HTRU2 VERİTABANINA 3 MAKİNE ÖĞRENİMİ ALGORİTMASI KULLANARAK PULSAR VEYA RFI OLARAK SINIFLANDIRMA

YAKUP ŞEKERCİ

İSTANBUL

ARALIK, 2022

High Time Resolution Universe (HTRU) Analiz Raporu

Bu raporun amacı HTR2'yi analiz etmektir. Makine öğrenimi algoritmalarının kullanımı yoluyla analiz edeceğiz ve bu değişkenlerin dengesiz bir şekilde nasıl davrandığını vurgulayacağız.

Veriseti Tanıtım

Veri seti, HTRU (Güney) tarafından toplanan pulsar adaylarının sinyallerinin istatistiksel özelliklerini içerir. Bu projenin amacı, "Orijinal Pulsar Adaylarını" ve "Gürültüyü" sınıflandırmak için makine öğrenimi modelleri yetiştirmektir. Verilerin pozitif:negatif sınıf dengesizliği 1:10'dur, bu da eğitim modellerinin pozitifleri doğru bir şekilde tanımlamasını zorlaştırır. Bu nedenle amaç, modellerin öngördüğü yanlış negatif sayısını azaltmaktır. Yanlış negatiflerin sayısı, yüksek örnekleme ve düşük örnekleme kullanılarak azaltıldı.

Attribute İnformation

Veri seti, entegre profilinin ve DM-SNR eğrisinin istatistiksel özelliklerini içerir. İstatistiksel özellikler şunları içerir:

1. Mean
2. Standart Deviation
3. Skewness
4. Kurtosis

Attribute	Range	Type
Mean of the integrated profile	5.8125~192.6171	Real
Standard deviation of the integrated profile	24.7720~98.7789	Real
Excess kurtosis of the integrated profile	-1.8760~8.0695	Real
Skewness of the integrated profile	-1.7918~68.1016	Real
Mean of the DM-SNR curve	0.2132~223.3921	Real
Standard deviation of the DM-SNR curve	7.3704~110.6422	Real
Excess kurtosis of the DM-SNR curve	-3.1392~34.5398	Real
Skewness of the DM-SNR curve	-1.9769~1191.0008	Real

Entegre Nabız Profili: Her pulsar, nabız profili olarak bilinen benzersiz bir nabız emisyon modeli üretir. Pulsarın parmak izi gibidir. Pulsarları yalnızca nabız profillerinden belirlemek mümkündür. Ancak nabız profili her dönemde biraz değişir. Bu, pulsarın tespit edilmesini zorlaştırır. Bunun nedeni, sinyallerinin tekdüze olmaması ve fazla mesai tamamen istikrarlı olmamasıdır. Bununla birlikte, bu profiller, binlerce dönüşün ortalaması alındığında kararlı hale gelir.

DM-SNR Eğrisi: Pulsarlardan yayılan radyo dalgaları, serbest elektronlarla dolu uzayda uzun mesafeler kat ettikten sonra dünyaya ulaşır. Radyo dalgaları doğası gereği elektromanyetik olduğu için bu elektronlarla etkileşime girer ve bu etkileşim dalganın yavaşlamasına neden olur. Önemli olan nokta, atarcaların geniş bir frekans aralığı yaymasıdır ve elektronların dalgayı yavaşlatma miktarı frekansa bağlıdır. Daha yüksek frekanslı dalgalar, daha yüksek frekanslı dalgalara kıyasla daha az ekilir. yani düşük frekanslar teleskopa yüksek frekanslardan daha sonra ulaşır. Buna dispersiyon denir.

Preprocessing

Bu veri kümesindeki veriler temiz ve eksiksizdir. Eksik veri yok ve sayı yok özellikleri küçüktür. Çok fazla olmadığı için özellik mühendisliği burada kullanışlı değil bu nedenle, ön işleme esas olarak iki sorunu içerir, biri aşırı uydurma sorunudur pozitif ve negatif örnek sayısının farkından kaynaklanır.

Z-Score Normalization

Özelliklerin sayısal dağılımı son derece dengesizdir. Varyasyon sekiz özellik arasında Skewness of the DM-SNR curve değişkeninde Büyük değerler 1191'e ulaşabilir ve küçük değerler sadece 0.2'dir. Bu tür dağımlar girdi ve eğitim için son derece düşmancadır. Yani normalleşme bu veri kümesi için gereklidir. Burada değişkenlerde belirli bir z değerin altında veya üstünde kalan alanıda bulmak için Z-Score Normalization tekniğini kullandım.

Resampling

Gerçekte çok fazla pulsar bulunmadığından, pozitif örnek veri hacmi sadece 1639, negatif örnek ise neredeyse on katı olan 16259'dur. Ham verilerle yapılan doğrudan eğitim, Negatif veri hacmı aşırı uyum sağlamaya neden olabilir. Bu sorunu çözmek bazı makine öğrenimi algoritmaları için aşırı örnekleme denenmiştir bu nedenle örnekleme sayısı aşırı örnekleme uygulanarak yükseltilmiştir.

Random Over Sampling

Aşırı Örnekleme, azınlık sınıfındaki örnek sayısını rastgele artırır azınlık sınıfının daha yüksek bir temsilini sunmak için bunları çoğaltmak gereklidir. Bu yöntem hiçbir bilgi kaybına yol açmaz. Örnekleme altında daha iyi performans gösterir. Aşırı uyum olasılığını artırır azınlık sınıfı olaylarını çoğaltır.

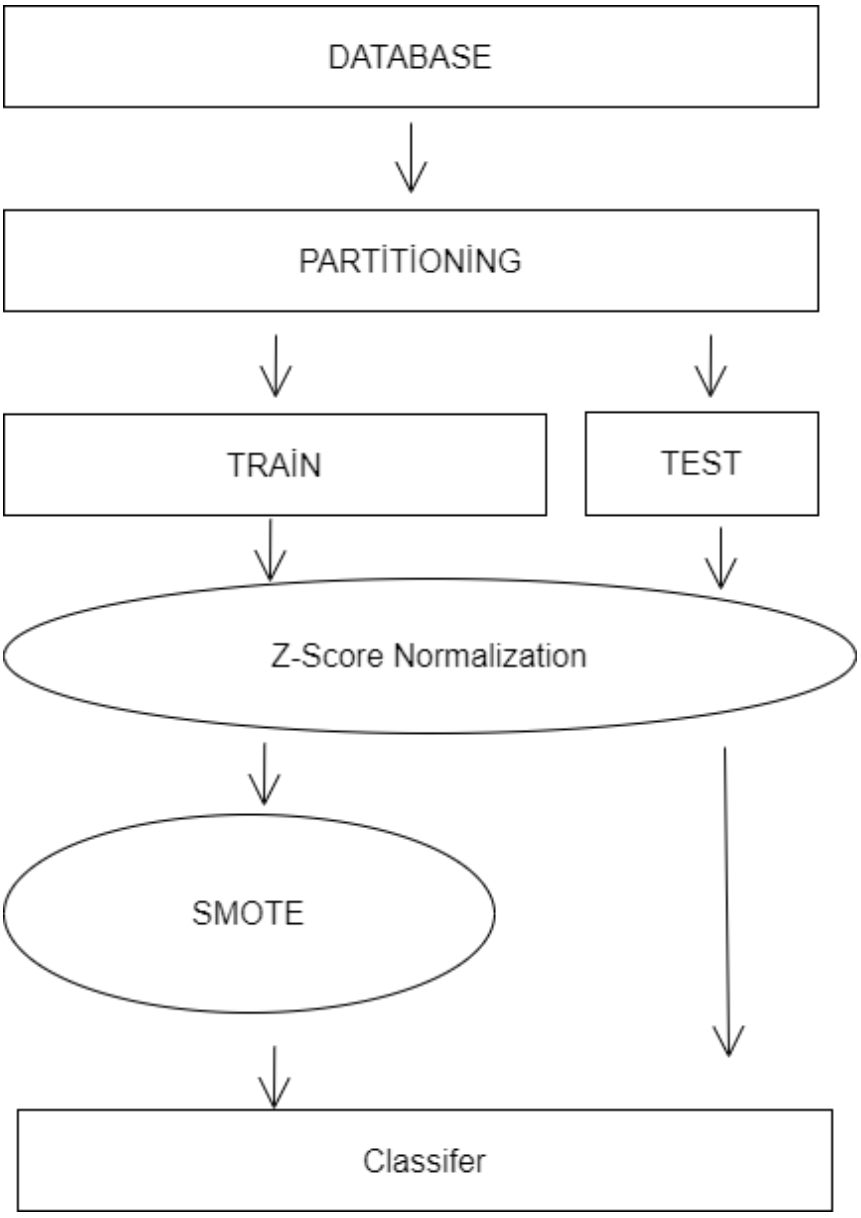
Synthetic Minority Over-Sampling Technique (SMOTE)

Aşırı Örnekleme Tekniği (SMOTE aşırı uyumu önlemek için takip uygulanır. azınlık örnekleri ana veri kümesine eklenir. Örnek olarak azınlık sınıfı ve ardından yeni sentetik benzer örnekler oluşturulur. Bu sentetik örnekler daha sonra orijinal veri kümesine eklenir. Yeni veri kümesi sınıflandırma modellerini eğitmek için örnek olarak kullanılır ve aşırı öğrenmenin önüne geçmiş olunur.

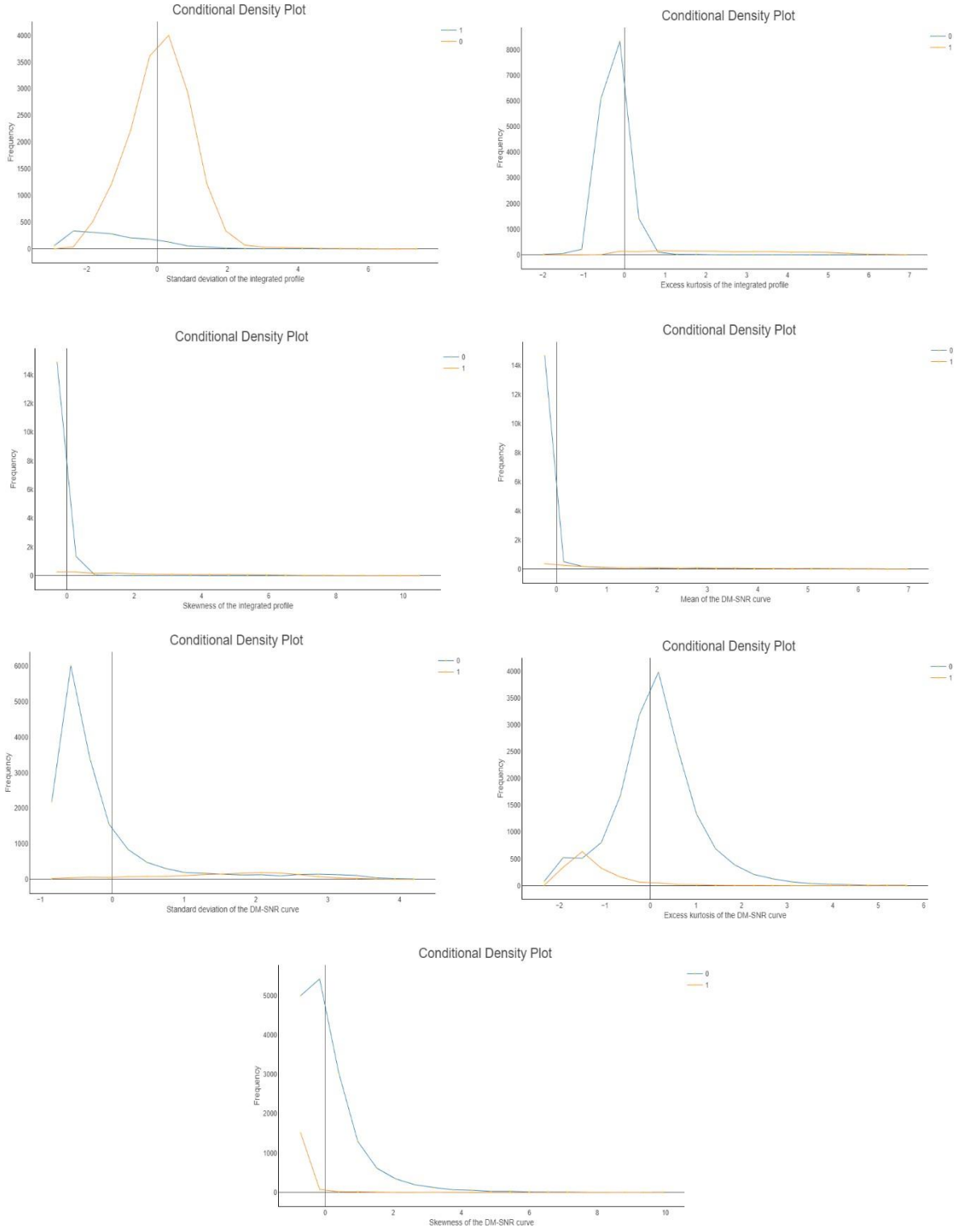
Column Filter

Birbirleriyle yüksek korelasyonlu olan Mean of the integrated profile değişkeni modelin overfitting problemi ile karşılaşmaması için silinmiştir.

Dataset Methodology



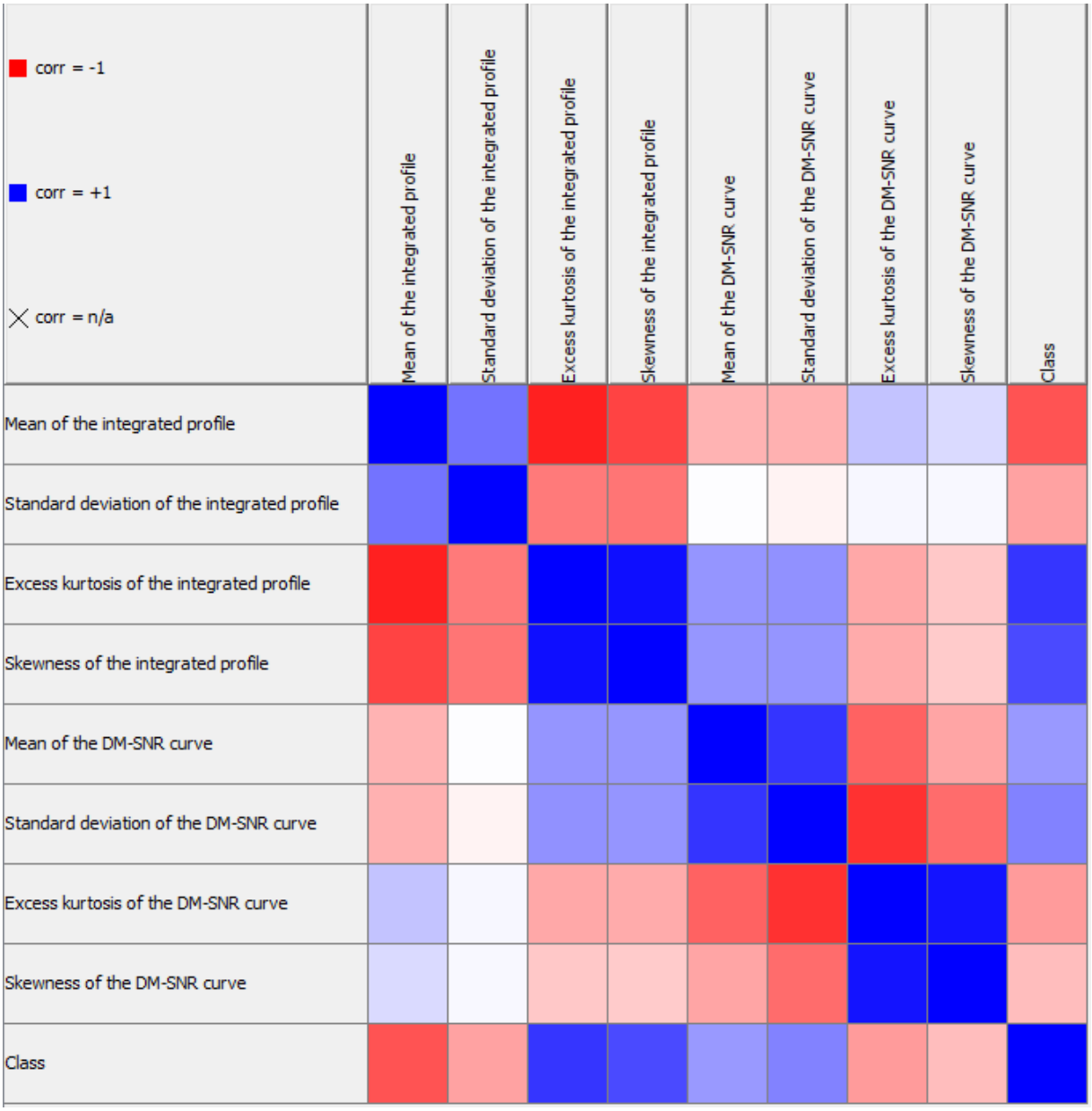
Z-Score Normalization Uygulanmış HTRU2 Değişkenlerinin Grafikleri



Şekil 1

Şekil 1'de gösterilen grafiklerden, değişkenlerin aykırı değerler olmadığını ve iyi dağılmış olduğunu gözlemleyebiliriz. Bu nedenle veri ön işleme tekniklerini kullanmayacağız.

Z-Score Normalization Uygulanmış HTRU2 değişkenlerinin Korelasyon Isı Haritası:



Değişkenlerin aralarında bir ilişki olduğu aşikardır.

Logistic Regression

Lojistik regresyon, iki veri faktörü arasındaki ilişkileri bulmak için matematikten yararlanan bir veri analizi tekniğidir. Lojistik regresyon, daha sonra diğerine dayalı bu faktörlerden birinin değerini tahmin etmek için bu ilişkiyi kullanır. Tahminin genellikle evet ya da hayır gibi sınırlı sayıda sonucu vardır.

Lojistik regresyon, matematikte *x* ve *y* arasındaki denklem olarak lojistik fonksiyonu veya logit fonksiyonu kullanan istatistiksel bir modeldir. Logit fonksiyonu, *y*'yi *x*'in sigmoid fonksiyonu olarak eşler.

$$f(x) = \frac{1}{1 + e^{-x}}$$

Lojistik regresyon, gürültü önleme performansı iyi olan bir modeldir. Bu problemdeki öznitelik sayısı fazla değildir ve bu nedenle lojistik regresyon bu soruna uygulanır.

LOGİSTİC REGRESSİON UYGULADIĞIMIZ HTRU2 VERİTABANININ PERFORMANSI

R^2:	0,768
Mean absolute error:	0,02
Mean squared error:	0,02
Root mean squared error:	0,141
Mean signed difference:	-0,009
Mean absolute percentage error:	NaN
Adjusted R^2:	0,768

CONFUSİON MATRİX

Class \ Pre...	0	1
0	4832	28
1	79	431

Correct classified: Wrong classified: 107

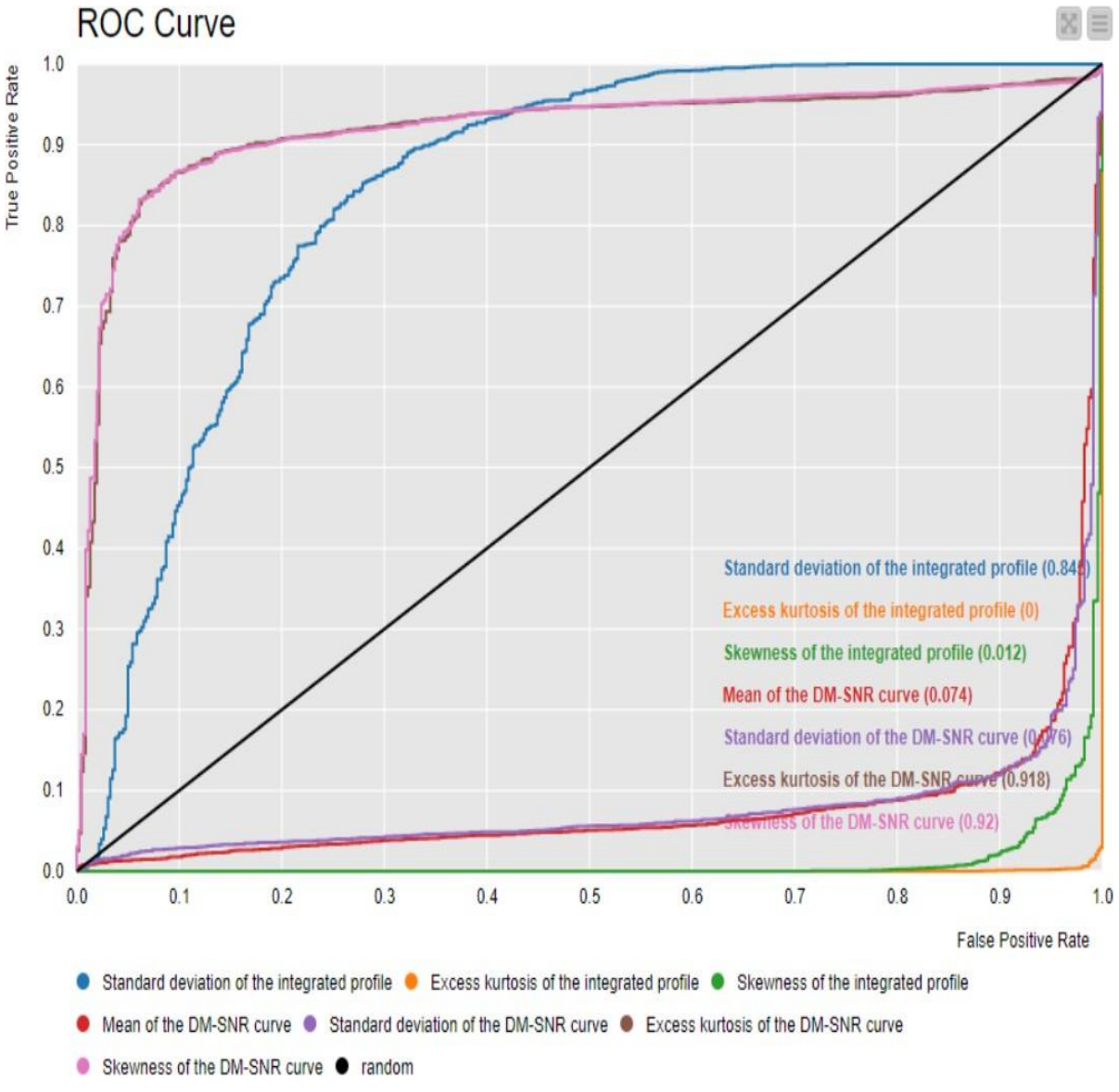
Accuracy: 98,007% Error: 1,993%

Cohen's kappa (κ):

ACCUARCY STATİSTİC

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
0	4832	79	431	28	0.994	0.984	0.994	0.845	0.989	?	?
1	431	28	4832	79	0.845	0.939	0.845	0.994	0.89	?	?
Overall	?	?	?	?	?	?	?	?	?	0.98	0.879

ROC CURVE



SUPPORT VECTOR MACHİNE (SVM)

Destek Vektör Makineleri (Support Vector Machine) genellikle sınıflandırma problemlerinde kullanılan gözetimli öğrenme yöntemlerinden biridir. Bir düzlem üzerine yerleştirilmiş noktaları ayırmak için bir doğru çizer. Bu doğrunun, iki sınıfının noktaları için de maksimum uzaklıkta olmasını amaçlar. Karmaşık ama küçük ve orta ölçekteki veri setleri için uygundur.

$$\hat{y} = \begin{cases} 0 \text{ if } \mathbf{w}^T \cdot \mathbf{x} + b < 0, \\ 1 \text{ if } \mathbf{w}^T \cdot \mathbf{x} + b \geq 0 \end{cases}$$

w; ağırlık vektörü (θ1), x; girdi vektörü, b; sapmadır (θ0). Yeni bir değer için çıkan sonuç 0'dan küçükse, beyaz noktalara daha yakın olacaktır. Tam tersi, çıkan sonuç 0'a eşit veya büyükse, bu durumda siyah noktalara daha yakın olacaktır.

SVM UYGULADIĞIMIZ HTRU2 VERİTABANININ PERFORMANSI

R^2:	0,732
Mean absolute error:	0,022
Mean squared error:	0,022
Root mean squared error:	0,148
Mean signed difference:	-0,012
Mean absolute percentage error:	NaN
Adjusted R^2:	0,732

CONFUSİON MATRİX

Class \ Pre...	0	1
0	4861	26
1	92	391

Correct classified: 5.252 Wrong classified: 118

Accuracy: 97,803% Error: 2,197%

Cohen's kappa (κ):

ACCUARCY STATİSTİC

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
0	4861	92	391	26	0.995	0.981	0.995	0.81	0.988	?	?
1	391	26	4861	92	0.81	0.938	0.81	0.995	0.869	?	?
Overall	?	?	?	?	?	?	?	?	?	0.978	0.857

PROBABILISTIC NEURAL NETWORK (PNN)

Olasılıklı bir sinir ağı (PNN), sınıflandırma ve örüntü tanıma problemlerini çözmek için kullanılan bir tür ileri beslemeli sinir ağıdır. PNN tekniğinde, her sınıfın ebeveyn olasılık dağılım fonksiyonu , bir Parzen penceresi ve parametrik olmayan bir fonksiyon kullanılarak tahmin edilir. Her sınıfın yeni girdi verilerinin sınıf olasılığını tahmin etmek için kullanılır ve Bayes kuralı, en yüksek arka olasılığa sahip sınıfı yeni girdi verilerine tahsis etmek için kullanılır. Bu yöntemle yanlış sınıflandırma olasılığı azaltılır. Bu tür YSA, bir Bayes ağı ve Kernel Fisher diskriminant analizi olarak bilinen istatistiksel bir yaklaşım kullanılarak oluşturulmuştur.

$$\hat{p}(x) = \frac{N_h(x)}{2hM}$$

P(x)'in belirli bir wi sınıfına ait çok boyutlu bir örneğe atıfta bulunduğu eğitim veri kümesini kullanarak denetimli bir model sınıflandırma sorununda sınıf koşullu yoğunlukları ("olasılıklar" olarak da bilinir) p(x|wi)I tahmin etme Parzen pencere tekniğinin önemli bir uygulamasıdır.

PNN UYGULADIĞIMIZ HTRU2 VERİTABANININ PERFORMANSI

R^2:	0,716
Mean absolute error:	0,023
Mean squared error:	0,023
Root mean squared error:	0,153
Mean signed difference:	-0,015
Mean absolute percentage error:	NaN
Adjusted R^2:	0,716

CONFUSİON MATRİX

Class \ Pre...	0	1
0	4866	21
1	104	379

Correct classified: 5.245 Wrong classified: 125

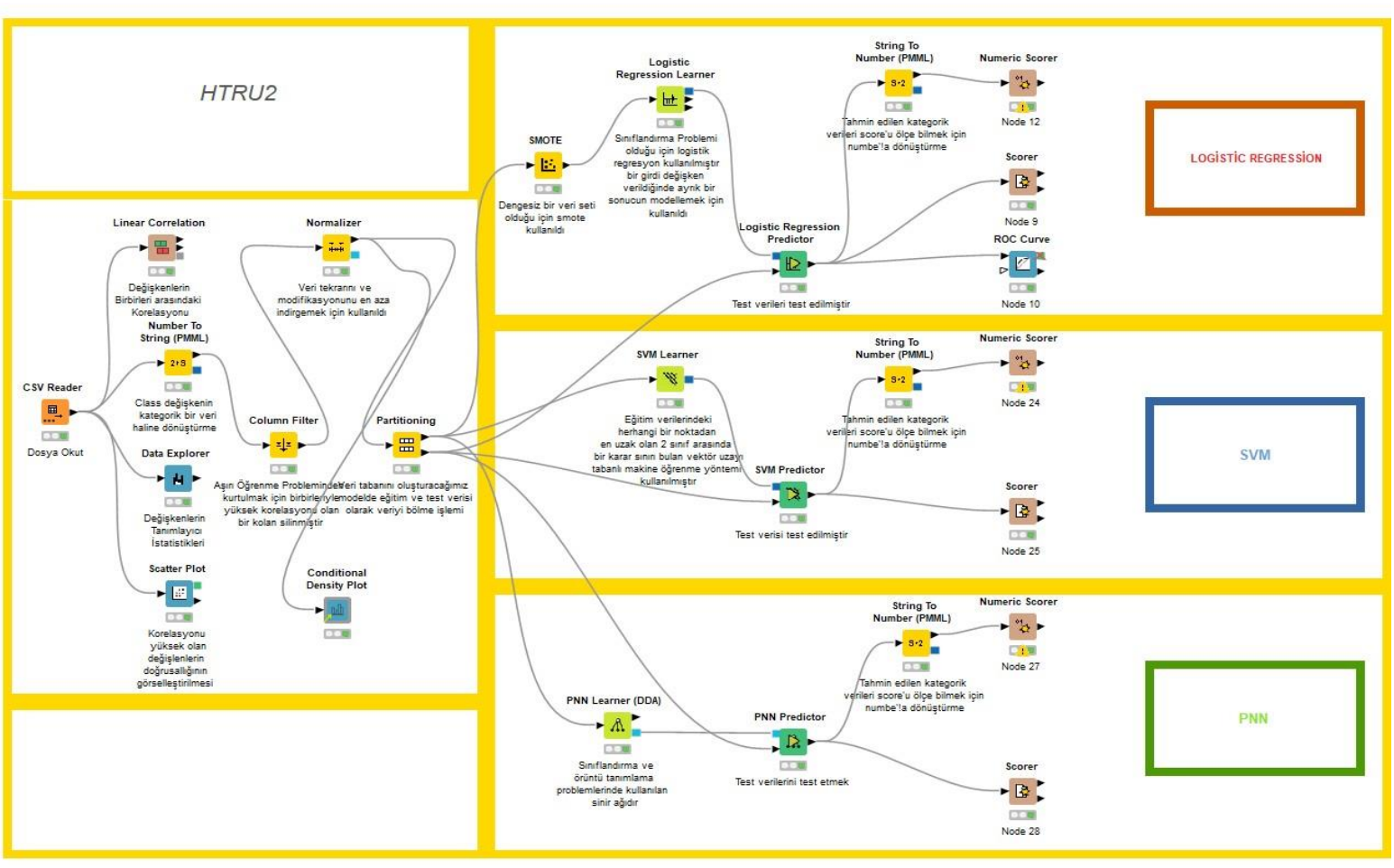
Accuracy: 97,672% Error: 2,328%

Cohen's kappa (κ): 0,846%

ACCUARCY STATİSTİC

Row ID	TruePo...	FalsePositives	TrueNegatives	FalseN...	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
0	4866	104	379	21	0.996	0.979	0.996	0.785	0.987	?	?
1	379	21	4866	104	0.785	0.948	0.785	0.996	0.858	?	?
Overall	?	?	?	?	?	?	?	?	?	0.977	0.846

KNIME İŞ AKIŞI



ÖZET VE SONUÇLAR

Bu projede HTRU2 veri setine dayalı olarak sınıflandırıcı oluşturmayı amaçladık.Öncelikle biz logistik regression algoritmasında dengesiz veri seti ile baş etmek için aşırı örnekleme yöntemini kullandık.Logistik regressionun yanı sıra SVM ve PNN algoritmalarını da kullandık.Nihai sonuç logistik regression diğer kurduğumuz algoritmadan daha iyi bir sonuç verdi.Astronomi ve fizik bilgimizin dışında bu veri seti modellenmiştir.