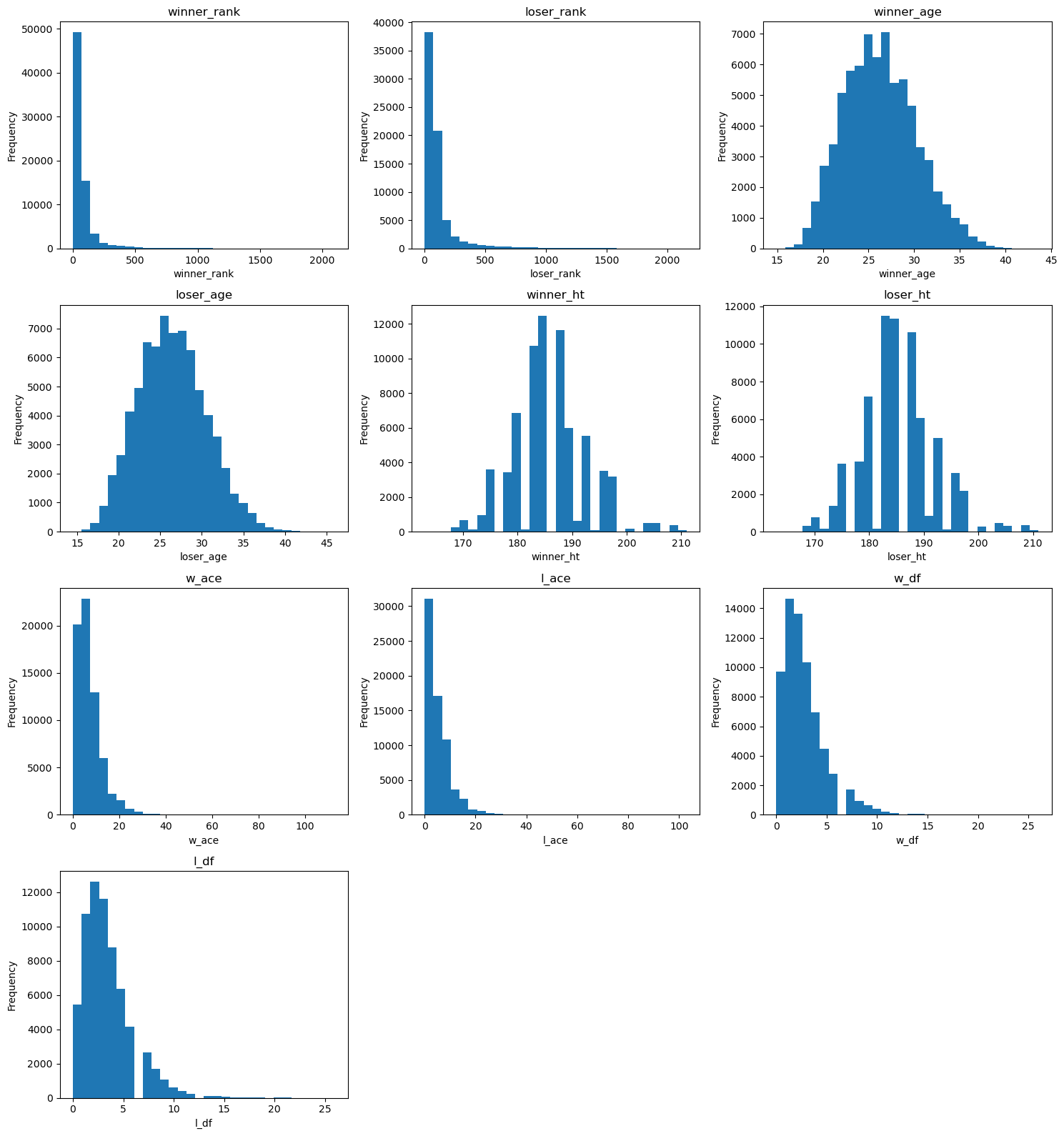
* + 1. Sayısal Olmayan Veriler

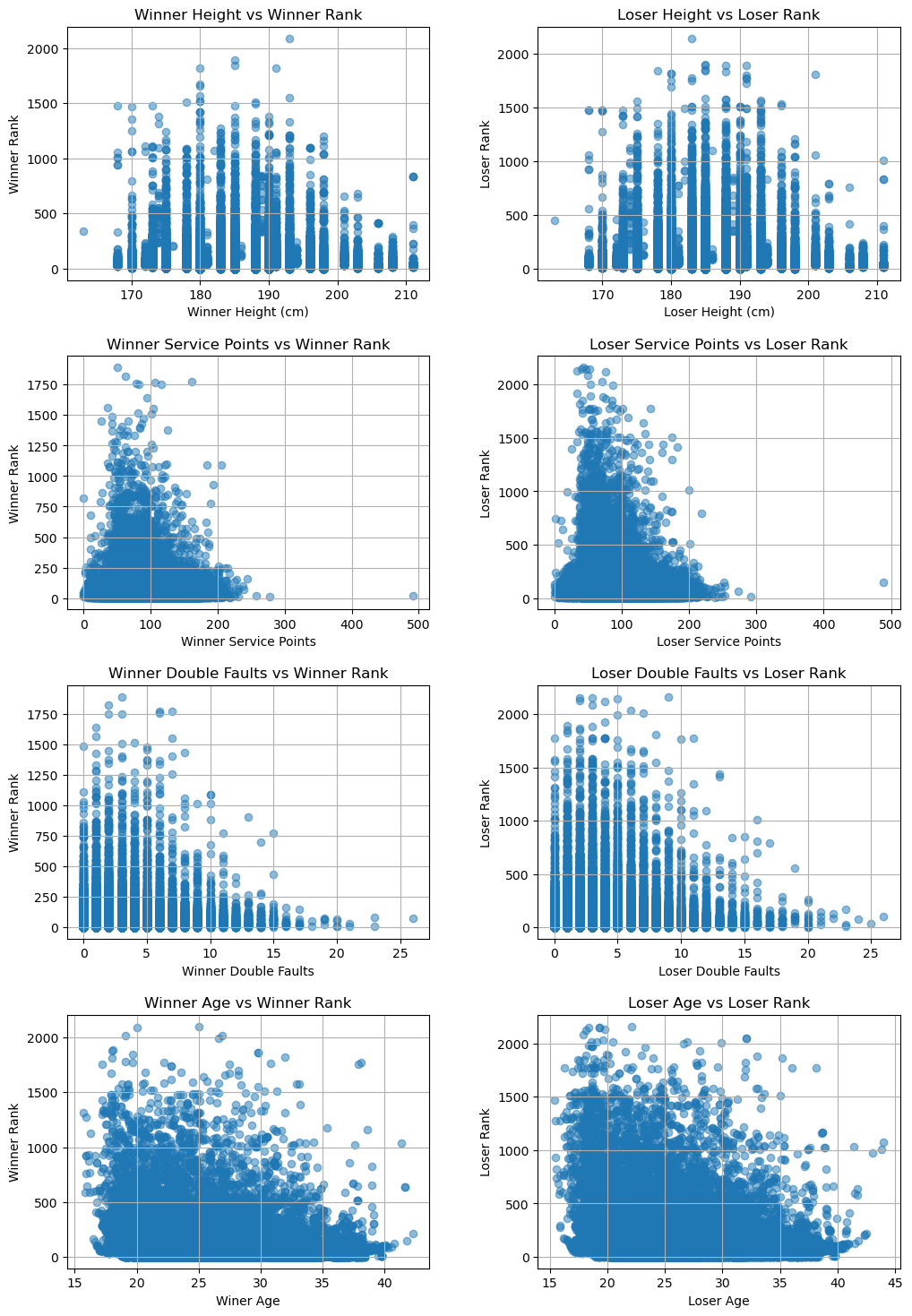
-oyuncunun hangi elini kullandığı

-surface

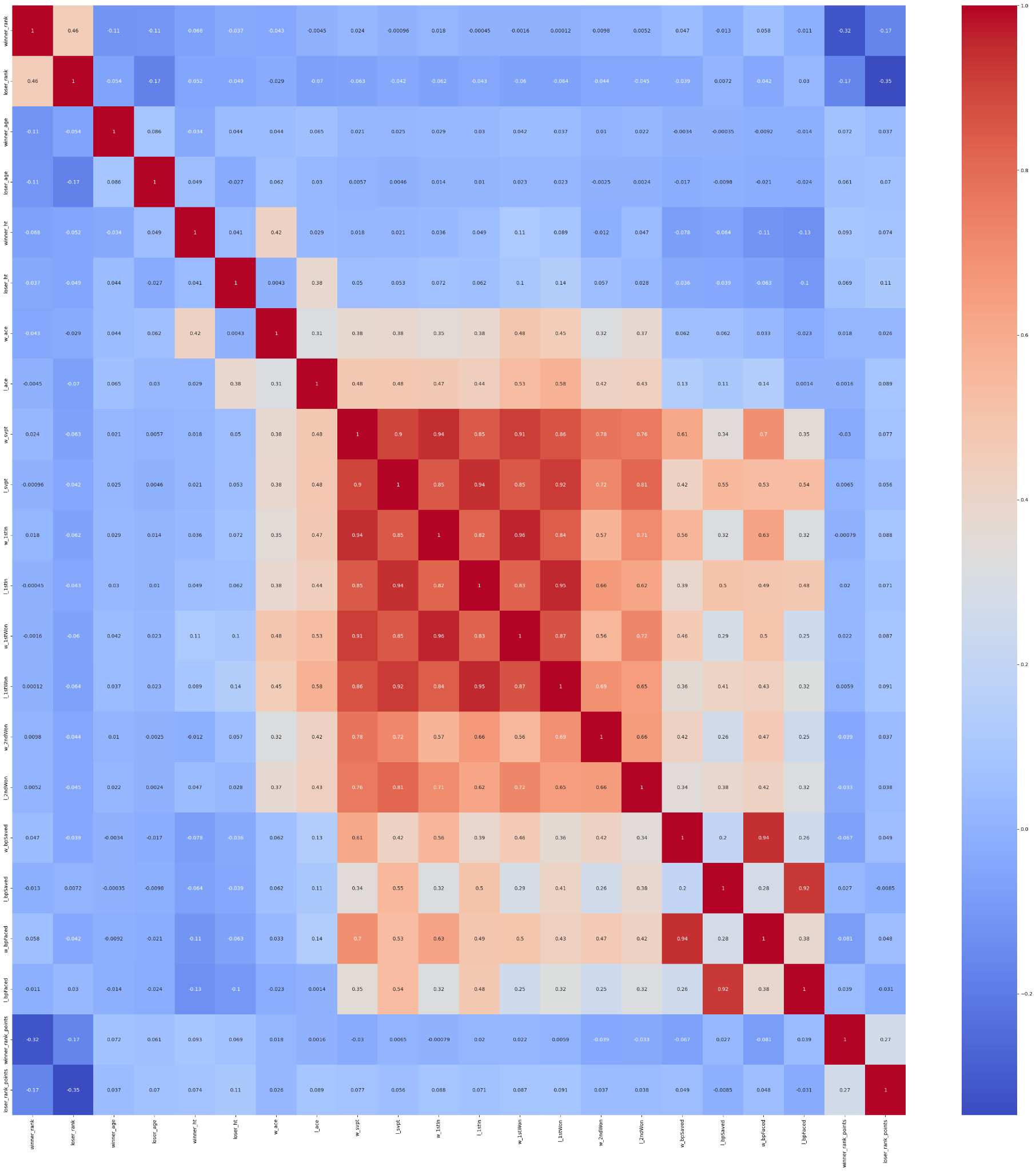
1. *Keşifsel Veri Analizi*

*Önemli featureların histogramları incelenmiştir.*

**



Heatmap: Değişkenler arasındaki korelasyonu göstermek için kullanılmıştır.



1. *Veri Ön İşleme*

* *Eğitimde gereksiz olan sütunlar droplanmıştır.*
* *Kayıp veriler oyuncu bazında özniteliğin ortalaması veya medianı alınarak doldurulmuştur. Kalan kayıp veriler için sütunun tamamının ortalamasına ve medianına bakılarak doldurulma gerçekleştirilmiştir. Kayıp veri sayısının 100’den küçük olduğu durumlarda bu veri satırları silinmiştir.*
* *surface, first\_hand ve second\_hand verilerine one-hot vector uygulanmıştır.*

IV. Kullanılan Modeller

1. *Eğitim ve Doğrulama*

Eğitim ve test verileri %80 ve %20 oranında ayrılmıştır. Çeşitli sınıflandırma modelleri uygulanmıştır. Modelin doğruluğunun ölçülmesi için çeşitli metrikler kullanılmıştır.

1. *Kullanılan Modeller*

* Random Forest
* XGBOOST
* Desicion Tree
* SVM

1. *Model Seçiminin Gerekçesi*

Random Forest, çok sayıda özellik içeren veri setlerinde, özellikler arasındaki karmaşık ilişkileri ve etkileşimleri tespit etmede etkili olduğu ve overfiti engellediği için seçilmiştir.

XGBoost, her iterasyonda en önemli özellikleri seçerek ve diğerlerini yavaş yavaş elemek suretiyle, modelin öğrenme sürecini optimize ettiği için seçilmiştir.

SVM, veri setindeki örnekler arasında optimal bir ayrım hiper-düzlemi bulmaya çalışır. Bu, modelin yeni verilere iyi genelleme yapmasını sağlar, yani bilinmeyen veriler üzerinde de iyi tahminlerde bulunabilir. Tenis gibi yüksek varyans gösteren spor dallarında, bu özellik önemli olduğu için seçilmiştir.

V. Model Performansları ve Karşılaştırılması

1. *Performans Metrikleri*

1-Random Forest

Macro Precision: 0.9107526517745157

Macro Recall: 0.9107540605652359

Macro F1 Score: 0.9107520546383785

Accuracy: 0.9107521005533165

Average Cross Validation Accuracy: 0.91506

2-XGBOOST

Macro Precision: 0.9466158279934371

Macro Recall: 0.9466142920318115

Macro F1 Score: 0.9466149263892196

Accuracy: 0.946615205956691

Average Cross Validation Accuracy:0.94731

3-SVM

Precision:0.95

Recall:0.95

F1 Score: 0.95

Accuracy: 0.95

Average Cross Validation Accuracy:

4-Desicion Tree

Precision:0.82

Recall:0.82

F1 Score: 0.82

Accuracy: 0.95187512808251930.

Average Cross Validation Accuracy:0.82127

1. *Modellerin Karşılaştırılması*

1-Random Forest: Birden fazla karar ağacının sonuçlarını birleştirerek genellikle yüksek doğruluk oranları sağlar. Tek bir karar ağacının overfitting eğilimi, çoklu ağaç yapısı ile dengelenir, bu da modelin genelleme yeteneğini artırır. Ancak, iç içe geçmiş çok sayıda karar ağacı içerdiği için, özellikle büyük ormanlar söz konusu olduğunda, yorumlaması zor olabilir.

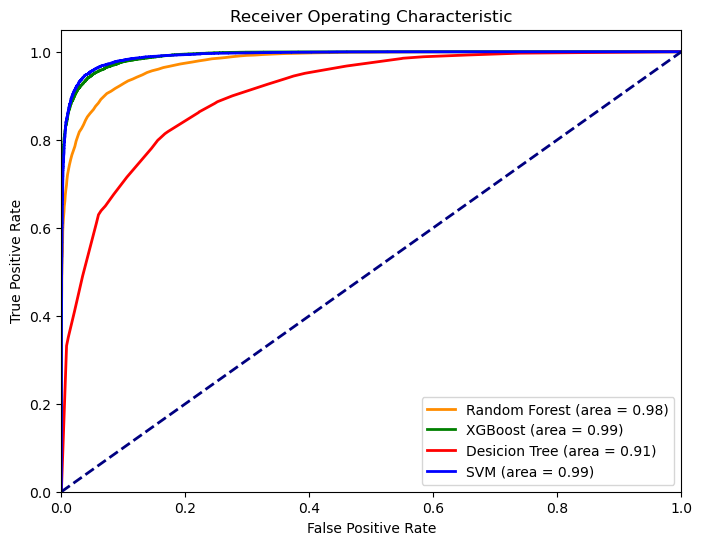
2-XGBOOST: XGBoost, veri setlerini hızlı işleyebilir, paralel işleme ve ağaç kesme (pruning) teknikleri sayesinde eğitim süreci hızlandırılmıştır.XGBoost, overfitting'i önlemek amacıyla L1 ve L2 düzenlileştirme tekniklerini içerir, bu da modelin genelleme yeteneğini artırır. Ancak, yüksek performans, hız, yüksek bellek ve işlem gücü gerektirir. Karar ağaçlarına dayalı olmasına rağmen, birleştirilmiş birçok ağacın sonuçlarını interpret etmek zor olabilir,bu sebeple anlaşılması güçtür.

3-SVM: Bu model farklı veri setleri üzerinde sağlam ve tutarlı sonuçlar üretir. Boyutların artışı, hesaplama karmaşıklığını önemli ölçüde artırmaz. Veri noktaları arasında en büyük marjı sağlayacak şekilde sınıflandırma çizgisini belirler, bu da modelin aşırı uyuma karşı dirençli olmasını sağlar. Ancak, hesaplama maaliyeti yüksek ve yorumlaması zor bir modeldir.

4-Desicion Tree: Karar ağaçları, basit kurallar dizisi şeklinde kararlar alır, bu nedenle modellerin nasıl çalıştığını anlamak ve yorumlamak kolaydır. Modeli oluştururken en informatif özellikleri otomatik olarak seçer. Ancak, eğitim verilerine çok iyi uyum sağlayabilir (aşırı uydurma), bu da genelleştirme kabiliyetini düşürebilir. Veri setindeki küçük değişiklikler modelin yapısını önemli ölçüde değiştirebilir.

1. *İstatistiksel Anlamlılık*

Forest, XGBoost, SVM ve Karar Ağaçları. Sonuçlar göstermektedir ki SVM en yüksek ortalama doğruluk oranına (accuracy) sahiptir, ardından XGBoost, Random Forest ve en son olarak Karar Ağaçları gelmektedir. SVM ve XGBoost modellerinin yüksek performansları, özellikle büyük ve karmaşık veri kümelerini işlerken etkin düzenleme teknikleri ve optimizasyon algoritmaları sayesinde elde edilmiştir. Buna karşın, Karar Ağaçları modelinin daha düşük performans göstermesi overfittingi önlemek için az depth ile oluşturulmuş olmasından kaynaklanmaktadır. İstatistiksel olarak anlamlı sonuçlar elde etmek adına, çeşitli testler ve çapraz doğrulama yöntemleri kullanılmıştır, bu sayede modellerin robustluğu (güvenilirliği) artırılmıştır.



VI. Sonuç

1. *Bulguların Özeti*

Bu projede, 4 farklı model ile ATP tenis maçı dataseti üzerinde yapılan eğitimler ve cross-validation doğrulamaları sonucunda SVM en iyi performansı gösteren modeldir. XGBOOST ikinci sırada yer almaktadır ve eğitim süresi SVM kadar uzun olmamıştır. Random Forest modeli %91,07 ve Desicion Tree modeli %82.12 başarım göstermiştir. Her modelin performansı, accuracy, precision, recall, F1 score ve ROC AUC skoru gibi metrikler kullanılarak ölçülmüştür. En önemli öznitelikler arasında karşılaşılan break pointler(bp\_faced), kazanılan ilk servisler (1stwon), yer almaktadır.

1. *Gelecek Çalışmalar*

Bu projede, öznitelik seçiminde daha farklı parametreler denenerek modelllerin performansına olan etkisi incelenebilir. Bunun yanı sıra farklı modeller denenerek başarımları ölçülebilir. Örneğin YSA modeli ile eğitim sağlanabilir.

Referanslar

1. <https://ejournal.um.edu.my/index.php/MJCS/article/view/17723>
2. <https://arno.uvt.nl/show.cgi?fid=158548>
3. <https://libstore.ugent.be/fulltxt/RUG01/002/945/727/RUG01-002945727_2021_0001_AC.pdf>
4. [Huge Tennis Database (kaggle.com)](https://www.kaggle.com/datasets/guillemservera/tennis)