Credit Card Fraud Detection Using State-of-the-Art Machine Learning Algorithms

Project - BLM5116 Data Mining and Knowledge Discovery

1st PhD Student Önder GÖRMEZ
Computer Engineering
Faculty of Electrical and Electronics
Yıldız Technical University
İstanbul, Türkiye
ondergormez@gmail.com

2th Prof. Dr. Songül VARLI
Computer Engineering
Faculty of Electrical and Electronics
Yıldız Technical University
İstanbul, Türkiye
svarli@yildiz.edu.tr

***Abstract*—Bu makalede dengesiz bir veri seti olan kredi kartı dolandırıcılığı yönteminin klasik makine öğrenmesi yöntemleri ile tespiti ve başarının ölçümlenmesi üzerine çalışmalar yapılacaktır.**

***Index Terms*—Decision Tree, Logistic Regression, KNN, SVM, Random Forest, Naive Bayes, Classification, Binary Classification, Imbalanced Data, Credit Card Fraud Detection**

I. Introduction

Makine öğrenmesi ve derin öğrenme yöntemleri ile alakalı makaleler [1], [3] incelenerek buralarda yapılan çalışmalarla kendi yaptığımız çalışmalar karşılaştırılmıştır. 2. bölümde kullanılacak olan temel konular hakkında kısa bilgiler verilecek, 3. bölümde performans metrikleri tanıtılacak, 4. bölümde veri seti tanıtılacak, 5. kullanılan klasik makine yöntemleri özet olarak verilecek, 6. bölümde Decision Tree, 7. bölümde Logistic Regression, 8. bölümde KNN, 9. bölümde SVM, 10. Bölümde Random Forest, 11. bölümde Naive Bayes algoritmaları ile yapılan çalışmalar verilecektir. 12. Bölümde sonuçlar ve gelecek çalışmalardan bahesidilecektir.

II. Temel Konular Hakkında Kisa Bilgi

A. Modelin Öğrenememesi, Öğrenmesi ve Aşırı Öğrenmesi Ne Demektir.?

Modelin öğrenememesi demek train veri seti üzerinde model çalıştırıldığında elde edilen sonuçların rastgelelikle yakın bir tahminde bulunabilmesi (%50) veya sonuçların bundan daha kötü kalması demektir.

Modelin öğrenmesi demek ise train veri seti üzerinde çalıştırıldığında %50'nin üzerinde bir accuracy elde etmesi anlamana gelmektedir. Bu öğrenme %100 e yaklaştıkça elde edilen modelin train setine çok daha uygun hale gelmesi mümkündür. Bu durumda validasyon seti üzerinde model iyi performans sergileyemeyebilir. Bu nedenle modelin train set üzerinde %100 başarıya ulaşması demek iyi çalıştığı anlamına gelmez.

Modelin aşırı öğrenmesi demek train set üzerinden %100 doğruluk sağlaması fakat validasyon ve test data setleri üzerinde çok kötü sonuçlar (%60 gibi) elde etmesi anlamına gelmektedir. Bu durum aşırı öğrenme (over-fitting) oluşmuş denir.

Yukarıda açıklanan kavramların görselleştirilmesi Fig. 1'dedir.

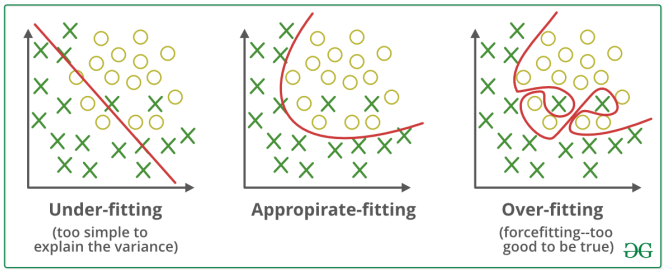

Fig. 1. Under-fitting vs. Appropriate-fitting vs. Over-fitting [4]

III. Performans Ölçüm Metrikleri

Çalışma yapılırken kullanılacak olan performans ölçüm metrikleri ve neleri ifade ettiklerine kısaca bir bakalım. Öncesinde kullanılacak olan tanımlardan bahsetmek gerekirse;

- True Positive (TP): Modelin pozitif olarak doğru tahmin ettiği örneklerdir.
- True Negative (TN): Modelin negatif olarak doğru tahmin ettiği örneklerdir.
- False Positive (FP): Modelin pozitif olarak yanlış tahmin ettiği örneklerdir.
- False Negative (FP): Modelin negatif olarak yanlış tahmin ettiği örneklerdir.

A. Accuracy

Accuracy, doğru sınıflandırılan örneklerin toplam örnek sayısına oranı ile hesaplanır. Modelin genel performansını ölçümlemek için kullanılan temel bir metriktir. Sınıflar arasında dengeli bir dağılım varsa accuracy iyi bir performans ölçütü olabilir. Fakat dengesiz veri setlerinde kullanılmak için uygun değildir. Örneğin fraud detection gibi unbalanced data

set ler üzerinde accuracy yüksek çıksa da modelin genel performansı hakkında doğru bilgi vermez.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

B. Precision (Kesinlik)

Precision, modelin pozitif tahminlerinin ne kadarının doğru olduğunu gösterir. Precision, yanlış pozitiflerin (False Positive) maliyetinin yüksek olduğu durumlarda özellikle önemlidir. Yanlış pozitifler modelin pozitif olarak tahmin ettiği ancak gerçekte negatif olan örneklerdir. Örneğin tıbbi sonuçları olan testlerde, precision yüksek olmalıdır çünkü yanlış pozitifler (sağlıklı kişilerin hasta olarak değerlendirilmesi) gereksiz yere endişe, korkuya neden olduğu gibi sonrasında yanlış tedavi ile sağlığın bozulmasına kadar giden kötü sonuçlar doğurabilir.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

C. Recall (Duyarlılık)

Recall, modelin pozitif sınıfı ne kadar iyi tespit edebildiğini gösteren bir metriktir. Yanlış negatiflerin (False Negative) maliyetinin yüksek olduğu durumlarda özellikle önemlidir. Örneğin tıbbi sonuçları olan testlerde, recall yüksek olmalıdır çünkü yanlış negatifler (hasta kişilerin sağlıklı olarak değerlendirilmesi) tedavinin gecikmesine ve hastanın sağlığının geri dönülmez bir şekilde kaybedilmesine ve hatta ölümlere bile yol açabilir. Geriye dönülmez sonuçlar doğurabilir.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

D. F1 Score

F1 Score, precision (kesinlik) ve recall (duyarlılık) metriklerinin harmonik ortalamasıdır. Özellikle dengesiz veri setlerinde (pozitif ve negatif sınıflar arasında büyük bir dengesizlik olduğunda) kullanışlıdır. Bu tarz durumlarda accuracy metriğine değil F1 Score metriğine bakılmalıdır.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

E. Support

Support, her bir sınıf için gerçek örnek sayısını ifade eder. Bu metrik, modelin performansını değerlendirirken sınıfların dağılımını anlamak için önemlidir. Özellikle dengesiz veri setlerinde, support metriği her bir sınıfın kaç örnekle temsil edildiğini gösterir ve bu da modelin hangi sınıflarda daha fazla veya daha az veriyle eğitildiğini anlamamıza yardımcı olur.

$$Support = \text{Her bir sınıf için gerçek örnek sayısı} \quad (5)$$

IV. Veri Seti

Kredi kartı dolandırıcılığı için hazırlanmış Kaggle [2] üzerinde yayınlanmış veri seti kullanılarak çalışmalar yapılmıştır.

A. Veri Setinin Özellikleri

Dataset Özellikleri:

- Worldline ve Université Libre de Bruxelles'deki Machine Learning Group işbirliğinde oluşturulmuş bir veri setidir.
- Eylül 2013'te Avrupa'da gerçekleşen kredi kartı işlemleri
- 2 günde gerçekleşen 284,807 işlemden 492 si fraud olarak kaydedilmiştir.

Yine Fig. 2 ve Fig. 2 te veri setini temsil eden kolonların değerleri görülmektedir. V1 ... V28 arası olan kolonlar kullanıcılara ait olan kişisel veri niteliğindeki bilgiler olduğu için data seti paylaşılmadan önce Principal Component Analysis (PCA) yöntemi ile ön işlemeden geçirilerek anonimleştirilmiştir. Time kolonu ilk transferin yapıldığı an ile ilgili sample arasında geçen süreyi göstermektedir. Amount kolonu yapılan harcamanın tutarını göstermektedir. Class kolonu ise işlemin fraud mu yoksa non-fraud mu olduğunu göstermektedir.

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...

5 rows × 31 columns

Fig. 2. Veri Setinin Kolonları

...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

Fig. 3. Veri Setinin Kolonları

Fig. 4 veri setinin dağılımını göstermektedir.

B. Verinin Training, Validation ve Test Setlerine Ayrılması

Veriseti %60 training, %20 validation ve %20 test seti olarak ayrılmıştır.

V. Klasik Makine Öğrenmesi Yöntemleri

Kredi kartı dolandırıcılık tespitinde kullanılacak klasik makine öğrenmesi teknikleri aşağıdaki gibidir.

- Decision Tree
- Logistic Regression
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Random Forest
- Naive Bayes

Bu klasik makine öğrenmesi yöntemleri aşağıda sırasıyla ele alınarak veri seti üzerinde çalıştırılmıştır.

VI. Descision Tree

Karar ağacı (Decision Tree), veriyi dallara ayırarak sınıflandırma veya regresyon yapabilen bir makine öğrenimi algoritmasıdır. Her bir düğümde, veri bir özelliğe göre

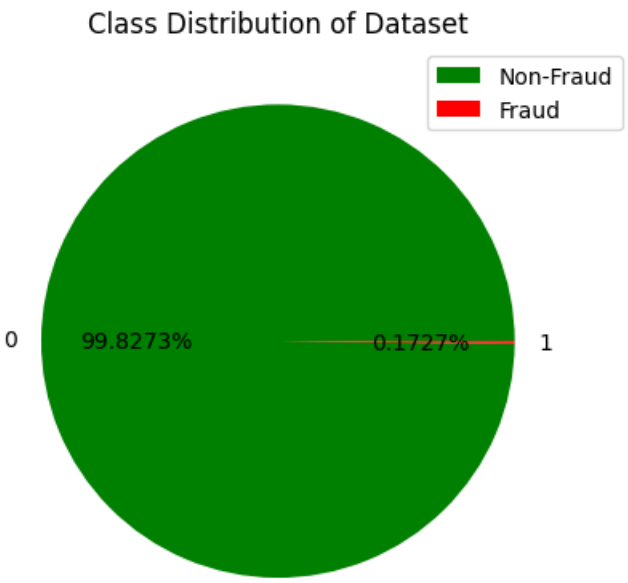

Fig. 4. Veri Setinin Dağılımmı

bölünür ve bu süreç yaprak düğümlere ulaşana kadar devam eder. Yaprak düğümler, sınıf etiketlerini veya tahmin edilen değerleri temsil eder.

Fig. 5'de confusion matrix Tablo I'de ise veri seti üzerinde elde edilen performans görülmektedir.

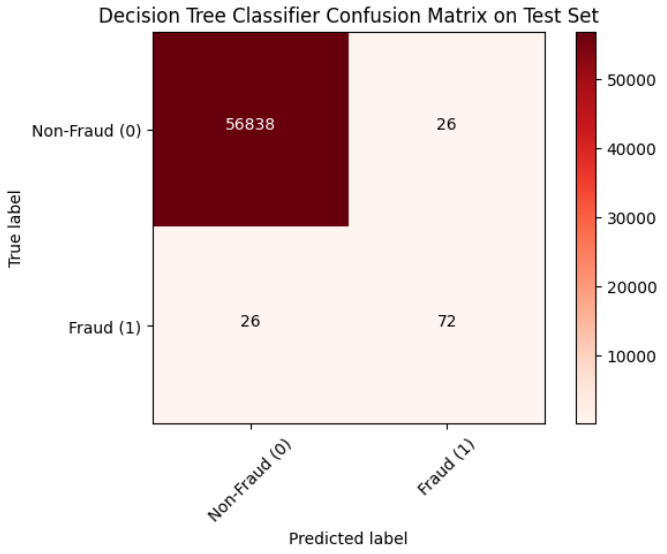

Fig. 5. DT Confusion Matrix

Model	Data	Accuracy	Precision	Recall	F1 Score	Support
DT	Train	1.0000	1.0000	1.0000	1.0000	394
DT	Test	0.9991	0.7347	0.7347	0.7347	98

TABLE I
DT PERFORMANCE METRICS

VII. LOGISTIC REGRESSION

Lojistik regresyon (Logistic Regression), ikili sınıflandırma problemlerinde kullanılan bir istatistiksel modeldir. Bu model, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi sigmoid fonksiyonu kullanarak tahmin eder ve sonuçları 0 ile 1 arasında bir olasılık olarak ifade eder. Lojistik regresyon, özellikle sınıflandırma problemlerinde yaygın olarak kullanılır ve verinin doğrusal olarak ayrılabilir olduğu durumlarda etkilidir.

Fig. 6'de confusion matrix Tablo II'de ise veri seti üzerinde elde edilen performans görülmektedir.

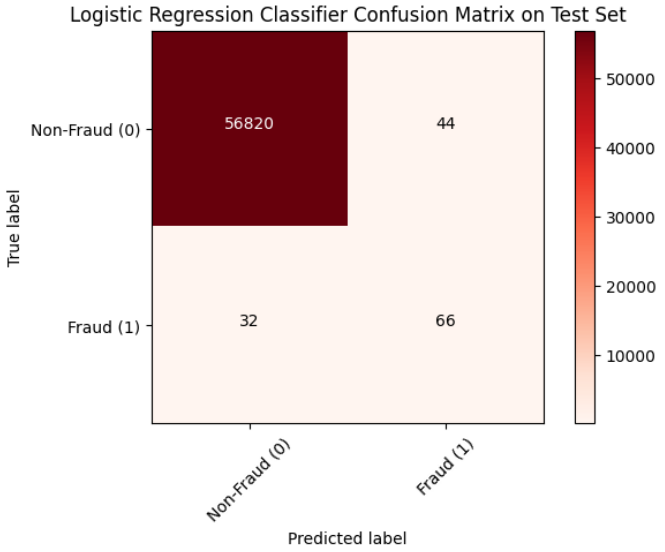

Fig. 6. LR Confusion Matrix

Model	Data	Accuracy	Precision	Recall	F1 Score	Support
LR	Train	0.9989	0.6915	0.7056	0.6985	394
LR	Test	0.9987	0.6000	0.6735	0.6346	98

TABLE II
LR PERFORMANCE METRICS

VIII. K-NEAREST NEIGHBORS (KNN)

K-En Yakın Komşu (K-Nearest Neighbors, KNN), sınıflandırma ve regresyon problemlerinde kullanılan basit ve sezgisel bir makine öğrenimi algoritmasıdır. KNN, bir veri noktasının sınıfını veya değerini belirlemek için en yakın K komşusunun etiketlerine veya değerlerine bakar. Bu algoritma, eğitim aşamasında herhangi bir model oluşturmaz ve tüm hesaplamaları tahmin aşamasında yapar, bu nedenle "tembel öğrenme" algoritması olarak da bilinir.

Fig. 7'de confusion matrix Tablo III'de ise veri seti üzerinde elde edilen performans görülmektedir.

IX. SUPPORT VECTOR MACHINE (SVM)

Destek Vektör Makineleri (Support Vector Machines, SVM), sınıflandırma ve regresyon analizinde kullanılan güçlü bir makine öğrenimi algoritmasıdır. SVM, veriyi en iyi şekilde

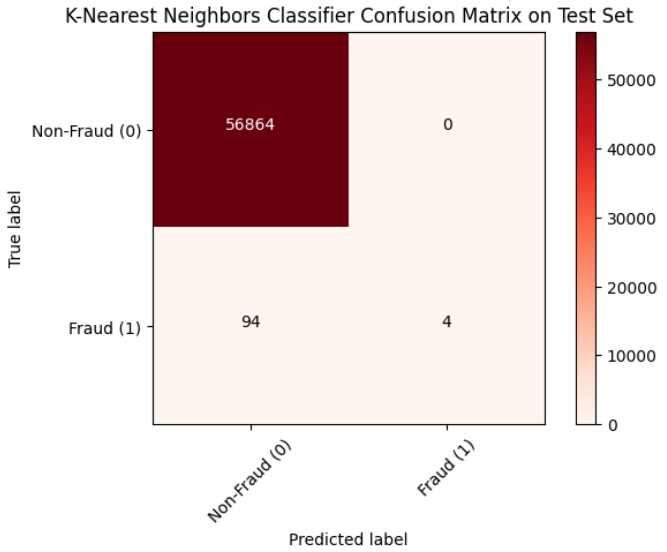

Fig. 7. KNN Confusion Matrix

Model	Data	Accuracy	Precision	Recall	F1 Score	Support
KNN	Train	0.9984	0.9750	0.0990	0.1797	394
KNN	Test	0.9983	1.0000	0.0408	0.0784	98

TABLE III
KNN PERFORMANCE METRICS

ayıran hiper düzlemi bulmaya çalışır ve bu hiper düzlemden en uzak mesafedeki destek vektörleri ile çalışır. Özellikle yüksek boyutlu veri setlerinde ve doğrusal olarak ayrılabilir olmayan verilerde etkili sonuçlar verir, çünkü çekirdek (kernel) triklerini kullanarak veriyi daha yüksek boyutlu uzaylara dönüştürebilir.

Fig. 8'de confusion matrix Tablo IV'de ise veri seti üzerinde elde edilen performans görülmektedir.

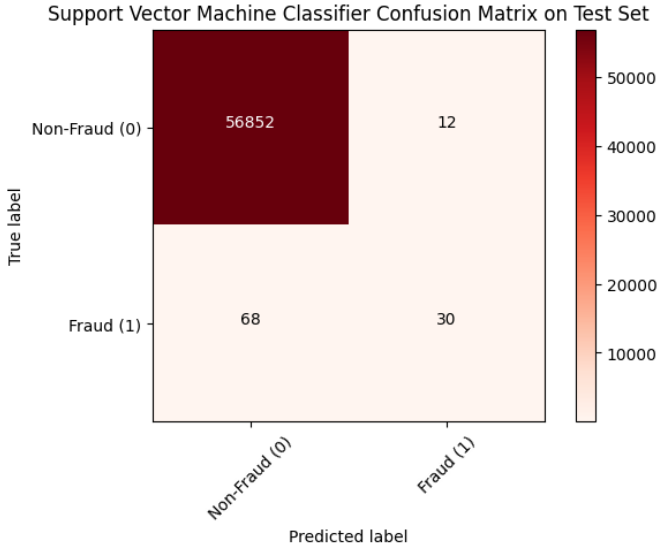

Fig. 8. SVM Confusion Matrix

Model	Data	Accuracy	Precision	Recall	F1 Score	Support
SVM	Train	0.9987	0.7674	0.3350	0.4664	394
SVM	Test	0.9986	0.7143	0.3061	0.4286	98

TABLE IV
SVM PERFORMANCE METRICS

X. RANDOM FOREST

Random Forest, birden fazla karar ağacının birleşiminden oluşan ve sınıflandırma veya regresyon problemlerinde kullanılan bir topluluk (ensemble) öğrenme yöntemidir. Her bir karar ağacı, veri setinin rastgele bir alt kümesi ve özelliklerin rastgele bir alt kümesi kullanılarak eğitilir, bu da modelin genelleme yeteneğini artırır ve aşırı öğrenmeyi (overfitting) azaltır. Sonuç olarak, sınıflandırma problemlerinde ağaçların çoğunluk oyu alınarak, regresyon problemlerinde ise ağaçların ortalaması alınarak nihai tahmin yapılır.

Fig. 9'de confusion matrix Tablo V'de ise veri seti üzerinde elde edilen performans görülmektedir.

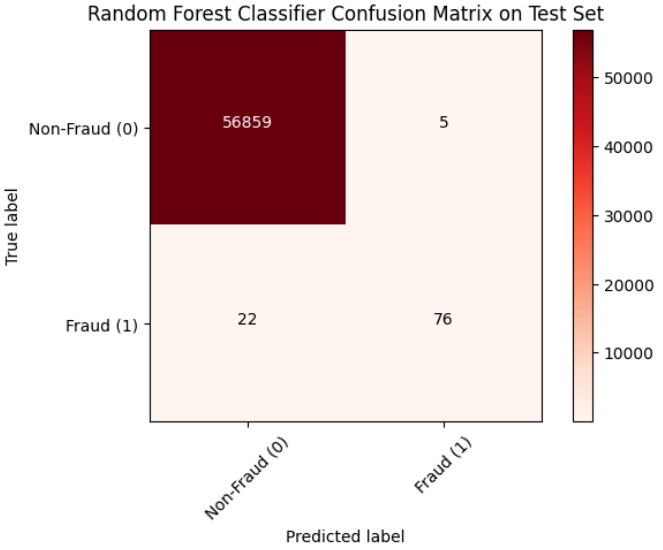

Fig. 9. RF Confusion Matrix

Model	Data	Accuracy	Precision	Recall	F1 Score	Support
RF	Train	1.0000	1.0000	0.9975	0.9987	394
RF	Test	0.9995	0.9383	0.7755	0.8492	98

TABLE V
RF PERFORMANCE METRICS

XI. NAIVE BAYES

Naive Bayes, olasılık teorisine dayanan ve özellikle metin sınıflandırma gibi yüksek boyutlu veri setlerinde etkili olan bir makine öğrenimi algoritmasıdır. Bu algoritma, her özelliğin sınıftan bağımsız olduğunu varsayar ve bu nedenle "naive" (saf) olarak adlandırılır. Naive Bayes, sınıflandırma problemlerinde hızlı ve hesaplama açısından verimli bir yöntemdir.

Fig. 10'de confusion matrix Tablo VI'de ise veri seti üzerinde elde edilen performans görülmektedir.

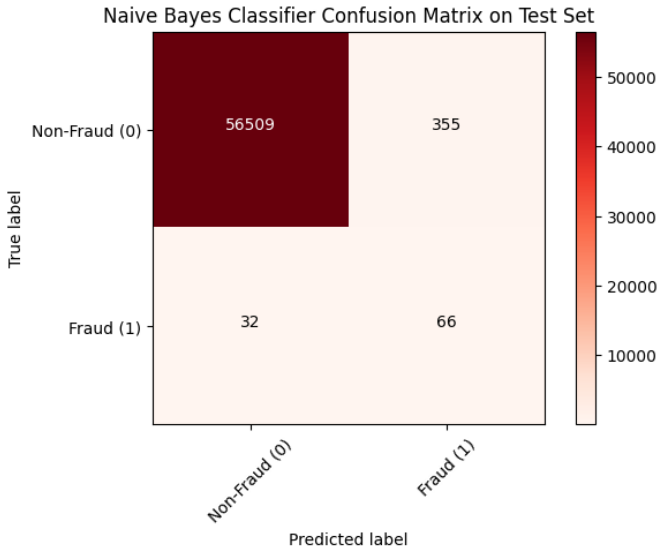

Fig. 10. NB Confusion Matrix

Model	Data	Accuracy	Precision	Recall	F1 Score	Support
NB	Train	0.9930	0.1470	0.6320	0.2385	394
NB	Test	0.9932	0.1568	0.6735	0.2543	98

TABLE VI
NB PERFORMANCE METRICS

XII. SONUÇLAR VE GELECEK ÇALIŞMALAR

Yukarıda elde edilen sonuçlara kümülatif olarak bakıldığında train ve test set ler için sırasıyla Fig. 11 ve Fig. 12 elde edilir.

Fig. 11. Accuracy and F1 Score of Classifiers on Train Set

Bu çalışmadan yapılan çıkarımlar aşağıdaki gibidir;

- Fig. 11 den anlaşılacağı üzere RF ve DT train set üzerinde %100 başarıya ulaşmışlardır. Bu de modellerin diğer yöntemlere göre daha iyi çalışacağının işaretidir.
- Fig. 12 den anlaşılacağı üzere en başarılı sınıflandırma yöntemleri sırasıyla RF, DT, LR, SVM, NB, KNN şeklindedir.
- Her makine öğrenmesi yöntemi her problem seti için uygun değildir.

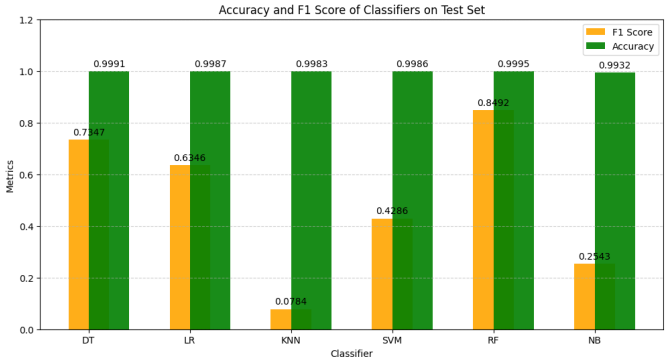

Fig. 12. Accuracy and F1 Score of Classifiers on Test Set

- Veri seti ön işleme modelin çalışabilmesi / performansı için önemlidir

Gelecekte yapılabilecek çalışmalar aşağıdaki gibidir;

- Derin öğrenme yöntemleri üzerinden başarım ölçümü yapılabilir.
- Imbalanced veri setini balanced bir veri seti haline getirerek sınıflandırma performansları ölçülebilir. (Random Oversampling, Random Undersampling, SMOTE (Synthetic Minority Over-sampling Technique))
- KNN'de yaşanan başarısızlığın düzeltilmesi için çalışma yapılabilir.
- SVM'de farklı kernel parametreleri ve Grid Search ile parametre optimizayonu yapılabilir.

ACKNOWLEDGMENT

Bu makalenin oluşturulmasında ve yayına hazır hale getirilmesinde bizi teşvik eden, geri bildirimlerini ve desteklerini esirgemeyen sayın hocamız Prof. Dr. Songül VARLI'ya teşekkürü bir borç bilirim.

REFERENCES

[1] "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms" https://ieeexplore.ieee.org [Online]. Available: https://ieeexplore.ieee.org/document/9755930. [Accessed: 30-Dec-2024].

[2] "Machine Learning Group - ULB - Credit Card Fraud Detection Dataset" https://www.kaggle.com [Online]. Available: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data. [Accessed: 30-Dec-2024].

[3] "Credit Card Fraud Detection Using Lightgbm Model" https://ieeexplore.ieee.org [Online]. Available: https://ieeexplore.ieee.org/document/9134072. [Accessed: 30-Dec-2024].

[4] "Overfitting," GeeksforGeeks. [Online]. Available: https://media.geeksforgeeks.org/wp-content/uploads/20190523171704/overfitting_21.png. [Accessed: 18-Nov-2024].

[5] "Machine Learning Equations in LaTeX," blmoistawinde.github.io. [Online]. Available: https://blmoistawinde.github.io/ml_equations_latex/. [Accessed: 18-Nov-2024].