FET445 Veri Madenciliği

Su Atıklarındaki Grip Sınıflandırma

Grup: Virus Vibe

Youtube Link: https://www.youtube.com/watch?v=l_Jmnp7glc8

Tarih: 21.12.2025

Su Atıklarındaki Grip Sınıflandırma

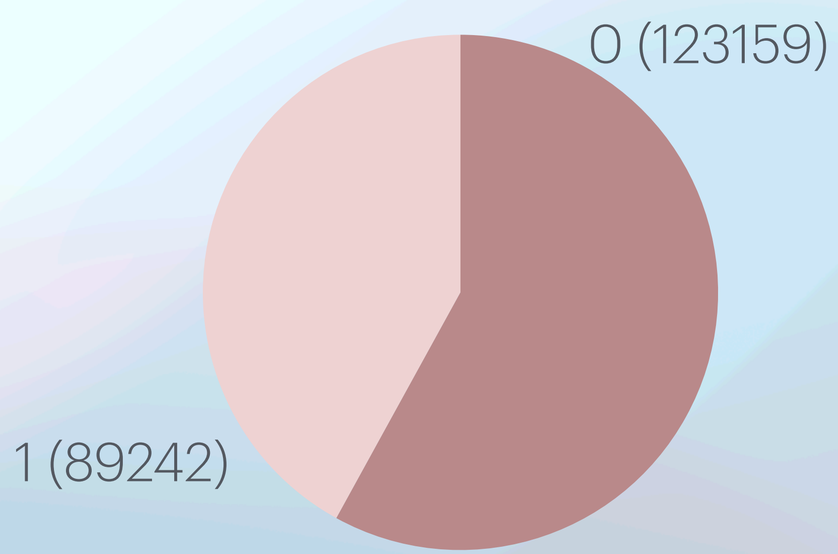
Problemin Açıklanması

Çalışmamızda atık su verileri kullanılarak toplum düzeyinde Influenza A varlığının tahmin edilmesidir. Klinik vaka verileri gecikmeli ve sınırlı olabildiğinden atık su tabanlı izleme erken uyarı açısından önemli bir alternatif sunmaktadır. CDC tarafından sağlanan atık su verilerinden elde edilen çevresel, demografik kullanılarak ikili bir sınıflandırma problemi ele alınmıştır. Hedef değişken, ölçülen viral RNA konsantrasyonuna bağlı olarak Influenza A'nın varlığı (1) veya yokluğu (0) şeklinde tanımlanmıştır. Geliştirilen modeller ile Influenza A yayılımının erken tespiti ve halk sağlığı kararlarının desteklenmesi amaçlanmaktadır.

Su Atıklarındaki Grip Sınıflandırma

Veri Seti Açıklaması

- **Link:** https://catalog.data.gov/dataset/cdc-wastewater-data-for-influenza-a
- **Size:** 212401 satır, 39 sütun
- **Özellik Tipi:** Sayısal (viral konsantrasyonlar, akış ve nüfus bilgileri), kategorik (tesis bilgileri, genetik hedefler) ve zamansal (tarih, ay, hafta) değişkenler
- **Class Distribution:** 0 (Influenza A yok): 123159, 1 (Influenza A var): 89242
- **Dengeli:** Sınıflar arasında ciddi bir dengesizlik bulunmamaktadır (%58 – %42)



Su Atıklarındaki Grip Sınıflandırma

Temel Feature Engineering ve Feature Extraction Teknikleri

Veri seti üzerinde zamansal, demografik ve operasyonel bilgileri daha anlamlı hale getirmek amacıyla çeşitli özellik mühendisliği adımları uygulanmıştır. Sayısal değişkenler üzerinde logaritmik ve polinom dönüşümler yapılmış, eksik değerler uygun istatistiksel yöntemlerle ele alınmıştır. Kategorik değişkenler modele uygun hale getirilmiş ve ek türetilmiş özellikler oluşturulmuştur. Model karmaşıklığını azaltmak ve performansı artırmak için özellik seçimi yöntemleri kullanılmıştır. Ayrıca boyut indirgeme teknikleri ile daha kompakt ve temsil gücü yüksek özellik uzayları elde edilmiştir.

Su Atıklarındaki Grip Sınıflandırma

Train-Test Split Oranı

Veri seti %80 eğitim ve %20 test olacak şekilde ayrılmıştır. Eğitim seti 176325 veri, test seti ise 36076 veri içermektedir. Eğitim ve test kümelerinde sınıf dağılımları benzer tutulmuş, her iki sınıfta da Influenza A var ve yok örneklerinin dengeli şekilde temsil edilmesine dikkat edilmiştir. Bu sayede modellerin sadece eğitim verisine ezber yapması engellenmiş ve daha önce görmediği veriler üzerindeki gerçek performansı daha doğru şekilde ölçülmüştür.

Su Atıklarındaki Grip Sınıflandırma

Performans Metrikleri

Modellerin başarımı birden fazla ölçüt kullanılarak değerlendirilmiştir. Sınıflandırma doğruluğunu ölçmek için Accuracy, pozitif sınıfı ne kadar doğru tahmin ettiğini görmek için Precision, gerçek pozitifleri yakalama oranını ölçmek için Recallve bu iki metriğin dengeli bir birleşimi olan F1-score kullanılmıştır. Ayrıca modellerin farklı eşik değerlerindeki performansını incelemek amacıyla ROC-AUC metriği hesaplanmış ve ROC eğrileri ile görselleştirilmiştir. Ek olarak, modellerin ürettiği olasılık tahminlerinin kalibrasyon kalitesini değerlendirmek için İBS hesaplanmış; daha düşük İBS değeri, daha iyi olasılık tahmini anlamına gelecek şekilde yorumlanmıştır.

Su Atıklarındaki Grip Sınıflandırma

Best Model 1: BaggingClassifier

- **Model Geliştirilirken Kullanılan Yaklaşımlar:** BaggingClassifier modelinde karar ağaçları temel öğrenici olarak kullanılmıştır. Model, farklı alt örneklemler ve özellik alt kümeleri üzerinde eğitilerek varyansın azaltılması ve daha kararlı tahminler yapılması hedeflenmiştir. Ayrıca GroupShuffleSplit ve StratifiedKFold ile modelin genelleme performansı kontrol edilmiştir.

- **Hyper Parametreler:** Model için ağaç sayısı (n_estimators), her ağacın kullandığı örnek oranı (max_samples), özellik oranı (max_features), ağaç derinliği (max_depth) ve yapraklardaki minimum örnek sayısı (min_samples_leaf) RandomizedSearchCV ile optimize edilmiştir.

- **Feature Set:** [sewershed_id, lod_sewage, pcr_target_avg_conc_lin, pcr_target_flowpop_lin, collection_week, log_flow_rate, jurisdiction_positive_count, log_pop_sq]

Su Atıklarındaki Grip Sınıflandırma

Best Model 1: BaggingClassifier

- PCA ve LDA uygulanmadan, SelectFromModel ile yapılan feature selection sonucunda model en yüksek performansına ulaşmıştır. Bu durumda model %94.20 accuracy ve 0.9776 ROC-AUC değeri elde ederek tüm deneyler arasında en başarılı sonuçları vermiştir.
- PCA (n=8) uygulandığında modelin performansı belirgin şekilde düşmüş, accuracy %78.02 ve ROC-AUC 0.8341 seviyesine gerilemiştir. Bu durum PCA'nın önemli bilgileri kaybettirdiğini göstermektedir.
- LDA uygulandığında model %70.86 accuracy ve 0.7640 ROC-AUC elde etmiştir. Tek bileşenli temsil, modelin sınıfları ayırt etme gücünü önemli ölçüde azaltmıştır.
- Hiperparametre optimizasyonu sürecinde ağaç sayısı, örnekleme oranları ve ağaç derinliği ayarlanarak modelin genelleme kabiliyeti artırılmış ve overfitting azaltılmıştır. Feature selection ile birlikte kullanıldığında model en kararlı ve yüksek performanslı yapıya ulaşmıştır.

Su Atıklarındaki Grip Sınıflandırma

Best Model 2: HistGradientBoosting

- **Model Geliştirilirken Kullanılan Yaklaşımlar:** HistGradientBoosting modeli, histogram tabanlı gradient boosting yaklaşımı ile eğitilmiştir. Bu yöntem sayesinde büyük veri setlerinde daha hızlı ve verimli öğrenme sağlanmıştır. Modelin aşırı öğrenmesini önlemek ve genelleme gücünü artırmak için çapraz doğrulama ve tuning adımları uygulanmıştır.

- **Hyper Parametreler:** Model için öğrenme oranı (learning_rate), maksimum ağaç derinliği (max_depth), iterasyon sayısı (max_iter) ve yapraklardaki minimum örnek sayısı (min_samples_leaf) RandomizedSearchCV kullanılarak optimize edilmiştir. En iyi parametreler ROC-AUC metriğine göre seçilmiştir.

- **Feature Set:** [sewershed_id, lod_sewage, pcr_target_avg_conc_lin, pcr_target_flowpop_lin, collection_week, log_flow_rate, jurisdiction_positive_count, log_pop_sq]

Su Atıklarındaki Grip Sınıflandırma

Best Model 2: HistGradientBoosting

- PCA ve LDA kullanılmadan, SelectFromModel ile seçilen özellikler üzerinde model %87.84 accuracy ve 0.9679 ROC-AUC değerine ulaşmıştır. Performans güçlü olmakla birlikte Bagging tabanlı modele kıyasla daha sınırlı kalmıştır.
- PCA (n=8) uygulandığında modelin accuracy değeri %78.12'ye, ROC-AUC değeri ise 0.8328'e düşmüştür. PCA sonrası performans kaybı, modelin orijinal feature yapısına daha duyarlı olduğunu göstermektedir.
- LDA uygulandığında model %71.07 accuracy ve 0.7648 ROC-AUC elde etmiştir. Boyutun tek bileşene düşmesi, modelin sınıflar arasındaki ayrımı yeterince yakalayamamasına neden olmuştur.
- Hiperparametre tuning sürecinde learning rate, ağaç derinliği ve minimum yaprak örnek sayısı optimize edilerek modelin daha dengeli öğrenmesi sağlanmıştır. Ancak feature selection sonrası elde edilen performans, Bagging tabanlı modele kıyasla daha düşük kalmıştır.

Modellerin Karşılaştırıldığı Tablo

Horizon	F1 Score	ROC AUC	PR AUC	IBS
Grip Yok(0)	LogisticRegression (Base): 0.9175 BaggingClassifier_Tuned (Tuned): 0.8726 HistGradientBoosting (Default): 0.8716 HistGradientBoosting_Tuned (Tuned): 0.8713 BaggingClassifier (Default): 0.8638 Gradient Boosting (Tuned): 0.8423 MLP C. (Default): 0.8395 ExtraTrees (Tuned): 0.8356 CatBoost (Default): 0.8339 FT-Transformer (Tuned): 0.8244 Wide & Deep (Tuned): 0.8182 Linear Regression (Default): 0.8170 AdaBoost (Default): 0.8072 KNN (Base): 0.7903 SGD (Weighted): 0.7798 SGD (Base): 0.7754 DT (Base): 0.7693 Naive Bayes (Base): 0.6613	HistGradientBoosting_Tuned (Tuned): 0.971860 HistGradientBoosting (Default): 0.971828 LogisticRegression (Base): 0.968998 BaggingClassifier_Tuned (Tuned): 0.959527 BaggingClassifier (Default): 0.932705 Gradient Boosting (Tuned): 0.8756 MLP C. (Default): 0.8733 CatBoost (Default): 0.8719 ExtraTrees (Tuned): 0.8689 FT-Transformer (Tuned): 0.8536 Wide & Deep (Tuned): 0.8479 Linear Regresyon (Default): 0.8327 AdaBoost (Default): 0.8268 SGD (Base): 0.8078 SGD (Weighted): 0.7994 KNN (Base): 0.781875 Naive Bayes (Base): 0.7543 DT (Base): 0.7315	HistGradientBoosting_Tuned (Tuned): 0.971860 HistGradientBoosting (Default): 0.971828 LogisticRegression (Base): 0.968998 BaggingClassifier_Tuned (Tuned): 0.959527 BaggingClassifier (Default): 0.950049 FT-Transformer (Tuned): 0.8716 Wide & Deep (Tuned): 0.8667 Gradient Boosting (Tuned): 0.8585 MLP C. (Default): 0.8544 CatBoost (Default): 0.8474 ExtraTrees (Tuned): 0.8353 SGD (Base): 0.8269 Linear Regresyon (Default): 0.8051 KNN (Base): 0.781875 AdaBoost (Default): 0.7775 DT (Base): 0.7315 SGD (Weighted): 0.7487 Naive Bayes (Base): 0.7030	LogisticRegression (Base): 0.078830 BaggingClassifier (Default): 0.109590 BaggingClassifier_Tuned (Tuned): 0.114560 HistGradientBoosting (Default): 0.125071 HistGradientBoosting_Tuned (Tuned): 0.126685 Gradient Boosting (Tuned): 0.1364 MLP C. (Default): 0.1378 ExtraTrees (Tuned): 0.1409 CatBoost (Default): 0.1410 FT-Transformer (Tuned): 0.1505 Wide & Deep (Tuned): 0.1533 Linear Regresyon (Default): 0.1605 SGD (Base): 0.1774 SGD (Weighted): 0.1786 KNN (Base): 0.187786 AdaBoost (Default): 0.2044 Naive Bayes (Base): 0.2366 DT (Base): 0.2613
Grip Var(1)	LogisticRegression (Base): 0.864956 HistGradientBoosting (Default): 0.856892 HistGradientBoosting_Tuned (Tuned): 0.855021 BaggingClassifier_Tuned (Tuned): 0.853732 BaggingClassifier (Default): 0.841297 ExtraTrees (Tuned): 0.7607ü Gradient Boosting (Tuned): 0.7563 MLP C. (Default): 0.7549 CatBoost (Default): 0.7473 FT-Transformer (Tuned): 0.7256 Wide & Deep (Tuned): 0.7253 Linear Regresyon (Default): 0.7120 SGD (Base): 0.7110 DT (Base): 0.6861 KNN (Base): 0.680881 AdaBoost (Default): 0.6733 SGD (Weighted): 0.6714 Naive Bayes (Base): 0.6547	HistGradientBoosting_Tuned (Tuned): 0.966960 HistGradientBoosting (Default): 0.966469 LogisticRegression (Base): 0.965323 BaggingClassifier (Default): 0.944520 BaggingClassifier_Tuned (Tuned): 0.935171 Gradient Boosting (Tuned): 0.8756 MLP C. (Default): 0.8733 CatBoost (Default): 0.8719 ExtraTrees (Tuned): 0.8689 FT-Transformer (Tuned): 0.8536 Wide & Deep (Tuned): 0.8479 Linear Regresyon (Default): 0.8327 AdaBoost (Default): 0.8268 SGD (Base): 0.8078 SGD (Weighted): 0.7994 KNN (Base): 0.790757 Naive Bayes (Base): 0.7543 DT (Base): 0.7315	LogisticRegression (Base): 0.966978 HistGradientBoosting_Tuned (Tuned): 0.966786 HistGradientBoosting (Default): 0.965907 BaggingClassifier_Tuned (Tuned): 0.935171 BaggingClassifier (Default): 0.923127 FT-Transformer (Tuned): 0.8239 Wide & Deep (Tuned): 0.8160 Gradient Boosting (Tuned): 0.8085 MLP C.(Defaul): 0.8061 ExtraTrees (Tuned): 0.8051 CatBoost (Default): 0.8474 Linear Regresyon (Default): 0.7763 AdaBoost (Default): 0.7775 SGD (Base): 0.7435 SGD (Weighted): 0.7487 KNN (Base): 0.737246 Naive Bayes (Base): 0.7030 DT (Base): 0.6062	LogisticRegression (Base): 0.078830 BaggingClassifier (Default): 0.109590 BaggingClassifier_Tuned (Tuned): 0.114560 HistGradientBoosting (Default): 0.125071 HistGradientBoosting_Tuned (Tuned): 0.126685 Gradient Boosting (Tuned): 0.1364 MLP C.(Defaul): 0.1378 ExtraTrees (Tuned): 0.1409 CatBoost (Default): 0.1410 FT-Transformer (Tuned): 0.1505 Wide & Deep (Tuned): 0.1533 Linear Regresyon (Default): 0.1605 SGD (Base): 0.1774 SGD (Weighted): 0.1786 KNN (Base): 0.187786 AdaBoost (Default): 0.2044 Naive Bayes (Base): 0.2366 DT (Base): 0.2613

Sonuç ve Değerlendirme

- HistGradientBoosting (özellikle Tuned) modelleri, hem Grip Yok (0) hem Grip Var (1) sınıflarında en yüksek ROC AUC ve PR AUC değerlerini elde etmiştir.
  - Bunun nedeni, doğrusal olmayan ilişkileri ve karmaşık örüntüleri güçlü şekilde modelleyebilmesidir.
- BaggingClassifier (Tuned) modeli, F1 Score ve ROC AUC metriklerinde üst sıralarda yer alarak genel olarak en dengeli modellerden biri olmuştur.
  - Birden fazla ağacın birleşimi, model varyansını azaltmış ve stabil sonuçlar üretmiştir.
- Hiperparametre optimizasyonu (Tuned) uygulanan modeller, çoğu durumda varsayılan (Default) modellere kıyasla daha yüksek ayırt edicilik sağlamıştır.
  - Özellikle ROC AUC ve PR AUC değerlerinde belirgin artış gözlenmiştir.
- Logistic Regression, ensemble modellere kıyasla daha basit olmasına rağmen en düşük IBS (Brier Score) değerini üretmiştir.
  - Bu durum, modelin olasılık tahminlerinin daha iyi kalibre edildiğini göstermektedir.
- FT-Transformer, Wide & Deep ve MLP gibi derin öğrenme tabanlı modeller, bazı metriklerde rekabetçi sonuçlar verse de ağaç tabanlı ensemble yöntemlerin gerisinde kalmıştır.
  - Tabular veri yapısı, ağaç tabanlı modellere daha uygun bulunmuştur.
- KNN, Naive Bayes ve Decision Tree gibi daha basit modeller, tüm metriklerde daha düşük performans göstermiştir.
  - Bu modeller karmaşık sınıf sınırlarını yeterince yakalayamamıştır.
- Genel olarak, ağaç tabanlı ensemble modeller, grip tahmini probleminde en başarılı yaklaşım olarak öne çıkmıştır

Bizi dinlediğiniz için Teşekkür ederiz :)