T.C
KONYA TEKNİ ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ
BİLİŞİM TEKNOLOJİLERİ UYGULAMASI
(DÖNEM PROJESİ) FİNAL RAPOR FORMU

Öğrencinin Adı-Soyadı:	Dilemre Ülkü
Numarası:	171213055
Danışmanın Adı-Soyadı:	Mesut Gündüz
Projenin Konusu: Konya'daki Tarım Eğilimleri ve Verimliliği	

Yapılan Çalışmaların Özeti:

Dönemin ilk yarısı Konya Büyükşehir Belediyesi'ne ait açık veri platformu olan acikveri.konya.bel.tr sitesinden, veri seti kategorileri altında bulunan tarım kategorisi altındaki 18 veri seti indirilmiştir. Bu veri setleri birleştirilerek ön işleme, keşifsel veri analizi ve görselleştirmeler yapılmıştır. Bu veri setiyle son olarak, TensorFlow kütüphanesi kullanarak tamamen bağlı 4 katmandan oluşan bir nöral ağ ile bir sınıflandırma gerçekleştirilmiştir. Sayısal veriler içeren SULAK_ALAN_MİKTARI, TRAKTÖR_SAYISI, BİREYSEL_ARAZİ_ORTALAMASI sütunları girdi olarak verilip ikili kategorik değişkenden oluşan İYİ_TARIM_UYGULAMALARI tahmin edilmeye çalışılmıştır. Model 15 epoch'ta 5 batch ile eğitildiği zaman %90 kesinliğe ulaşılmıştır.

Ayrıca ALTERNATİF_YENİ_ÜRÜN_ÇEŞİTLERİ_ÜRETİMİ sütununda bulunan çok kategorili veriler kullanılarak çek etiketli sınıflandırma denenmiştir. Scikit-learn kütüphanesi ile %77'si (240) eğitim, %33'ü (129) test olacak şekilde iki veri setine ayrıldıktan sonra tüm veri setleri TensorFlow kütüphanesinin tensor veri tipine dönüştürülmüştü Tamamen bağlı 2 katmandan oluşan bir nöral ağ oluşturulmuş, giriş katmanının aktivasyon fonksiyonu he_uniform, çıkış katmanı aktivasyon fonksiyonu sigmoid seçilmiştir. Optimizasyon algoritması olarak Adam (Adaptif Momentum), kayıp fonksiyonu olarak Binary Crossentropy seçilmiştir. Model 100 epoch'ta eğitildiği zaman %0.09 kesinliğe ulaşılmıştır.

Veri setinin sadece 2016 yılına ait veri içermesi, üç veri dışındaki diğer verilerin sadece kategorik veriden oluşması, kategorik verilerde baskın olarak "YOK" değerinin olması gibi sebepler ile beraber nöral ağdan elde edilen sonuçların güvenilir olmamasından dolayı farklı veri setlerine yönelinmiştir.

Türkiye İstatistik Kurumu'nun (TÜİK) veri tabanı olan https://biruni.tuik.gov.tr/ sitesinden bitkisel üretim istatistikleri iller bazında indirilmiştir. Bu sitede bulunan bitkisel ürün istatistikleri şunlardır:

1. Tarım alanı
2. Tahıllar ve diğer bitkisel ürünler
3. Sebzeler
4. Meyveler, içecek ve baharat bitkileri
5. Süs bitkileri
6. Örtü altı sebzeler
7. Örtü altı meyveler
8. Örtü altı süs bitkileri
9. Örtü altı tarım alanı
10. (Kuru/Sulu) -(1. Ekiliş/2. Ekiliş) ürünleri

Bu istatistiklerden tamamı indirilmiştir. Elde edilen istatistikler ile üretim tahmini yapılmasına kararlaştırılmıştır. Fakat her istatistikte her il ve her yıl için verileri bulunmamasından dolayı veriler düzenlendikten sonra bir milyonu bulan eksik veri ile karşılaşılmıştır. Veri setinin aşırı büyümesi performansı azalttığından ve tüm verilerin mevcut problemde kullanılmasının modeli doğru etkileyeceğinden emin olunamamasından dolayı daha küçük bir veri seti ile çalışılmaya karar verilmiştir. Yeni veri seti sadece tahıl verilerinden oluşturulmuştur. Yeni veri setinde bulunan veriler şunlardır:

- Tahıl ekilen al (dekar)
- Tarım alanı (dekar)
- Tahıl yıllık üretim (ton)

Tahıl yıllık üretim ve ekilen alan veri setinde bulunan ürünler aşağıdaki tablodadır.

Durum Buğdayı	Sorgum	Darı
Buğday (Durum Buğdayı Hariç)	Kanola veya Kolza Tohumu	Patates (Tatlı Patates Hariç)
Mısır	Kaplıca	Susam Tohumu
Arpa (Biralık)	Kuş Yemi	Yerfıstığı, Kabuklu
Çavdar	Mahlut	Kenevir Tohumu
Yulaf	Triticale	Haşhaş Tohumu
Tatlı Patates	Çeltik	Keten Tohumu
Yer Elması	Adaçayı	Aspir Tohumu
Ayçiçeği Tohumu (Yağlık)	Ayçiçeği Tohumu (Çerezlik)	Pamuk Çekirdeği (Çiğit)
Salep	Şeker Pancarı	Şeker Pancarı Tohumları
Pamuk, Çırçırlanmamış (Kütlü)	Pamuk, Çırçırlanmış (Lifli)	Tütün, İşlenmemiş
Keten, Lif	Kenevir, Lif	Fiğ (Yeşilot)
Fiğ (Adi) (Yeşil Ot)	Fiğ (Macar) (Yeşil Ot)	Fiğ (Diğer) (Yeşil Ot)
Burçak (Yeşilot)	Yonca (Yeşilot)	Korunga (Yeşilot)
Üçgül (Yeşilot)	Yulaf (Yeşilot)	Sorgum (Yeşilot)
Triticale (Yeşilot)	Mürdümük (Yeşilot)	Mısır (Hasıl)
Mısır (Slaj)	Hayvan Pancarı	Yem Şalgamı
Buğday (Hasıl/Yeşilot)	Bakla, Kuru (Yemlik)	Sudan Otu (Yemlik)
Çayır Otu (Yeşilot)	Arpa (Yeşilot)	Çavdar (Yeşilot)
Bezelye (Yemlik)	İtalyan Çimi (Yemlik)	Üçgül Tohumu
Korunga Tohumu	Fiğ Tohumu	Fiğ (Adi) Tohumu
Fiğ (Macar) Tohumu	Fiğ (Diğer) Tohumu	Çim Tohumu
Haşhaş Kapsülü (Haşhaş Kellesi)	Yonca Tohumu	Isırgan Otu
Lavanta	Oğul Otu (Melisa)	Gül, Yağlık

Tablo 1: Veri setindeki ürünler

Tarım alanı veri setinde şunların bilgisi verilmektedir:

1. Meyveler, İçecek ve Baharat Bitkileri Alanı
2. Nadas Alanı
3. Sebze Alanı
4. Süs Bitkileri Alanı
5. Tahıllar ve Diğer Bitkisel Ürünlerin Alanı

Veriler Türkiye'deki 81 il için 2004 ile 2021 arasındaki istatistiklerden oluşan 1053 satırdan oluşmaktadır.

Tüm veri setlerinde İl ve Tarih bulunmaktadır fakat il bilgisi her bir il-tarih kombinasyonunun sadece ilk satırında bulunmaktadır ve bir sonraki il-tarih kombinasyonuna geçene kadar diğer satırlarında nan vardır. İl sütunundaki nan bilgileri doldurulmak için Pandas'ın ffill fonksiyonu kullanılmıştır. Bu fonksiyon nan veya null satırları bir önceki dolu satırla değiştirir. İl sütunundaki eksik verileri doldurduktan sonra veri setindeki tamamı nan veriden oluşan unnamed satırlar kaldırılmıştır. Bu işlemlerin sonucunda elde edilen veri setleri full outler join yöntemi ile birleştirilmiştir.

Veri setindeki eksik veriler bulundukları satırların ortalama değerleri ile doldurulmuştur. Ardından modele tahmin edilmesi için verilecek her bir ürünün yıllık üretim verisi tek bir satırda sıralanacak şekilde tüm il-yıl çiftleri ile kombinasyonunu oluşturmak için Pandas'ın melt fonksiyonu kullanılmıştır. Yapılan işleme ait örnek aşağıdaki şekillerde gösterilmiştir.

	Person	House	Age	Books	Movies
0	Alan	A	32	100	10
1	Berta	B	46	30	20
2	Charlie	A	35	20	80
3	Danielle	C	28	40	60

Şekil 1: Melt fonksiyonu uygulanmammış veri seti

	Person	House	variable	value
0	Alan	A	Age	32
1	Berta	B	Age	46
2	Charlie	A	Age	35
3	Danielle	C	Age	28
4	Alan	A	Books	100
5	Berta	B	Books	30
6	Charlie	A	Books	20
7	Danielle	C	Books	40
8	Alan	A	Movies	10
9	Berta	B	Movies	20
10	Charlie	A	Movies	80
11	Danielle	C	Movies	60

Şekil 2: Age, Books ve Movies sütunlarına melt fonksiyonu uygulandıktan sonra veri seti

Veri seti 1458 satır ve 179 sütundan oluşurken bu işlem sonucunda 125388 satır ve 95 sütun olmuştur. İl, tarih ve ürün üretim verisi sütunları makine öğrenmesi algoritmasına girdi olarak verilebilsin diye one-hot-encoding ile bu satırların ikili temsillerini içeren 186 yeni sütun eklenmiştir. Ayrıca yıllar içindeki değişimi vurgulamak için label-encoding ile tarih satırının sayısal temsilini içeren fazladan bir satır daha eklenmiştir.

Veriler 3'e ayrılmıştır:

1. 2014'ten önceki veriler eğitim, 69660 satır
2. 2013'ten sonra ve 2018'ten önceki veriler doğrulama, 27864 satır
3. 2017'den sonra ve 2021'den önceki veriler test, 27864 satır

Uygulanan modeller ve sonuçları aşağıdaki tablodadır.

Model	Eğitim	Doğrulama	Test
LinearSVR	- 1.149	-0.983	-1.125
LinearRegression	0.991	-4206	-2480
OrthogonalMatchingPursuit	0.496	0.413	0.371
XGBRegressor	1.000	0.894	0.883
LGBMRegressor	0.978	0.820	0.816
PLSRegression	0.950	0.742	0.832
ARDRegression	0.962	0.950	0.943
Ridge	0.970	0.706	0.888
RidgeCV	0.990	0.720	-1.671
Lasso	0.990	0.525	-1.844
LassoCV	0.990	0.525	-1.844
ElasticNet	0.829	0.770	0.777
LassoLars	0.988	0.948	-0.069
BayesianRidge	0.990	0.795	-1.305
TweedieRegressor	0.143	0.126	0.131
TweedieRegressor (power=1)	0.936	-1.490	-2.230
SGDRegressor	0.955	0.777	0.885
PassiveAggressiveRegressor	0.405	0.259	0.164
TheilSenRegressor	0.978	0.481	0.868

Tablo 2: Kullanılan modeller ve sonuçları

En başarılı 3 model sırası ile şunlardır:

1. ARDRegression
2. XGBRegressor
3. LGBMRegressor