**YAZ / 2017**

**İSTANBUL TEKNİK ÜNİVERSİTESİ**

**Bilgisayar ve Bilişim Fakültesi**

**STAJ**

**Kadir Emre Oto**

**150140032**

**Makine Öğrenmesi: Film Öneri Sistemi**

**İstanbul Teknik Üniversitesi**

**Bilgisayar ve Bilişim Fakültesi**

**STAJ RAPORU**

Akademik Yıl: 2017

Staj yapılan dönem:☑Yaz ☐Bahar ☐Güz

**Öğrenci ile ilgili bilgiler**

|  |  |
| --- | --- |
| Adı ve Soyadı: | Kadir Emre Oto |
| Öğrenci Numarası: | 150140032 |
| Bölüm: | Bilgisayar Mühendisliği |
| Program: | %30 İngilizce |
| E-posta Adresi: | [otok@itu.edu.tr](mailto:otok@itu.edu.tr) |
| (Cep) Tel No: | 0 (507) 713 83 90 |
| ÇAP öğrencisi misiniz? | ☐Evet (ÇAP yaptığınız Fakülte/Bölüm: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)  ☑ Hayır |
| Mezuniyet  durumunda mısınız? | ☐ Evet  ☑ Hayır |
| Yaz okulunda ders alıyor musunuz? | ☐Evet (Ders sayısı: \_\_)  ☑ Hayır |

**Öğrencinin çalıştığı kurum ile ilgili bilgiler**

|  |  |
| --- | --- |
| İsmi: | Erstream Video Delivery Company |
| Birimi: | Yazılım Departmanı |
| Web Adresi: | <http://erstream.com> |
| Kısa Adresi: | Defterdar Mah. Otakçılar Cad. 80/9, Eyup | 34050 Istanbul | Turkey |

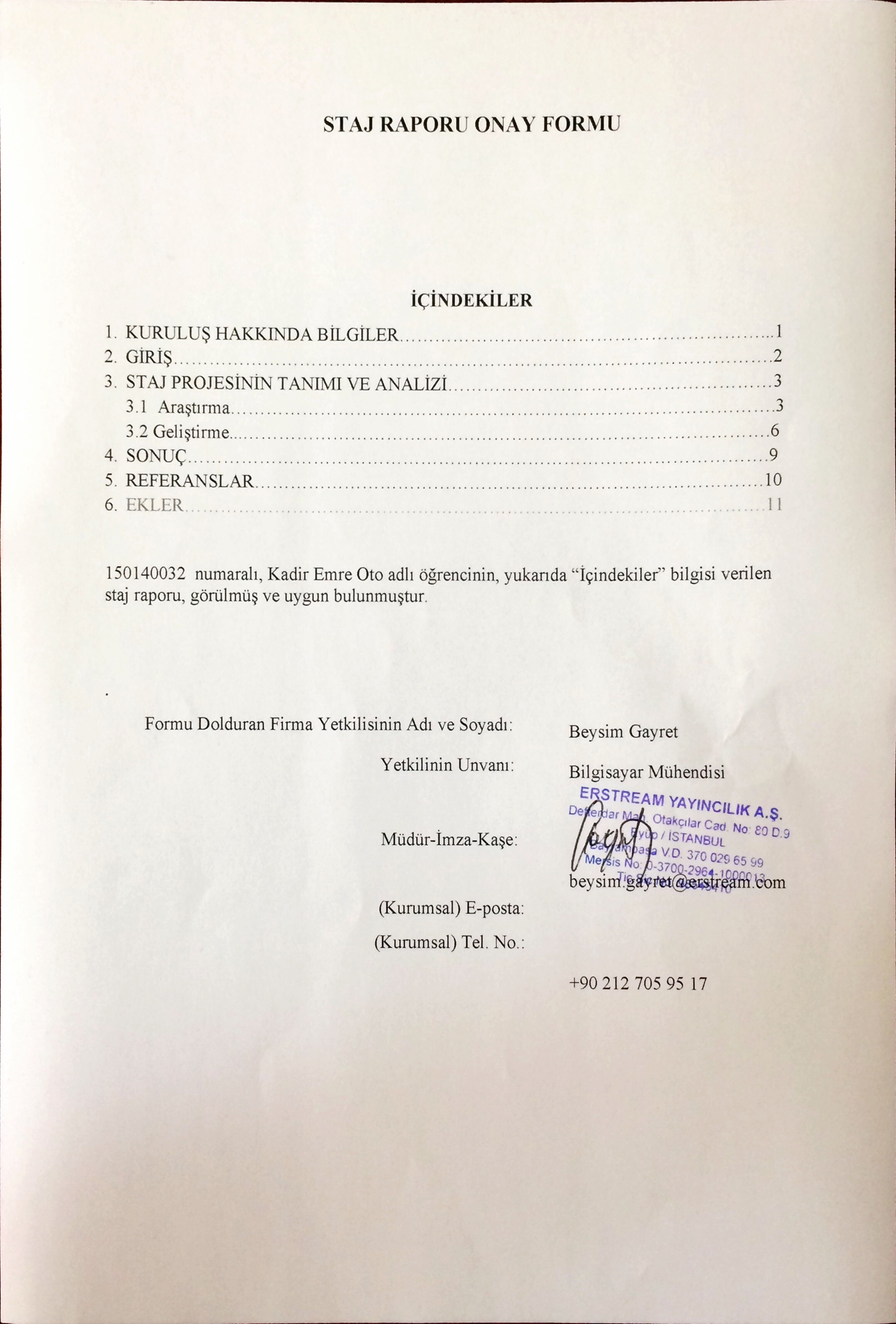
**Yetkili kişi ile ilgili bilgiler**

.

|  |  |
| --- | --- |
| Bölümü: | Yazılım Departmanı |
| Unvanı: | Bilgisayar Mühendisi |
| Adı ve Soyadı: | Beysim Gayret |
| (Kurumsal) E-posta: | [beysim.gayret@erstream.com](mailto:beysim.gayret@erstream.com) |
| (Kurumsal) Tel. No.: | +90 212 705 95 17 |

**Yapılan iş ile ilgili bilgiler**

|  |  |
| --- | --- |
| Staj yeri | ☑Türkiye  ☐Yurtdışı |
| Staj başlangıç tarihi | 19.06.2017 |
| Staj bitiş tarihi | 21.07.2017 |
| Stajda çalışılan net **gün** sayısı | 20 |
| Staj süresince sigortanız var mıydı? | ☑Evet, İTÜ tarafından sigortalandım.  ☐Evet, kurum tarafından sigortalandım.  ☐Hayır, yurtdışı stajı yaptım.  ☐Hayır. |

****

**KURULUŞ HAKKINDA BİLGİLER**

Uğur Kalaba, Radoslav Raychev ve Yağız Buran tarafından 2003 tarihinde kurulmuş olan Erstream Video Delivery Com pany internet video teknolojileri alanında ürettiği ürünler ve sunduğu hizmetler ile bu alanda öncü firmalardan biri olmuştur. Dünya çapında ün yapmış içerik sahipleri ve yayıncılar ile birlikte Türkiye’de yayıncılık yapan bir çok firma Erstream ile işbirliği içerisinde çalışmaktadır.

Erstream’in dünya çapında yüzlerce internet ve televizyon kanalına yayın sağlayabilmesinin temel nedeni Avrupa’daki bir çok ülkenin veri merkezlerinde encoding ve downlink hizmetlerine sahip olmasıdır.



**Şekil 1** – Erstream Logosu

**GİRİŞ**

Ersteam staj programına kabul ettiği stajyerlere şirketin üzerinde çalıştığı bir çok ürün ve sağladığı hizmetler üzerinde çalışma imkanı sağlamakta ve tecrübelerini aktarmayı hedeflemektedir. Bunlara ek olarak çeşitli araştırma-geliştirme (AR-GE) projelerinde de çalışma imkanı sunmaktadır.

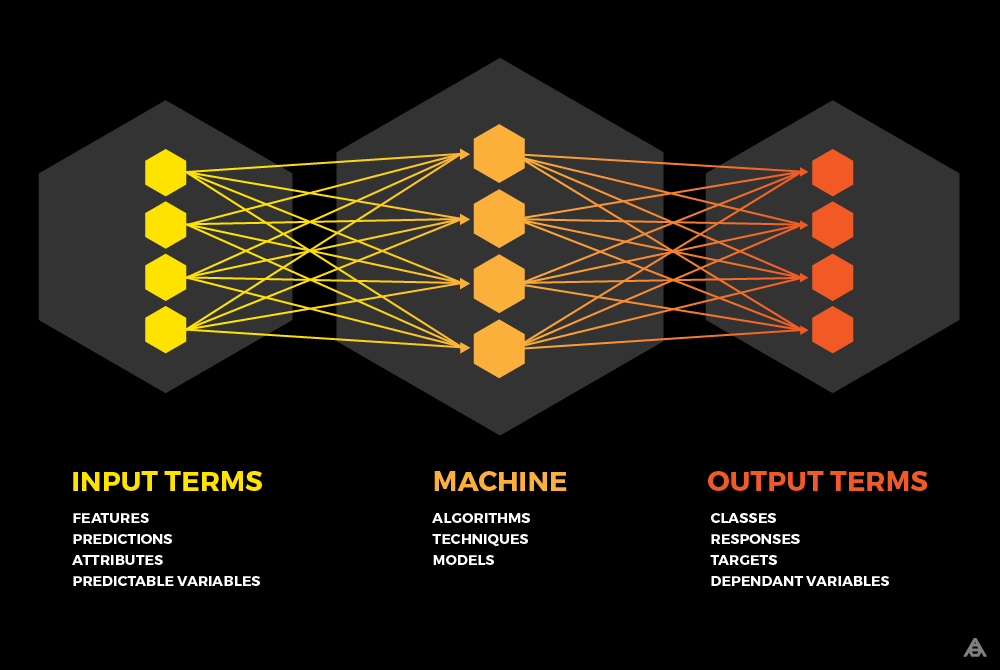
Stajımda bir AR-GE projesi olarak makine öğrenmesi ile kullanıcılara izledikleri filmlere göre yeni filmler öneren bir sistem geliştirdim. Proje üzerinde çalışan başka kimse olmadığı için makine öğrenmesi ve film öneri sistemleri üzerinde detaylı araştırmalar ve denemeler yapılması gerekiyordu. Böylelikle uzmanı olmadığım bir alanda nasıl proje geliştirebileceğim hakkında tecrübe kazanma fırsatı elde ettim.

**STAJ PROJESİNİN TANIMI VE ANALİZİ**

Stajda, kullacıların filmlere verdikleri puanlar, filmi ne kadar izledikleri gibi verileri kullanarak makine öğrenmesi ile yeni filmler öneren bir araştırma ve geliştirme (**AR-GE**) projesi yapmam bekleniyordur. Stajımı bu doğrultuda araştırma ve geliştirme süreçleri olmak üzere iki kısıma ayırmak mümkündür.

* 1. **Araştırma**

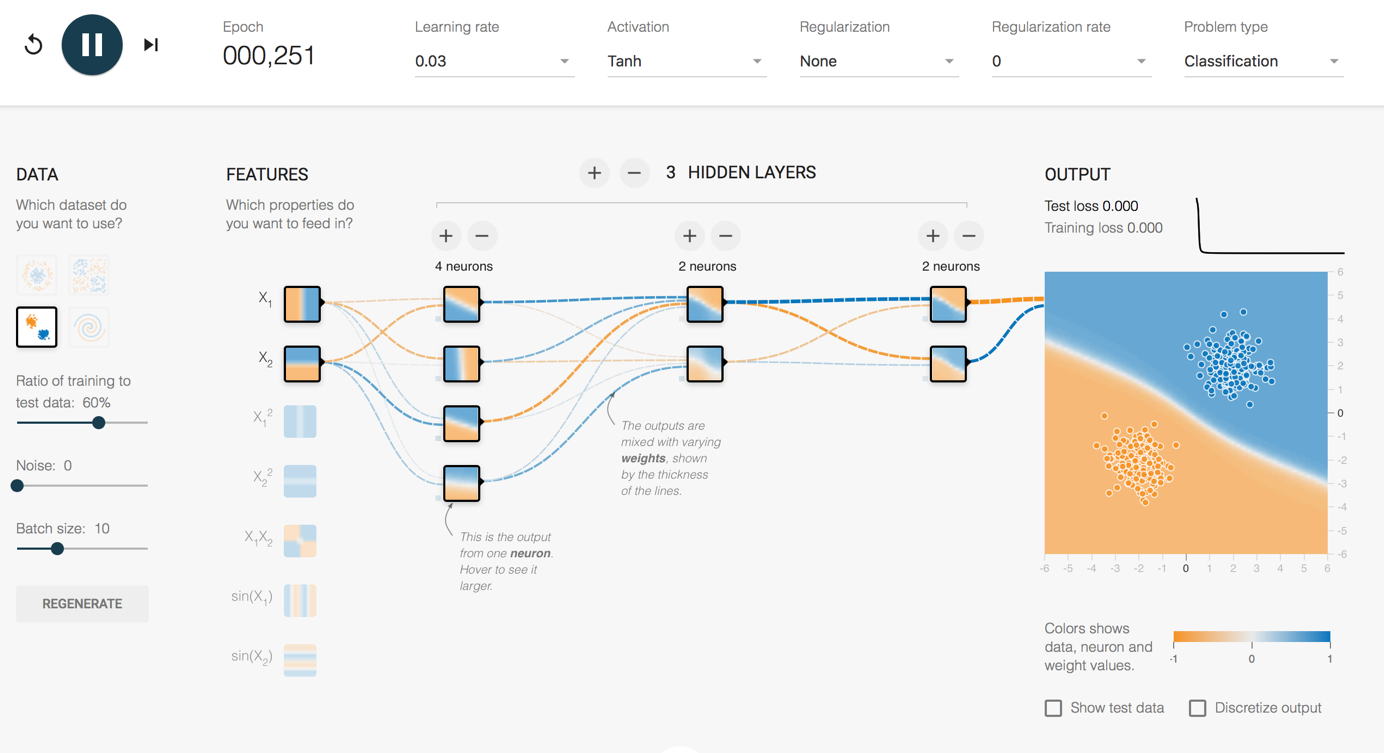
Projenin ilk aşaması makine öğrenmesinin araştırılması ve projeye uygun kütüphanenin belirlenmesinden oluşuyor. Araştırma aşamasında çeşitli internet

**Makine öğrenmesi**, bilgisayarların açık bir şekilde programlanmadan önceden elde edilmiş verileri kullanarak çıkarımlar yapmasını sağlayan bir bilim tekniğidir. Son 10 yılda popülerliği oldukça artmış ve bir çok alanda kullanılmaya başlanmıştır. Örneğin gereksiz veya saldırı amacı taşıyan e-postaların belirlenmesi ve kullanıcıların uyarılmasında makine öğrenmesi kullanılarak yazılmış filtreler kullanılır ve kullacıların güvenliklerinin arttırılması amaçlanır. Çeşitli vücut verilerinin incelenip insanların kansere yakalanma olasılığın belirlenmesi gibi sağlık alanlarında makine öğrenmesi kullanılabilir. Film sektöründe önceden izlenen filmleri kullarak izleyicilere beğenebilecekleri yeni filmlerin önerilmesinde makine öğrenmesi algoritmaları da kullanılmaya başlanmıştır.

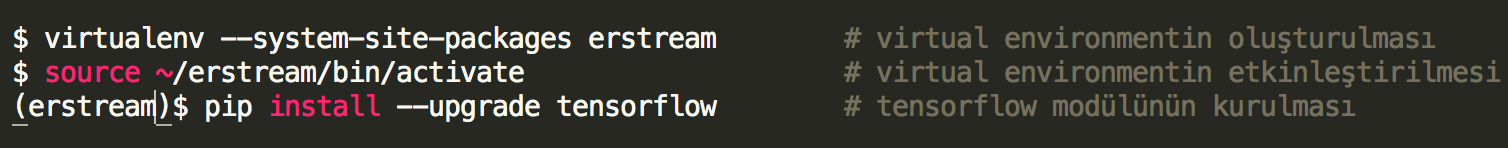
**Şekil 2** – Makine Öğrenmesi

Makine öğrenmesi ya da makine öğrenimi temel olarak 2’e ayrılır:

* **Gözetimli Öğrenme (*Supervised Learning*):** Makinenin sınıflandırılmış veya etiketlendirilmiş veriler kullanılarak eğitilmesidir. Veriler etiketlendirilmiş olduğu için sonuç değerlerinin doğruluğunun karşılaştırılmasında ve makinenin doğruluk yüzdesinin de belirlenmesinde kullanılabilmektedir.
* **Gözetimsiz Öğrenme (*Unsupervised Learning*):** Gözetimli öğrenmenin tam aksine makinenin sınıflandırılmamış veya etiketlendirilmemiş veriler kullanılarak eğitilmesidir.

Makine öğrenmesi ile ilgili bilgilerimi pekiştirdikten sonra bu alanda en çok kullanılan kütüphaneleri araştırdım ve Google’ın geliştirdiği açık kaynak kodlu (*open-source*) makine öğrenme kütüphanesi olan **TensorFlow**’u kullanmaya karar verdim. Python, C++, Java, Go gibi bir çok dilde kullanılabilmesi, aynı zamanda GPU desteğinin olması ve çok detaylı dökümantasyonlara sahip olması TensorFlow’u seçmemin temel sebepleri oldu. Python bilgimin diğer dillere kıyasla daha iyi olması nedeniyle denemelerimi **python** programlama dili ile yapmaya karar verdim.

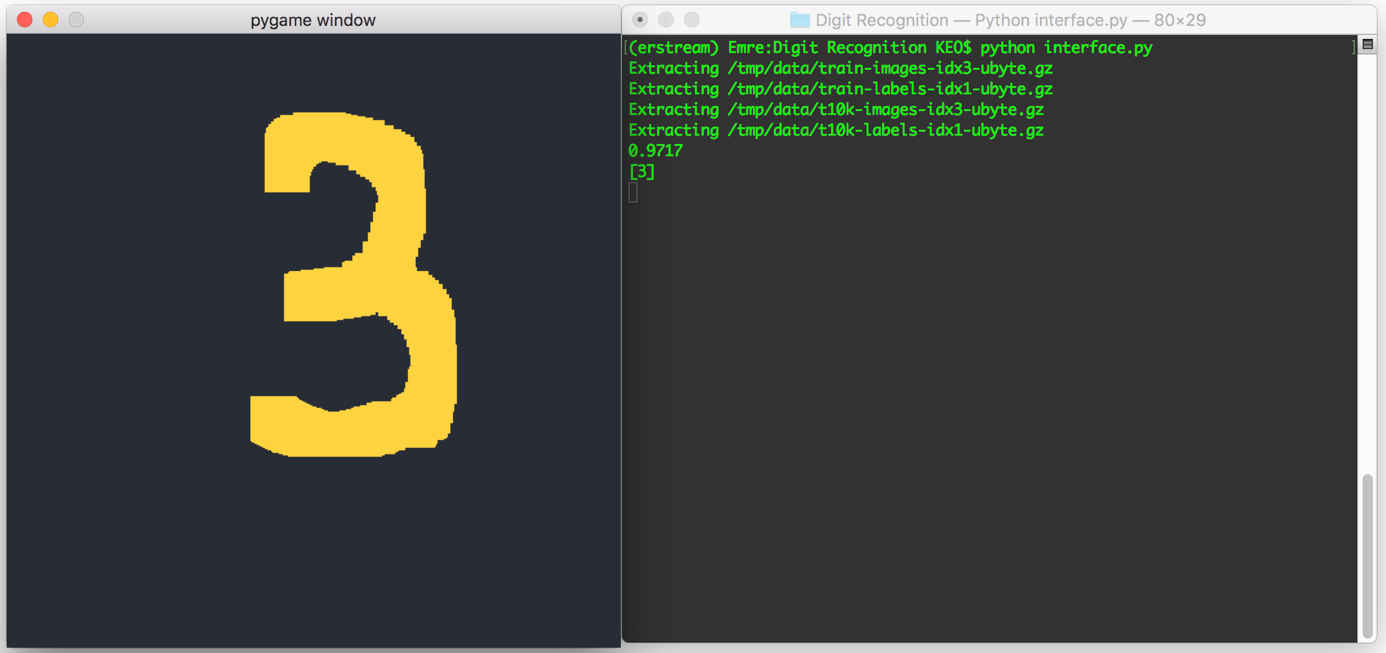
**Şekil 3** – TensorFlow Playground Classification Probleminin Görselleştirilmesi

TensorFlow’un Python modülünü kurmadan önce kendi sistemimin etkilenmemesi için stajdaki projelerimde kullanabileceğim bir *virtual environment* oluşturdum. Bilgisayarımda TensorFlow’un GPU desteği verdiği bir donanım olmadığı için sadece CPU’da çalışan versiyonunu kurdum.

**Şekil 4** – TensorFlow’un kurulumu

Yeni bir programlama dilini öğrenirken ilginç bir gelenek vardır; ilk olarak ekrana nasıl “*Hello world*” yazdırılacağı öğrenilir. Makine öğrenmesinde de **MNIST** (multinomial logistic) gelenekselleşmiştir. MNIST bilgisayarın anlayabileceği bir veri kümesidir ve 70000’e yakın el yazısı ile yazılmış rakam resmi içermektedir.

Tensorflow bilgimi ve makine öğrenmesinin nasıl çalıştığını pekiştirmek için TensorFlow dökümanlarınından faydalanarak **Digit Recognition** isimli uygulama geliştirdim. Python’un **PyGame** kütüphanesini kullarak (arayüz işlemleri için) çizilen rakamın hangi rakam olduğunu tahmin etmeye çalıştım. Uygulamayı Python için avantajlı özellikleri olan **PyCharm** editöründe geliştirdim. Programa ait örnek kodları **Ek-1** ve **Ek-2** de bulabilirsiniz.



**Şekil 5** – Digit Recognition Programının Örnek Kullanımı

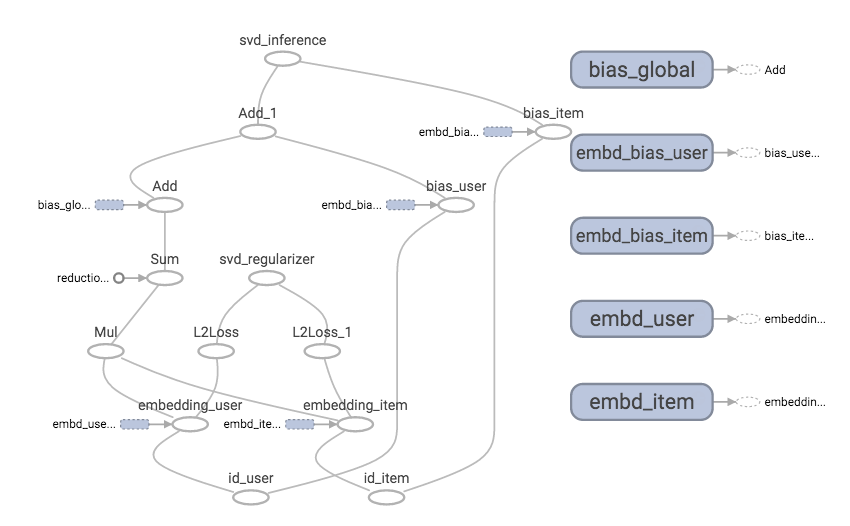
Digit Recognition programı ile birlikte makine öğrenmesi ve TensorFlow kütüphanesi ile ilgili bilgilerim oldukça gelişti. Bu aşamadan sonra asıl proje olan film öneri sistemleri hakkında araştırmaya başladım.

Öneri sistemlerinde kullanılan algoritmaları iki kısımda inceleme incelemek mümkün:

* **İçerik Bazlı Algoritmalar (*Content-base Algorithms)***: Bu algoritmalarda filmler arasındaki ilişkiler hesaplanır ve bu ilişkiler kullanılarak yeni filmler önerilir.
* **İşbirliğine Dayalı Filtreleme (*Colloborative Filtering*)**: Bu algoritmalarda kullanıcı ile benzer filmleri izleyen kişiler bulunur ve onların izledikleri filmler kullanıcıya önerilir.

**YouTube** internet sitesinde yayınlanan Google Cloud Platform User Group Singapur konferansının videosunda [7] TensorFlow ile öneri sistemlerinin nasıl yazılabileceği ile ilgili fikir sabibi oldum ve **Şekil-6**’daki modeli inceledim.

**GELİŞTİRME**

****Araştırma kısmı bittikten sonra projeyi geliştirme süreci başladı. Bu süreçte araştırma aşamasında edindiğim **Şekil-6**’daki modeli kendim inşa ettim.

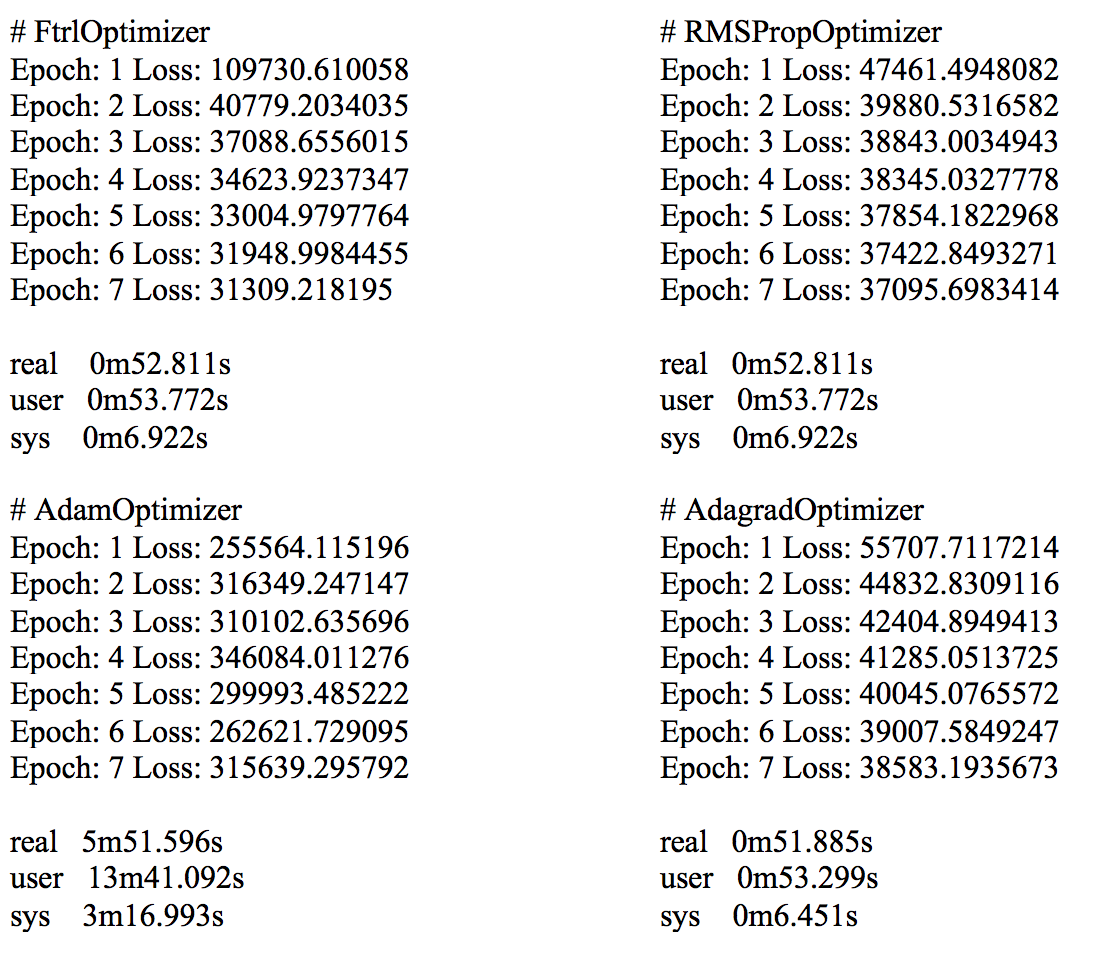
**Şekil 6** – Film Öneri Sistemi Modelinin Grafiği

**Şekil 7** – Film Öneri Sistemi Modeli

Kurduğum bu modeli eğitmek için **MovieLens** şirketinin sunduğu **kullanıcı, film, filme verilen puan** bilgilerini içeren verilerini kullandım. Veriler **csv** formatında sunulduğu için csv dosyaları için kullanışlı fonksiyonları olan pythondaki **pandas** modülü ile verileri okudum ve eğitme fonksiyonu olan **train** fonksiyonunda bu verileri kullandım. Bu aşamada 2 temel problem ile karşılaştım.İlki verilerin boyutunun çok yüksek olması nedeniyle verinin tamamını tek seferde train fonksiyonuna vermek mümkün değildi. Bu sorunu veriden rasgele daha küçük parçalar oluşturup onları train fonksiyonuna göndererek çözdüm. Karşılaştığım diğer sorun ise büyük csv dosyalarında pandas kütüphanesinin RAM’de çok fazla hafıza kullanmasından dolayı çok yavaş çalışması oldu. Bu sorunu da okuma işlemini yapan kodu kendim yazarak çözdüm. Train fonksiyonunu ve csv dosyasından veri okuma kodlarını **Ek-3** ve **Ek-4** de bulabilirsiniz.

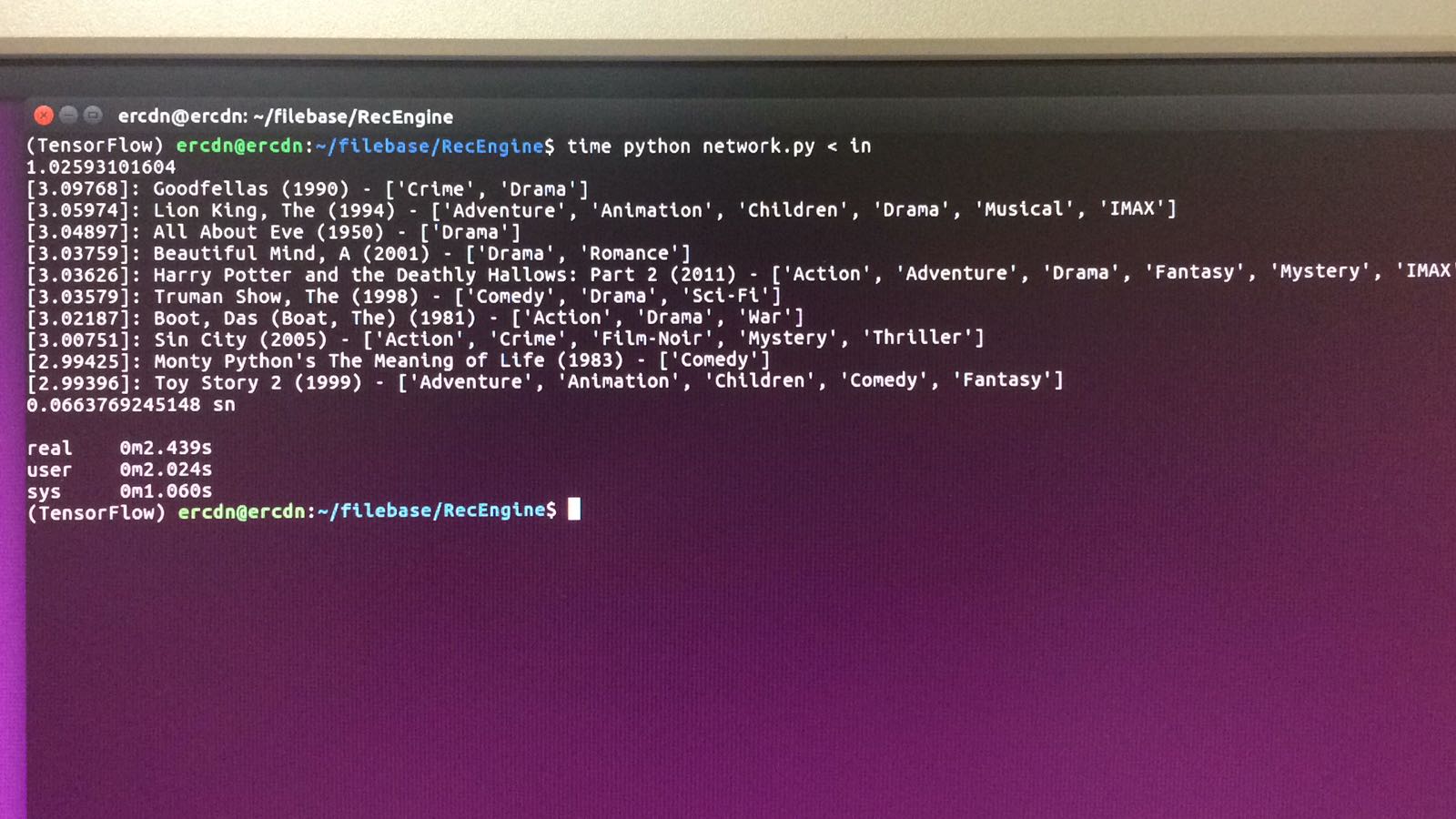
TensorFlow modellerin eğitilmesi ve asıl öğrenme işlemini yapan bir çok ***optimizerı*** hazır olarak sunmaktadır. Optimizerlar modellere ve eğitilirken kullanılan verilere göre birbirlerine üstünlük kurabilmekteler. Bu nedenle optimezerları tek tek test ettim ve hızlı ve daha iyi sonuç veren optimizerı bulmaya çalıştım. Yaptığım testlere göre FtrlOptimezer diğerlerinden daha hızlı çalışmakta va daha iyi sonuçlar üretmektedir.

Optimzerların kaybetme sayıları ve ne kadar süre çalıştığı aşağıda verilmektedir.

****

**Şekil 8** – Optimizerların Sonuçları

Eğitirken kullandığım veriler çok büyük olduğu için programın çalışma süresi oldukça uzun sürmekteydi. Bu sebeple şirketten TensorFlow’u GPU’da çalıştırabileceğim bir makine talep ettim ve programı bu ortamda test etmeye başladım. Ortalama bir GPU ile programda yaklaşık olarak 5 kat hızlanma gözlemlendi.

****

**Şekil 9** – Film Öneri Sistemi Örnek Çıktısı

**SONUÇ**

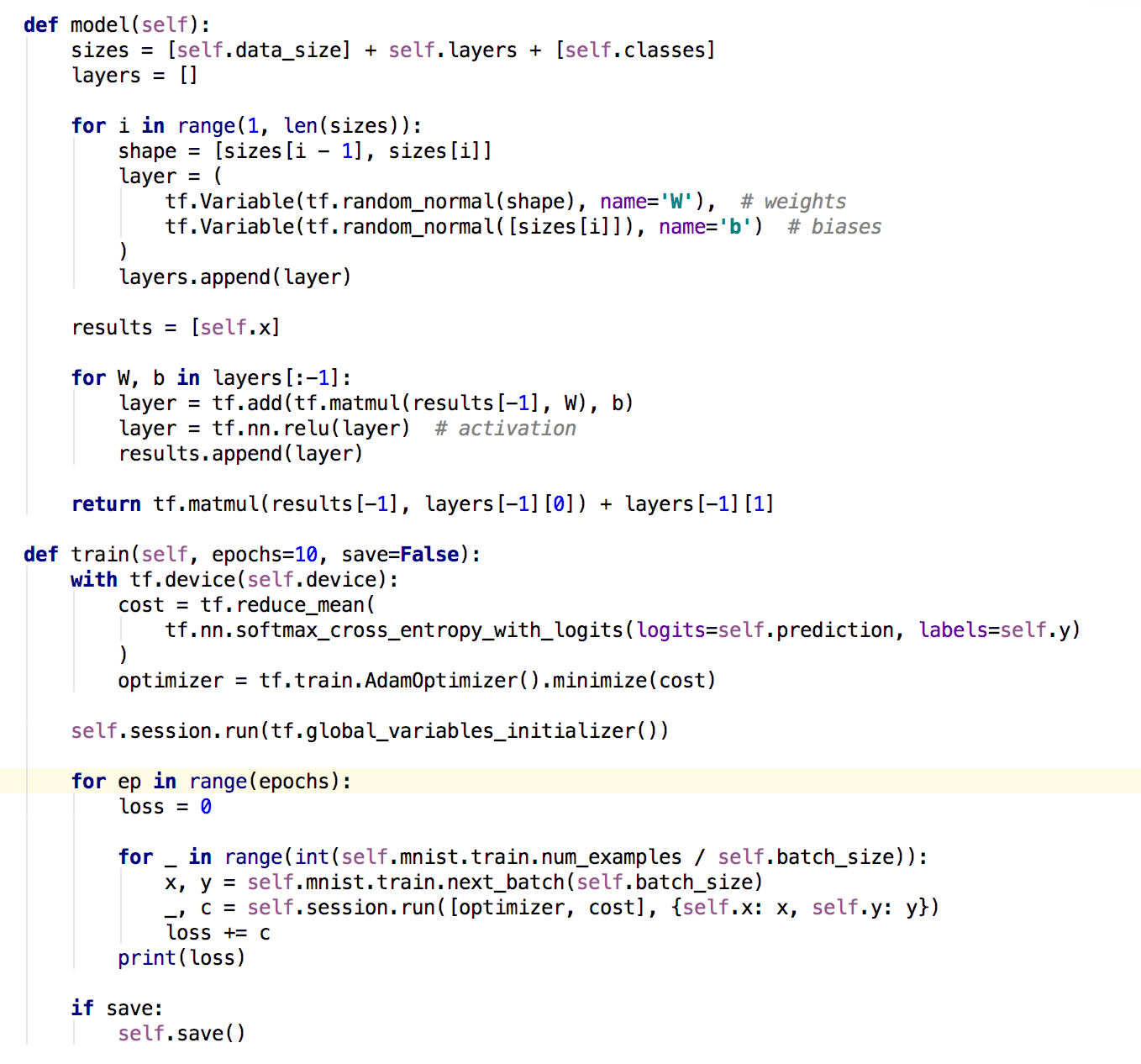
Yaptığım bu stajda daha önceden çok merak ettiğim makine öğrenmesi hakkında detaylı bilgi edinmiş, örnek projeler geliştirmiş oldum. Dünya çapında bir çok şirketin kullandığı Google’ın makine öğrenmesi alanında geliştirdiği TensorFlow kütüphanesini inceleme ve proje yapma fırsatı bularak teknik anlamda gelişme fırsatı buldum.

Şirket ortamının nasıl olduğunu, sektöre yönelik ürünlerin ve hizmetlerin nasıl üretildiğini gözlemleyebilmek sektöre yönelik bilgimi arttırdı. Alanında uzman kişilerin yardımı ile araştırma-geliştirme projelerinde nasıl bir yol izlenmesi gerektiği hakkında çok değerli fikirler edindim. Sağladığı bu staj imkanından dolayı Ersteam şirketine teşekkür ederim.

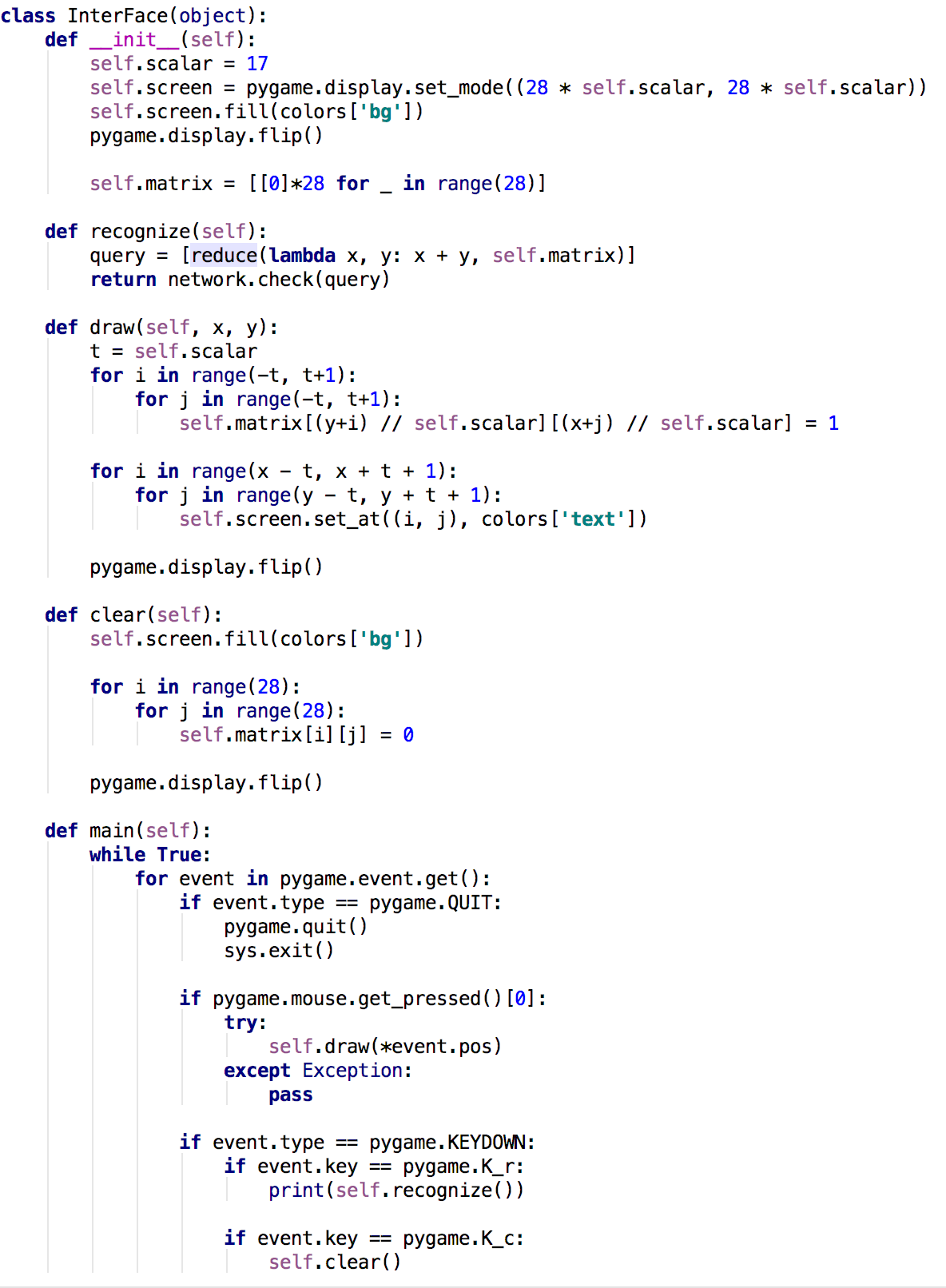
**REFERANSLAR**

1. *Experts System. Makine Öğrenmesi*. http://www.expertsystem.com/machine-learning-definition/
2. *Analytics VidhyaMakine Öğrenmesi*. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
3. *Tecleer. Makine Öğrenmesi*. https://www.techleer.com/articles/203-machine-learning-algorithm-backbone-of-emerging-technologies/
4. *Şekil 2*. https://dzone.com/articles/understanding-machine-learning
5. *GroupLens*. https://grouplens.org/datasets/movielens/
6. *Google*. http://tensorflow.com
7. *Youtube. Recommendation Engine.* https://www.youtube.com/watch?v=TNiWwaMGYzo

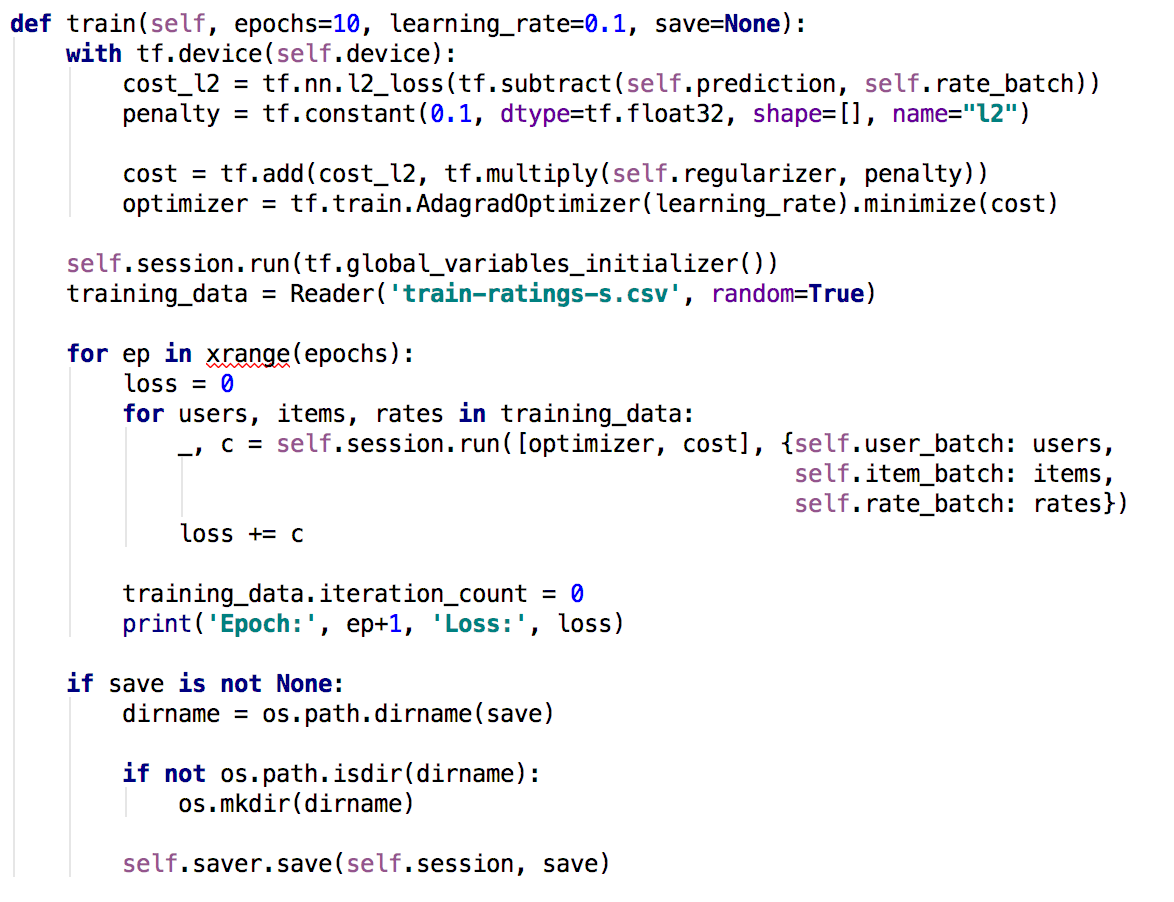
**EKLER**

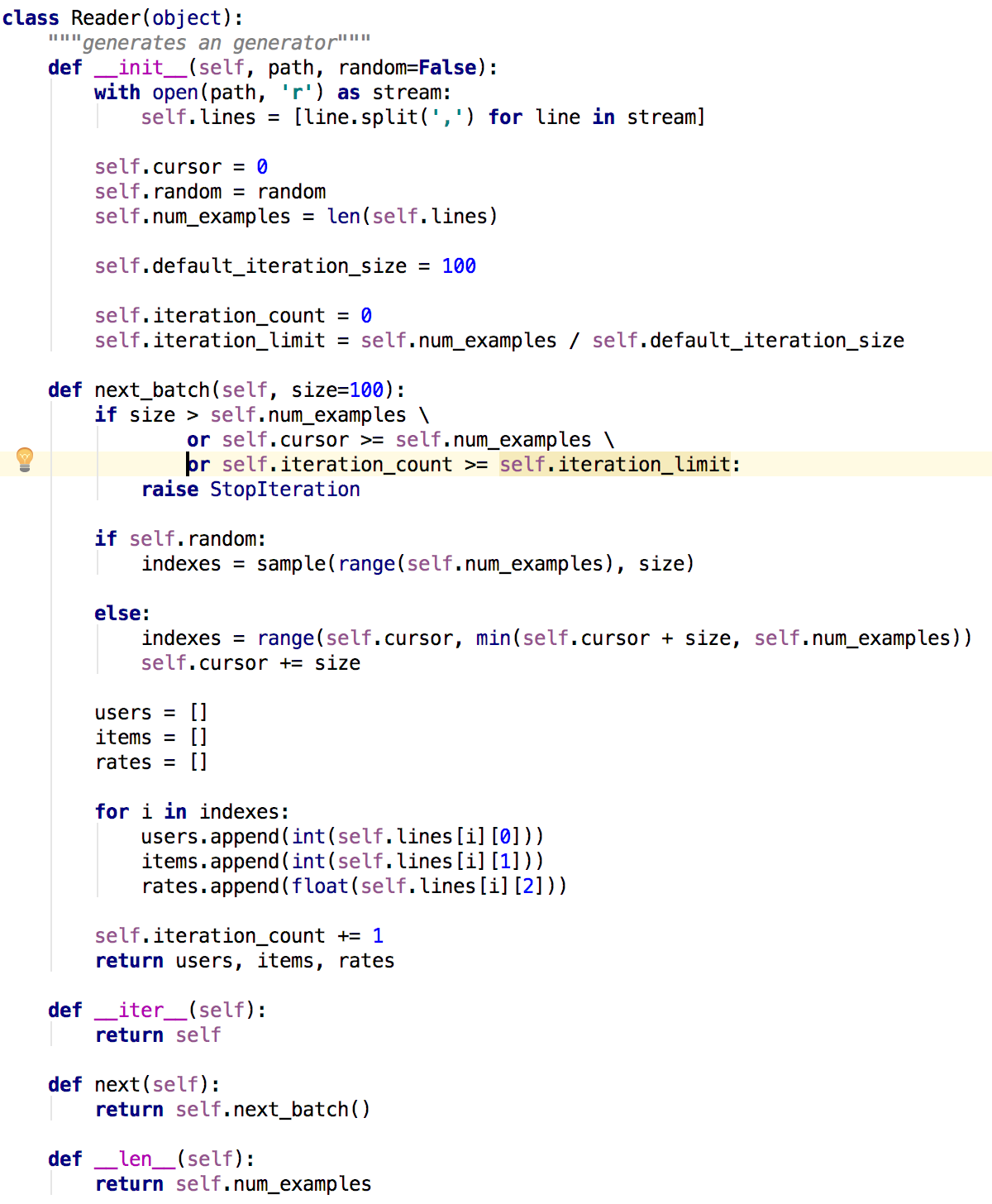
**Ek 1:** El yazısı ile yazılmış rakamların tanınması projesindaki model ve MNIST verisi ile eğitilmesi

**Ek 2:** El yazısı ile yazılmış rakamların tanınması projesindeki arayüzün PyGame ile oluşturulması

****

**Ek-3:** Film öneri sisteminde kullanılan train fonksiyonu

****

**Ek-4:** Film öneri sisteminde kullanılan csv dosyasını okuyan yapı