# **AR-GE Mühendisi Mülakat Projesi**

## **Giriş**

Bir sınıflandırma algoritmasında, algoritmanın genel başarımını bu algoritmanın iç yapısı kadar veri setine uygun hiperparametreler de belirlemektedir.

Bu projede, hiperparametre optimizasyonunun otomasyonu amaçlanmaktadır. Projenin sonraki aşamaları önceki aşamaların gerçekleştirilmesini gerektirmektedir.

Projede kullanılacak programlama dili ve kütüphaneler, adayın tercihine bırakılmıştır.

## **Veri Kümesi**

Projede kullanılacak olan veri kümesine bu adresten ulaşabilirsiniz:

<https://s3.amazonaws.com/benchm-ml--main/train-0.1m.csv>

## **Aşamalar**

### Sağlanan veri kümesi üzerinde, bir sınıflandırma algoritmasını eğitiniz (son kolon hedef değişkendir). Algoritmanın hiperparametreleri, kullandığınız kütüphanedeki algoritmanın varsayılan hiperparametreleri olmalıdır.

Sınıflandırma algoritması olarak;

* Logistic Regression
* *Random Forest*,
* *XGBoost*,
* Naif Bayes,

algoritmaları arasından bir seçim yapılmalıdır.

1. *Grid Search* ve *Randomized Search* yöntemlerini kullanarak, hiperparametre optimizasyonu gerçekleştiriniz. Bunu yaparken, uygun bir skor metriği seçerek neden bu metriği seçtiğinizi açıklayınız.
2. Üst-sezgisel (*meta-heuristic*) yöntemlerden birini kullanarak, hiperparametrelerden biri üzerinde optimizasyon yapınız. Bu optimizasyon yöntemlerine örnek olarak *Basin Hopping* gösterilebilir.
3. 3. maddedeki yöntemi, algoritmanın bütün hiperparametreleri için genelleştiriniz.
4. *Grid Search*, *Random Search* ve kullandığınız üst-sezgisel yöntemin sonuçlarını aşağıdaki açılardan karşılaştırınız, aldığınız sonuçları detaylıca yorumlayınız:

* İterasyon sayısı
* Bulunan hiperparametrelerin benzerliği
* Bulunan hiperparametrelerde algoritmanın skoru

Tüm geliştirme süreci aşağıdaki maddeleri içerecek basit bir rapor halinde teslim edilmelidir.

* Sınıflandırma için kullanılan yöntemin kısa ve anlaşılır özeti
* Geliştirme sırasında uyguladığınız optimizasyon yöntemleri ve amaçları
* Geliştirme sırasında eksik kaldığını düşündüğünüz yerler (var ise)

**BONUS:** Sınıflandırma aşamasında dağıtık makine öğrenmesi kütüphanelerinden olan Spark, H2O veya Mahout’u kullanarak; bütün bu adımları tekrarlayınız. Sonuçları başta elde ettiğiniz sonuçlarla karşılaştırınız ve yorumlayınız.