**T.C.**

**BALIKESİR ÜNİVERSİTESİ**

**MÜHENDİSLİK FAKÜLTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

**YAPAY SİNİR AĞLARI KULLANARAK HABER KATEGORİZE ETMEK VE YORUMLANDIRMAK**

**201813709047 – İrem Bahar KOÇ**

**BMM4101 YAPAY ZEKA TEKNİKLERİ VİZE ÖDEVİ**

**Danışman: Dr. Öğr. Üyesi Kadriye ERGÜN**

**BALIKESİR, 12 – 2022**

Ödev için veri seti konusunda baya araştırma yaptığımı söyleyebilirim. İlk önce ne yapacağımızı tam olarak anlayamadım daha sonrasında haber kategorizasyonu yapmamız gerektiğini düşündüm ve bunun için haber verileri bulabilmek için önerdiğiniz bütün sitelere tek tek girdim, inceledim. Kemik sayfasındaki veriler çok güzeldi fakat verilerin çoğu ham ve .arff şeklinde oldukları için csv ye çevirmek zaman alacaktı. Araştırmaya kaggle ile devam ettim. önerdiğiniz gibi arama kelimelerini değiştirerek veri buldum. Verilerin .csv formatında olduğunu fark edince orada bulduğum haber verilerini aldım.

**Verileri aldığım site:** [**https://www.kaggle.com/search?q=text+classification+turkish**](https://www.kaggle.com/search?q=text+classification+turkish)

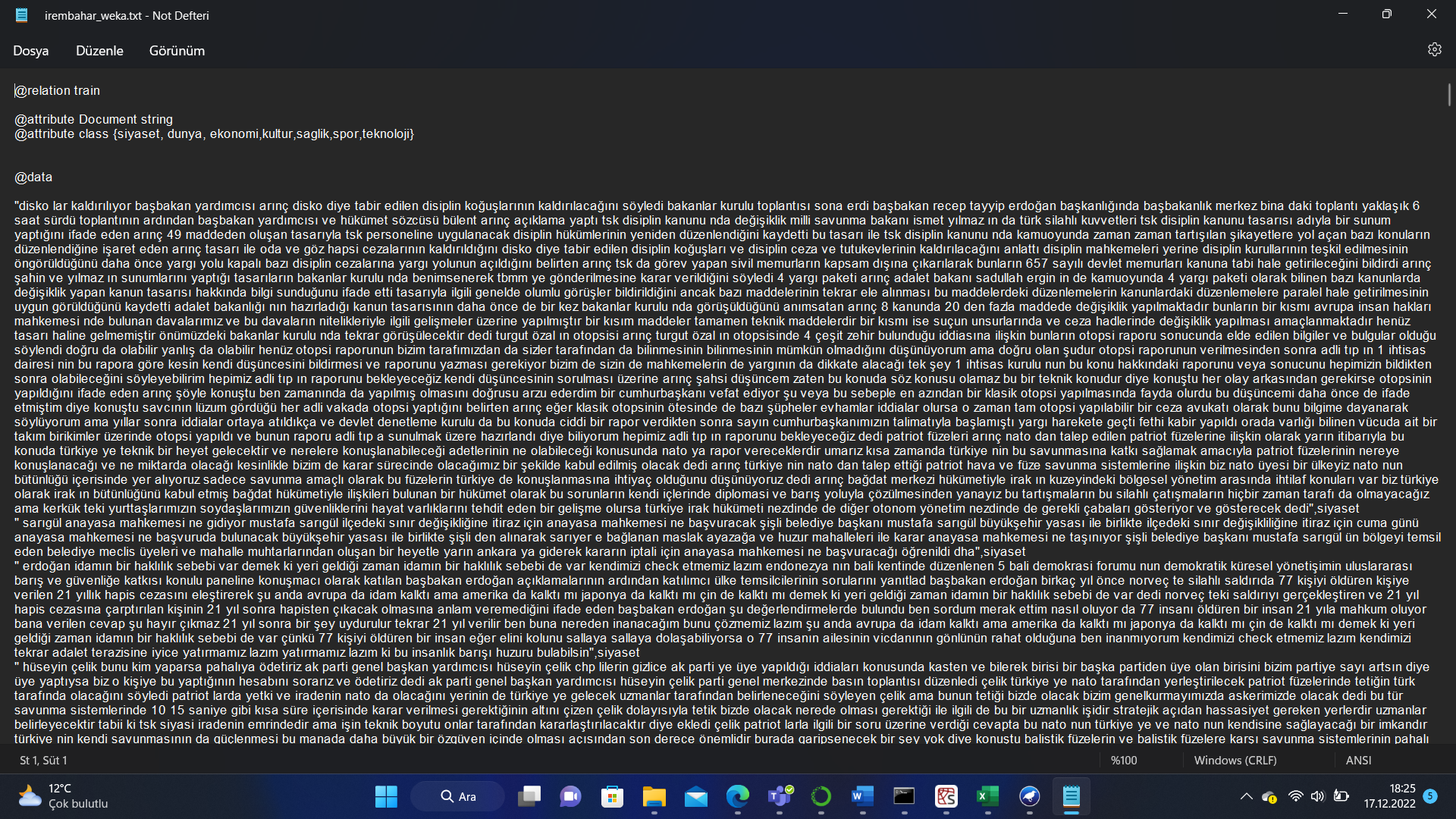
Aşağıda tokenize ettiğim, her kategoriden bir tane olmak üzere örnek haberler bulunmaktadır.

|  |  |
| --- | --- |
| kılıçdaroğlu na el maliki nin istismarına alet olma uyarısı mhp genel\_başkanı bahçeli türkiye ye karşı hasmane tavırlar sergileyen irak\_başbakanı\_maliki nin davetini kabul eden chp lideri kılıçdaroğlu nu üstü kapalı uyardı | siyaset |
| endonezya da yanardağ aktif hale geldi endonezya nın sulawesi adasındaki lokon yanardağı aktif hale geldi lokon dağından çıkan dumanlar 600 metre çapındaki alanı olumsuz etkiliyor | dunya |
| tuşa dalgın tüik çi bastı tüik in kasım ayı enflasyon verilerini 17 dakika önce açıklamasının sırrı çözüldü verilerin 17 dakika önce açıklanmasında iddia edildiği gibi hackerlerin değil dalgın uzmanın rolü olduğu ortaya çıktı hüseyin özay piyasalarda polemiğe yol açan tüik in kasım ayı enflasyon verilerini 17 dakika önce açıklamasının sırrı çözüldü verilerin 17 dakika önce açıklanmasında iddia edildiği gibi hackerlerin değil dalgın uzmanın rolü olduğu ortaya çıktı söz konusu verilerin ayrıca siteye yaklaşık 30 45 dakika önce yüklendiği de belirtildi türkiye\_istatistik\_kurumu nun tüik kasım ayı enflasyon verilerinin 17 dakika önce açıklanması ile ilgili ön incelemesi tamamlandı olayda ihmali bulunan personelle ilgili soruşturma ise devam ediyor soruşturmanın sonuçlanmasının ardından ise ilgili personele verilecek cezalar netleşecek uzman şifre koymayi unutmuş tüik in yaptığı ön incelemeye göre enflasyon rakamlarının erken açıklamasında uzmanların dalgınlığı etkili oldu star ın tüik yöneticilerinden edindiği bilgiye göre enflasyon gibi veriler resmi internet sitesine stabil veriler ve bülten olarak iki aşamada yükleniyor verilerin sisteme yüklenmesi uzun sürdüğü için de stabil veri olarak nitelendirilen tabloların siteye yüklenmesi açıklanmadan çok önce gerçekleştiriliyor tüm araştırmaların stabil verileri de bu şekilde siteye yükleniyor uzmanlar verilerin tüm oyunculara aynı anda ulaşması için de stabil verilere şifre koyarak siteyi yüklüyorlar saat geldiğinde ise şifre kaldırılıyor ve verilere tüm oyuncuların ve müşterilerin aynı anda ulaşması sağlanıyor kasım ayı enflasyon rakamları da yaklaşık saat 09 00 dan itibaren siteye girilmeye başlandı ancak bu kez uzman stabil olarak nitelendirilen verileri şifre koymayı unuttu ve saat 10 00 da açıklanması gereken veriler böylece erişilebilir hale geldi yapılan ilk araştırmada şifreleme işleminin gerçekleştirilmemiş olmasında uzmanların dalgınlığının etkili olduğu görüldü ancak tüik yine de bir kasıt olup olmadığını araştırıyor spk beklemede öte yandan enflasyon verileri ile borsada alım satım yapıldığı yönündeki iddialar üzerine gözler spk ya çevrildi spk ise konuyu ve iddiaları değerlendirdi yapılan değerlendirmede inceleme için tüik in soruşturmasının sonuçlanmasının beklenmesi kararlaştırıldı rapora göre spk işlemlerle ilgili inceleme başlatabilecek | ekonomi |
| afife balesi 14 yıl aradan sonra yeniden sahnede ilk müslüman\_türk kadın tiyatro oyuncusu afife\_jale nin hayatını konu alan afife balesi 14 yıl aradan sonra yeniden sahnelenecek devlet\_opera ve balesi koreografisi ve librettosu beyhan murphy e müzikleri turgay erdener e ait afife adlı modern bale eserini ankara devlet\_opera ve balesi nde sergilenişinden 14 yıl sonra yeniden izleyiciyle buluşturacak müslüman\_türk kadınlarının sahneye çıkmasının yasak olduğu 1900 lü yılların başında tüm olumsuzluklara karşın sahneye çıkan ilk müslüman\_türk kadın tiyatro oyuncusu afife\_jale nin hayatı üzerine kurulu bir modern bale prodüksiyonu olan afife nin prömeyeri 22 kasım da süreyya operası nda yapılacak balede afife rolünü tülay yalçınkaya ile zuhal balkan ebru cansız deniz zirek ve ilke kodal canlandıracak afife nin hayatında önem teşkil eden ziya karakterini erhan güzel mehmet\_ali karakterini berk sarıbay dr suat karakterini egemen kement ve selahattin pınar karakterini ise arkın zirek hayata geçirecek aynı zamanda dans drama olan ve çağdaş türk bale repertuvarında özel bir yere sahip afife afife\_jale nin hayatındaki altın kırmızı mor ve gümüş olarak tanımlanan dört ayrı dramatik dönemi 25 kişilik bir ensemble hareketli ve lirik danslar ile canlandıracak kostüm tasarımları bahar korçan a ait eserin dekor tasarımıysa adnan öngün imzasını taşıyor eser 22 24 27 29 kasım ve 1 aralık ta süreyya operası nda izlenebilecek | kultur |
| fazla oturmak ölüm riskini artırıyor ingiltere de yapılan bir çalışma fazla oturmanın diyabet kalp hastalıkları ve ölüm riskini artırdığını ortaya koydu leicester ve loughborough üniversitelerinin diabetologia dergisinde yayımlanan araştırmaya göre uzun süre hareketsiz kalan kişiler iş çıkışı spor salonuna ya da havuza giderek egzersiz yapsalar dahi hareketsiz kaldıkları sürenin olumsuz etkisini tam olarak ortadan kaldıramıyor uzun süre oturanların oturmayanlara göre diyabet ve kalp rahatsızlığı geçirme riskinin daha fazla olduğunu vurgulayan bilim adamları televizyon izlemek ya da bilgisayar kullanmak gibi insanı hareketsizliğe iten alışkanlıkların modern toplumlarda oldukça yaygın hale gelmeye başladığına da işaret ediyor araştırmacılar hareketsizliğin önüne geçebilmek için bazı iş toplantılarının ayakta düzenlenmesi öğle yemeği için ayrılan sürede yürüyüş yapılması ya da akşamları televizyon izleme süresinin azaltılması gibi önerilerde bulunuyorlar | saglik |
| kıran kırana mücadele 2012 2013 sezonunda 10 haftanın tamamlandığı ptt 1 lig de oynanan maçlarda 14 takım geriye düştükleri müsabakalardan puan çıkarmayı başardı çaykur\_rizespor un 21 puanla lider tamamladığı ptt 1 lig de 10 haftaya kadar oynanan 90 karşılaşmanın 62 si en az bir takımın galibiyetiyle biterken geri kalan 28 karşılaşma berabere tamamlandı ev sahibi ekiplerin 37 galibiyetinin bulunduğu ligde deplasman takımları da 25 kez sahadan 3 puanla ayrıldı ligde şimdiye kadar oynanan karşılaşmalarda 14 takım maç içinde ilk golü kalelerinde görmelerine rağmen müsabakadan puanla ayrılmayı bildi 1461 trabzon maçlarda yenik duruma düşmesine rağmen en fazla puan alan ekip 1461 trabzon 9 olurken karadeniz temsilcisini kayseri\_erciyesspor 8 boluspor çaykur\_rizespor ve gaziantep\_büyükşehir\_belediyespor 6 takip etti adanaspor denizlispor göztepe ve manisaspor ise şimdiye kadar yenik duruma düştükleri müsabakalardan puan çıkaramadı en başarısız ekip olan göztepe yenik duruma düştüğü 6 maçtan da puan alamadı 1461 trabzon un 3 kez yenik durumdan maçı çevirmeyi başardığı şimdiye kadarki maçlarda çaykur\_rizespor gaziantep\_büyükşehir\_belediyespor ve kayseri\_erciyesspor 2 boluspor bucaspor ve kartalspor da birer kez mağlup duruma düşmesine rağmen müsabakayı 3 puanla tamamladı | Spor |
| sony playstation 4 için 4k verdi bu flaş gelişme ile ilgili olarak sony electronics başkanı phil molyneux çok önemli açıklamalar yaptı bu flaş gelişme ile ilgili olarak sony electronics başkanı phil molyneux çok önemli açıklamalar yaptı playstation 4 film servisi olacağı ancak zamanı hakkında açıklama yapılmadı bu servisin ne zaman açılacağı hakkında bilgi vermeyen yetkili yine de 4k filmlerin bir gün mutlaka ps4 için satılacağını belirtti phil molyneux un bir diğer önemli açıklaması ise 4k filmlerin boyutunun 100 gb ve üstü olacağını söylemesiydi bu kadar büyük boyutlu filmleri müşteriye online olarak sunmanın problem olacağının farkında olan sony nin bu sorunu çözmek için çeşitli fikirleri varmış bakalım bu konu hakkında sony electronics başkanı phil molyneux başka açıklamalar yapacak mı | teknoloji |

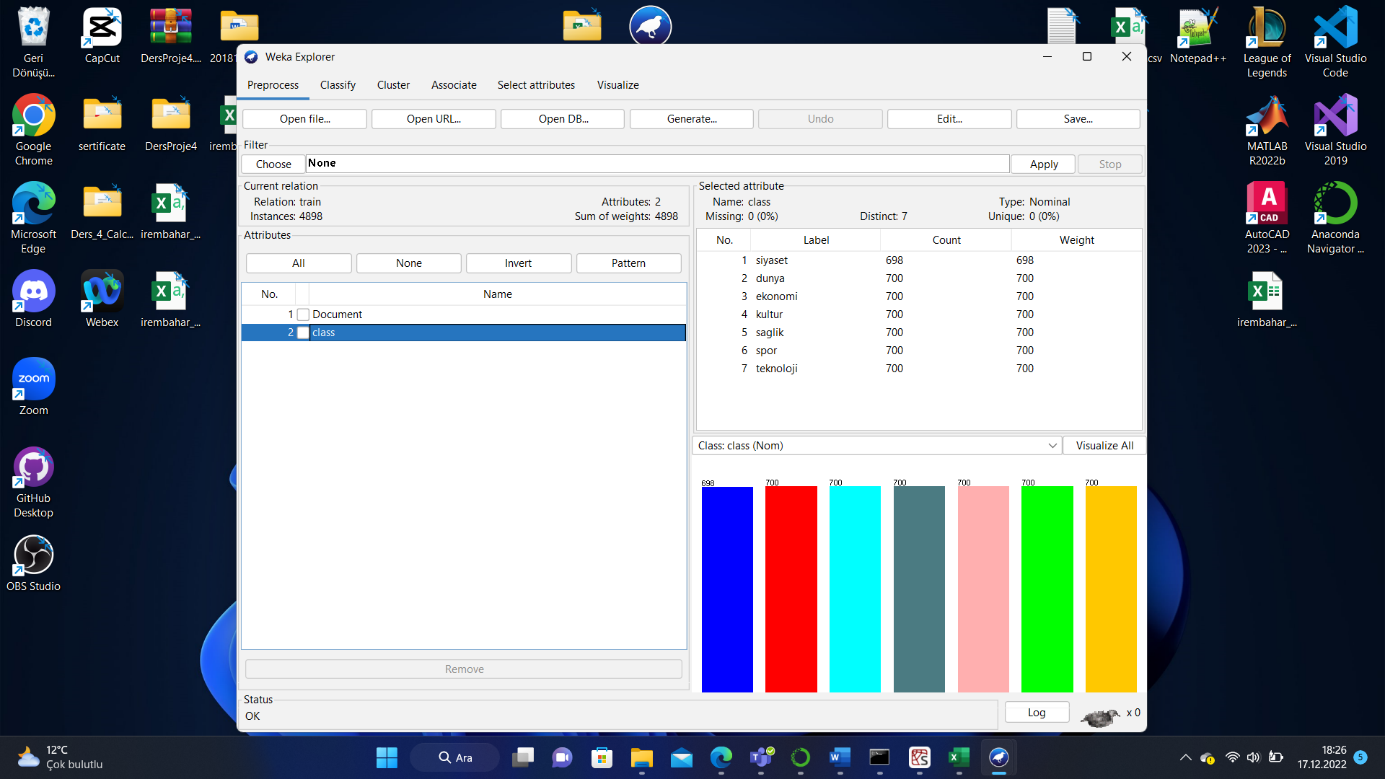
Veriler kategori ve haber olmak üzere iki sütunda tek excel dosyasında bulunuyordu(irembahar\_news.csv). Python’da yazdığım (irembahar\_remove\_puncc.py) kod sayesinde yorumlarda bulunan noktalama işaretlerini kaldırdım ve kod ile bir dosya oluşturup haberlerin noktalamasız hallerini yeni oluşturduğum dosyaya aktardım(irembahar\_removepunc\_news.csv).

Daha sonra yazdığım diğer kodla (irembahar\_tokenize.py) bir önceki dosyada( irembahar\_remove\_puncc.py) bulunan cümlelerin harflerini küçülttüm, tırnak ekledim. Notepad++ üzerinden birkaç düzenleme de yapmam gerekti. Böylelikle tokenize işlemini yapmış oldum.( irembahar\_tokenize.csv)

Tokenize işleminden sonra artık sıra verileri sayısallaştırma işlemine geldi. Bunun için öncelikle haberlerin ve kategorilerin bulunduğu excel dosyasını derste gösterdiğiniz şekilde Not Defteri üzerinde düzenleme yaptım. Örneğin boşluk sileceksem sileceğim alanı seçip ctrl+h veya düzenle>değiştir kısmından seçilen boşluğun yerine gelecek boşluğu ayarlayarak örnek bir düzenleme yaptım. Weka’nın dosyayı okuyabilmesi için üzerine datasını ve kategorilerini yazdıktan sonra bu işlemi bitirmiş oldum.



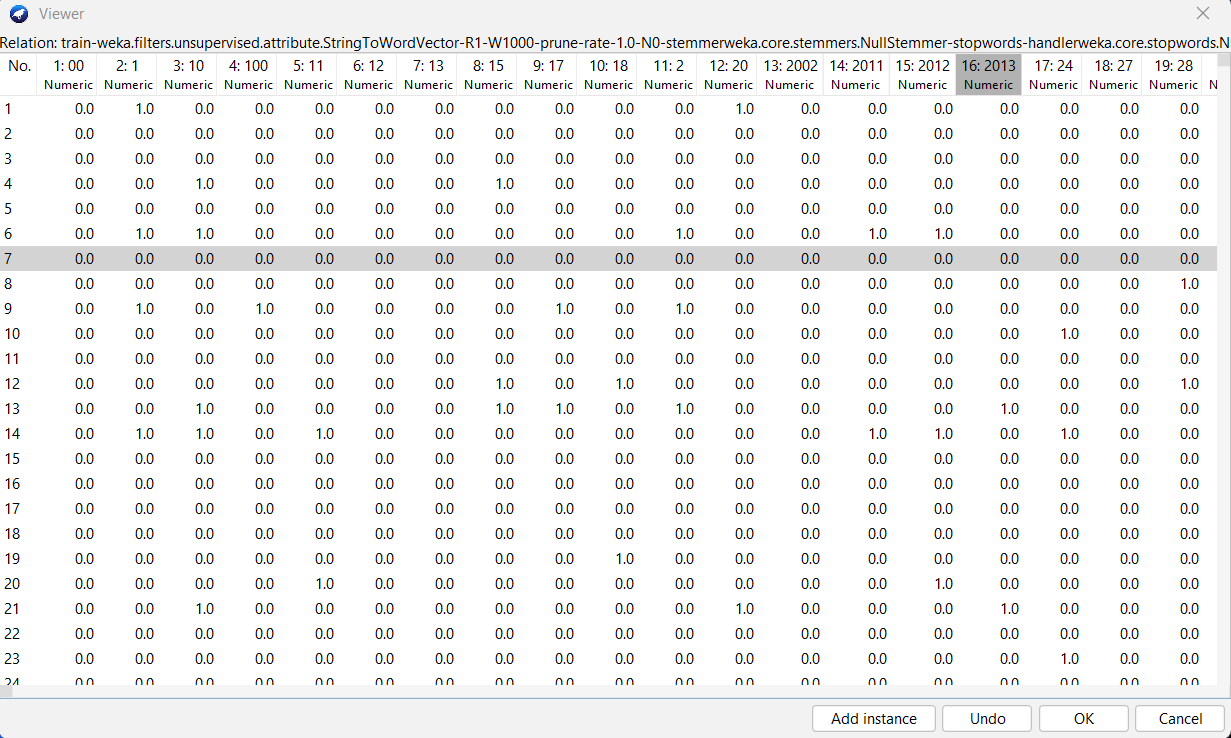
Haberleri düzenledikten irembahar\_tokenize.csv dosyasını arff formatına çevirecektim fakat sonrasında program içinde dosya seçerken weka’nın artık csv dosyasını okuyabildiğini gördüm. Bu sebeple direkt dosyayı yükledim. Wekaya dosyayı yükledikten sonra verileri sayısallaştırmasını yani vektöre çevirmesini beklediğim için gereken işlemlere şöyle devam ettim:



Choose’a tıklayarak StringToWordVector filtresini seçtim. Bu filtre ile veriler vektöre dönüştürülecekti.

StringToWordVector’un üzerine tıkladım ve ayarlarını açtım. Stemmer ayarını NullStemmer’den SnowballStemmer’a çevirdim çünkü SnowballStemmer’in Türkçe karakterler için kullanılıyor, bizim verilerimiz de Türkçe olduğu için bu seçeneği işaretledim.

Daha sonra Apply diyerek StringToWordVector’ü aktif hale getirdim. Edit sekmesine girdim ve oluşturduğu vektörleri gördüm. Üst sekmesine sağ tıklayarak attribute as class seçeneğine tıkladım.



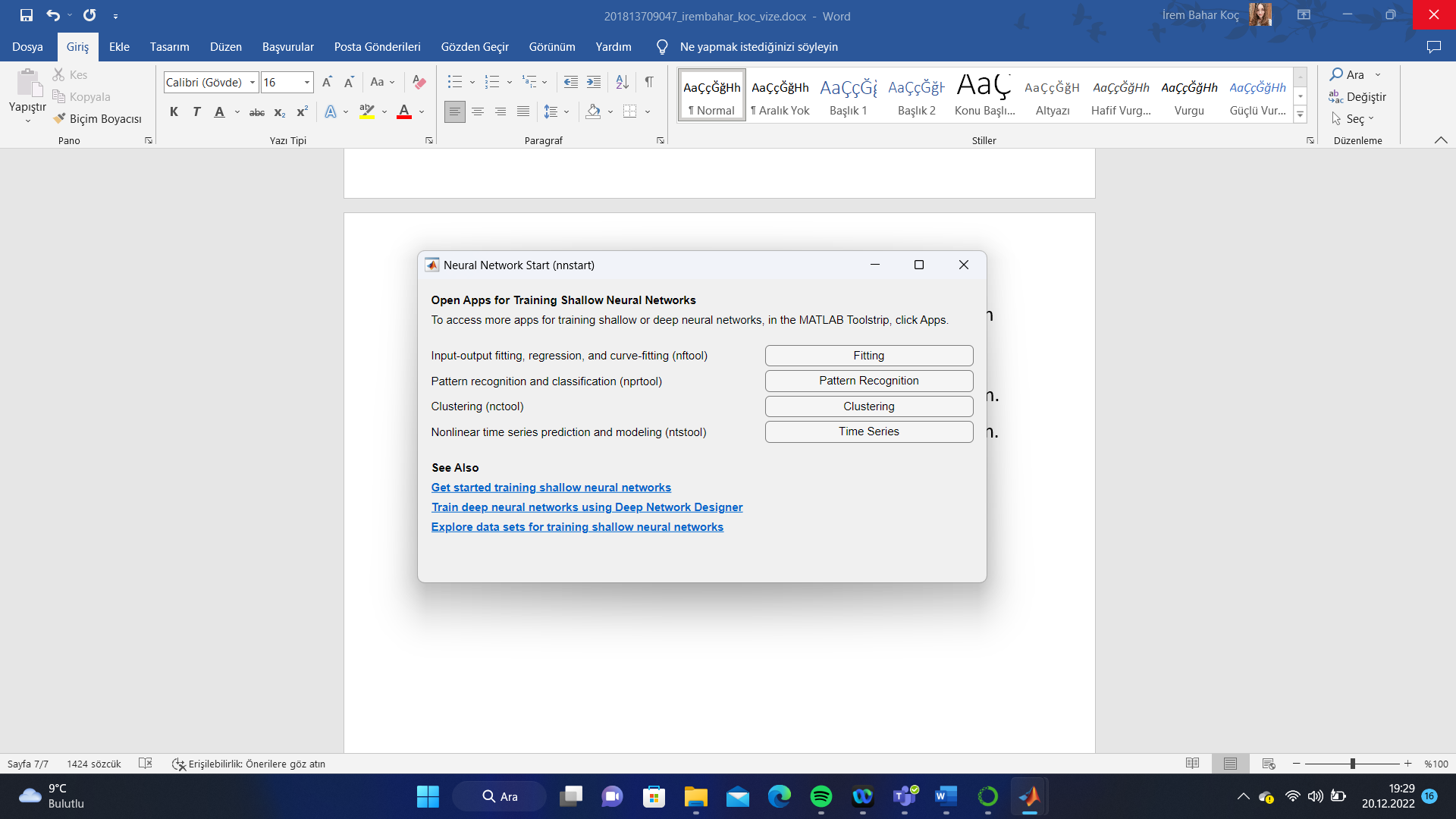
Üstte de görüldüğü üzere her bir kelime vektöre çevrildi. Fakat burda şöyle bir sorunla karşılaştım. Kategorilerim olan kelimeler(siyaset,dunya vb.) vektöre çevirilmedi. Bunun üzerine haberleri ve kategorileri iki ayrı Excel dosyası haline getirdim (irembahar\_just\_news , irembahar\_just\_cats).

irembahar\_just\_cats dosyasındaki verileri nasıl kategorize edebilirim diye biraz araştırma yaptıktan sonra verileri nasıl kategoreize edebilirim diye bir Python kodu yazdım sonucunda kategorilere tek tek 0’dan 6 ya kadar indeks numarası verdiğini gördüm(cats\_vector.py). Dosyanın kopyasını alarak kategorilere tek tek 0 dan başlayarak sayı verdim(0=siyaset,1=dunya,2=ekonomi,3=kultur,4=saglik,5=spor,6=teknoloji olacak şekilde). Sonrasında bu vektör verilerini irembahar\_output olarak kaydettim.

Weka’nın çevirdiği haberleri ise irembahar\_input şeklinde kaydettim.

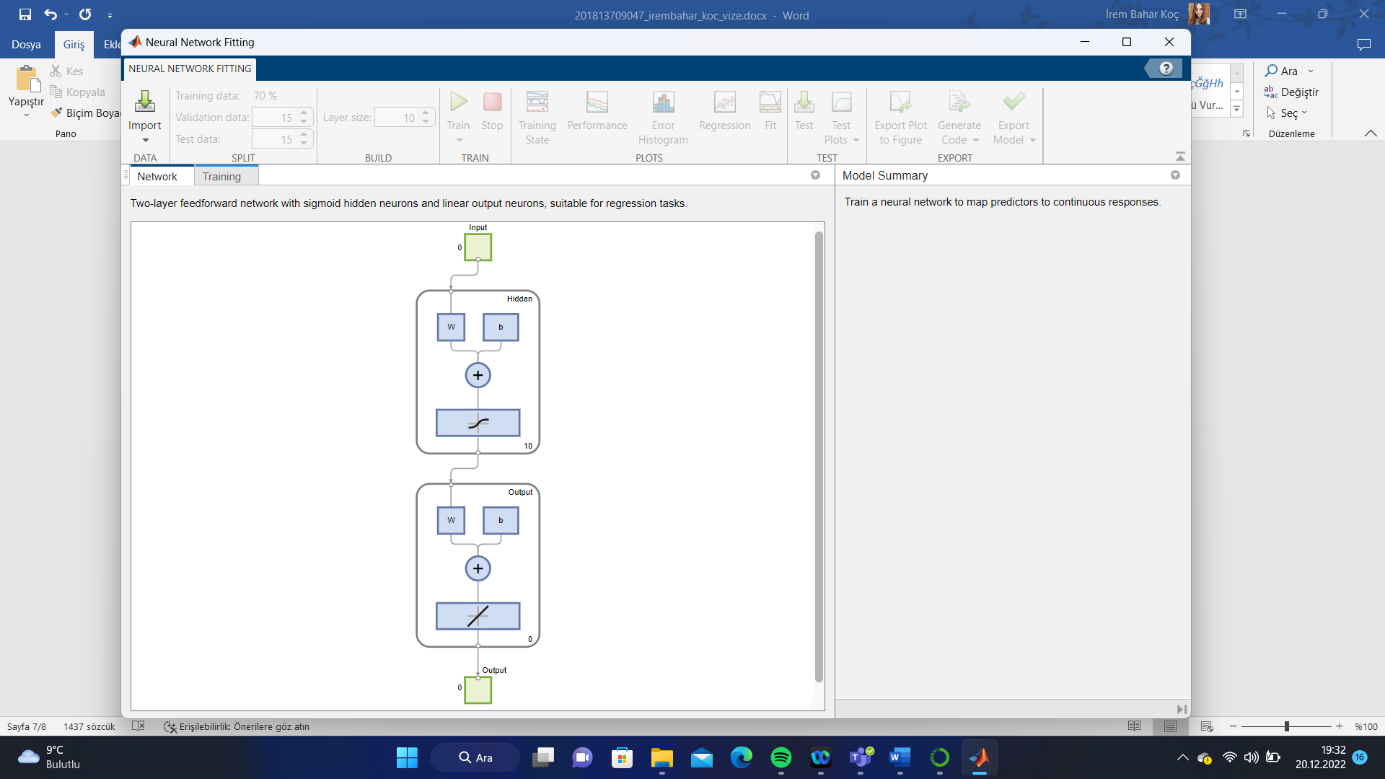
Artık sıra geldi matlab üzerinde sinir ağı oluşturarak sonuçlara bakmaya.

Matlab’ı açtım ve nnstart (neural network start) yazarak yapay sinir ağı oluşturma işlemini başlattım. Karşıma gelen ekranda Fitting’i seçerek datayı ekleyeceğim ve yapay sinir ağını görebileceğim ekranın karşıma gelmesini bekledim. Burada kullandığımız fitting’i verilerimizi modellemek ve anlamlı veriler çıkarabilmek için kullanıyorum.

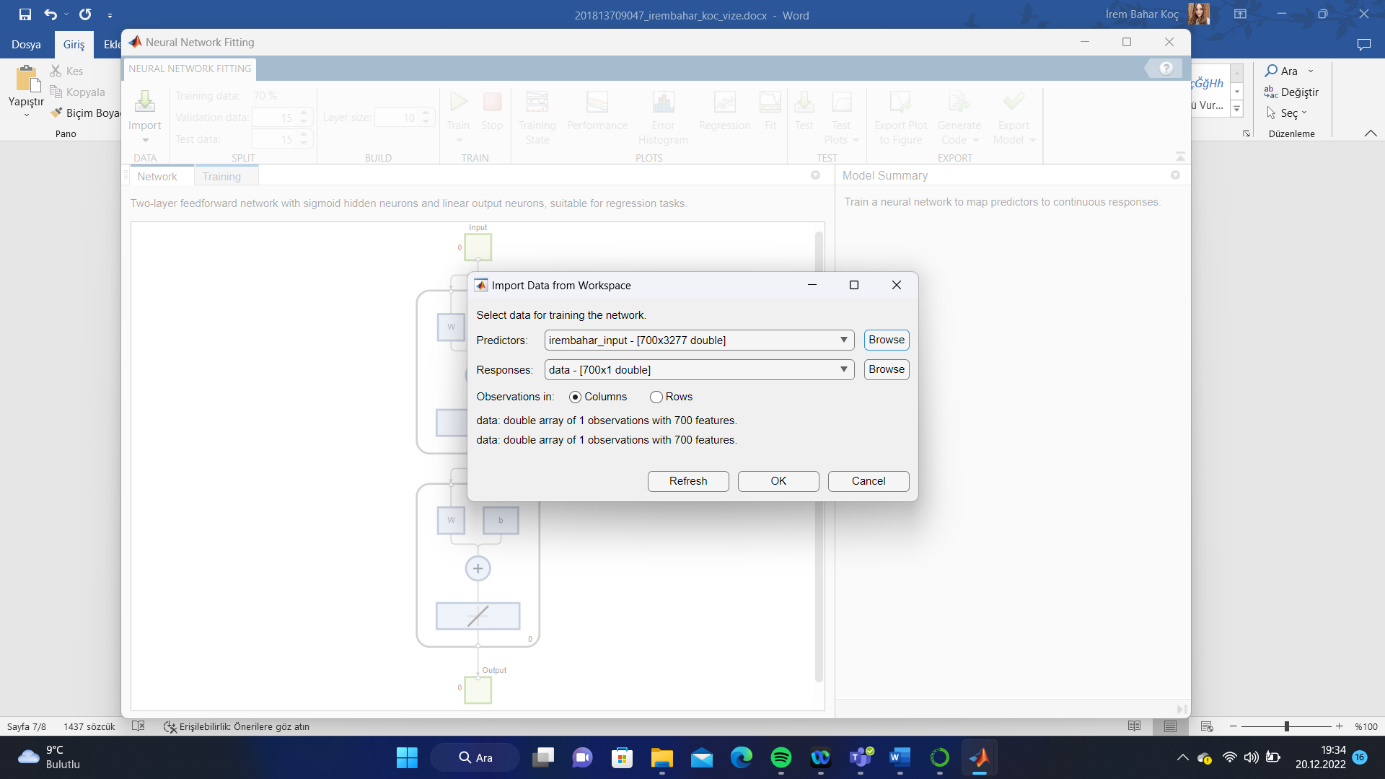


Karşıma aşağıdaki ekran geldi. Burda yapay sinir ağımın örnek bir gösterimi mevcut. Sol üstteki import kısmından datamı seçtim.

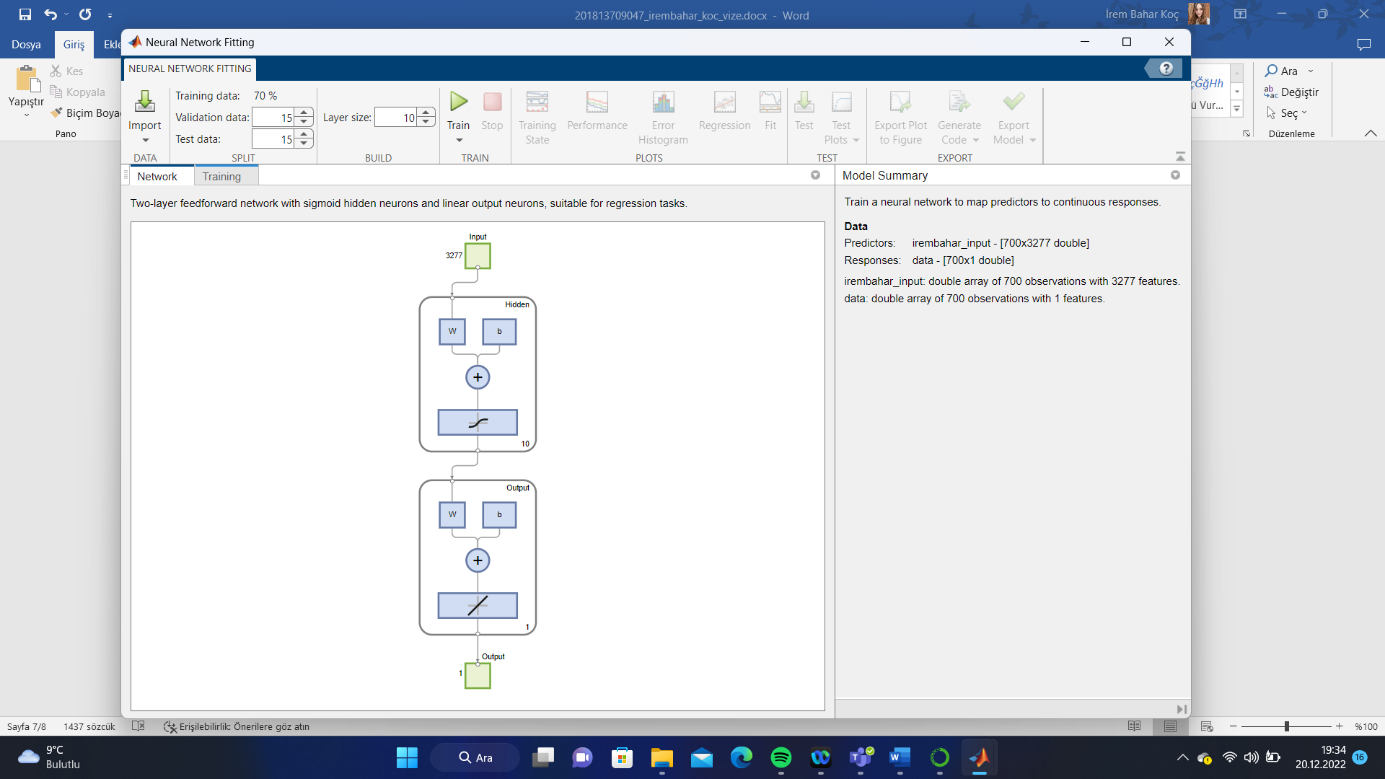
Burada input verilerimin yani haberlerimin vektör matrisinin istenilen maksimum matris büyüklüğünden(8GB) fazla olduğunun hatasını aldım. Böyle olunca işlemleri baştan sona tekrar yaptım. Her kategoriden 700’er tane haber vardı. Birkaç kere denedim. Her seferinde büyük geldiği için verilerimi çokca kırptım. Kırpmalarım sonucunda her kategoriden 100er tane haber kaldı. Dosyaları excel ve notepad de düzenledim ve tekrardan matlaba yüklenebilir hale getirdim.



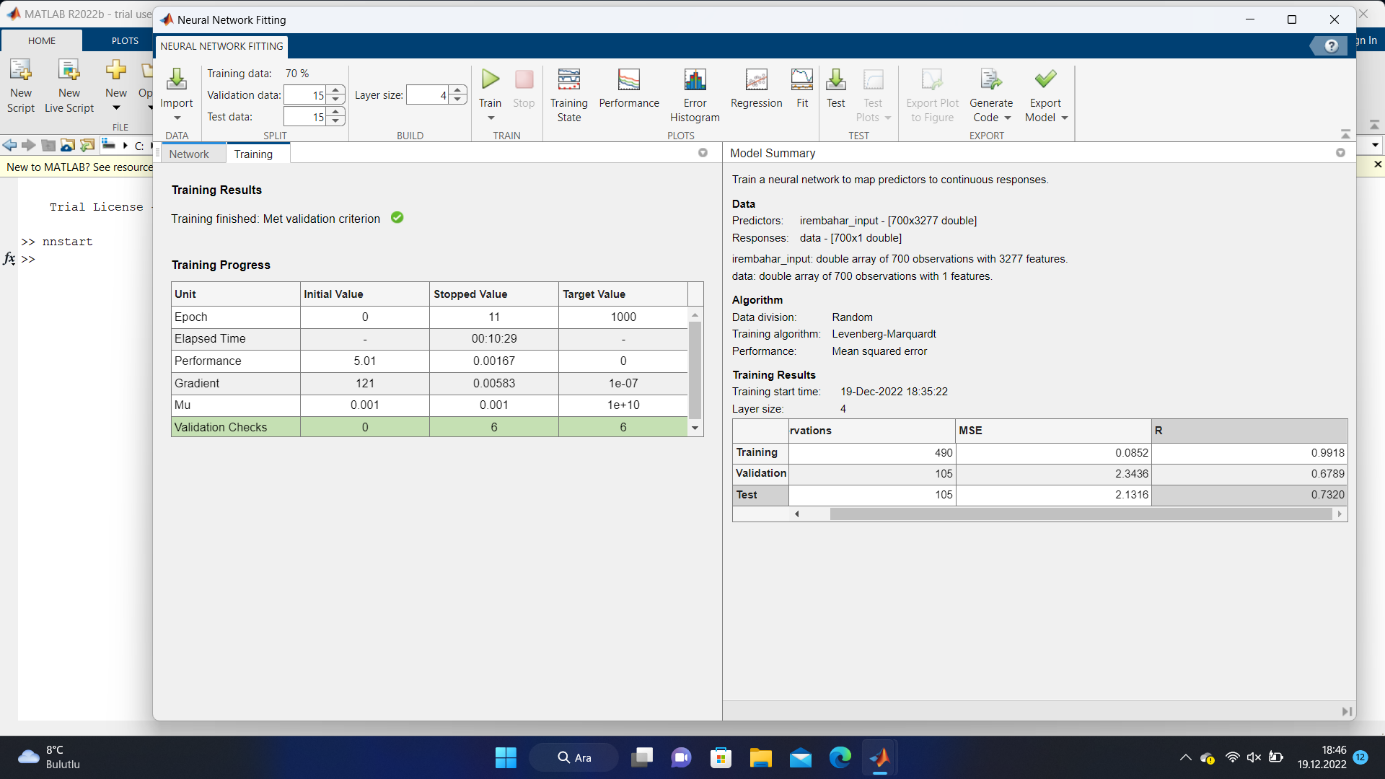
Ok ile gösterdiğim kısımlardan irembahar\_input ve irembahar\_output dosyalarımı girdi ve çıktı vektörleri olarak ekledim. Observation kısmından da row’ u seçerek datayı ağıma yükledim. Matlab vektörleri hep satır vektörü olarak okur fakat benim vektörlerim sütun vektörü olduğu için row seçeneğini işaretlemek beni matrislerin transpozesini almaktan kurtarmış oldu. Burada bir diğer önemli olan şeyler satır sayılarının eşit olmasıdır buna dikkat ettim aksi takdirde veriyi ağa yükleyemiyordum.



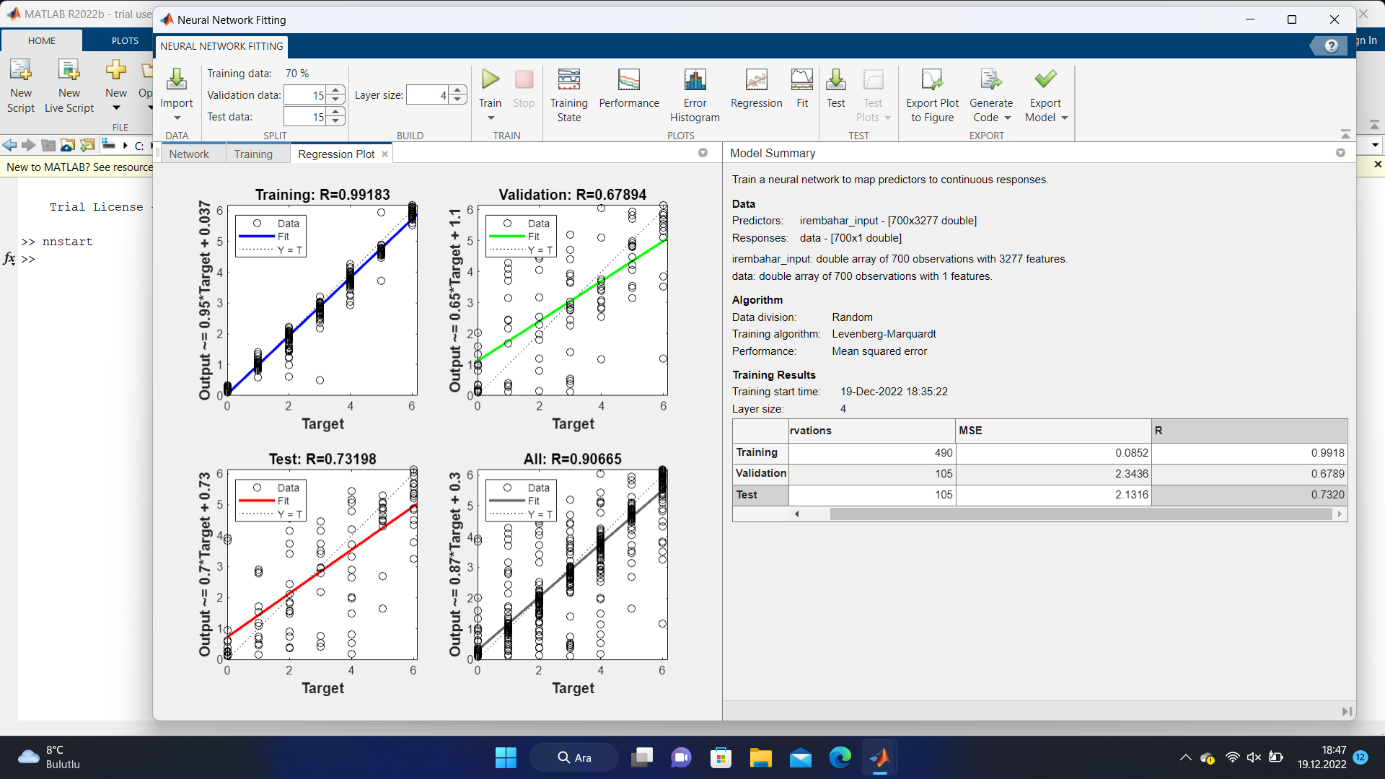
Aşağıda da görüldüğü üzere 3277 input girdi ve 1 çıktı olarak yapay sinir ağımız geldi. Sıradaki işlem train yani eğitim işlemi bunu da ok ile gösterdiğim train kısmına basarak yapıyoruz. Tabi önce test datasını, katman sayısını ayarlamak gerekiyordu. Kayman sayısını ilk başta 10 yaptım fakat bilgisayarım ciddi kasmaya başladı. Sonrasında sırayla katman sayısını 2 3 4… şeklinde arttırdım.



Bizden optimum iterasyonun olduğu veriyi istediğiniz için 4 gizli katman ile verimizi eğittiğimizde aşağıda gözüktüğü gibi bu modelle %90 oranında verilerin hangi kategoride olduğunu tahmin edebilecektim. Burada trainlerken levenberg-Marquardt algoritmasını kullandım.



Yukarıda gördüğümüz epoch iterasyon sayısını gösteriyor. Yani biz 8 iterasyonda sonucu bulmuşuz demek. Gradient, varsayılan olarak alının değeri, hedef değeri ve ortadaki sütunla bu hedefe ne kadar yaklaştığımızı gözlemleyebiliriz. Mu, modelin öğrenme hızını belirtir. Öğrenme hızı, nöronların giriş verilerine nasıl yanıt vereceğini belirleyen bir değerdir. Eğer mu değeri düşürülürse yapay sinir ağı daha yavaş öğrenme gerçekleştirir ancak daha emin bir öğrenme gerçekleştirir. Eğer mu değeri yükseltilirse, yapay sinir ağı daha hızlı öğrenir. Kısaca mu değerinin doğru seçilmesi, yapay sinir ağımızın performansını ve güvenilirliğini etkileyeceğinden önemlidir. Burada görünen epoch, gradiant, mu değerleri varsayılan değerleridir. Bu değerleri değiştirmek istersek garklı bir algoritma seçebiliriz.



Üstte gördüğümüz sayfa bizim beklenen değerimiz ile gerçekleşen çıkıt değerimiz arasındak, regresyonu gösteriyor. Burada görüdüğü gibi train yani eğitim setinde sonucu doğru bulma oranımız %99 iken validationda yani doğrulama setinde %67 gibi bir oran gözlemlenmekte. Nihayetinde genel olarak baktığımızda yukarıda da söylediğim gibi %90 oranında bir doğru tahmin söz konusu olmuş oldu.

Daha sonrasında verilerin karşılaştırmasını yapacaktım. Fakat tekrar train ettiğimde saatlerce beklememe ve katman sayısı 2 olmasına rağmen bilgisayarım bir sonuç üretmedi. Birkaç kere denedim, hepsinde aynı şekilde bilgisayarım dondu ve kapandı. Çıktı verilerine göre birebir karşılaştırma yapamadım fakat şöyle olacaktı. Öncelikle kategorilerimdeki vektörler 0 dan büyük olduğu için bunlar çok yakın yanıtlar vermeyecekti. Yakın yanıtlar vermesi için min max normalizasyonu yapılırsa yani sonuçlar 0 ile 1 arasında bir değere dönüştürülürse daha optimal bir sonuç bulunabilirdi.