SEMA YALÇIN – 20120205034

BIL 366 VERİ MADENCİLİĞİ PROJE RAPORU

PROJE AMACI:

Bu projenin amacı, veri madenciliği tekniklerini kullanarak veri seti içerisindeki toksik yorumları doğru bir şekilde tespit etmektir.

VERİ SETİNİN OKUNMASI:

Projeye başlarken öncelikle veri seti okundu ve kopyası alındı. Sonra id, text ve toksiklik seviyesini gösteren sütun dışındaki diğer sütunlar çıkarıldı.

Toksiklik seviyesini gösteren değerler göz önüne alınarak class_id sütununa toksik değeri 0'a eşit olan yorumlara 0, toksik değeri 0'dan büyük olanlara 1 olacak şekilde değerler atandı.

Toksik değerler için bir threshold(0.05) belirleyerek, class_id_2 sütununa toksik değeri 0'a eşit olan yorumlara 0, toksik değeri 0.05'ten küçük olanlara 1, diğer yorumlara da 2 olacak şekilde değerler atandı. Bu sınıflandırmada derecelendirme şu şekildedir:

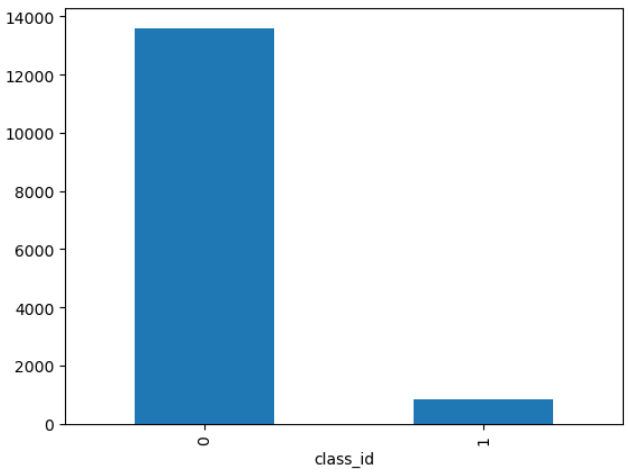
0 -> toksik değil

1 -> hafif toksik

2 -> aşırı toksik

SAMPLİNG:

Sampling işleminde class_id sütunundan 0.008 oranında her iki sınıf için de örnekler alındı.



0 13602
1 837
Name: class_id, dtyp:int64

PRE-PROCESSİNG İŞLEMİ:

Bu aşamada verideki text ifadeleri için bazı pre-processing işlemleri uygulandı.

Bu işlemler:

- Noktalama işaretlerinin kaldırılması
- Sayıların kaldırılması
- Stopwordslerin çıkarılması
- Stemming yapılması
- En fazla 2 harf uzunluğunda olan kelimeler çıkarılması.
- Nadir kelimelerin kaldırılması
- Harflerin hepsini küçük harfe çevirme
- Tokenization yapılması

TF-IDF MATRİX:

Bu adımda temizlenen ve tokenlara ayrılan yorumlardan her kelime için tf-idf değerleri hesaplandı ve matris haline getirildi.

samples: 14439, features: 24487

FEATURE SELECTİON:

Oluşturulan matriste feature sayısı çok fazla olduğu için feature sayısını azaltmak için Chi-Squared yöntemiyle en iyi 5000 feature seçildi.

MODEL OLUŞTURMA VE EĞİTME YÖNTEMLERİ:

- CLASSİFİCATİON MODELLLERİ:

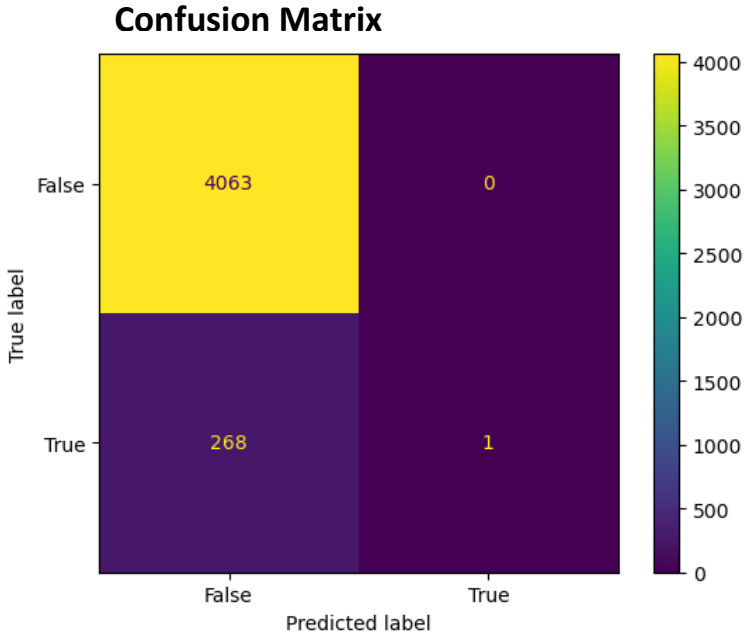
    - CLASS SINIFI: class_id

      Veri seti train ve test setine ayrıldı. Test size 0.3 olarak belirlendi. Verinin classı da class_id sütunu olarak belirlendi.

- ✓ X_train: (10107, 5000)
- ✓ Y_train: (10107, 1)
- ✓ X_test: (4332, 5000)
- ✓ Y_test: (4332, 1)

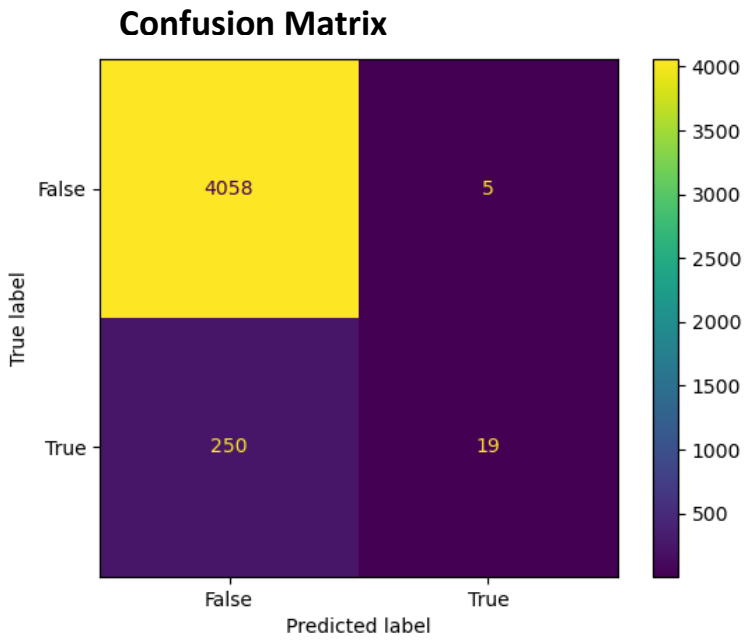
- KNN:

KNN modeli veriler arasındaki benzerlikleri kullanarak tahmin yapmayı amaçlar. KNN modelinde en önemli parametre “k” değeridir. “k” değeri, tahminlerde kullanılacak en yakın komşu sayısıdır. Bu projede “k” değeri 5 olarak seçildi. Ayrıca uzaklık ölçme metriği olarak minkowski kullanıldı.



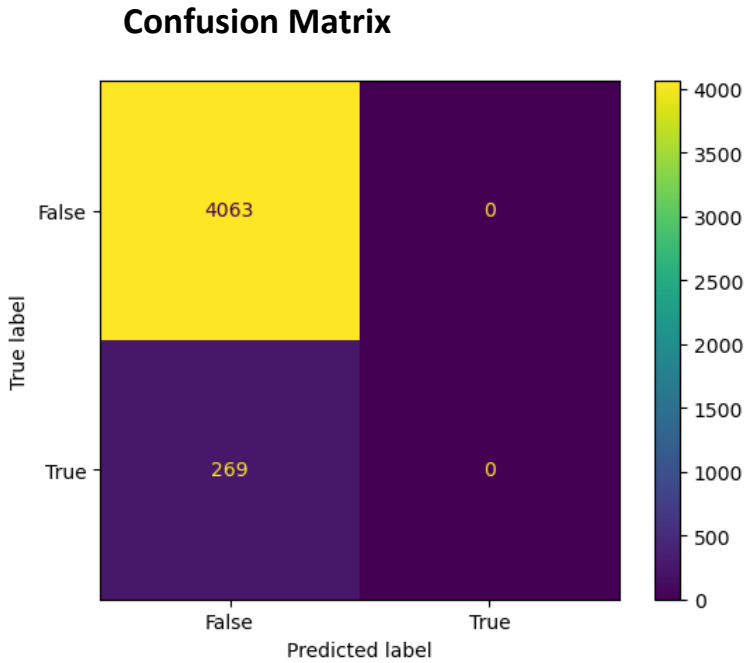
- SVM:

SVM modeli veriler arasındaki benzerlikleri kullanarak tahmin yapmayı amaçlar. SVM modelinde, en önemli parametreler “c” değeri ve Kernel fonksiyonudur. “c” değeri, hatayı azaltmaya çalışırken modelin ne kadar sıkı olması gerektiğini belirler. Kernel fonksiyonu ise, veriler arasındaki benzerlikleri nasıl ölçeceğini belirler. C değeri 1.0, Kernel değeri ise linear olarak seçilmiştir.



- NAİVE BAYES:

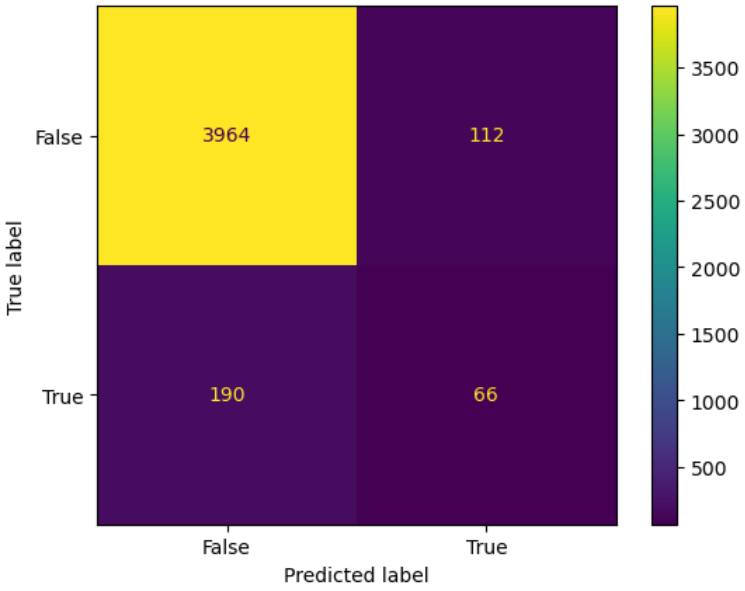
  Naive Bayes modeli veriler arasındaki benzerlikleri kullanarak tahmin yapmayı amaçlar. Naive Bayes modelinde, parametre değeri yoktur. Ancak, modelin hangi tahmin yöntemi kullanılacağı belirlenebilir. Bu projede Multinomial Naive Bayes tahmin yöntemi kullanılmıştır.



- DECİSİON TREE:

  Decision Tree modeli veriler arasındaki benzerlikleri kullanarak tahmin yapmayı amaçlar. Decision Tree modelinde, diving criteria olarak Entropy kullanılmıştır.

Confusion Matrix



Class_İd için Classification Yöntemlerinin Performans Değerlendirmeleri Tablosu:

	Accuracy Score	F-measure	Precision	Re-call	Cross Validation
KNN	0.938	0.908	0.942	0.938	0.942
SVM	0.941	0.130	0.792	0.071	0.947
Naive Bayes	0.938	0.968	0.938	1.000	0.942
Decision Tree	0.928	0.265	0.306	0.234	0.927

- CLASS SINIFI: class_id_2

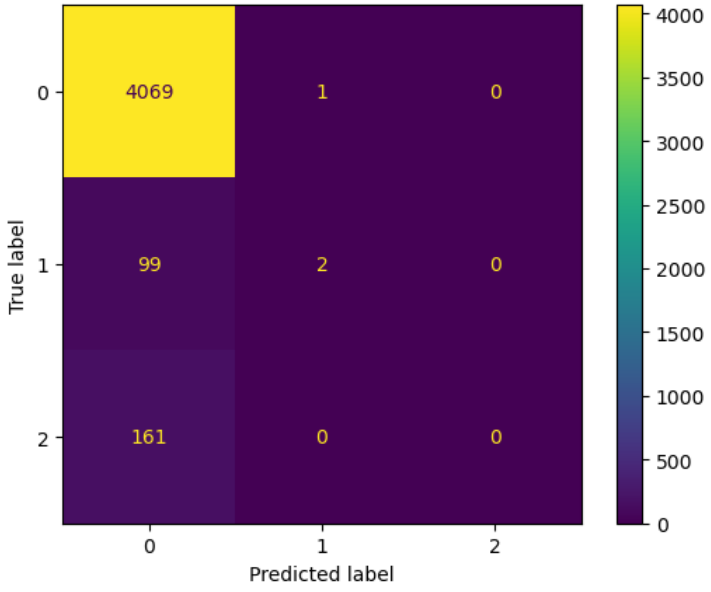
Veri seti train ve test setine ayrıldı. Test size 0.3 olarak belirlendi. Verinin classı da class_id_2 sütunu olarak belirlendi.

- ✓ X_train: (10107, 5000)
- ✓ Y_train: (10107, 1)
- ✓ X_test: (4332, 5000)
- ✓ Y_test: (4332, 1)

- KNN:

KNN modeli veriler arasındaki benzerlikleri kullanarak tahmin yapmayı amaçlar. KNN modelinde en önemli parametre “k” değeridir. “k” değeri, tahminlerde kullanılacak en yakın komşu sayısıdır. Bu projede “k” değeri 5 olarak seçildi. Ayrıca uzaklık ölçme metriği olarak minkowski kullanıldı.

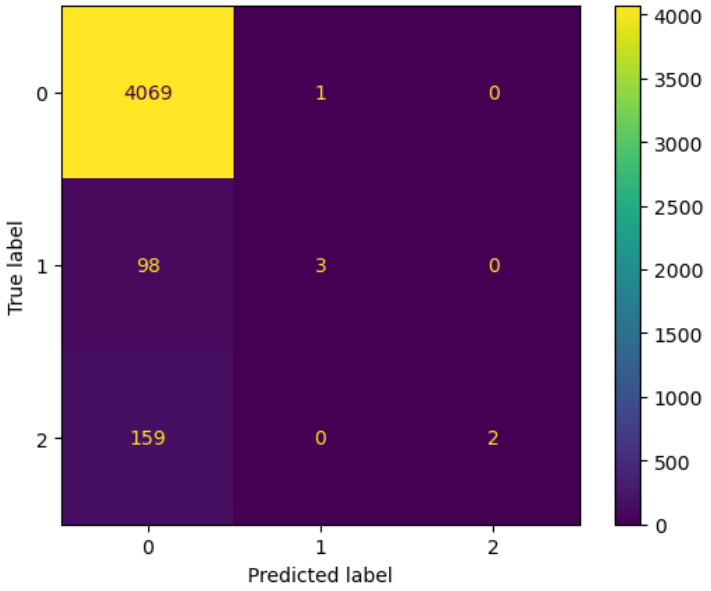
Confusion Matrix



- SVM:

  SVM modeli veriler arasındaki benzerlikleri kullanarak tahmin yapmayı amaçlar. SVM modelinde, en önemli parametreler “c” değeri ve Kernel fonksiyonudur. “c” değeri, hatayı azaltmaya çalışırken modelin ne kadar sıkı olması gerektiğini belirler. Kernel fonksiyonu ise, veriler arasındaki benzerlikleri nasıl ölçeceğini belirler. C değeri 1.0, Kernel değeri ise linear olarak seçilmiştir.

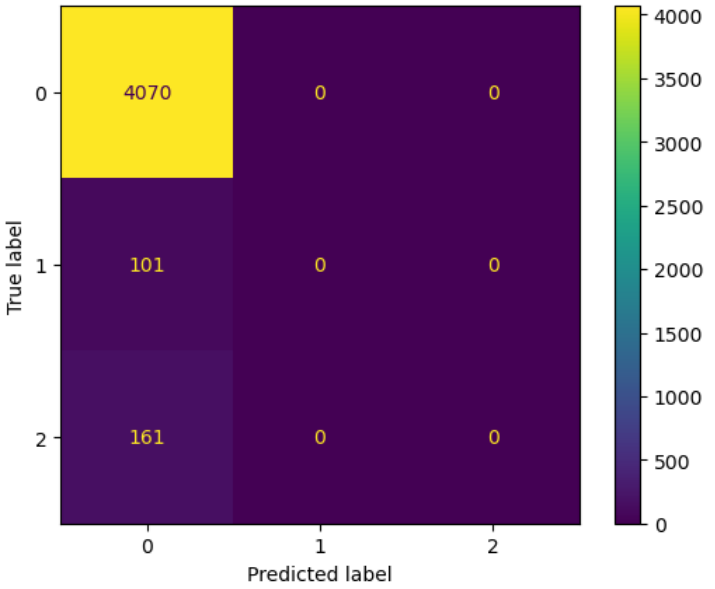
Confusion Matrix



- NAİVE BAYES:

  Naive Bayes modeli veriler arasındaki benzerlikleri kullanarak tahmin yapmayı amaçlar. Naive Bayes modelinde, parametre değeri yoktur. Ancak, modelin hangi tahmin yöntemi kullanılacağı belirlenebilir. Bu projede Multinomial Naive Bayes tahmin yöntemi kullanılmıştır.

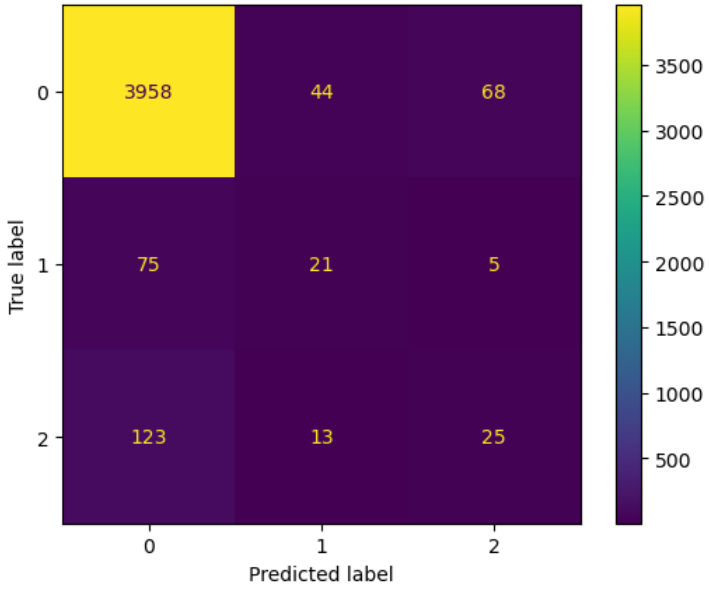
Confusion Matrix



- DECİSİON TREE:

  Decision Tree modeli veriler arasındaki benzerlikleri kullanarak tahmin yapmayı amaçlar. Decision Tree modelinde, diving criteria olarak Entropy kullanılmıştır.

Confusion Matrix



Class_İd_2 için Classification Yöntemlerinin Performans Değerlendirmeleri Tablosu:

	Accuracy Score	F-measure	Precision	Re-call	Cross Validation
KNN	0.940	0.946	0.933	0.976	0.942
SVM	0.940	0.913	0.938	0.940	0.944
Naive Bayes	0.940	0.969	0.940	1.000	0.942
Decision Tree	0.924	0.917	0.911	0.924	0.923

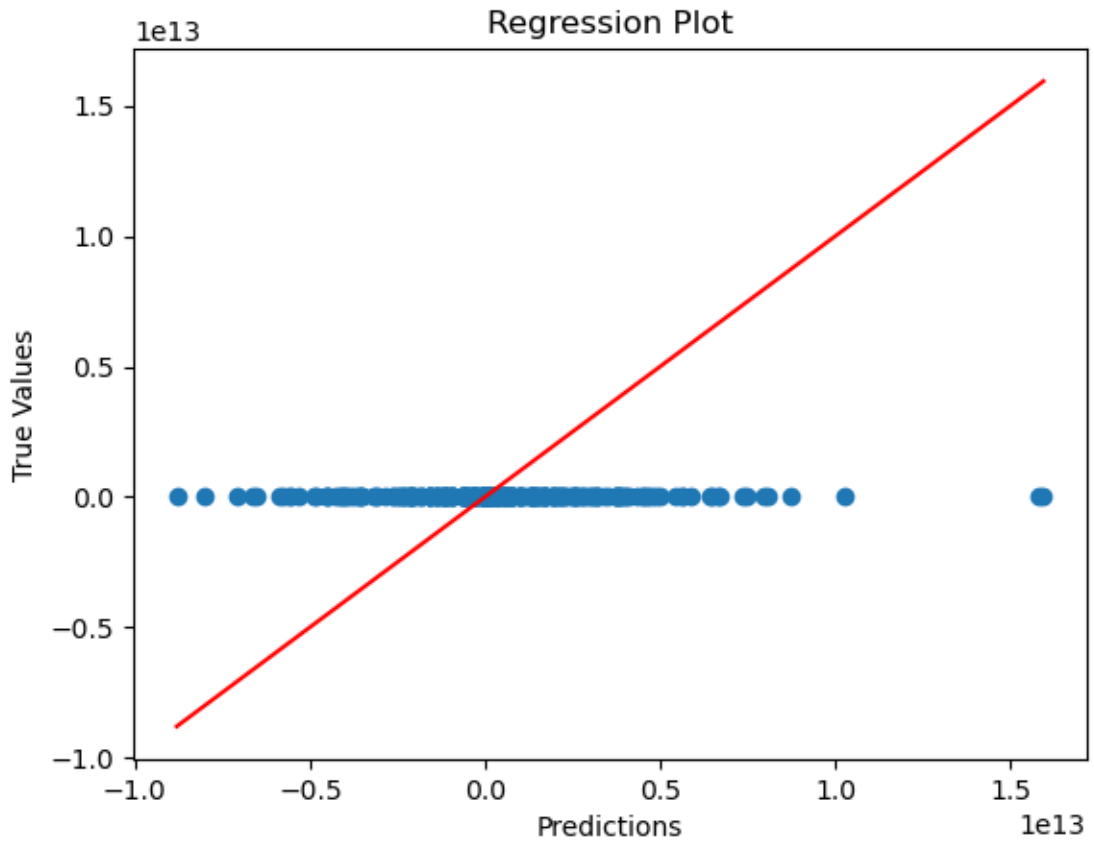
- REGRESSİON MODELİ:
    - CLASS SINIFI: class_id

Regresyon öğrenme yöntemlerinin birden fazla modeli vardır. Bu projede Linear Regresyon modeli kullanıldı. Linear Regresyon modeli veriler arasındaki ilişkiyi kullanarak değişkeni diğer değişkenlerden tahmin etmeye çalışır.

Class_İd için Linear Regresyon Yönteminin Performans Değerlendirmeleri Tablosu:

	R2 Score	Mean Squared Error	Mean Absolute Error	Cross Validation
Linear Regresyon	-120956239613145 75972827136.000	63050861701346387 3486848.000	132233483034.666	-0.082

Linear Regresyon Grafiği:



- CLASS SINIFI: class_id_2

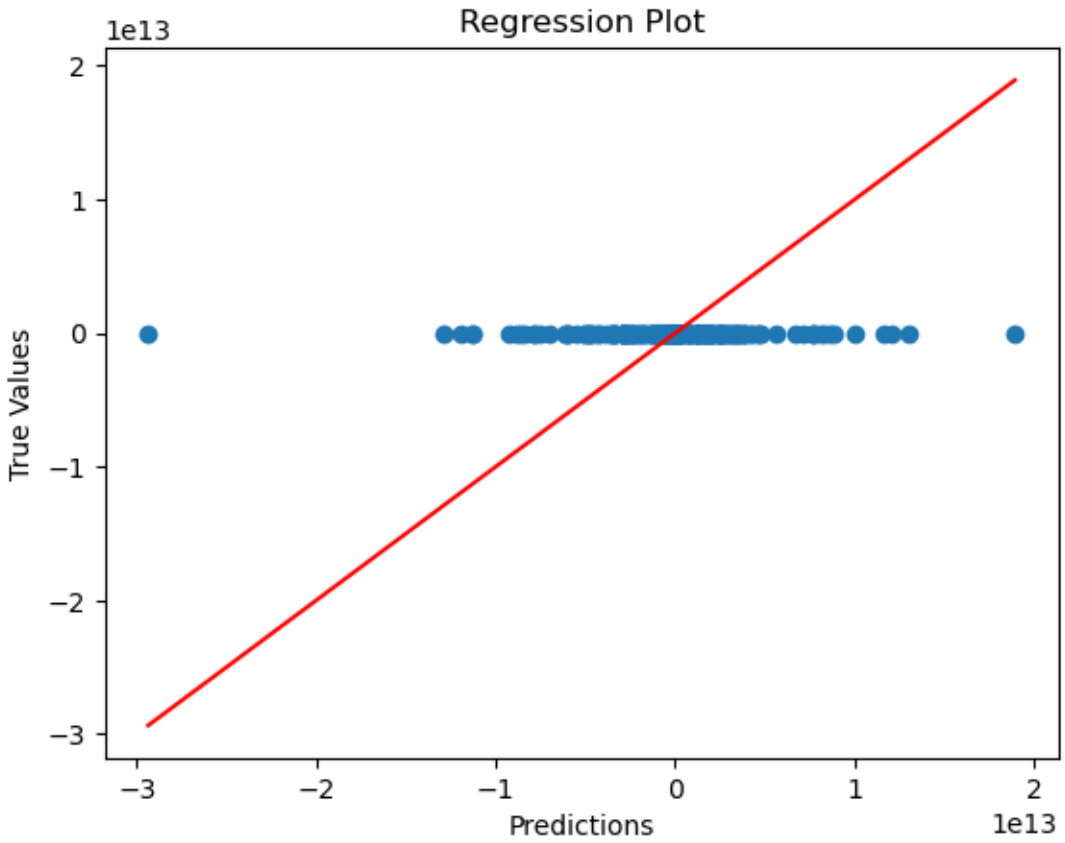
Regresyon öğrenme yöntemlerinin birden fazla modeli vardır. Bu projede Linear Regresyon modeli kullanıldı.

Linear Regresyon modeli veriler arasındaki ilişkiyi kullanarak değişkeni diğer değişkenlerden tahmin etmeye çalışır.

Class_İd_2 için Linear Regresyon Yönteminin Performans Değerlendirmeleri Tablosu:

	R2 Score	Mean Squared Error	Mean Absolute Error	Cross Validation
Linear Regresyon	-644226451690382 4772825088.000	10464902272422681 80348928.000	144574473757.783	-0.073

Linear Regresyon Grafiği:



- CLUSTERİNG MODELLERİ:

  Clustering modeli bir kümeleme modelidir. Class bilgisine ihtiyaç duyulmadan verileri benzerliklerine göre kendi aralarında kümler. Birçok clustering modeli vardır. Bu projede K-means ve Hierarchical Clustering modelleri kullanılmıştır.

  - K-MEANS CLUSTERİNG:

    K-Means modeli, beriler arasındaki benzerlikleri kullanarak verileri belirli sayıda gruplara ayırmayı amaçlar. K-Means modelinde en önemli parametre “k” değeridir. “k” değeri, oluşturulacak kümelerin sayısını belirler. Bu projede küme sayısı 2 olarak alınmıştır.

    Silhouette Score, bir clustering modelinin kümeler arasındaki benzerlikleri ve farklılıkları ölçen bir metriktir. Bu metrik kümeler arasındaki benzerliğin düşük ve kümeler arasındaki farklılığın yüksek olduğu durumlarda yüksek değerler, kümeler arasındaki

benzerliğin yüksek ve kümeler arasındaki farklılığın düşük olduğu durumlarda düşük değerler verir. Bu projede K-Means modeli için Silhoutte Score,

Silhouette Score: 0.018

olarak bulunmuştur. Silhouette Score düşük bir değer gelmesi kümeler arasındaki benzerliğin yüksek olduğunu gösterir.

- HİERARCHİCAL CLUSTERİNG:

  Hierarchical Clustering modeli, veriler arasındaki benzerlikleri kullanarak belirli sayıda gruplara ayırmayı amaçlar. Ancak, K-Means Clustering modeline göre daha farklı bir yöntem kullanır.

  Hierarchical Clustering modelinde en önemli parametreler "k" değeri, "linkage" ve "distance metric" tir. "k" değeri K-Means modelindeki gibi küme sayısını belirlemek için kullanılır. "Linkage" parametresi, verilerin nasıl birleştirileceğini belirler. "Distance metric" parametresi ise, veriler arasındaki benzerlikleri nasıl ölçeceğini belirler. Bu projede "k" değeri 2, "linkage" parametresi "average linkage", "distance metric" ise "euclidean similarity" olarak alınmıştır.

  Silhoutte Score, bir clustering modelinin kümeler arasındaki benzerlikleri ve farklılıkları ölçen bir metriktir. Bu metrik kümeler arasındaki benzerliğin düşük ve kümeler arasındaki farklılığın yüksek olduğu durumlarda yüksek değerler, kümeler arasındaki benzerliğin yüksek ve kümeler arasındaki farklılığın düşük olduğu durumlarda düşük değerler verir. Bu projede Hierarchical Clustering modeli için Silhoutte Score,

  Silhouette Score: 0.164

  olarak bulunmuştur. Silhouette Score düşük bir değer gelmesi kümeler arasındaki benzerliğin yüksek olduğunu gösterir.