Logo

Description automatically generated

**ATATÜRK ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ**

**DERİN ÖĞRENME VE EVRİŞİMLİ SİNİR AĞLARI VİZE ÖDEVİ**

**Bilgisayar Mühendisliği– Yüksek Lisans**

**21040201052**

**İsa KULAKSIZ**

**Leaf Classification (Yaprak Sınıflandırma)**

Bu çalışma, yaprakların üç özelliğine yani şekil, kenar boşluğu ve dokuya dayanan 99 bitki türünü tanımlamaktadır.

Gereksinimler:

* Kullanılacak Makine Öğrenmi Algoritmaları:

1. Naive Bayes
2. Support Vector Machiene (SVM)
3. Logistic Regression
4. KNN
5. Decision Tree

* Yapay sinir ağı ve oluşturulacak bir CNN sınıflandırıcıların veri seti üzerinde çalıştırılması.
* Çapraz doğrulama yapılması.
* Tahmin skorlarının elde edilmesi (Accuracy, Recall vb.)
* En uygun parametrelerin tespit edilmesi
* Boyut azaltmanın araştıtılması.
* En iyi tahminleme sonucunu hangi Makine öğrenimi algoritması buldu.

Yukarıda yer alan gereksinimlerin hesaplanması beklenmektedir.

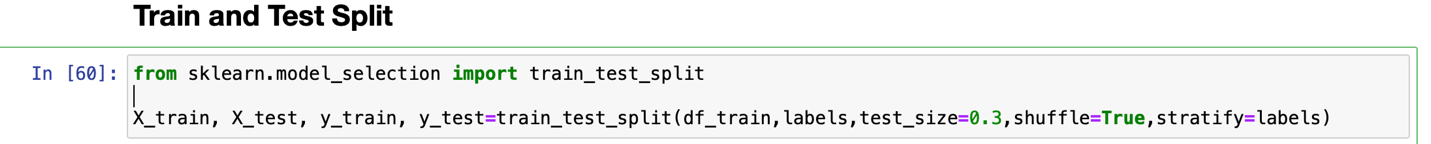
Kaggle’da yer alan veri setini incelendiğinde test, eğitim gibi parçalara ayrılmış olduğu fark edildi.

A picture containing table

Description automatically generated

Veri seti üzerinde doğrudan Makine öğrenimi algoritmalarını çalıştırmadan önce LabelEncoder() metoduyla sayısal hale getirilme işlemi yapıldı.

Veri setinde üzerinde bölme işlemi yapılarak veri setinde %30 kısım ayrıldı.



*Train and test split*

**MLOPS**

Diagram

Description automatically generated

Denklemde FP yanlış pozitiflerin sayısı yanlış negatiflerin sayısı ile aynı ise (FP == FN)

Recall, Precision ve F1 değerlerin aynı değere sahip olur.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F-measure |
| Naive Bayes | 48,82 | 48,81 | 48,82 | 48,82 |
| SVM | 91,58 | 91,58 | 91,58 | 91,58 |
| Logistic Regression | 59,26 | 59,25 | 59,25 | 59,25 |
| KNN | 87,21 | 87,20 | 87,20 | 87,20 |
| Decision Tree | 65,99 | 65,99 | 65,99 | 66,99 |

*MLOPS*

**En uygun parametreler:**

SVC(probability=True),

KNeighborsClassifier(n\_neighbors=3),

DecisionTreeClassifier(max\_features=7, min\_samples\_leaf=5)

**K Fold Cross Validation (Çapraz Doğrulama)**

Diagram

Description automatically generated

Elimizde çok az veri varsa test için ayırmanız halinde çok küçük bir veri seti ile karşı karşıya kalabiliriz. Bunun sonucu olarak da; test verileri hangi verinin test verisi hangi verinin eğitim olduğuna bağlı olarak değişecektir. Bu durumu önlemek içim eldeki veriyi K parçaya genelde (K=5 veya K=10) ayrılır. K tane aynı model oluşturulup her biri K-1 parça veriyle eğitilip kalanı da değerlendirmeye tabi tutulur.

Makine öğrenimi algoritmalarından Logistic Regression algoritması ile 10 parçaya bölünerek ortalama Cross Validation skoru elde edilmiştir.

Graphical user interface, text, application

Description automatically generated

**PCA (Principle Component Analysis)**

Temel amacı yüksek boyutlu verilerde en yüksek varyans ile veri setini tutmaktır. Bunu yaparken boyut azaltmadan faydalanır.

Boyut azaltmada bir miktar verinin kaybolacağı kesindir. Fakat amaç kaybolan bu verilerin popülasyon hakkında çok az bilgi içeriyor olmasıdır.

Graphical user interface, text

Description automatically generated

*PCA Result*

Makine öğrenimi algoritmalarının sonıçları Tablo 1’de gösterilmiştir.

*Tablo 1*

En iyi tahminleme sonucunu Support Vector Mechiene (SVM) algoritması % 91,58 ile tahmin edebilmiştir.

**Artificial Neural Network (ANN)**

A picture containing text, building, window

Description automatically generated

Veri ön işleme aşamasında metinsell ifadelerin sayısal ifadelere LabelEncoder() dönüştürme işlemi yapıldı.

ANN (Yapay Sinir Ağları) üzerinde Jupyter Notebook IDE kullanıldığında “Kernel Died” problemi ile karşılaşıldı. Jupyter Notebook, Pycharm ve Spider kullandığımız makine üzerinde işlem yaparak bizlere sonuç çıkarmaya çalışır. Kullandığımız makinenin GPU iyi olmadığı zaman yukarıdaki hata ile veya işlem süresinin çok uzun olması ile mutlaka karşı karşıya kalırız.

Bu problemlerden sıyrılmak için “Google Colab” üzerinde işlem yapmak avantaj sağlar.

StratidiedShuffleSplit, çapraz doğrulama nesnesi kullanılarak test için %30’luk kısım ayrıldı.

Text

Description automatically generated

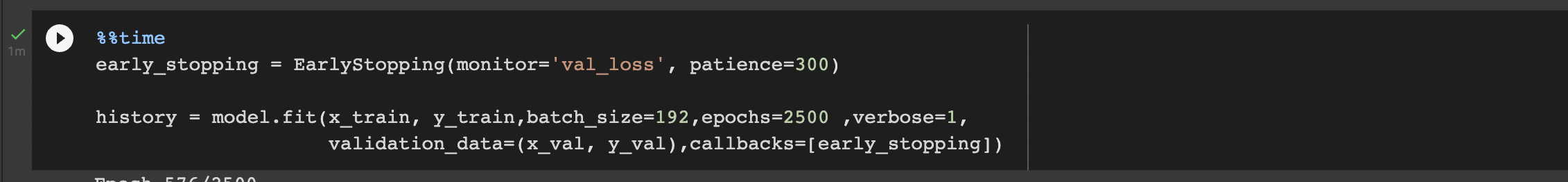
Text

Description automatically generated

ANN için Hidden Layer eklendi. Burada “Dropout” overfit’i önlemek için eklendi.

Epoch: Tüm veri setinin sinir ağları boyunca bir kere gidip gelmesine denir. Parametre olarak 2500 verilmiştir.

Batch size: Bir yinelemede kullanılan eğitim örneklerinin sayısını ifade eder. Parametre olarak 192 verildi.



Accuracy (Doğruluk) sonuçları:

Text

Description automatically generated

**Sonuçlar :**

* **Model Loss:**

Chart

Description automatically generated with medium confidence

* Model Accuracy:

A picture containing text

Description automatically generated

Projedeki çalışmalarım versiyon kontrol sistemi (git) kullanılarak GitHub hesabımda yer almaktadır.

<https://github.com/isakulaksiz/LeafClassification>