Öğrenci No: 21011004, 20011502

Ad-Soyad : Yavuz Çetin, Osman Berkay Sukas

Öğrenci E-Postası: yavuz.cetin1@std.yildiz.edu.tr

berkay.sukas@std.yildiz.edu.tr

Bölüm: Bilgisayar Mühendisliği

Biyoenformatiğe Giriş
Proje Ödevi

Ders Yürütücüsü
Prof.Dr. Nizamettin Aydın
Mayıs/2024

Projenin Tanımı
Proje kapsamında bir veri seti seçilmesi ve seçilen veri setine çeşitli ön işleme yöntemleri uygulandıktan sonra seçilen 3 adet makine öğrenmesi yöntemiyle sonuçlar elde edilmesi ve sonrasında bu sonuçların karşılaştırılması istenilmiştir.

Projenin Çözümü
Projenin çözümünde Support Vector Machines, K nearest neighbors ve Logistic Regression yöntemleri kullanılmıştır.

Karşılaşılan Problemler
Veri ön işleme aşamasında dosya isimlerinden kaynaklı bütün dosyalarda aynı anda sıkıntı çıkmıştır. Bu soruna çözüm olarak dosyaların isimlerini değiştiren bir kod parçası yazılmıştır.

Veri Setindeki Hastalıklar
Veri setinde bulunan hastalıklar bu bölümde anlatılmıştır.

Influenza Virus
İnfluenza, influenza virüslerinin neden olduğu bulaşıcı bir solunum yolu hastalığıdır. A, B ve C tipleri bulunan influenza virüsü solunum yollarının parçası olan burun, boğaz ve akciğerleri enfekte eder. Daha çok mevsimsel olarak ortaya çıkan influenza halk arasında grip olarak bilinen hastalığa neden olan virüstür. İnfluenza'nın A ve B tipi yaygın olan türü olup, A türü influenza salgınlara neden olabilenidir. C tipi ise daha hafif belirtilere neden olan türüdür. İnfluenza virüsünde kişilerde ateş, vücut ağrısı, boğaz ağrısı, titreme ve yorgunluk belirtileri görülür.

Solunum sistemine saldıran ve vücudu olumsuz etkileyen bir virüs olan influenza, genellikle bağışıklık sistemi düşük olan hasta, çocuk ve yaşlı bireyleri daha çok etkiler. Diğer yandan influenza insanla birlikte domuz, at, kuş ve deniz memelileri üzerinde de hastalıklara sebep olabilir. [1]

Rhino Virus
Soğuk algınlığının en yaygın nedeni olan Rhinovirüs gergedan gribi olarak da bilinir. Bu virüs genelde burun ve boğaz bölgesinde enfeksiyonlara yol açarak tipik soğuk algınlığı belirtilerini tetikler. Rhinovirüs daha çok sonbahar ve kış aylarında görülür. Hava yoluyla ve özellikle damlacık yoluyla yayılır bu da genellikle hapşırma veya öksürme yoluyla havaya saçılan virüsün solunumla vücuda giriş yapmasını sağlar. Ayrıca elle temas sonucu da bulaşabilir. [2]

Hastalığı Seçme Sebebi
Bu hastalığı seçme sebebimiz olarak hastalığın yaygın ve bulaşıcı bir hastalık olması, toplum sağlığı için önemli etkileri olması verilebilir. Ayrıca bu iki hastalığa aşı geliştirilse bile mutasyon gibi sebeplerle bu hastalıkların araştırılması ve kalıcı çözümlere ulaşılması için araştırma yapmanın önemli olduğu düşünülmüştür.

Kullanılan Veri Seti

Kullanılan veri setinin ismi 'A Systems Approach to Understanding Human Rhinovirus and Influenza Virus Infection'dır. Bu veri setinde toplam örnek sayısı 186'dır. Aşağıda her bir örnekten kaç adet olduğu verilmiştir.

Uninfected -> 51

Influenza Virus Infected -> 45

Rhino Virus Infected -> 45

Influenza Virus & Rhino Virus Infected -> 45

Görülebileceği gibi sadece 1 hastalığa sahip olan örnekler dışında her 2 hastalığa sahip örnekler de veri setinde bulunmaktadır.

Ön İşleme Adımları

Veri Setinin Alınması

Bu aşamada kullanılacak olan veri seti GEOquery ve BiocManager kütüphaneleri kullanılarak alınmıştır.

Dosya adlarının düzenlenmesi

Kullanılan veri setinde örneklerin isimlendirmesinde karmaşıklık olduğu için bu aşamada dosya isimleri belirli bir formata getirilmiştir.

Phenodata hazırlanması

Bu aşamada fenotip verileri dataframe haline getirilerek dosya adlarıyla eşleştirilmiştir. Fenotip verileri kaydedilmiştir.

CEL dosyalarının okunması

Bu aşamada CEL dosyaları okunmuştur.

Normalizasyon

Bu aşamada CEL verileri RMA(Robust Multi-array Average) kullanılarak normalizasyon gerçekleştirilmiştir.

Anotasyon

Bu aşamada aynı gen sembolüne sahip tekrar eden satırların ortalamaları alınarak tek bir satıra kaydedilmiştir.

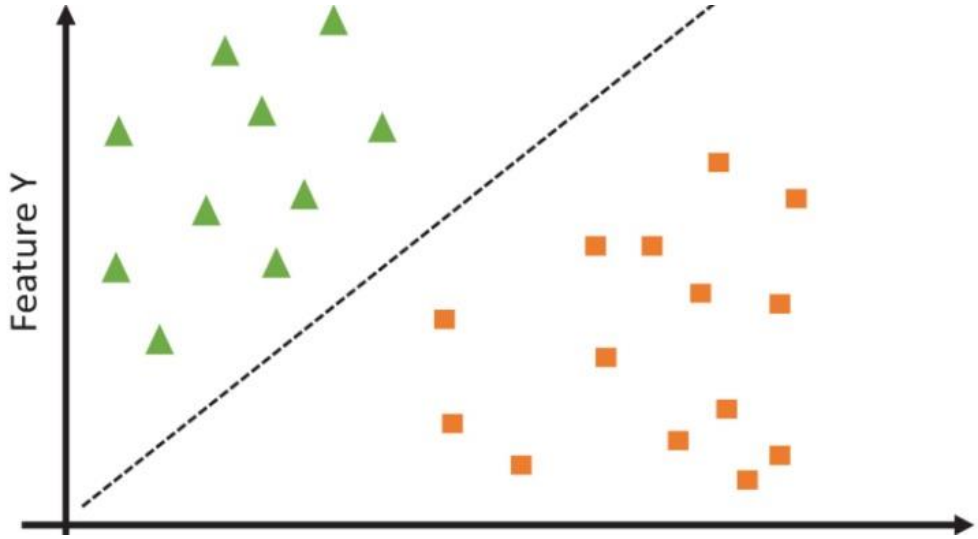
Verilerin Kaydedilmesi

Ön işleme adımı tamamlanmış veriler 'GSE71766.csv' isimli dosyaya kaydedilir.

Kullanılan Makine Öğrenmesi Yöntemleri

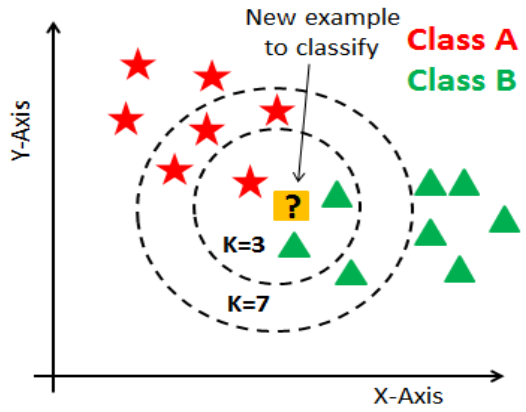
Support Vector Machines(Destek Vektör Makineleri)

Makine öğrenimi ve veri madenciliğinde yaygın olarak kullanılan güçlü bir sınıflandırma algoritmasıdır. SVM, verileri iki sınıfa ayırmak için en iyi hiper düzlemi bulmayı amaçlar. Bu hiper düzlem, iki sınıf arasındaki mesafeyi maksimize eden, yani sınıflar arasındaki "marjini" en geniş tutan bir sınır çizer. SVM, bu sınıflandırma görevini gerçekleştirmek için bir eğitim kümesi kullanır ve her veri noktası için bir sınıf etiketi gerektirir. Algoritma, doğrusal olarak ayrılabilen veriler için mükemmel sonuçlar verirken, doğrusal olmayan veriler için de çekirdek (kernel) hileleri kullanarak bu verileri daha yüksek boyutlu bir uzaya projeksiyon yapabilir. Bu sayede, doğrusal olmayan veri setlerinde de etkili bir şekilde çalışabilir. SVM'nin en büyük avantajlarından biri, yüksek boyutlu veri setlerinde bile başarılı sonuçlar verebilmesidir. Ayrıca, aşırı öğrenmeyi (overfitting) engellemek için düzenleme (regularization) parametreleri sunarak modelin genel performansını iyileştirir. Bu nedenle, SVM'ler genellikle yüz tanıma, metin sınıflandırma ve biyoinformatik gibi çeşitli alanlarda tercih edilen bir yöntemdir.



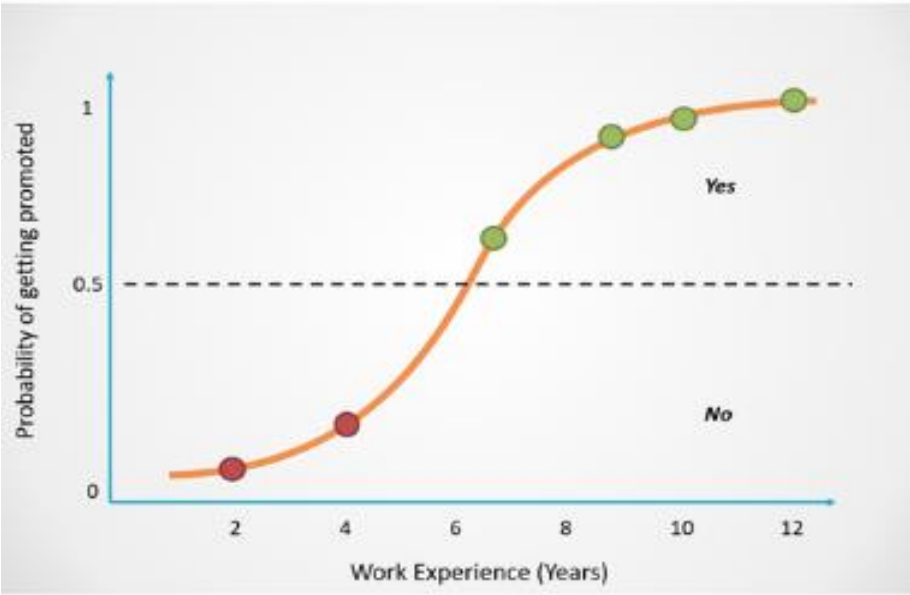
K-Nearest Neighbors(K En Yakın Komşular)

Bu algoritma sınıflandırmayı gerçekleştirirken, sınıflandırılacak yeni verinin, daha önce sınıflandırılmış k adet veriye olan uzaklıklarının hesaplanarak en yakın uzaklığa sahip olan sınıfa dahil edilmesi mantığına göre çalışmaktadır. Karşılaştırmada yeni verinin komşuluk mesafesinin hesaplanmasında genellikle Öklid Bağıntısı kullanılmaktadır. Hesaplamalar sonucunda yapılacak tahminler belirlenirken komşu sınıfların örnek sayılarının çokluğu dikkate alınmaktadır. Bu yöntem, eğitim aşamasının olmayışı ve gürültü verilerine karşı dayanıklı olması nedeniyle tercih edilen bir sınıflandırma algoritmasıdır.



Lojistik Regresyon

Lojistik regresyon, istatistiksel bir modelleme tekniğidir ve bağımlı değişkenin kategorik olduğu durumlar için kullanılır. Bu yöntem, bir veya daha fazla bağımsız değişkenin kategorik bir sonuç üzerindeki etkisini tahmin etmeyi amaçlar. Genellikle sınıflandırma problemlerinde kullanılır ve sonuç, bir olayın gerçekleşme olasılığı veya belirli bir kategoriye ait olma olasılığı olarak yorumlanabilir. Burada bir degişkenin bağımlılığının birden fazla olması durumunda elde edilecek sonuçlar 0 ve 1'e indirgenerek gösterilmektedir. Lojistik regresyon, lojistik fonksiyonu adı verilen bir sınır fonksiyonu kullanarak doğrusal olmayan ilişkileri modelleyebilir.



SVM ile Elde Edilen Sonuçlar

Kernel lineer olarak ayarlandığında ve train-test split 0.8 ve 0.2 olacak şekilde model çalıştırıldığında 0.68 gibi bir accuracy elde edilmiştir. Aşağıda elde edilen sonucun görseli verilmiştir.

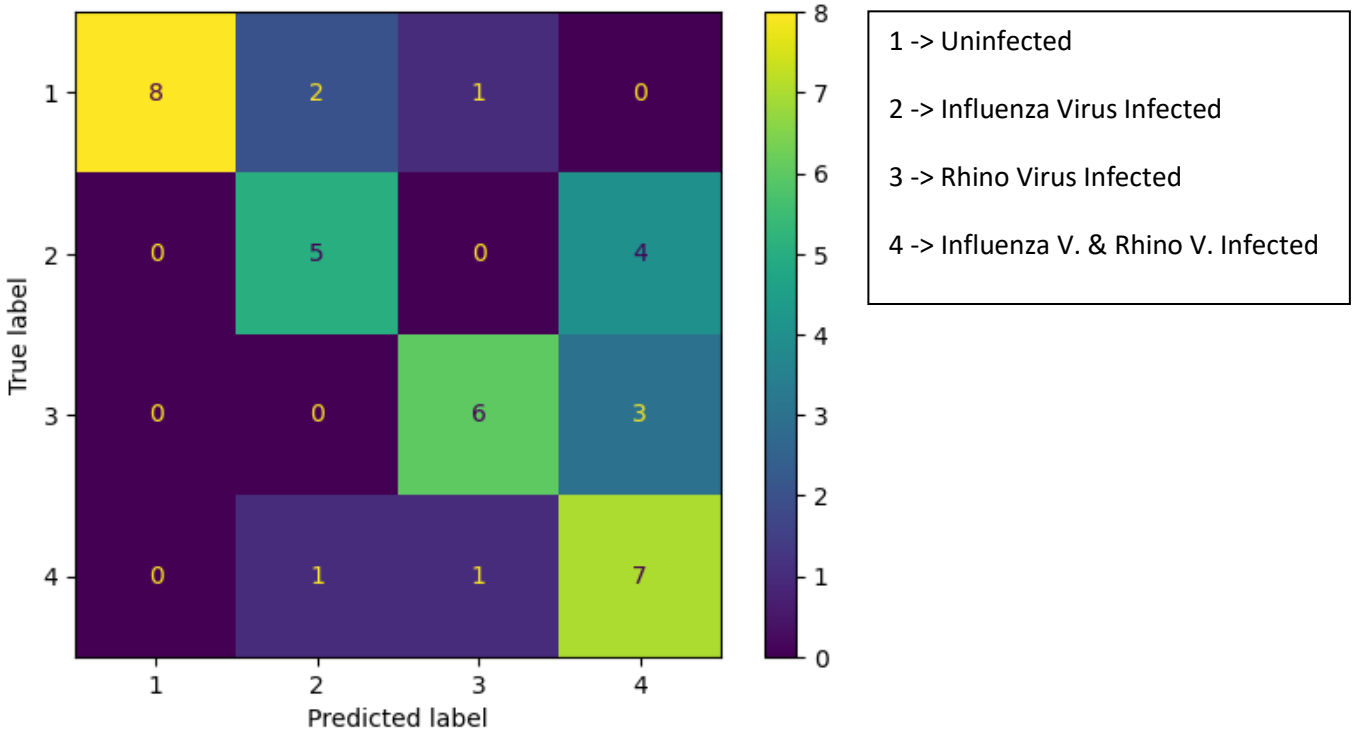
```
Accuracy: 0.6842105263157895
Macro Precision: 0.71875
Macro Recall: 0.6818181818181818
```

Modelin precision, recall ve f1-score bilgileri aşağıdaki görselde verilmiştir.

	precision	recall	f1-score	support
Uninfected	1.00	0.73	0.84	11
Influenza virus infected	0.62	0.56	0.59	9
Rhino virus infected	0.75	0.67	0.71	9
Influenza & Rhino virus infected	0.50	0.78	0.61	9
accuracy			0.68	38
macro avg	0.72	0.68	0.69	38
weighted avg	0.73	0.68	0.69	38

Görülebileceği üzere sağlıklı örneklerde çok yüksek bir f1-score elde edilmiştir.

Aşağıda ise bu modele ait confusion matrix verilmiştir.



Şekilde görülebileceği üzere tek hastalığa sahip olan hastalar, her iki hastalığa sahip olan hastalarla karışmaktadır.

Daha etkili bir sınıflandırma için ayrıca veri seti Uninfected-Infected olacak şekilde 2 sınıflı hale getirilmiştir. Bu durumda 51 Uninfected, 135 Infected örnek bulunmaktadır. Aşağıda elde edilen sonuçlar verilmiştir.

	precision	recall	f1-score	support
Uninfected	1.00	0.80	0.89	10
Infected	0.93	1.00	0.97	28
accuracy			0.95	38
macro avg	0.97	0.90	0.93	38
weighted avg	0.95	0.95	0.95	38

Cross validation sonucu aşağıdaki gibidir.

```
0.92 accuracy with a standard deviation of 0.07
```

GridSearch kullanıldığında verilen parametreler aşağıdaki gibidir.

Kernel: linear, rbf

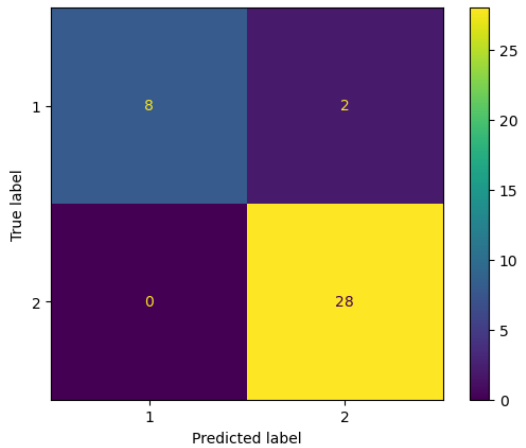
C: 1, 10

Bu parametreler ile elde edilen sonuçlara bakıldığında C değeri 1 ve kernel linear olarak seçilmiştir. Aşağıda modelin elde ettiği sonuçlar verilmiştir.

	precision	recall	f1-score	support
Uninfected	1.00	0.80	0.89	10
Infected	0.93	1.00	0.97	28
accuracy			0.95	38
macro avg	0.97	0.90	0.93	38
weighted avg	0.95	0.95	0.95	38

Görülebileceği üzere çok yüksek precision, recall ve f1-score elde edilmiştir.

Aşağıda ise bu modele ait confusion matrix verilmiştir.



1 -> Uninfected

2 -> Infected

KNN ile Elde Edilen Sonuçlar

Default parametreler ile sınıflandırma yapıldığında aşağıdaki gibi bir sonuç elde edilmiştir.

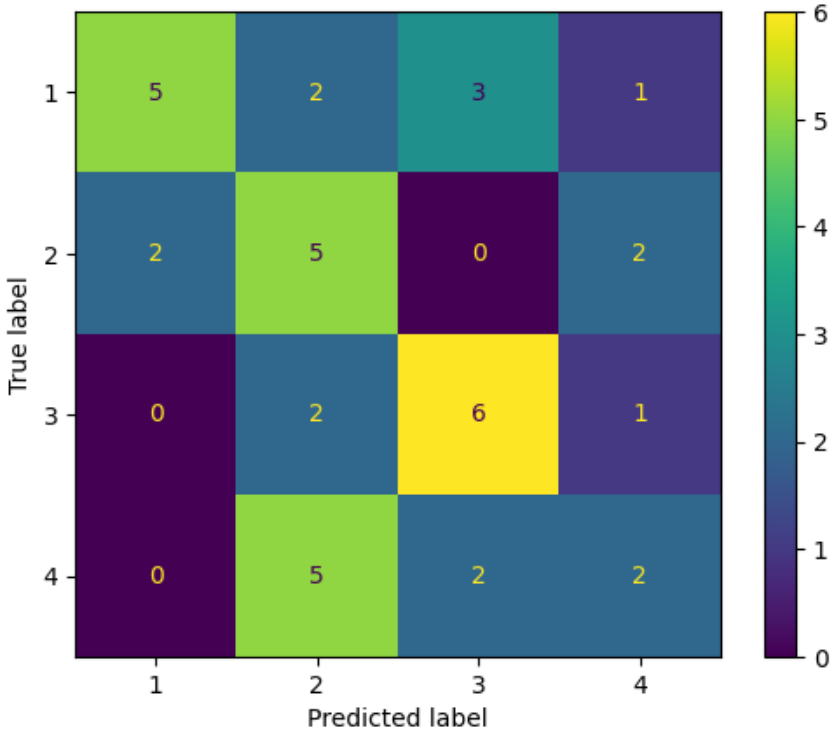
```
Accuracy: 0.47368421052631576
Macro Precision: 0.4875541125541125
Macro Recall: 0.4747474747474748
```

Modelin precision, recall ve f1-score bilgileri aşağıdaki görselde verilmiştir.

	precision	recall	f1-score	support
Uninfected	0.71	0.45	0.56	11
Influenza virus infected	0.36	0.56	0.43	9
Rhino virus infected	0.55	0.67	0.60	9
Influenza & Rhino virus infected	0.33	0.22	0.27	9
accuracy			0.47	38
macro avg	0.49	0.47	0.46	38
weighted avg	0.50	0.47	0.47	38

Modelin performansının, SVM'ye göre çok daha düşük olduğu gözlemlenmiştir.

Bu modele ait confusion matrix aşağıdaki gibidir.



Daha etkili bir sınıflandırma için ayrıca veri seti Uninfected-Infected olacak şekilde 2 sınıflı hale getirilmiştir. Bu durumda 51 Uninfected, 135 Infected örnek bulunmaktadır. Aşağıda elde edilen sonuçlar verilmiştir.

	precision	recall	f1-score	support
Uninfected	0.78	0.70	0.74	10
Infected	0.90	0.93	0.91	28
accuracy			0.87	38
macro avg	0.84	0.81	0.82	38
weighted avg	0.87	0.87	0.87	38

Görülebileceği üzere model performansı 2 sınıf olduğunda ciddi bir biçimde artış göstermiştir.

Cross validation sonucu elde edilen sonuç aşağıda verilmiştir.

```
0.84 accuracy with a standard deviation of 0.07
```

GridSearch kullanıldığında verilen parametreler aşağıdaki gibidir.

n_neighbors: 3, 5, 7

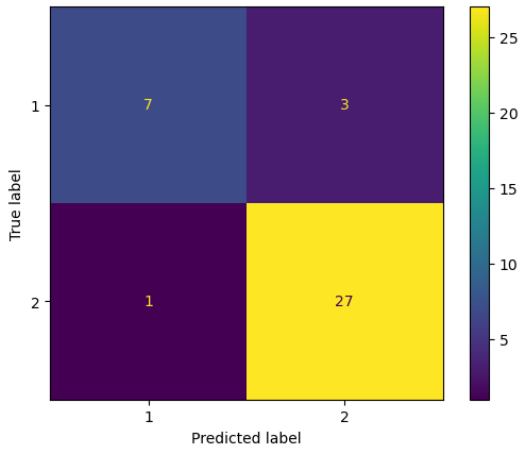
weights: uniform, distance

Bu parametreler ile elde edilen sonuçlara bakıldığında n_neighbors değeri 3 ve weights değeri uniform olarak seçilmiştir.

Aşağıda modelin elde ettiği sonuçlar verilmiştir.

	precision	recall	f1-score	support
Uninfected	0.88	0.70	0.78	10
Infected	0.90	0.96	0.93	28
accuracy			0.89	38
macro avg	0.89	0.83	0.85	38
weighted avg	0.89	0.89	0.89	38

Aşağıda modele ait confusion matrix verilmiştir.



Lojistik Regresyon ile Elde Edilen Sonuçlar

Lojistik regresyon yöntemini kullanabilmek için modelin max_iter değeri 10000 olarak ayarlanmıştır.

Aşağıda elde edilen sonuçlar verilmiştir.

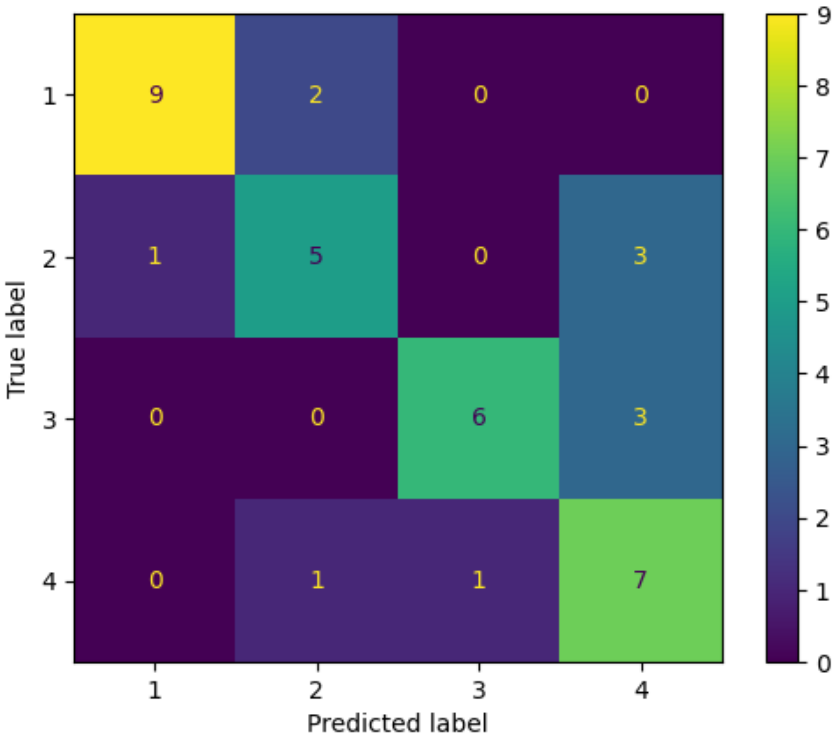
```
Accuracy: 0.7105263157894737
Macro Precision: 0.7301510989010989
Macro Recall: 0.7045454545454545
```

Modelin precision, recall ve f1-score bilgileri aşağıdaki görselde verilmiştir.

	precision	recall	f1-score	support
Uninfected	0.90	0.82	0.86	11
Influenza virus infected	0.62	0.56	0.59	9
Rhino virus infected	0.86	0.67	0.75	9
Influenza & Rhino virus infected	0.54	0.78	0.64	9
accuracy			0.71	38
macro avg	0.73	0.70	0.71	38
weighted avg	0.74	0.71	0.72	38

Modelin performansının KNN'den yüksek SVM'den düşük olduğu gözlemlenmiştir.

Bu modele ait confusion matrix aşağıdaki gibidir.



Daha etkili bir sınıflandırma için ayrıca veri seti Uninfected-Infected olacak şekilde 2 sınıflı hale getirilmiştir. Bu durumda 51 Uninfected, 135 Infected örnek bulunmaktadır. Aşağıda elde edilen sonuçlar verilmiştir.

	precision	recall	f1-score	support
Uninfected	1.00	0.80	0.89	10
Infected	0.93	1.00	0.97	28
accuracy			0.95	38
macro avg	0.97	0.90	0.93	38
weighted avg	0.95	0.95	0.95	38

Görülebileceği üzere model performansı 2 sınıf olduğunda ciddi bir biçimde artış göstermiştir.

Cross validation sonucu elde edilen sonuç aşağıda verilmiştir.

```
0.92 accuracy with a standard deviation of 0.07
```

GridSearch kullanıldığında verilen parametreler aşağıdaki gibidir.

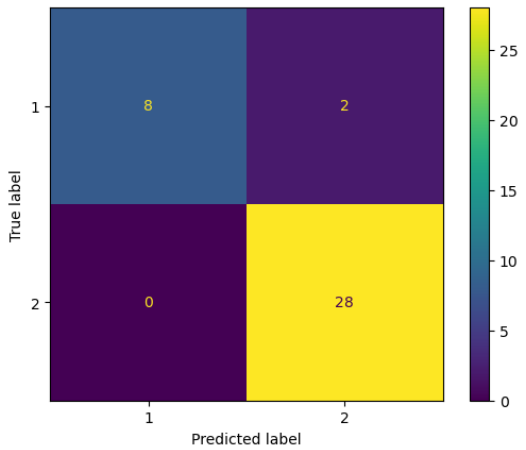
C: 0.1, 1, 10

Bu parametreler ile elde edilen sonuçlara bakıldığında C değeri 1 olarak seçilmiştir.

Aşağıda modelin elde ettiği sonuçlar verilmiştir.

	precision	recall	f1-score	support
Uninfected	1.00	0.80	0.89	10
Infected	0.93	1.00	0.97	28
accuracy			0.95	38
macro avg	0.97	0.90	0.93	38
weighted avg	0.95	0.95	0.95	38

Aşağıda modele ait confusion matrix verilmiştir.



Kaynakça

[1] https://www.memorial.com.tr/hastaliklar/influenza-nedir

[2] https://www.buyukanadoluhastanesi.com/haber/2341/rhinovirus-nedir-nasil-tedavi-edilir