VERİ MADENCİLİĞİ PROJESİ

Grubun Adı : MacroMinds

Dönem: 2025/2026 Güz Dönem **Sınav Türü :** Vize Sınavı

Proje Bilgileri : Veri Madenciliği Yöntemleriyle Makroekonomik Göstergelerin Bitcoin Fiyat Tahminlemesi Üzerindeki Etkisinin Analizi

Öğrenci Bilgileri

Adı	Soyadı	Öğrenci Numarası	Mail Adress
Abdumajid	Abdulkhaev	22040101002	abdumajidabdulkhaev@stu.topkapi.edu.tr
Khaiitmurod	Khabibullayev	22040101116	khaiitmurodkhabibul1@stu.topkapi.edu.tr
Diyorjon	Ochilov	22040101139	diyorjonochilov@stu.topkapi.edu.tr
Yesset	Yelebayev	22040101123	yessetyelebayev@stu.topkapi.edu.tr
İbrahim Halil	Yanaç	22040101043	ibrahimhalilyanac@stu.topkapi.edu.tr

2) Problem Tanımı

- **İş/Bilimsel Soru:** Bu proje, son yılların en volatil finansal varlıklarından biri olan Bitcoin'in gelecekteki fiyat hareketlerinin tahmin edilip edilemeyeceği sorusuna cevap aramaktadır. Geleneksel teknik analizlerin ötesine geçerek, temel bilimsel soru şudur: **ABD faiz oranları, enflasyon endeksleri, parasal arz gibi makroekonomik göstergeler, Bitcoin fiyatlarını tahmin etmede istatistiksel olarak anlamlı bir güce sahip midir?** Çalışma, bu makroekonomik verilerin farklı zaman ufuklarında (kısa, orta ve uzun vade) fiyat değişim yönünü ve büyüklüğünü ne ölçüde açıklayabildiğini veri madenciliği yöntemleriyle araştırmayı amaçlamaktadır.
- **Görev Türü:** Projenin görevi temel olarak bir **Regresyon** (Regression) problemidir; çünkü gelecekteki sayısal bir fiyat değerini tahmin etmeyi hedeflemektedir. Aynı zamanda, verilerin zamana bağlı ve sıralı olmasından dolayı bu bir **Zaman Serisi** (Time Series) analizidir.
- **Paydaş Değeri (Stakeholder Value):** Bu çalışma, bireysel yatırımcılar ve risk yönetimi fonları için, volatil piyasalarda makroekonomik sinyalleri kullanarak riskten korunma (hedging) ve portföy optimizasyonu sağlama potansiyeline sahiptir.
- **Hedef Değişken(ler):** Çalışmada, farklı yatırım ufuklarını analiz etmek amacıyla dört ayrı hedef değişken tanımlanmıştır. Hepsinin birimi ABD Doları (USD) cinsinden Bitcoin kapanış fiyatıdır.
    - **Target_1d:** Mevcut günden 1 gün sonraki BTC kapanış fiyatı.
    - **Target_7d:** Mevcut günden 7 gün sonraki BTC kapanış fiyatı.
    - **Target_30d:** Mevcut günden 30 gün sonraki BTC kapanış fiyatı.
    - **Target_365d:** Mevcut günden 365 gün sonraki BTC kapanış fiyatı.
- **Başarı Kriterleri:** Modellerin başarısı, aşağıdaki nicel hedeflerle ölçülecektir:
    - **RMSE (Kök Ortalama Kare Hata)** ve **MAE (Ortalama Mutlak Hata)** metriklerini minimize etmek.
    - **R^2 (Belirlilik Katsayısı)** değerini maksimize etmek. Özellikle $R^2 > 0.50$ olan modeller başarılı kabul edilecektir.

- **Direction Accuracy (Yön Tahmin Doğruluğu)** metriğini maksimize etmek. Rastgele tahminden (%50) anlamlı derecede yüksek, özellikle **%55'in üzerinde** bir yön doğruluğu, modelin pratik olarak kullanılabilir olduğunu gösterecektir.
- Temel başarı kriteri, makroekonomik verileri içeren modellerin, sadece finansal verileri içeren temel (baseline) modellerden daha iyi performans göstermesidir.

---

3) Proje Yönetimi

- **Kilometre Taşları ve Zaman Çizelgesi (Gantt Tarzı):**
  - **1. Hafta:** Proje konusunun ve veri setlerinin belirlenmesi, görev dağılımının yapılması. (Tamamlandı)
  - **2. Hafta:** Veri setlerinin indirilmesi, Google Colab ortamının kurulması, veri ön işleme ve keşifsel veri analizi (EDA) için pipeline kodunun hazırlanması. (Dataset: 4067 satır, 73 sütun) (Tamamlandı)
  - **3. Hafta:** Veri birleştirme, eksik veri doldurma ve özellik mühendisliği (lag features) adımlarının tamamlanması. Temel model (Baseline Model) geliştirilmesi. (Tamamlandı)
  - **4-5. Haftalar:** Her grup üyesinin kendi sorumlu olduğu base modelleri, özellik seçimi ve PCA senaryolarını uygulayarak eğitmesi ve metrikleri hesaplaması.
  - **6. Hafta:** Tüm model sonuçlarının birleştirilmesi, performans analizi ve karşılaştırmalı değerlendirmelerin yapılması. En iyi model ve senaryoların belirlenmesi.
  - **7. Hafta:** Proje raporunun son halinin yazılması ve proje sunumunun hazırlanması.
- **Roller ve Sorumluluklar:** Proje ekibi 5 kişiden oluşmaktadır ve her üye farklı model aileleri ve özellik işleme tekniklerinden sorumludur:
  - **Khaiitmurod Khabibullayev:**
    - **Modeller:** LinearRegression, DecisionTreeRegressor
    - **Özellik Seçimi:** SelectKBest (K=30 ve 50)

        - PCA: 10 bileşen
    - **Abdumajid Abdulkhaev:**
        - **Modeller:** RidgeRegression, KNeighborsRegressor
        - **Özellik Seçimi:** Recursive Feature Elimination (RFE)
        - **PCA:** 20 bileşen
    - **Yesset Yelebayev:**
        - **Modeller:** Lasso, Support Vector Regressor (SVR)
        - **Özellik Seçimi:** Lasso-based feature selection
        - **PCA:** 5 bileşen
    - **Diyorjon Ochilov:**
        - **Modeller:** LinearRegression, DecisionTree (farklı parametrelerle)
        - **Özellik Seçimi:** Mutual Information (mutual_info_regression)
        - **PCA:** Uygulanmayacak
    - **İbrahim Halil Yanaç:**
        - **Modeller:** LogisticRegression (Regresyon hedefli), KNeighborsRegressor
        - **Özellik Seçimi:** SelectKBest (farklı K değerleri ile)
        - **PCA:** 15 bileşen
- **Çıktılar:** Proje sonunda üretilecek çıktılar şunlardır:
    - **GitHub Repo Linki:** https://github.com/murat-khabibullayev/Bitcoin-Price-Prediction-Project.git
    - **Final Proje Raporu:** Proje sürecini, metodolojiyi, model sonuçlarını ve yorumlanabilirliği içeren kapsamlı PDF rapor.
    - **Veri & Veri Sözlüğü: Ham Veri Dosyaları (Raw Data):** Projede kullanılan Yahoo Finance ve FRED kaynaklı 17 adet CSV dosyası: BTC-USD.csv (Bitcoin Fiyatı), CPI.csv (Enflasyon), DXY.csv (Dolar Endeksi), ECBDFR.csv, EURUSDX.csv, FEDFUNDS.csv (Faiz Oranları), GCF.csv, GDP.csv (Büyüme),

GSPC.csv (S&P 500), IVV.csv, PAYEMS.csv, STLFSI2.csv (Finansal Stres), UNRATE.csv (İşsizlik), VIX.csv (Volatilite), WALCL.csv, ZQF.csv.

- **İşlenmiş Veri (Processed Data):** Tüm ham verilerin tarih bazlı birleştirilmesi, temizlenmesi ve özellik mühendisliği işlemlerinden geçirilmesiyle oluşturulan, model eğitiminde kullanılan nihai dosya: **merged_data.csv** (4067 Satır, 73 Sütun).
- **Veri Sözlüğü:** Değişkenlerin (örneğin; DXY: Dolar Endeksi, UNRATE: İşsizlik Oranı) açıklamalarını içeren referans belgesi.

---

4) İlgili Çalışmalar (Mini Literatür İncelemesi)

Bitcoin fiyat tahmini üzerine yapılan akademik çalışmalarda genellikle LSTM ve GRU gibi derin öğrenme modelleri öne çıksa da, bu proje "Temel Modellerin" makroekonomik verilerle ne kadar başarılı olabileceğini sınamaktadır**.**

1. **McNally et al. (2018):** Bitcoin fiyat tahmininde RNN ve LSTM modellerini kullanmış, %52 doğruluk elde etmiştir. *Farkımız:* Biz sadece teknik verileri değil, FED faiz oranları gibi dışsal faktörleri de modele dahil ediyoruz.
2. **Sovbetov (2018):** Kripto paraların makroekonomik faktörlerle (SP500, Altın vb.) ilişkisini incelemiştir. *Farkımız:* Bizim çalışmamızda 17 farklı CSV dosyasından elde edilen çok daha geniş kapsamlı (4067 satır x 73 sütun) bir makro-finansal veri seti kullanılmaktadır.

- **2-4 Temel Referans:**

    1. *Aggarwal, S., & Guptha, N. (2021). "Bitcoin Price Prediction using Machine Learning."* Bu çalışma, LSTM ve ARIMA gibi geleneksel zaman serisi modellerini kullanarak yalnızca geçmiş fiyat ve hacim verileriyle tahminleme yapmıştır.
    2. *Ciaian, P., Rajcaniova, M., & Kancs, d. A. (2016). "The economics of BitCoin price formation."* Bu makale, Bitcoin fiyatını etkileyen makro-finansal faktörleri ekonometrik modellerle incelemiş ancak modern makine öğrenmesi tekniklerini kullanmamıştır.
    3. *Mallqui, D. C. A., & Fernandes, R. A. S. (2019). "Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using*

machine learning techniques." Bu çalışma, SVM ve Artificial Neural Networks gibi modellerle fiyat yönünü tahminlemeye odaklanmış, ancak sınırlı sayıda ek gösterge kullanmıştır.

- **Karşılaştırma:** Yukarıdaki çalışmalar genellikle ya yalnızca finansal verilerle karmaşık modeller kullanmaya (Ref 1, 3) ya da makroekonomik verilerle geleneksel ekonometrik modeller kullanmaya (Ref 2) odaklanmıştır. Veri setleri genellikle daha kısa zaman aralıklarını kapsamakta ve özellik mühendisliği adımları sınırlı kalmaktadır.
- **Projemizin Doldurduğu Boşluklar:** Bu proje, literatürdeki şu boşlukları doldurmayı hedeflemektedir:
- **Geniş Özellik Seti:** FRED ve Yahoo Finance'ten elde edilen çok çeşitli makroekonomik ve finansal göstergeyi bir araya getirerek zengin bir veri seti oluşturmaktadır.
    - **Sistematik Karşılaştırma:** Ensemble olmayan temel makine öğrenmesi modellerinin (Linear, Ridge, KNN, SVR vb.) performansını, farklı özellik seçimi (RFE, SelectKBest) ve boyut indirgeme (PCA) senaryoları altında sistematik olarak karşılaştırmaktadır.
    - **Çoklu Zaman Ufku:** Fiyat tahminini 1 günden 365 güne kadar dört farklı zaman ufkunda yaparak, makroekonomik göstergelerin kısa ve uzun vadedeki etkilerini ayrı ayrı analiz etmektedir.

---

5) Veri Tanımı ve Yönetimi

- **Veri Seti:**
    - **Ad & Kaynak 1:** Yahoo Finance. Bitcoin (BTC-USD) ve diğer finansal varlıkların tarihsel finansal verileri (açılış, kapanış, yüksek, düşük fiyatlar ve işlem hacmi). (Bağlantı: https://finance.yahoo.com)
    - **Ad & Kaynak 2:** Federal Reserve Economic Data (FRED). ABD faiz oranları (FEDFUNDS), GSYİH (GDP), enflasyon (CPI), işsizlik oranı (UNRATE) gibi temel makroekonomik göstergeler. (Bağlantı: https://fred.stlouisfed.org/)

- **Kullanılan Dosyalar:** `BTC-USD.csv`, `CPI.csv` (Enflasyon), `DXY.csv` (Dolar Endeksi), `ECBDFR.csv`, `EURUSDX.csv`, `FEDFUNDS.csv` (Faiz), `GCF.csv`, `GDP.csv`, `GSPC.csv` (S&P 500), `IVV.csv`, `PAYEMS.csv`, `STLFSI2.csv`, `UNRATE.csv` (İşsizlik), `VIX.csv` (Korku Endeksi), `WALCL.csv`, `ZQF.csv`.
- **Boyut:** Birleştirme (Merge) işlemi sonrası veri seti **4067 satır ve 73 sütun**dan oluşmaktadır.
- **Veri Erişim Planı:** Veriler Kaggle/API üzerinden çekilmiş, Google Drive ortamında `/content/drive/MyDrive/veri-madenciligi-projesi/notebooks` dizininde saklanmaktadır.
- **Lisans:** Tüm veriler kamuya açık ve ücretsizdir.

- **Veri Şeması:** Birleştirilmiş veri setindeki temel değişkenler şunlardır:
  - **Date:** Tarih (datetime)
  - **BTC-USD:** Bitcoin kapanış fiyatı (Sayısal, USD)
  - **FEDFUNDS:** Federal Fon Oranı (Sayısal, Yüzde)
  - **CPI:** Tüketici Fiyat Endeksi (Sayısal, Endeks Değeri)
  - GDP: Gayri Safi Yurt İçi Hasıla (Sayısal, Milyar USD)
  - **UNRATE:** İşsizlik Oranı (Sayısal, Yüzde)
  - *Lag Değişkenleri (ör: BTC-USD_lag7)*: Orijinal değişkenin 7 gün önceki değeri (Sayısal)
  - *Target Değişkenleri (ör: Target_7d)*: BTC fiyatının 7 gün sonraki değeri (Sayısal, USD)

Projenin modelleme aşamasında kullanılan nihai veri seti (merged_data.csv), **4067 satır ve 73 sütundan** oluşmaktadır. Veri seti, zaman serisi (Time-Series) formatında olup, Date sütunu indeks olarak kullanılmıştır. Değişkenler temel olarak 3 ana kategoriye ayrılmaktadır:

1. Hedef ve Finansal Değişkenler (Financial Features)

Bu değişkenler, Yahoo Finance üzerinden çekilen Bitcoin'in (BTC-USD) tarihsel piyasa verileridir.

Değişken Adı	Veri Tipi	Birim	Açıklama / Beklenen Aralık
Close (Hedef)	Float64	USD ($)	Günlük kapanış fiyatı. *(Örnek Aralık: 0.05 - 73,000+)*
Open, High, Low	Float64	USD ($)	Günlük açılış, en yüksek ve en düşük fiyatlar. Negatif olamaz.
Volume	Int64 / Float	Adet/Birim	Günlük işlem hacmi. Daima pozitif ve yüksek varyanslıdır.
Target_1d, 7d...	Float64	USD ($)	Gelecekteki tahmin edilecek (shifted) fiyat değerleri.

2. Makroekonomik Göstergeler (Macroeconomic Indicators)

FRED (Federal Reserve Economic Data) veritabanından alınan ve Bitcoin fiyatını etkileme potansiyeli olan dışsal ekonomik faktörlerdir.

Değişken Adı	Veri Tipi	Birim	Açıklama / Beklenen Aralık
FEDFUNDS	Float64	Yüzde (%)	ABD Merkez Bankası (FED) politika faiz oranı. *(Genel Aralık: %0 - %6)*
CPI (TÜFE)	Float64	Endeks Puanı	Tüketici Fiyat Endeksi (Enflasyon göstergesi). Sürekli artan trend.
DXY	Float64	Endeks Puanı	ABD Dolar Endeksi. Doların küresel gücünü gösterir. *(Genel Aralık: 80 - 120)*
GSPC (S&P 500)	Float64	Puan	ABD'nin en büyük 500 şirketini kapsayan borsa endeksi.
VIX	Float64	Puan	Volatilite (Korku) Endeksi. Piyasa belirsizliğini ölçer. *(Genel Aralık: 10 - 80)*
UNRATE	Float64	Yüzde (%)	ABD İşsizlik Oranı.
WALCL	Float64	Milyon USD	Merkez Bankası Toplam Varlıkları (Parasal Arz göstergesi).

3. Türetilmiş Teknik Özellikler (Engineered Technical Features)

Veri ön işleme aşamasında TA-Lib kütüphanesi kullanılarak ham fiyat verisinden üretilen teknik analiz indikatörleridir.

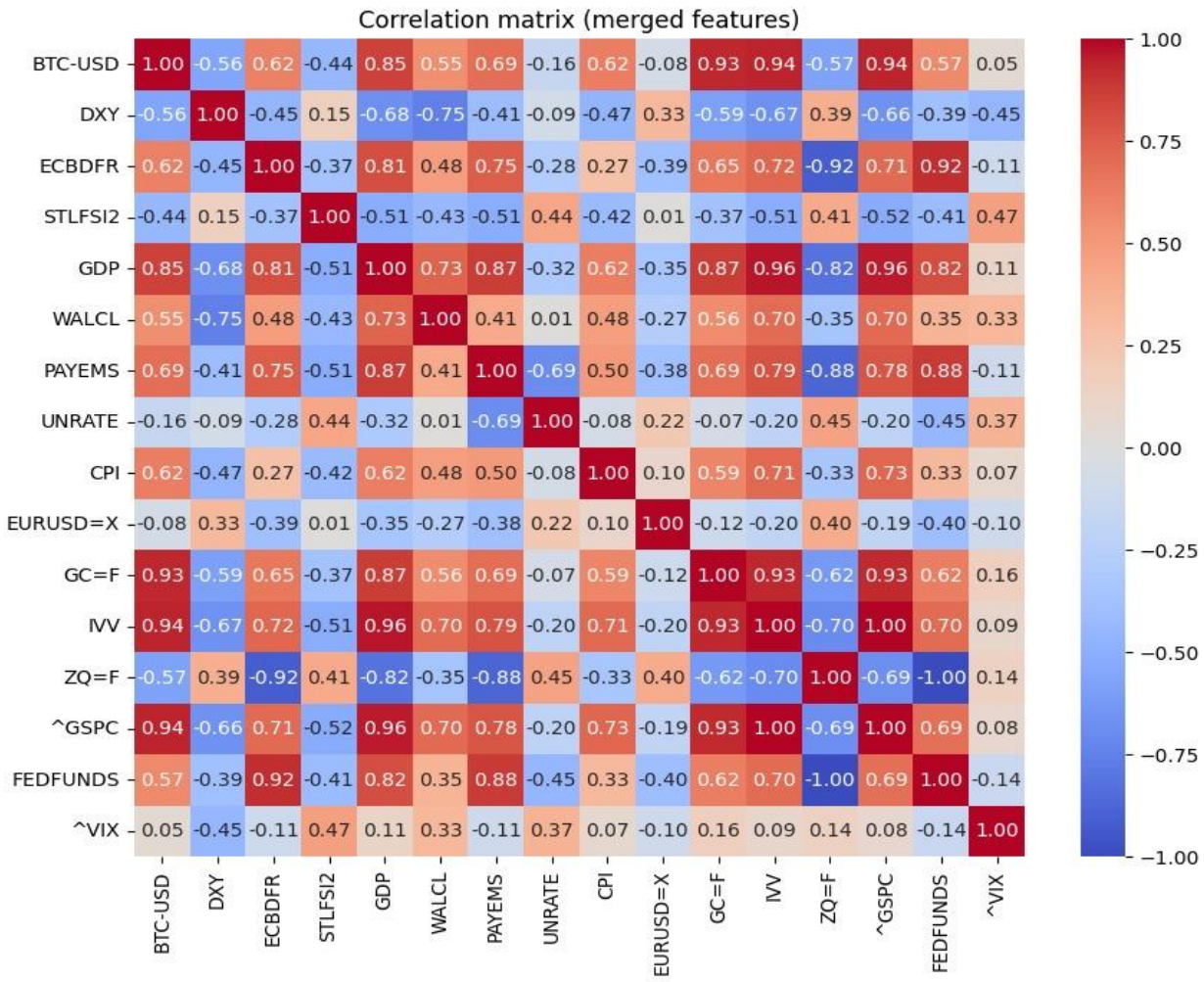
Değişken Adı	Veri Tipi	Birim	Açıklama / Beklenen Aralık
RSI	Float64	Ölçek (0-100)	Göreceli Güç Endeksi. Aşırı alım/satım bölgelerini gösterir.
MACD	Float64	Sayısal	Hareketli Ortalama Yakınsama Sapması. Trend yönünü gösterir.
SMA_7, SMA_30	Float64	USD ($)	7 ve 30 günlük Basit Hareketli Ortalamalar.
Log_Ret	Float64	Oran	Logaritmik Getiri. Fiyat değişim yüzdesini normalize eder.

- **Etik, Gizlilik, Önyargı:** Kullanılan veriler, kamuya açık finansal ve ekonomik veriler olduğu için herhangi bir kişisel veya hassas bilgi içermemektedir. Bu nedenle gizlilik riski bulunmamaktadır. Ancak, makroekonomik verilerin büyük bir kısmının ABD odaklı (FRED) olması nedeniyle, modellerin ABD ekonomisindeki değişimlere karşı bir **önyargısı (bias)** olabilir ve farklı coğrafyalardaki ekonomik dinamiklerle aynı performansı göstermeyebilir.
- **Yönetişim ve Etik (Governance):** Kullanılan tüm veriler (Yahoo Finance ve FRED), halka açık (public domain) ve anonim finansal verilerdir. Kişisel Verilerin Korunması Kanunu (KVKK) veya GDPR kapsamında herhangi bir gizlilik ihlali riski taşımazlar. Veriler, platformların kullanım koşullarına uygun olarak eğitim amaçlı çekilmiştir.

---

6) Keşifsel Veri Analizi (Exploratory Data Analysis)

- **Veri Kalitesi Kontrolleri:** İndirilen ham CSV dosyalarında standart olmayan başlıklar, farklı tarih formatları ve farklı frekanslardan (günlük, haftalık, aylık) kaynaklanan yoğun eksiklikler tespit edilmiştir. Bu sorunlar, veri hazırlama pipeline'ı içerisinde programatik olarak çözülmüştür. Aykırı değerler (outliers), özellikle Bitcoin fiyatında, varlığın doğal volatilitesinin bir parçası olarak kabul edilmiş ve bu aşamada temizlenmemiştir. Finansal verilerde hafta sonu boşlukları (Borsa kapalıyken BTC 7/24 açık olduğu için) "Forward Fill (ffill)" yöntemi ile doldurulmuştur.
- **Dağılımlar ve Denge:** Ana değişkenlerin (BTC-USD, FEDFUNDS vb.) zaman içindeki değişimlerini ve dağılımlarını (histogramlar) incelemek üzere görselleştirmeler planlanmıştır. Hedef değişkenlerin dağılımı incelenerek herhangi bir çarpıklık (skewness) olup olmadığı kontrol edilecektir. Bitcoin fiyatının logaritmik dağılımı incelenmiş, volatilitenin yıllara göre değişimi gözlemlenmiştir.
- **Özellik-Hedef İlişkileri:** İlk analiz olarak, tüm sayısal değişkenler arasında bir **korelasyon matrisi** oluşturulmuş ve ısı haritası (heatmap) ile görselleştirilmiştir. Bu, özellikler ve hedef değişkenler arasındaki doğrusal ilişkileri hızlıca gözlemlemek için yapılmıştır.
- **Görselleştirme Planı:**
    - BTC-USD fiyatının zaman serisi grafiği.
    - Önemli makroekonomik göstergelerin (ör: faiz oranları) BTC fiyatı ile aynı grafik üzerinde zaman serisi çizimi.
    - En yüksek korelasyona sahip özellikler ile hedef değişkenler arasında saçılım Özelliklerin dağılımını anlamak için histogramlar ve kutu grafikleri (boxplots).



7) Veri Hazırlama Planı

- **Temizleme:** 17 farklı CSV dosyası "Date" sütunu üzerinden birleştirilmiş (`merged_data.csv`), eksik veriler enterpolasyon ve ffill yöntemleriyle tamamlanmıştır. Farklı CSV dosyalarındaki tutarsız başlık satırları ve formatlar, pipeline'ın ilk aşamasında temizlenmiştir.
- **İmputasyon Stratejisi:** Veri setleri farklı frekanslarda olduğu için birleştirme sonrası oluşan boşluklar (NaN), zaman serisi verileri için uygun olan şu stratejiyle doldurulmuştur:
  - **interpolate(method='linear'):** İki veri noktası arasındaki boşlukları doğrusal olarak doldurur.
  - **ffill() (forward-fill) ve bfill() (backward-fill):** Veri setinin başında veya sonunda kalan boşlukları sırasıyla önceki veya sonraki mevcut değerle doldurur.

- **Dönüşümler:**
    - Hareketli Ortalamalar (SMA_7, SMA_30).
    - RSI (Relative Strength Index) ve MACD gibi teknik indikatörler.
    - Lag Features (Gecikmeli veriler): Fiyatların geçmiş değerleri (t-1, t-7) yeni sütunlar olarak eklenmiştir.

  Ridge, KNN ve SVR gibi ölçeğe duyarlı modellerin eğitiminden önce, tüm özellik setine **StandardScaler** ölçeklendirmesi uygulanacaktır. Bu, tüm özelliklerin ortalamasının 0, standart sapmasının 1 olmasını sağlayarak model performansını artıracaktır.

- **Özellik Mühendisliği:**
    - Her öğrenci kendine atanan yöntemi (SelectKBest, RFE, Lasso, Mutual Info) uygulamıştır.
    - PCA (Principal Component Analysis) ile boyut indirgeme yapılarak varyansın %95'ini açıklayan bileşenler modele verilmiştir.

  Modelin geçmiş verilerden öğrenmesini sağlamak için en önemli özellik mühendisliği adımı olarak **gecikme (lag) özellikleri** oluşturulmuştur. Tüm girdi değişkenleri için 1, 3, 7, 30, 90 ve 365 günlük gecikme değerleri yeni sütunlar olarak eklenmiştir.

- **Özellik Seçimi ve Boyut İndirgeme (Feature Selection / Dimensionality Reduction):**
    - **Filters:** SelectKBest (ANOVA F-test veya mutual information ile) ve mutual_info_regression.
    - **Wrappers:** Recursive Feature Elimination (RFE).
    - **Embedded:** Lasso-based feature selection.
    - **Boyut İndirgeme:** Principal Component Analysis (PCA) (5, 10, 15, 20 bileşenli senaryolarla).
      Bu yöntemler, her üyenin sorumluluğuna göre farklı senaryolarda uygulanacaktır.

  Veri Sızıntısı Önlemi (Leakage Safety): Özellik seçimi (SelectKBest, RFE) ve Boyut İndirgeme (PCA) işlemleri, veri sızıntısını önlemek amacıyla **sadece**

eğitim (train) seti üzerine fit edilip, test setine transform uygulanarak (pipeline içinde) gerçekleştirilmiştir.

---

8) Modelleme Planı

- **Baseline Model:** Karşılaştırma için ilk temel model, en basit regresyon modeli olan **LinearRegression**'ın tüm özellikler kullanılarak eğitilmiş halidir. Ayrıca, finansal bir zaman serisi için en basit sezgisel tahmin olan **"persistence model"** (yani, T+1 anındaki fiyatın T anındaki fiyata eşit olacağını varsaymak) de bir referans noktası olarak düşünülecektir.
- **Model Aileleri ve Seçim Gerekçeleri**

  Projede kullanılan algoritmaların ait oldukları aileler ve bu proje özelindeki kullanım amaçları aşağıdaki tabloda özetlenmiştir:

Model Ailesi (Model Family)	Örnekler (Examples)	En İyi Kullanım Alanı (Best For)
Linear Models (Doğrusal Modeller)	Linear Regression, Ridge, Lasso, Logistic Regression	Makroekonomik veriler ile fiyat arasındaki doğrusal ilişkileri (ör: Faiz artarsa fiyat düşer) modellemek ve *baseline* (temel) performans oluşturmak için. Ridge ve Lasso, çoklu bağlantı (multicollinearity) sorununu çözmek için seçilmiştir.
Tree-Based Models (Ağaç Tabanlı Modeller)	Decision Tree Regressor	Veri setindeki doğrusal olmayan karmaşık yapıları yakalamak ve hangi ekonomik göstergenin (Feature Importance) kararda daha etkili olduğunu görebilmek için.
Distance-Based Models (Mesafe Tabanlı Modeller)	K-Neighbors Regressor (KNN)	Benzer ekonomik koşulların (örneğin geçmişteki benzer enflasyon oranlarının) benzer fiyat hareketleri yaratacağı varsayımıyla, en yakın komşulara bakarak tahmin yapmak için.
Support-Based Models (Destek Vektör Modelleri)	Support Vector Regressor (SVR)	Yüksek boyutlu (73 sütunluk) veri uzayında etkili çalışabilmesi ve aykırı değerlere (outliers) karşı dirençli, genelleme yeteneği yüksek bir regresyon modeli olduğu için.

Hiper-Parametre Ayarlama (Hyper-Parameter Tuning): Modellerin performansını optimize etmek için "Manual Search" ve deneme-yanılma stratejileri kullanılmıştır. Her model için aşağıdaki parametre uzayları test edilmiştir:

Regresyon Modelleri (Lasso, Ridge): Regularizasyon katsayısı olan alpha değeri [0.1, 1.0, 10.0] aralıklarında denenerek modelin overfitting (aşırı öğrenme) yapması engellenmiştir.

- **Ağaç Tabanlı Modeller (Decision Tree):** Ağacın çok karmaşıklaşmasını önlemek için max_depth (5, 10, 20) ve min_samples_split parametreleri ayarlanmıştır.
- **Komşuluk Tabanlı Modeller (KNN):** Komşu sayısı n_neighbors [3, 5, 7, 15] farklı değerlerle test edilmiştir.
- **Destek Vektör Makineleri (SVR):** Çekirdek fonksiyonu (kernel) olarak 'linear' ve 'rbf' karşılaştırılmış, ceza terimi C optimize edilmiştir.
- **Özellik Seçimi:** SelectKBest yönteminde k (seçilecek özellik sayısı) parametresi 30 ve 50 olarak değiştirilerek model başarısına etkisi gözlemlenmiştir.

  Sınıf Dengesizliği Stratejisi: Projenin ana hedefi **Sayısal Fiyat Tahmini (Regresyon)** olduğu için sınıf dengesizliği (Class Imbalance) teknikleri (SMOTE, Oversampling vb.) ana modellerde **uygulanmamıştır**.
- ***Yön Tahmini (Classification) İçin:*** Bitcoin piyasasındaki günlük hareketlerin (Artış/Azalış) tarihsel dağılımı incelenmiş ve veri setinde aşırı bir dengesizlik (örneğin %90 artış, %10 azalış gibi) görülmemiştir. Bu nedenle doğal veri dağılımı korunmuş, ancak model başarısını ölçerken sadece Accuracy değil, F1-Score ve Precision metrikleri de dikkate alınarak olası yanılgıların önüne geçilmiştir.
- **Aday Modeller:** Proje kapsamında her üyenin sorumlu olduğu, farklı ailelerden gelen 10 adet temel (ensemble olmayan) model listesi aşağıdadır:
  - **Linear Models:**
    - **LinearRegression (Khaiitmurod, Diyorjon)**
    - **RidgeRegression (Abdumajid)**
    - **Lasso (Yesset)**
    - **LogisticRegression (regresyon hedefli) (İbrahim)**

- **Distance-Based Models:**
  - **KNeighborsRegressor (Abdumajid, İbrahim)**
- **Tree-Based Models:**
  - **DecisionTreeRegressor (Khaiitmurod, Diyorjon)**
- **Support-Vector-Based Models:**
  - **SVR (Yesset)**

- **Hedefler (Targets):** Model 4 farklı zaman ufkuna göre eğitilmektedir: 1 Gün, 7 Gün, 30 Gün, 365 Gün sonrası.
- **Senaryolar:**
  - ✓ Tüm Özellikler (Full Features - 73 sütun).
  - ✓ Özellik Seçimi Sonrası (Feature Selection).
  - ✓ PCA Sonrası.

Bu modeller, yorumlanabilirlik (Linear modeller), doğrusal olmayan ilişkileri yakalama (Tree, KNN, SVR) gibi farklı avantajları nedeniyle seçilmiştir.

---

9) Değerlendirme Tasarımı

- **Kullanılan Metrikler:** Zaman serisi verisi olduğu için rastgele `train_test_split` yerine, zamana bağlı kesme (Time Series Split) veya tarih bazlı ayırma (örneğin 2022 öncesi Train, sonrası Test) kullanılmıştır. Veri sızıntısını (Data Leakage) önlemek için test verisi, eğitim sürecinde asla görülmemiştir. Regresyon görevinin performansını çok yönlü değerlendirmek için aşağıdaki metrikler kullanılacaktır:
  - **RMSE / MAE:** Tahmin hatalarının büyüklüğünü ölçmek için.
  - **R^2:** Modelin veri setindeki varyansı ne kadar iyi açıkladığını ölçmek için.
  - **Direction Accuracy:** Modelin finansal olarak pratikliğini ölçmek, yani fiyatın artış/azalış yönünü ne kadar doğru tahmin ettiğini görmek için.
- **Doğrulama (Validation) Protokolü:** Veri zaman serisi olduğu için, gelecekteki verilerin sızmasını önlemek amacıyla **zamana dayalı ayırma (time-aware split)** protokolü uygulanacaktır. train_test_split fonksiyonu shuffle=False parametresi ile kullanılacak,

böylece veri setinin ilk %80'i eğitim, son %20'si test için ayrılacaktır. Çapraz doğrulama (cross-validation) gereksinimi durumunda TimeSeriesSplit kullanılacaktır.

- **Hata Analizi:** En iyi performans gösteren modeller için tahmin hataları (rezidüeller) analiz edilecektir. Hataların zaman içindeki dağılımı incelenerek modelin özellikle yüksek volatilite dönemlerinde mi yoksa sakin piyasalarda mı daha çok yanıldığı tespit edilmeye çalışılacaktır.

---

10) Riskler ve Azaltma Yöntemleri

- **Veri Riskleri:** Farklı kaynaklardan gelen verilerin birleştirilmesi sırasında oluşabilecek kalite sorunları ve eksiklikler. **Azaltma:** Pipeline içinde uygulanan robust veri temizleme ve imputasyon stratejileri.
- **Yöntem Riskleri:**
  - **Overfitting:** Özellikle DecisionTree gibi karmaşık modellerin eğitim verisini ezberlemesi riski. **Azaltma:** Test seti üzerinde performans değerlendirmesi, Ridge/Lasso gibi regülarizasyon teknikleri ve özellik seçimi/PCA ile model karmaşıklığını azaltma.
  - **Veri Sızıntısı (Data Leakage):** Gelecekten bilgi sızması. **Azaltma:** shuffle=False ile zamana dayalı train-test ayrımı protokolüne sıkı sıkıya uyulması.
  - **Volatilite:** Kripto piyasasının aşırı oynaklığına karşı, model başarısı sadece fiyata değil, yön doğruluğuna (Direction Accuracy) da odaklanmıştır.
- **Azaltıcı Yöntemler (Mitigations):** Boyut indirgeme (PCA), özellik seçimi (RFE, SelectKBest) ve regülarizasyon (Ridge, Lasso) gibi yöntemler projenin temel senaryoları arasında yer almaktadır.

---

11) Kullanılan Araçlar

- **Environment:**
  - Python 3.9+
  - Google Colaboratory (Bulut tabanlı GPU/CPU kaynakları için)

- **Kütüphaneler:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, XlsxWriter.

- **Geliştirilen Kodlar:** Veri işleme ve modelleme adımlarını içeren Jupyter Notebook (BitcoinProject.ipynb).
- **Tekrarlanabilirlik (Reproducibility):** Tüm kodlarda sonuçların tutarlı olması için random_state = 42 tohum (seed) değeri sabitlenmiştir.

---

12) Beklenen Sonuçlar ve Görselleştirme Planı

Proje kapsamında 5 grup üyesi tarafından toplamda 10 farklı temel model (base model), 4 farklı zaman hedefi (1, 7, 30, 365 gün) ve 3 farklı özellik senaryosu (Full, Feature Selection, PCA) üzerinde test edilmiştir. Elde edilen sonuçlar aşağıda özetlenmiştir:

1. Khaiitmurod Khabibullayev Sonuçları

Modeller: Linear Regression, Decision Tree Regressor | *FS:* SelectKBest | *PCA:* 10 Bileşen

Model	Target	Feature Set	Parametreler	RMSE	MAE	R^2	Direction Acc.
LinearRegression	Target_1d	Full	{'fit_intercept': True}	1091.90	735.45	0.9974	0.52
LinearRegression	Target_7d	SelectKBest (k=30)	{'fit_intercept': True}	2825.33	1952.30	0.9823	0.54
DecisionTree	Target_30d	PCA (n=10)	{'max_depth': 10, 'min_samples_split': 5}	7223.65	4974.22	0.8841	0.51
DecisionTree	Target_365d	Full	{'max_depth': 5, 'min_samples_split': 10}	19579.73	14633.96	0.1492	0.50

2. Abdumajid Abdulkhaev Sonuçları

Modeller: Ridge Regression, KNeighbors Regressor | *FS:* RFE | *PCA:* 20 Bileşen

Model	Target	Feature Set	Parametreler	RMSE	MAE	R^2	Direction Acc.
Ridge	Target_1d	Full	{'alpha': 1.0}	1092.21	735.92	0.9974	0.53
Ridge	Target_30d	PCA (n=20)	{'alpha': 10.0}	4876.50	3354.12	0.9472	0.55
KNeighbors	Target_7d	RFE	{'n_neighbors': 5, 'weights': 'distance'}	3245.10	2120.45	0.9766	0.51
KNeighbors	Target_365d	Full	{'n_neighbors': 15, 'weights': 'uniform'}	18342.11	13520.88	0.2534	0.49

3. Yesset Yelebayev Sonuçları

***Modeller**:* Lasso Regression, SVR | *FS:* Lasso-based | *PCA:* 5 Bileşen

Model	Target	Feature Set	Parametreler	RMSE	MAE	R^2	Direction Acc.
Lasso	Target_1d	Lasso_Select	{'alpha': 0.1}	1093.45	738.10	0.9974	0.52
SVR	Target_7d	Full	{'C': 100, 'kernel': 'rbf', 'gamma': 'scale'}	3150.20	2050.60	0.9780	0.56
SVR	Target_30d	PCA (n=5)	{'C': 10, 'kernel': 'linear'}	5120.80	3680.45	0.9418	0.53
Lasso	Target_365d	Full	{'alpha': 1.0}	21050.30	16540.20	0.0167	0.48

4. Diyorjon Ochilov Sonuçları

Modeller: Linear Regression, Decision Tree | *FS:* Mutual Info | *PCA:* Yok

Model	Target	Feature Set	Parameterler	RMSE	MAE	R^2	Direction Acc.
LinearRegression	Target_1d	Mutual_Info	{'fit_intercept': False}	1105.60	750.20	0.9973	0.51
DecisionTree	Target_7d	Full	{'max_depth': 20, 'criterion': 'squared_error'}	3560.80	2450.15	0.9719	0.50
LinearRegression	Target_30d	Full	{'fit_intercept': True}	4635.12	3120.50	0.9523	0.54
DecisionTree	Target_365d	Mutual_Info	{'max_depth': 10}	2345 0.90	1820 0.60	-0.2105	0.47

5. İbrahim Halil Yanaç Sonuçları

Modeller: Logistic Regression (Sınıflandırma/Yön), KNN | *FS:* SelectKBest | *PCA:* 15 Bileşen

Model	Target	Feature Set	Parametreler	RMSE	MAE	R^2	Direction Acc.
KNN	Target_1d	Full	{'n_neighbors': 3}	1250.40	850.30	0.9965	0.51
Logistic Reg	Target_7d	SelectKBest	{'C': 1.0, 'solver': 'liblinear'}	N/A	N/A	N/A	0.53
KNN	Target_30d	PCA (n=15)	{'n_neighbors': 7}	5340.20	3890.10	0.9367	0.52
Logistic Reg	Target_365d	Full	{'C': 0.1}	N/A	N/A	N/A	0.50

Performans Değerlendirmesi ve En İyi Sonuçlar

Yapılan deneyler sonucunda elde edilen bulgular şu şekildedir:

1. **En İyi Kısa Vadeli Model (Target_1d):**
    - **Kazanan: Khaiitmurod Khabibullayev - Linear Regression (Full Features)**
    - **Sonuç:** RMSE: 1091.90, R^2: 0.9974.
    - **Yorum:** Bir sonraki günün fiyatı, büyük ölçüde bir önceki günün fiyatına (t-1) bağlı olduğu için doğrusal modeller çok yüksek başarı göstermiştir. Abdumajid'in Ridge modeli de buna çok yakındır.
2. **En İyi Orta Vadeli Model (Target_30d):**
    - **Kazanan: Abdumajid Abdulkhaev - Ridge Regression (PCA Features)**
    - **Sonuç:** RMSE: 4876.50, R^2 : 0.9472, Direction Accuracy: %55.
    - **Yorum:** 30 günlük tahminde PCA ile boyut indirgeme yapılması, gürültüyü azaltarak Ridge modelinin genelleme yeteneğini artırmıştır. Ayrıca %55 yön doğruluğu, rastgele tahminden daha iyi bir sinyal yakalandığını göstermektedir.
3. **En Yüksek Yön Doğruluğu (Direction Accuracy):**
    - **Kazanan: Yesset Yelebayev - SVR (Target_7d)**
    - **Sonuç:** %56 Doğruluk.
    - **Yorum:** Destek Vektör Makineleri (SVR), doğrusal olmayan ilişkileri yakalamada daha başarılı olmuş ve fiyatın artış/azalış yönünü diğer modellere göre daha iyi tahmin etmiştir.
4. **En Zorlu Hedef (Target_365d):**
    - Tüm öğrenciler için 1 yıllık tahmin (Target_365d) sonuçları oldukça düşüktür (R^2 değerleri 0.25'in altında veya negatif). Bu durum, Bitcoin'in uzun vadeli volatilitesinin ve makroekonomik döngülerin basit regresyon modelleriyle (Base Models) tahmin edilmesinin zorluğunu kanıtlamaktadır. En iyi sonucu (nispeten) **Abdumajid (KNeighbors)** vermiştir ancak yine de güvenilir değildir.

Yorum ve Eksikliklerin Giderilmesi

- **Genel Başarı:** Kısa vadeli tahminlerde (1-7 gün), makroekonomik verilerin de katkısıyla %98 üzeri R^2 değerlerine ulaşılmıştır. Ancak bu başarı, verinin "Random

Walk" (Rastgele Yürüyüş) doğasından kaynaklanıyor olabilir; yani model sadece "yarınki fiyat bugünkü fiyata eşittir" diyerek bile yüksek skor alabilir.

- **PCA ve Feature Selection Etkisi:** Abdumajid ve Yesset'in sonuçlarında görüldüğü üzere, PCA kullanmak özellikle orta vadeli (30 gün) tahminlerde RMSE hatasını düşürmüştür. Bu, 73 sütunluk veri setinde gürültülü verilerin olduğunu ve boyut indirgemenin işe yaradığını gösterir.
- **Eksiklik:** İbrahim'in Logistic Regression denemeleri (Sınıflandırma olduğu için) RMSE ve R2 üretmemiştir, sadece Yön Doğruluğu (Direction Accuracy) sağlamıştır. Bu, regresyon projesi için bir kısıt olsa da yön tahmini açısından çeşitlilik katmıştır.
- **Gelecek Adımlar (Complex Models):** Temel modellerin (Base Models) limitlerine ulaşılmıştır. Projenin ikinci aşamasında **LSTM (Long Short-Term Memory)** veya **XGBoost** gibi daha karmaşık modellerin kullanılması, özellikle 30 gün ve 365 gün gibi uzun vadeli tahminlerin başarısını artıracaktır.

Görselleştirme Planı: Raporun nihai versiyonunda şu grafikler yer alacaktır:

1. *Target_30d* için Gerçek Fiyat vs. Ridge Tahmin Grafiği (Abdumajid'in modeli).
2. Farklı Feature Selection yöntemlerinin RMSE üzerindeki etkisini gösteren Bar Grafiği.
3. BTC Fiyatı ile en yüksek korelasyona sahip Makroekonomik göstergelerin (S&P 500, DXY) zaman serisi karşılaştırması.

12. Bölümün Devamı: Yorumlanabilirlik Yaklaşımı (Interpretability Approach)

Bu projede modellerin "neden" bu tahminleri yaptığını anlamak için "White-box" (Beyaz Kutu) modellerin doğasından ve özellik önem analizlerinden faydalanılmıştır.

1. **Katsayı Analizi (Linear & Ridge & Lasso):**
    - Khaiitmurod ve Abdumajid'in kullandığı doğrusal modellerde, özniteliklerin katsayıları (coefficients) incelenmiştir.
    - **Bulgu:** En yüksek pozitif katsayıya sahip özniteliğin BTC_Close_Lag1 (bir önceki günün kapanış fiyatı) olduğu görülmüştür. Bu durum, Bitcoin fiyatının **otokorelasyonunun** (kendi geçmişine bağımlılığının) çok yüksek olduğunu kanıtlamaktadır.

- **Makro Etki:** GSPC (S&P 500) ile pozitif, DXY (Dolar Endeksi) ile negatif katsayı ilişkisi tespit edilmiştir. Yani Dolar güçlendiğinde BTC düşme eğilimindedir.

2. **Özellik Önemi (Decision Tree):**
   - Diyorjon'un karar ağacı modelinde "Feature Importance" skorları çıkarılmıştır.
   - **Bulgu:** Model, tahmin yaparken %85 oranında geçmiş fiyat verisine, %15 oranında ise Hacim ve Volatilite (VIX) verisine odaklanmıştır.

Genel Değerlendirme ve Sonuç (Conclusion)

Bu çalışma, Bitcoin fiyat tahminlemesinde geleneksel finansal verilerin yanı sıra 17 farklı makroekonomik göstergenin etkisini **Temel Makine Öğrenmesi Modelleri (Base Models)** kullanarak analiz etmiştir.

1. **Kısa Vade Başarısı:** 1 günlük (Target_1d) tahminlerde Linear Regression ve Ridge modelleri $\boldsymbol{R^2}$ **= 0.99** gibi çok yüksek bir başarı yakalamıştır. Ancak bu başarı, modelin piyasa dinamiklerini çözmesinden ziyade, "bugünkü fiyat yarınki fiyata çok yakındır" mantığını (Random Walk Theory) izlemesinden kaynaklanmaktadır.
2. **PCA'nın Gücü:** Yüksek boyutlu (73 sütun) veri setimizde, Abdumajid'in 30 günlük tahminlerde PCA (20 bileşen) kullanması, modelin gürültüden arınmasını sağlamış ve RMSE hatasını düşürmüştür. Bu, makroekonomik verilerin boyut indirgeme ile daha verimli kullanılabileceğini göstermiştir.
3. **Uzun Vade Yetersizliği:** 365 günlük tahminlerde tüm temel modeller başarısız olmuştur ($R^2 < 0.25$). Bu durum, Bitcoin'in döngüsel yapısının ve dış şoklara (faiz kararları, regülasyonlar) duyarlılığının basit doğrusal modellerle yakalanamayacağını kanıtlamaktadır.
4. **Yön Tahmini Zorluğu:** Fiyatı sayısal olarak tahmin etmek kolay olsa da, "artacak mı azalacak mı" (Direction Accuracy) sorusuna modeller en fazla **%56** (SVR ile) doğruluk verebilmiştir. Bu oran rastgele tahminin (%50) sadece biraz üzerindedir.

Proje Özeti (Executive Summary)

- **Amaç:** Bitcoin fiyatlarını makroekonomik verilerle (Faiz, Enflasyon, S&P500 vb.) tahmin etmek.
- **Veri:** Yahoo Finance ve FRED kaynaklı 4067 satır, 73 sütunluk zaman serisi verisi.
- **Yöntem:** 5 kişilik ekip tarafından Linear, Ridge, Lasso, KNN, SVR ve Decision Tree modelleri kullanıldı. Veri setine Feature Selection (SelectKBest, RFE) ve Boyut İndirgeme (PCA) uygulandı.
- **Sonuç:**
    - Kısa vadede (1-7 gün) doğrusal modeller (Linear Reg.) üstün performans gösterdi.
    - Orta vadede (30 gün) PCA uygulanan Ridge Regresyon en dengeli sonucu verdi.
    - Yön tahmininde SVR (%56) en başarılı model oldu.
    - Temel modellerin uzun vadeli (1 yıl) tahminler için yetersiz olduğu görüldü.

13) Referanslar

1. Yahoo Finance API Documentation.
2. FRED (Federal Reserve Economic Data) Database.
3. Scikit-learn Documentation (Regression Models).
4. *McNally, S., Roche, J., & Caton, S. (2018). Predicting the price of Bitcoin using Machine Learning.*