**PIMA INDIANS DIABETES DATASET İÇİN MULTILAYER PERCEPTRON İLE SINIFLANDIRMA**

**1.GİRİŞ**

Bu dökümanda, diyabet rahatsızlığı olan bir hastanın rahatsızlığının tespit edilmesinde, yaş, kan basıncı, vücut-kitle endeksi gibi bazı verilerin, sonuçları nasıl etkilediği sınıflandırılmış ve Weka programı üzerinde tespit edilmiştir.

Kullanacağımız verisetinde 768 kişi ve bu kişilerin nitelikleri verilmiştir bunları kullanarak bir yapay sinir ağları üretip, herhangi bir kişinin diyabet hastası olup olmadığını tahmin etmeye çalışacağız.

**2.MULTI-LAYER PERCEPTRON**

Çok Katmanlı algılayıcıları (MLP) daha iyi anlamak için XOR problemini anlamalıyız. Öncelikle XOR fonksiyonu doğrusal ayrılabilir değil yani Single Perceptron ile bunu yapamayız fakat çok katmanlı olarak ayrılabilir. Denetimli öğrenme(Supervised Training) algoritmasıdır.

Birden çok giriş için birden çok nöron gereklidir. Paralel işlem yapılması istenildiğinde ise birden çok katman devreye girer. Veriler giriş katmanından gelir, ara katmanlardan geçerek sonuç döndürür. Her nöron bir sonraki tüm nöronlara bağlıdır bu da çeşitliliği arttırır.

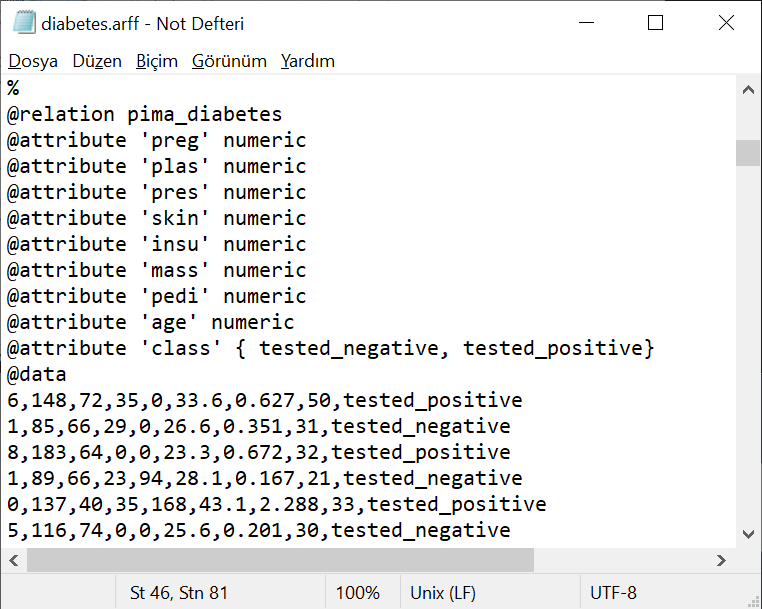
Çok katmanlı algılayıcılar tasarlanırken, probleme göre katman sayısı(Hidden Layer), her bir katmandaki(Layer) perceptron sayısı, Aktivasyon fonksiyonu ( Activation Function (step, sigmoid)), öğrenme oranı (Learning Rate) ve Çevrim sayısı (Iteration) değişebilir.

**3. PIMA INDIANS DIABETES DATASET**

Bu çalışmada Weka programına National Institute of Diabetes and Digestive and Kidney Diseases (Ulusal Diyabet ve Sindirim ve Böbrek Hastalıkları Enstitüsü) tarafından 9 mayıs 1990 tarihinde bağışlanan Diabeter dataset üzerinde örneklendirmeler ve sınıflandırma geliştirmeleri yapılacaktır. Bu örneklerin seçiminde çeşitli kısıtlamalar yapılarak daha büyük bir veritabanandan çıkartılmıştır. Bu verisetindeki tüm hastalar 21 yaşından büyük kadınlardır ve hiçbir değer boş(Missing Attribute Values) değildir.

Veri setine göre, bir kişinin diyabet hastası olmasında etken olabilecek durumlar toplanılarak veri seti haline getirilmiştir. Bu veri setinde 768 kişinin (Number of Instances), 8 farklı niteliği (Number of Attributes) ele alınmıştır. Bunlar,

1. Hamile kalma sayısı (Number of times pregnant)
2. Kandaki glikoz plazma oranı, bir glikoz çözeltisi içtikten sonra (Plasma glucose concentration a 2 hours in an oral glucose tolerance test)
3. Kan basıncı (Diastolic blood pressure (mm Hg) )
4. Deri kalınlığı (Triceps skin fold thickness (mm))
5. İnsülin oranı (2-Hour serum insulin (mu U/ml))
6. Vücut kitle indeksi (Body mass index (weight in kg/(height in m)^2))
7. Diabet soyağacı (Diabetes pedigree function)
8. Yaş (Age (years))
9. Diabet olup olmadığı 0 ise değil 1 ise diyabet (Class variable (0 or 1))

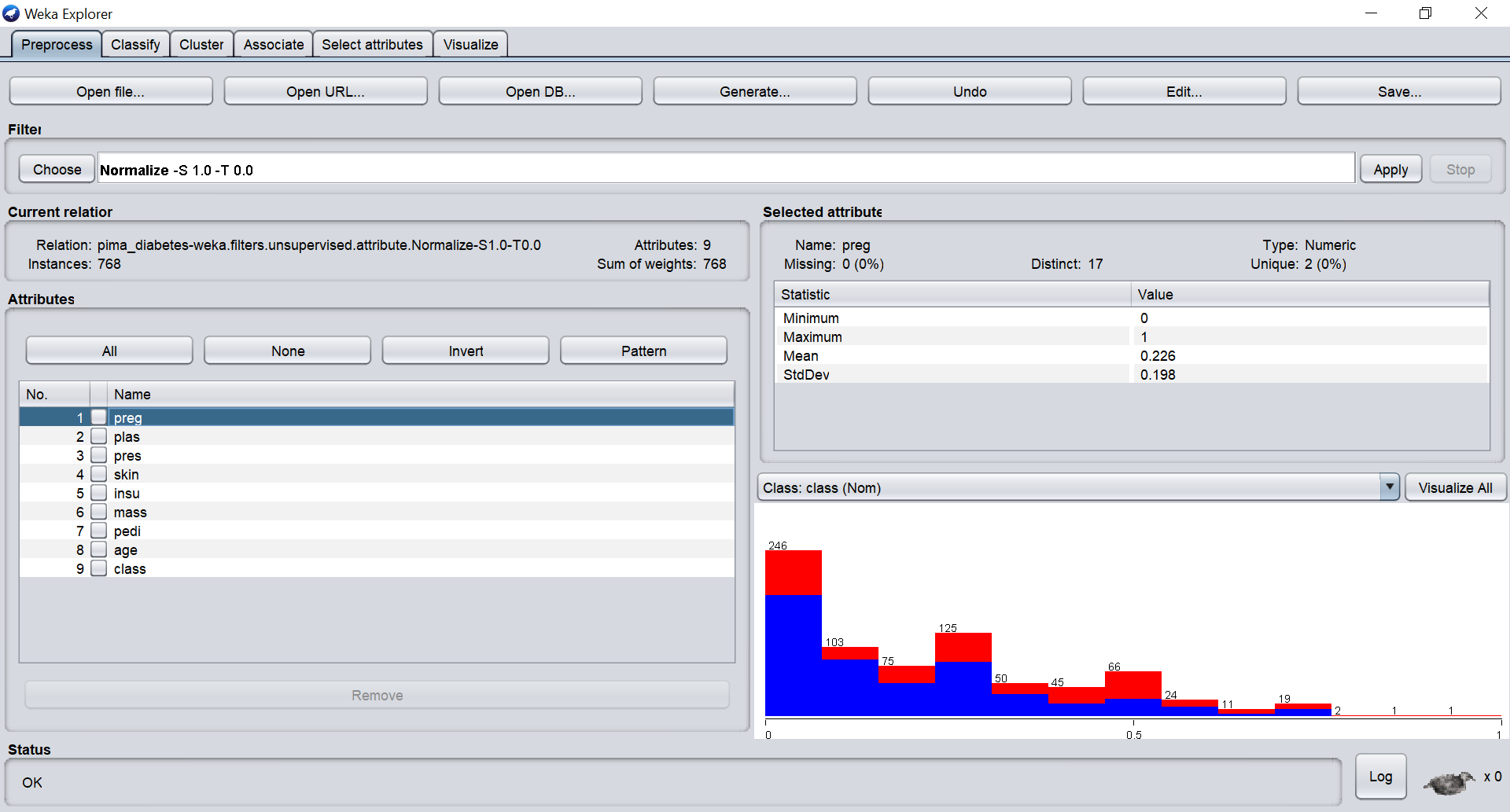


Şekil 1

Şekil 1’ de verisetinin bir bölümü örnek olarak gösterilmiştir.

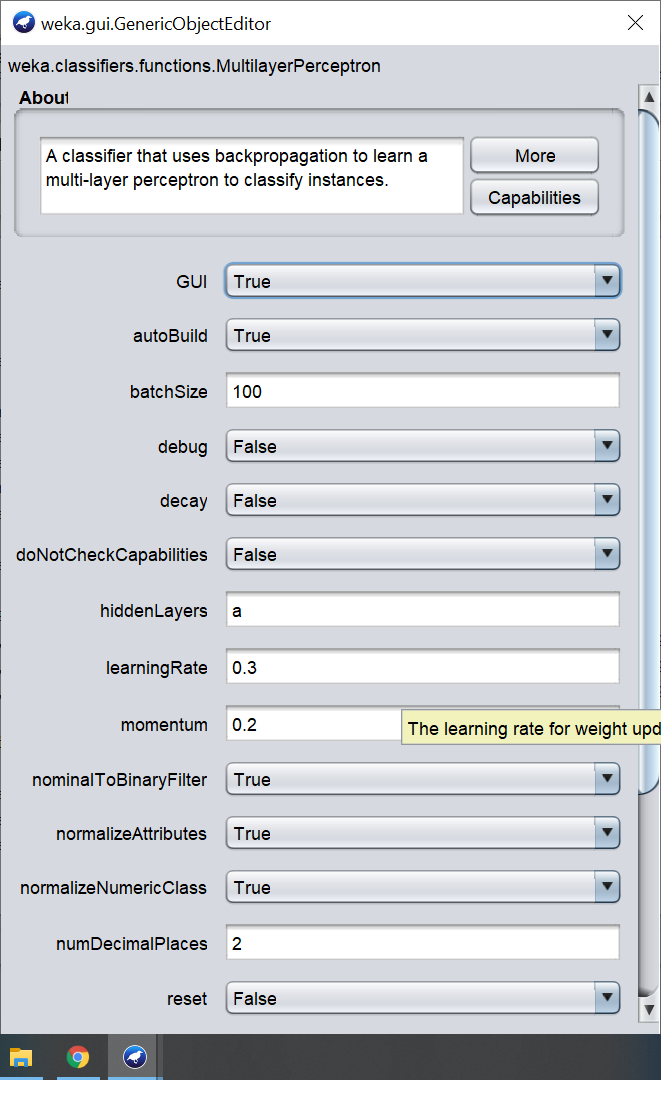
Weka programına geçmeden önce bir veri ile işlem yaparken daha doğru sonuçlar almak için bazı adımları gerçekleştiriyoruz. Bunlar Normalize(Min-Max method- z-score vb), olmayan değerleri ortalama değerler ile değiştirmek (Replace missing value), veri temizleme (Data Clearing) gibi bazı methodlar uygulanır.

Bu çalışmada kayıp değerler olmadığı için sadece normalize yapılacaktır.



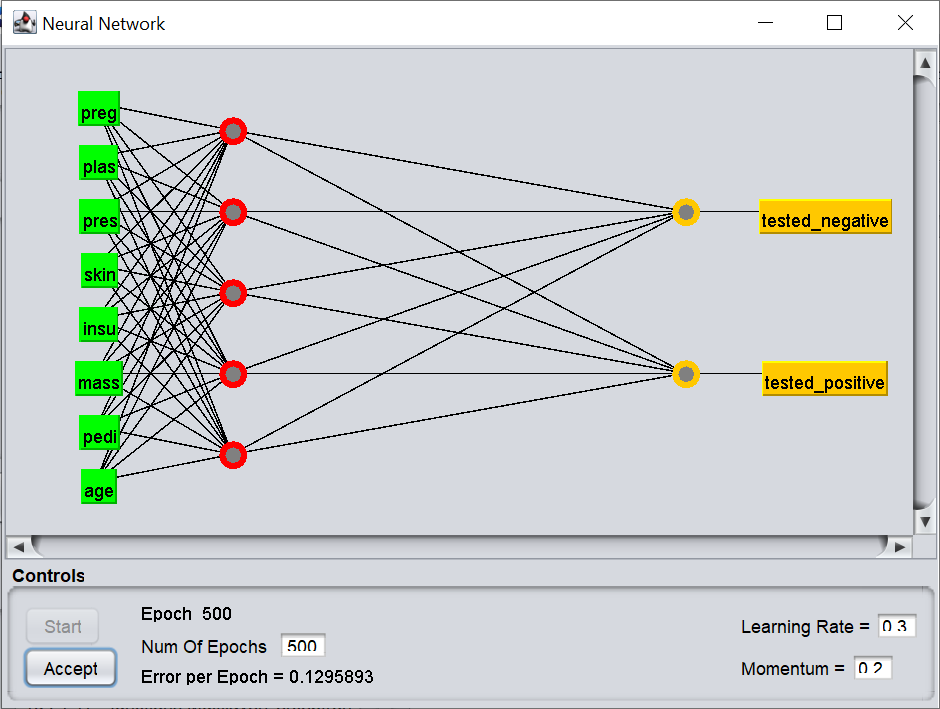
Şekil 2

Şekil 2’ de gösterildiği gibi veriseti weka programına yüklenmiş olup değerler normalize edilmiştir. Görüldüğü üzere ilk attribute preg değeri 1 ile 0 aralığındadır. Şimdi Classify bölümüne geçiyoruz.



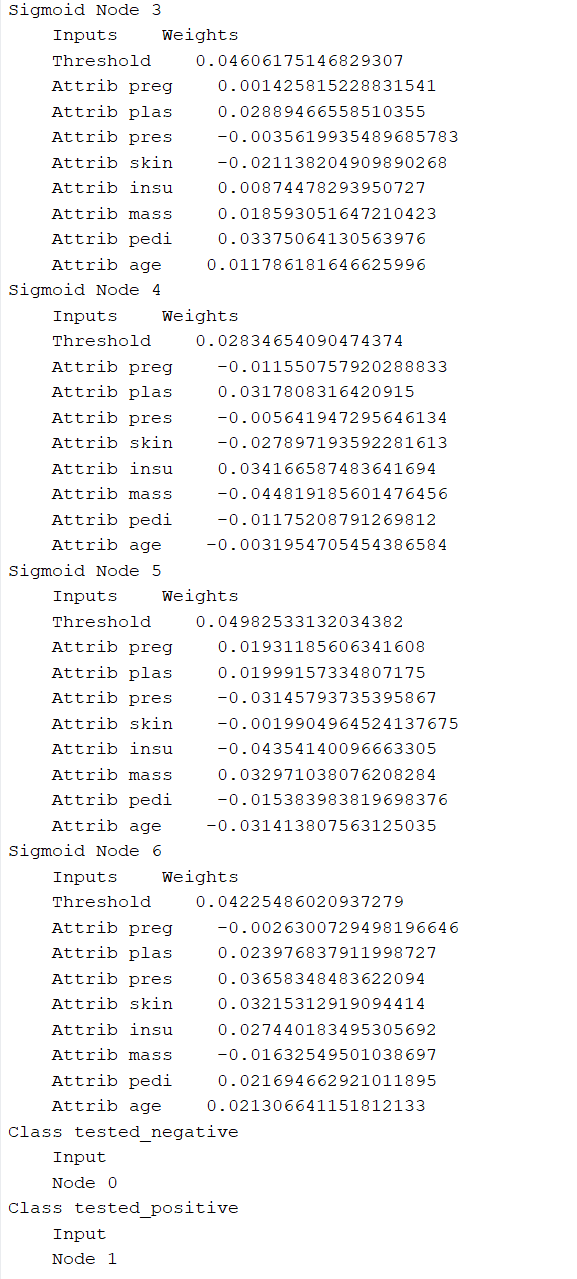
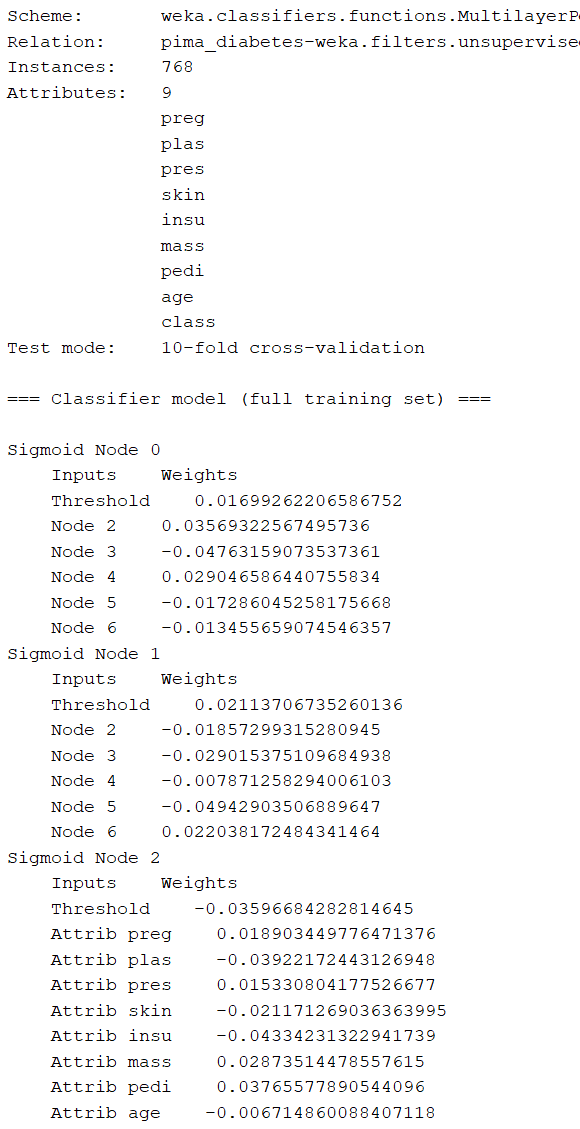
Şekil 3

Şekil 3’ de wekanın MLP için bize sunduğu bazı seçenekler görüyoruz buradan herhangi bir değeri değiştirerek çalışma değiştirilebilir. İlk olarak ilk seçenek olan GUI(Graphical User Interface) seçeneğini true yapıyorum ve Cross- Validation ile fold sayısını 10 vererek başlatıyorum.



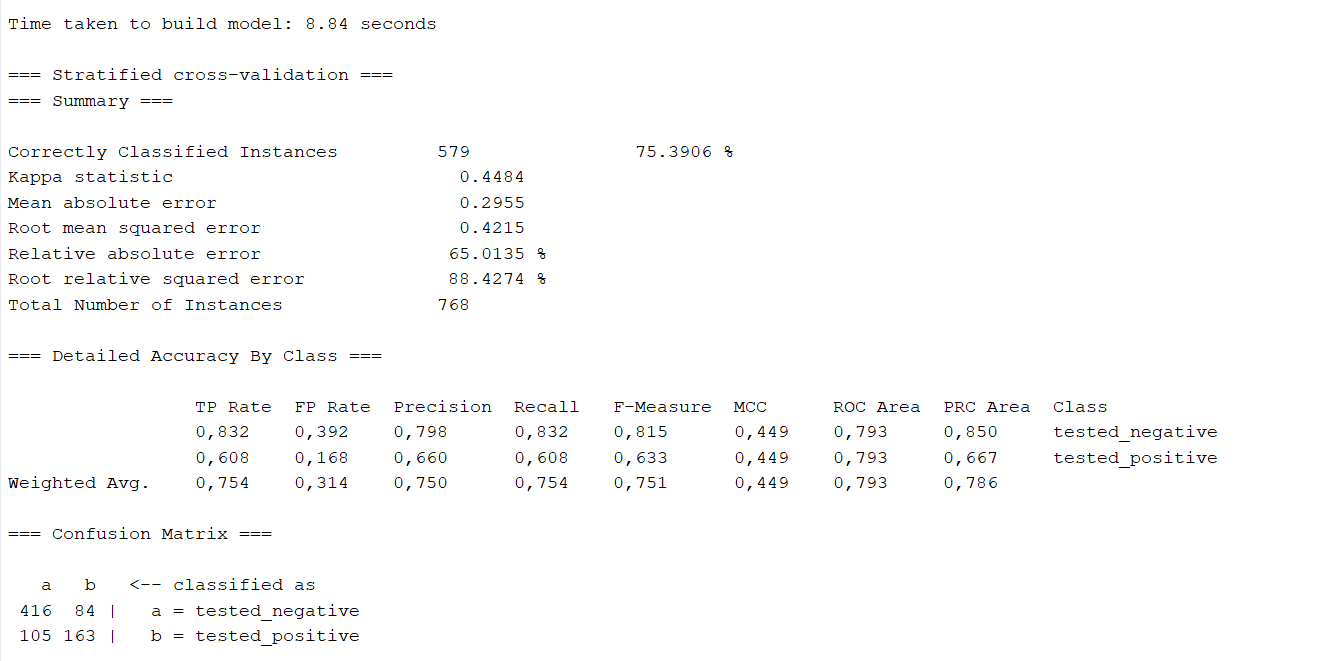
Şekil 4

Şekil 4’de gördüğümüz üzere yapay sinir ağı oluşturmuş olduk. Ekranda gördüğümüz Epoch 500 (num of Epochs) çevrim sayısıdır, Error per Epoch (0.1295893) ise hata sayısıdır. Learning rate’i default olarak 0.3 verilmiştir.



Şekil 5

Şekil 5’ de her node’a giden değerleri göstermiştir sonuç olarak tested\_negative yada tested\_positive vermektedir. Bu değerlerin bu şekilde görünmesinin sebebi en başta normalize yapmamızdır yani 0 ile 1 aralığında değer verdik.

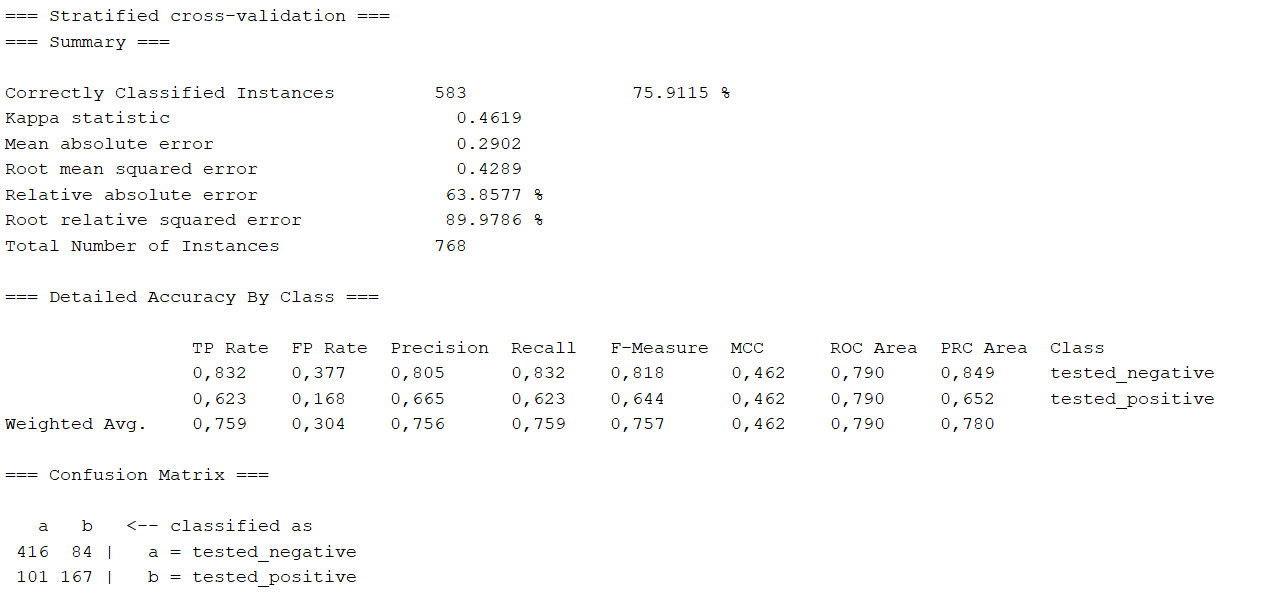


Şekil 6

Şekil 6’da gördüğümüz üzere doğru sınıflandırılan örnekler (Correctly Classified Instances) sayısı 579 ve oranı ise %75.3906 (768/579)’ dur. F1 score değeri, precision, recall değerleri ve karmaşıklık matrisi şekilde gösterildiği gibidir.

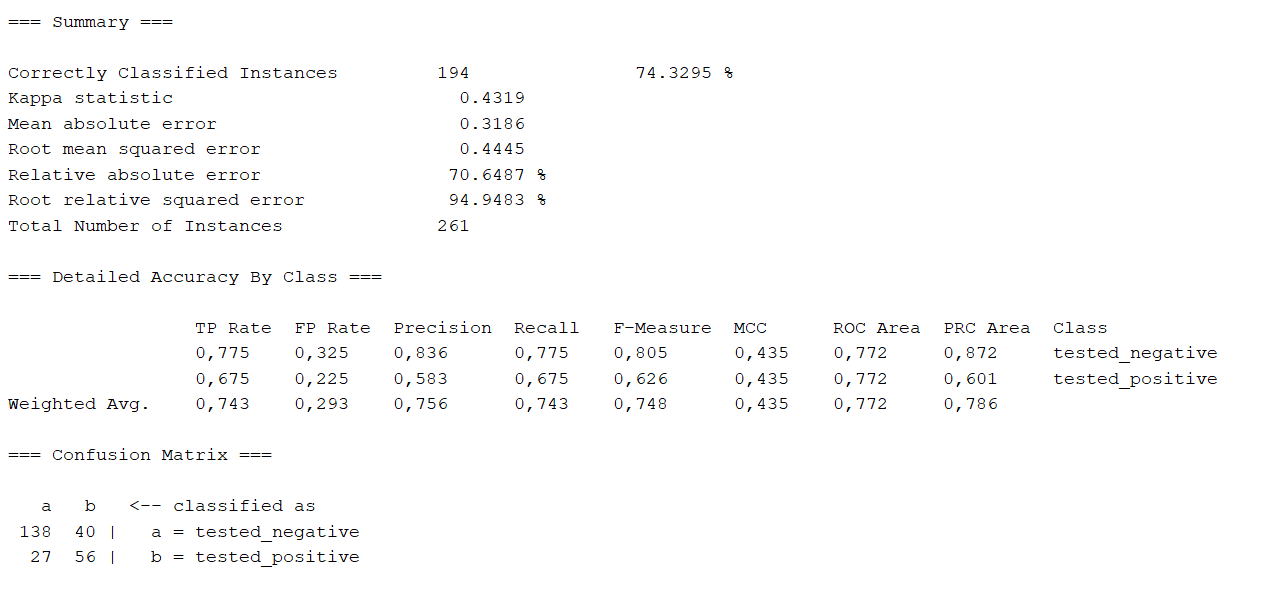
Karmaşılık matrisi ise a ve b için 4 farklı değer barındırır. Bunlar True-Negative(TN), True-Positive(TP), False-Negative(FN), False-Positive(FP) değerlerini ifade eder. Programın bize verdiği değerlerden tested\_nagitive TP Rate(True-Positive) değeri 0.832 olmuş yani 500 tane sağlıklı kişinin 416 tanesine negative demiş ve 84 tane kişiye de positive demiştir. Diğer bir değer tested\_positive TP Rate değerlerine bakarsak, gerçekten hasta olanların 268 kişinin 163 kişisini doğru bilmiş yani 0.608 oranında hasta demiştir. Öğrenme süresi 8.84 saniyedir.

Precision değeri TP/TP+FN olarak, Recall değeri TP/TP+FP olarak, F-Measure değeri 2/(1/p+1/R) olarak hesaplanmıştır. Mean Absolute Error (MAE) ise tüm hataların ortalamasıdır ve Root Mean Squared error (RMSE) ise hataların karesinin, ortalamasının kökü ile bulunur.



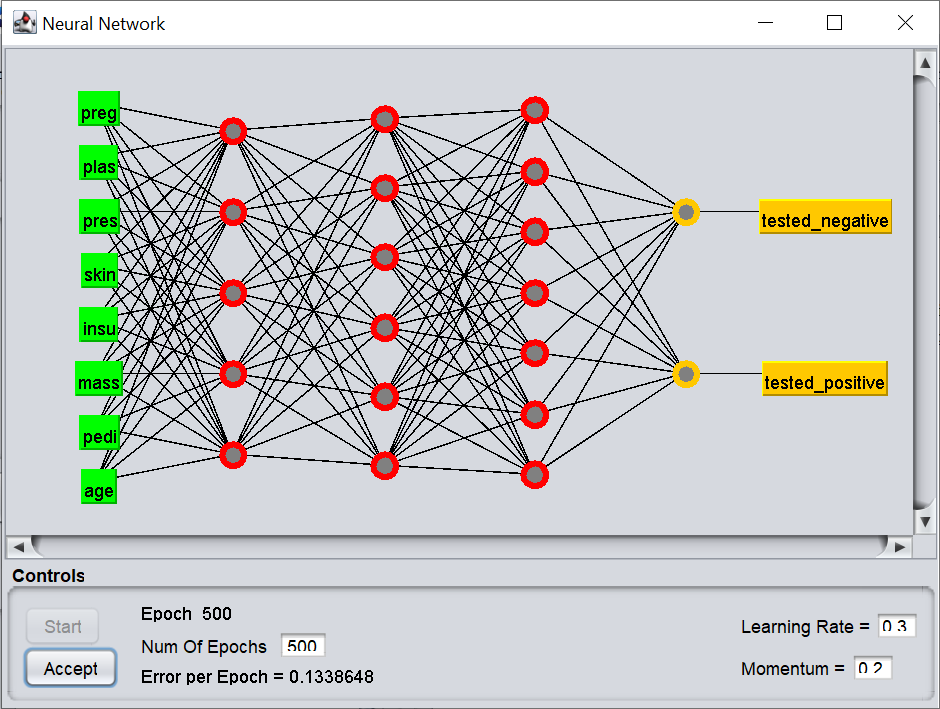
Şekil 7

Şekil 7’de Çevrim sayısı (Training Time) 1000 ve Öğrenme hızı (Learning Rate) 0.5 olarak değiştirilmiş ve hidden layer sayısı, perceptron sayısı ve cross validation fold 10 aynı kalmıştır. Sonuç olarak çok küçük bir artış sağlanmış %0.5 iyi tahminde bulunmuştur. Ve öğrenme süresi kayda değer düşerek 4.13 saniye olmuştur.



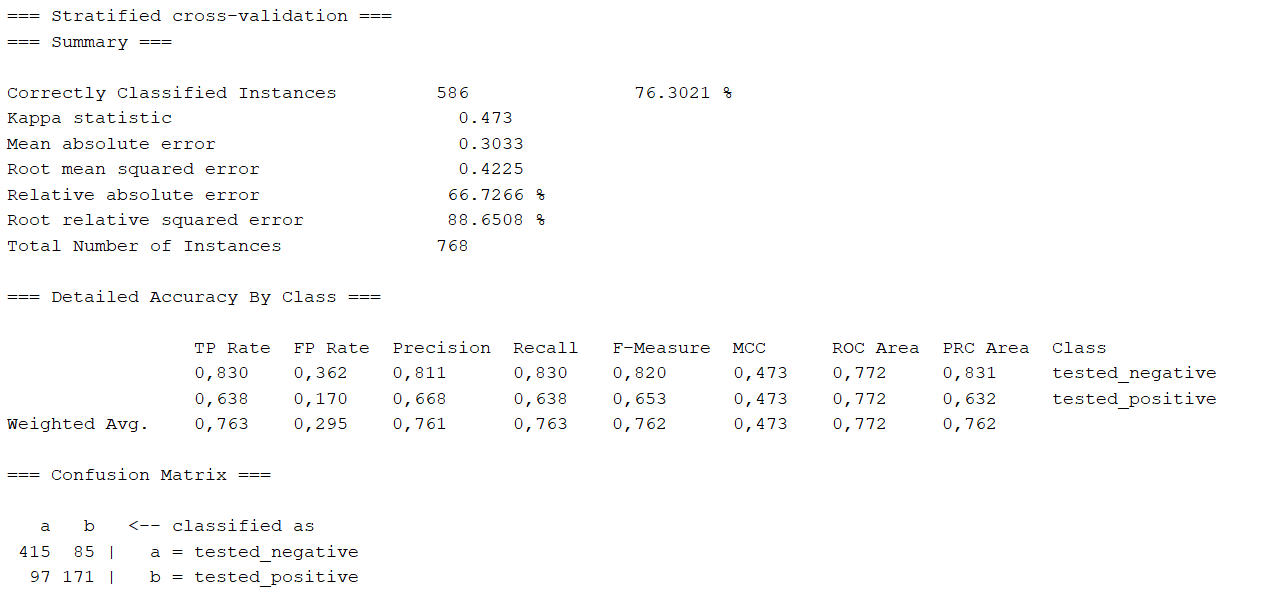
Şekil 8

Şekil 8’de görünen sonuçlar default ayarlar ile percentage split algoritması kullanılarak alınmıştır. Percentage Split verilen oranda kümeyi ikiye böler ve ilk kümeyi eğitim ikinci kümeyi ise test amaçlı kullanır %66 bölümün anlamı, kümenin ilk %66 kısmının eğitim sonraki %34 kısmının ise test amaçlı kullanılacağıdır. Sonuç olarak doğru bilme yüzdesi düşmüştür. Öğrenme süresi 7.5 saniyedir.



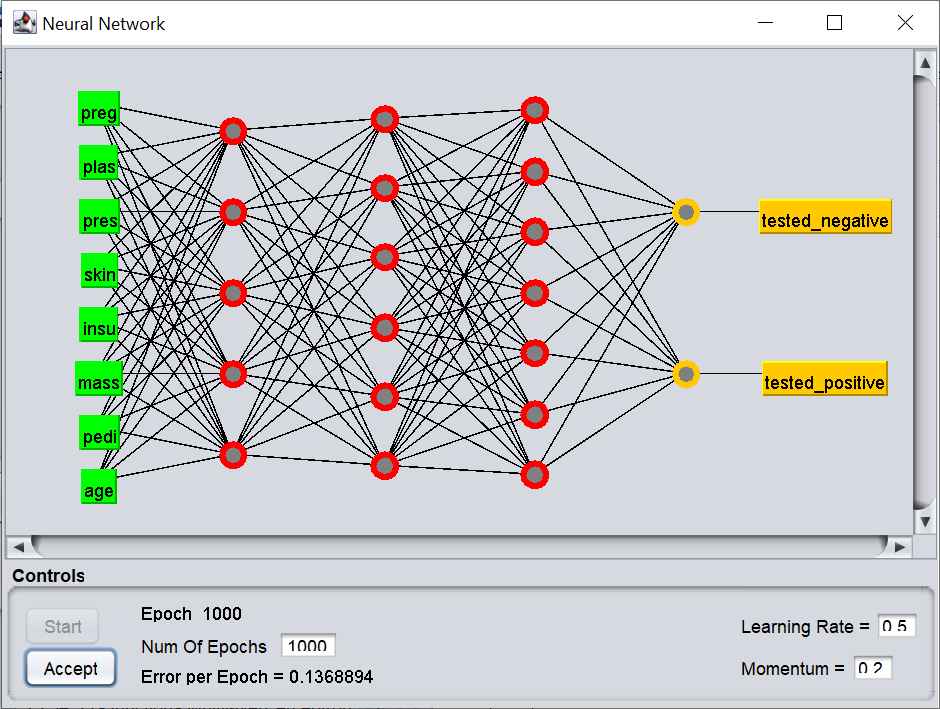
Şekil 9

Şekil 9’da her değer default iken Hidden layer sayısı ve her layerdaki perceptron sayısı elle 5,6,7 girilerek 3 katmandan oluşturulmuş ve oluşturma süresi ilk YSA (ANN)’lere göre çok daha uzun sürerek 34.88 saniye çıkmıştır.



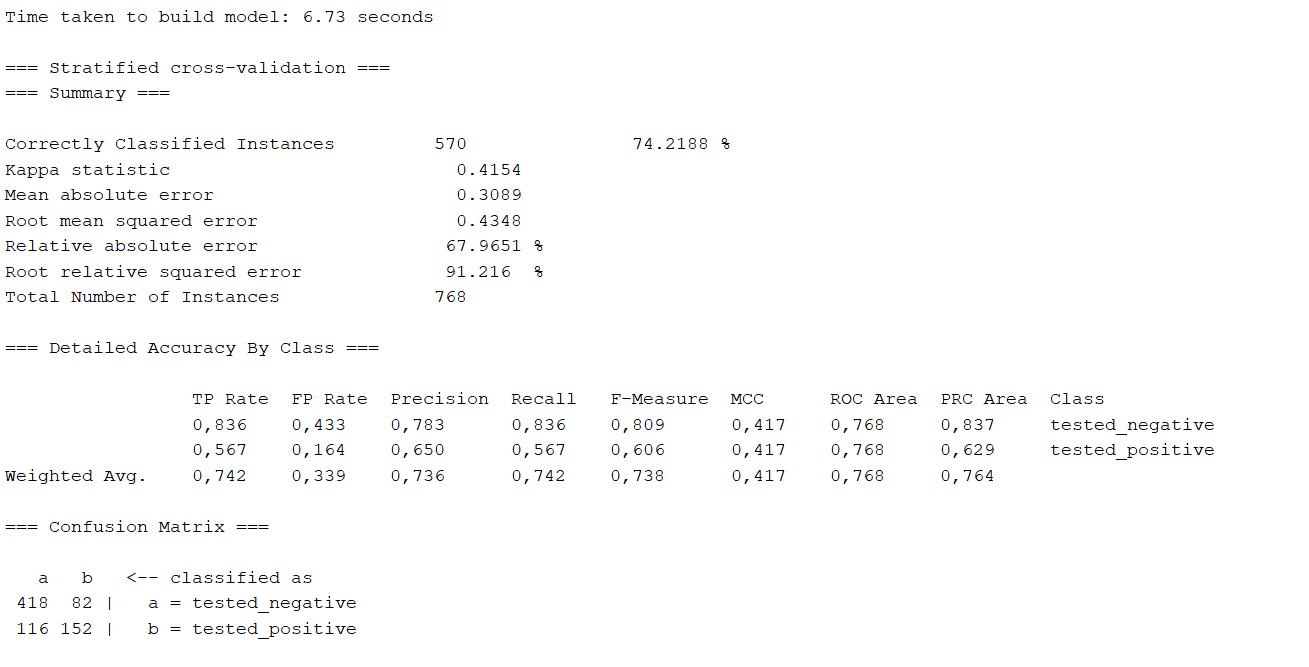
Şekil 10

Şekil 10’da üç katmanlı ysa ile şuana kadar aldığımız en iyi sonuçları almış olduk %76.3021 doğru tahmin oranı ve 586 kişi doğru bilmiştir. Burada dikkan çeken kısım Karmaşıklık matrisinde a sutununda TP Rate değeri neredeyse aynı kalmışken b sutunundaki TN Rate değeri kaydadeğer artmıştır.



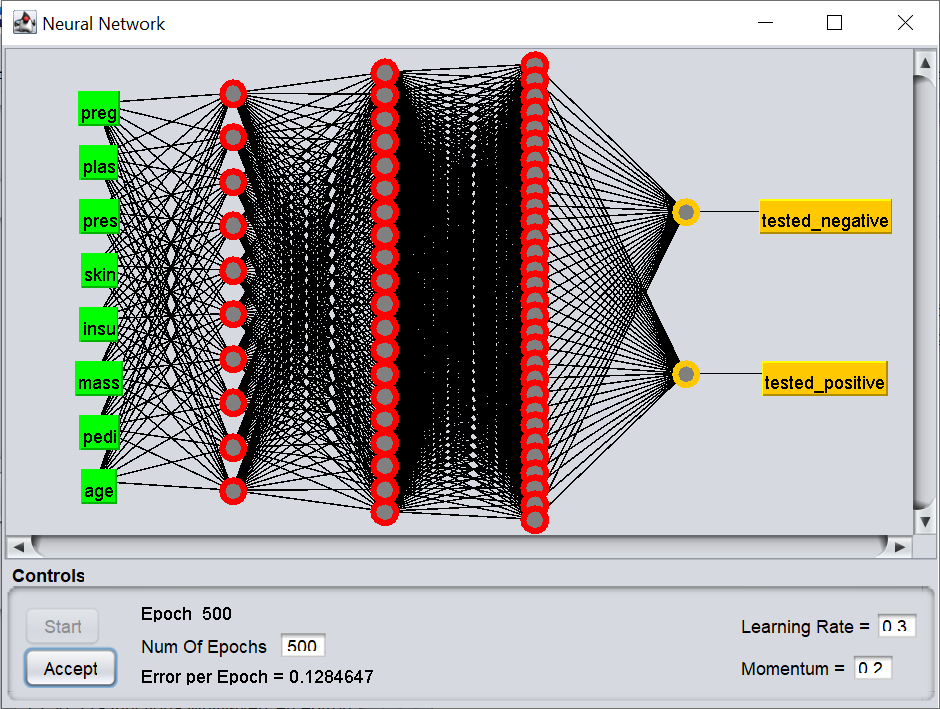
Şekil 11

Şekil 11’de hem katman sayısı hem de daha önce yaptığımız ve daha iyi sonuçlar aldığımız Learning Rate 0.5 , Epoch 1000 olarak alındı ve işlem yapıldı.



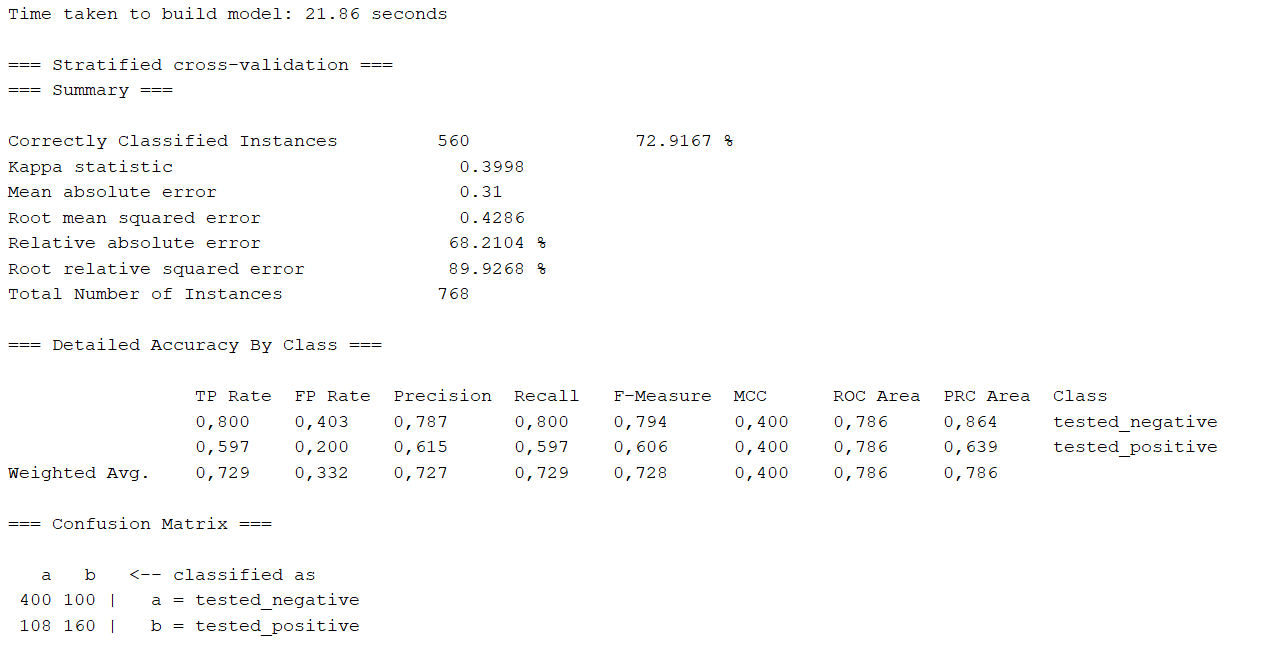
Şekil 12

Şekil 12’de görüldüğü üzere daha önce yaptığım 2 farklı ysa yolunu birleştirdiğimde aldığım sonuç daha düşük olmuştur tahmin oranı düşmüştür.



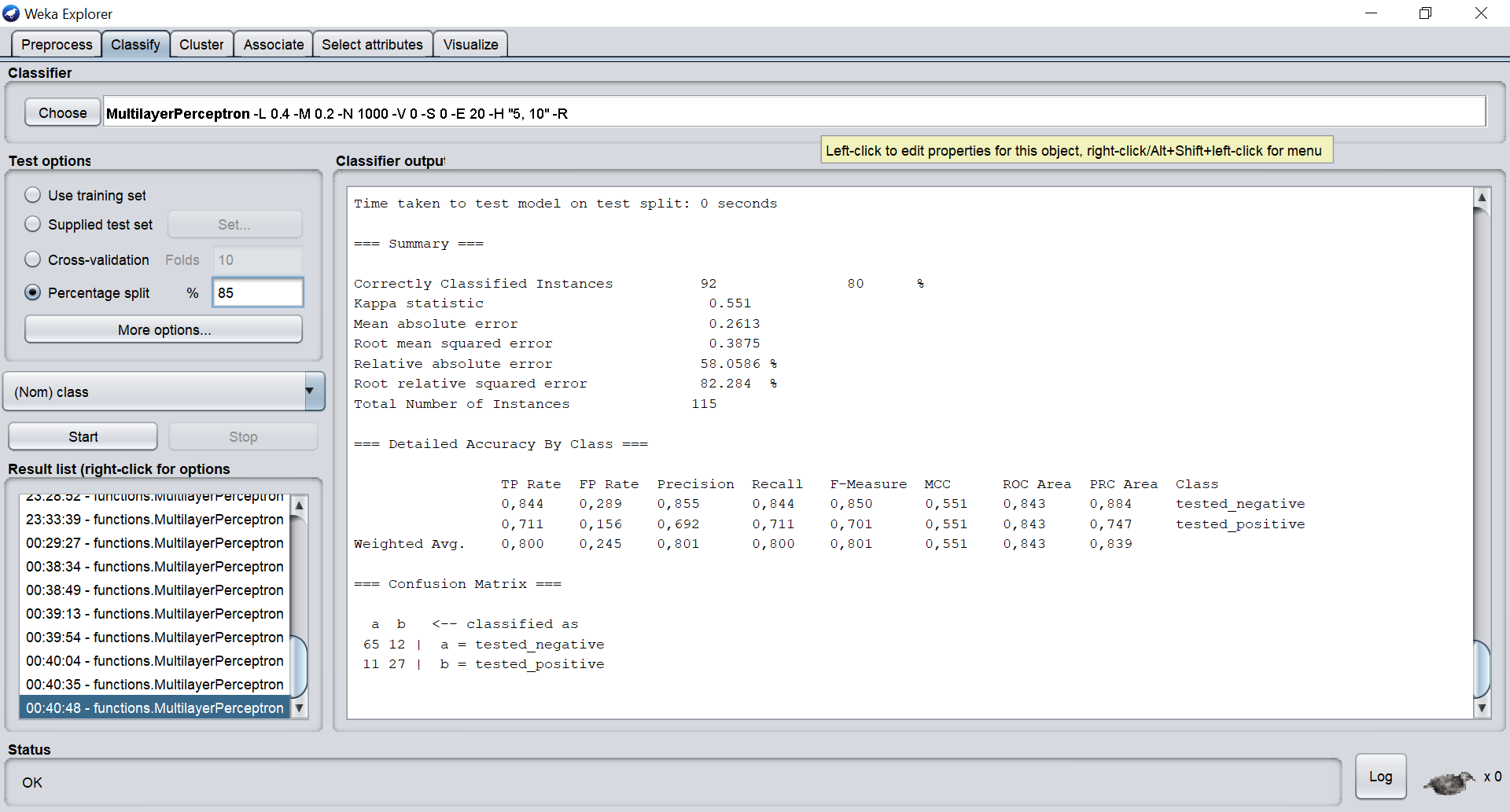
Şekil 13

Şekil 13’de görüldüğü üzere katman sayısı 3 ve sırasıyla 10,20,30 perceptron dan oluşmaktadır diğer değerler farkı anlamak adına default bırakılmıştır.



Şekil 14

Şekil 14’de görüldüğü üzere sadece katman ve perceptron sayısını arttırmak ile daha doğru sonuçlar elde edilemiyor. Her biri denenerek en mantıklı çözüm ortaya çıkabilir. Doğru sınıflama oranı %72.9167’ye düşmüş oldu ve önceki çalışmalar daha yararlı bulunabilir.



Şekil

Şekil 15’de görüldüğü üzere Learning rate 0.4 , TrainingTime 1000, hidden layer katmanı 2 ve sırayla 5, 10 perceptron içermektedir. Percentage split %85 verilerek aldığım en iyi sonuca ulaştım Doğru sınıflandırma oranı %80 çıkmıştır.

**4.KAYNAKLAR**

1. <https://en.wikipedia.org/wiki/Multilayer_perceptron>
2. <https://en.wikipedia.org/wiki/Artificial_neural_network>
3. <http://bilgisayarkavramlari.sadievrenseker.com/2011/09/19/weka-ile-svm/>
4. <https://www.youtube.com/watch?v=Bo3i9cyZ5U4>
5. <https://medium.com/@isikhanelif/multi-layer-perceptron-mlp-nedir-4758285a7f15>
6. <https://kod5.org/yapay-sinir-aglari-ysa-nedir/>
7. <https://ceaksan.com/tr/weka-nedir/>