**GRUP 2 PROJE RAPORU**

**IAN 512**

**ARABA KAZALARINDA YAŞANAN YARALANMA SEVİYESİNİN TAHMİNİ**

**(KNIME)**

**MERT AÇIKGÖZ**

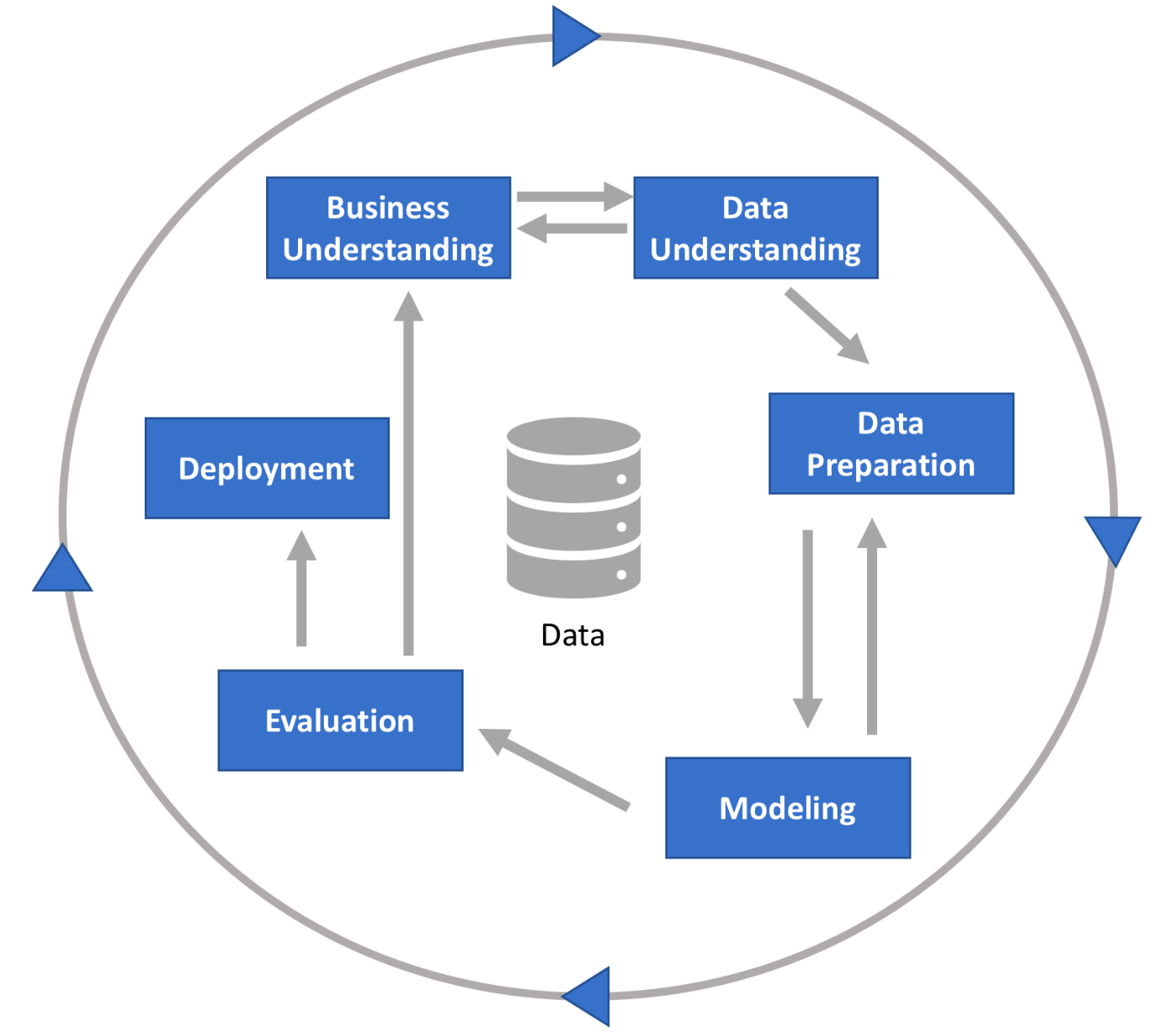
**SEMRA ÇAĞAN**

**BEKİR EREN**

**ŞEYMA KAPICIOĞLU**

Araba kazalarında yaşanan yaralanma seviyelerini **accident.sas7bdat, person.sas7bdat, vehicle.sas7bdat** verilerini kullanarak knime’da tahminlemeye çalıştık. Bu çalışmada CRISP-DM metodolojisini kullandık.

**CRISP-DM**

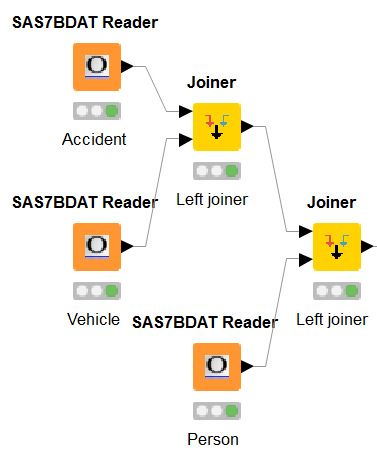


**1-) Businnes Understanding**

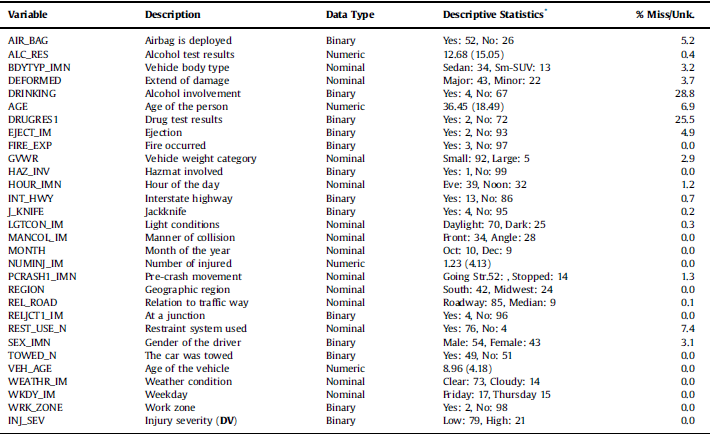
Projede CRISP-DM metodolojisini baz alıp, elimizdeki verileri kullanarak, kazalardaki yaralanma seviyelerini tahmin etmek istedik. Bunu yaparken amaç, elimizdeki verilerdeki değişkenleri etkin bir şekilde kullanarak doğru bir model oluşturabilmek.

**2-Data Understanding**

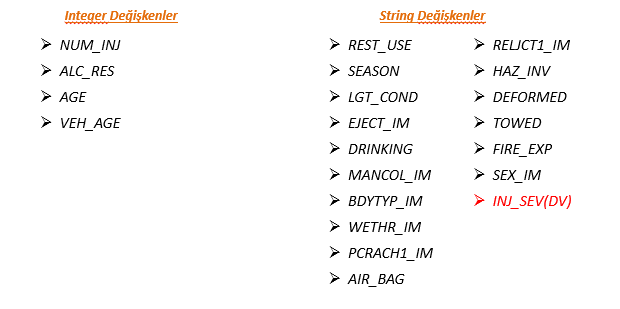
Bu kısımda **accident.sas7bdat, person.sas7bdat, vehicle.sas7bdat** sas formatındaki verileri joiner noduyla birleştirdik.



Çalışmamızda kulandığımız değişkenlerin listesi aşağıdaki gibidir. Elimizdeki verilerde bulunan değişkenlerin bir kısmını, modele daha az etki edeceğini düşünerek ve modeli biraz daha basitleştirebilmek adına dışarda bıraktık. Modele daha fazla etki edeceğini düşündüğümüz değişkenleri ise modele dahil ettik.

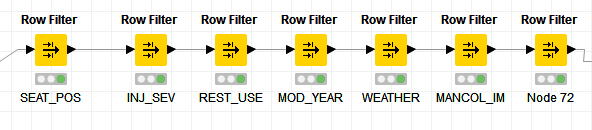


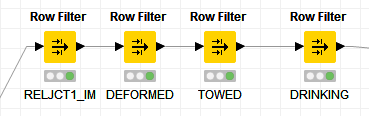
Modele dahil ettiğimiz **21** adet değişken bulunuyor, bu değişkenlerin **4** tanesi **Integer, 17** tanesi ise **string** formatındadır.



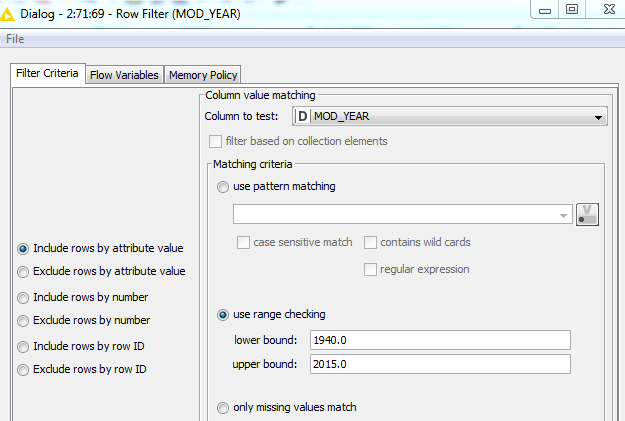
**3-) Data Preparation**

Değişkenlerden bazılarını **row filter** kullanarak grupladık.

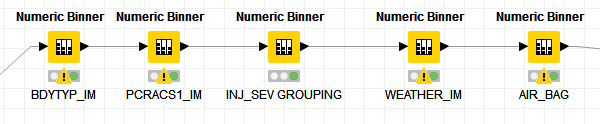




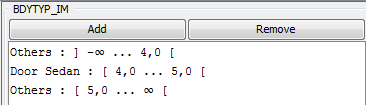
Row Filter kullanılırken, M**OD\_YEAR** değişkeninde 1940-2015 arası değerleri aldık.

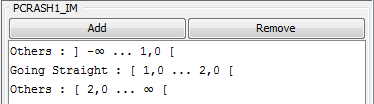


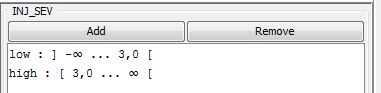
Bazı değişkenleri istediğimiz sayısal değerlerle gruplayabilmek için **numeric binner** kullandık.

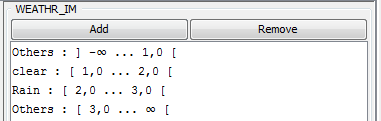


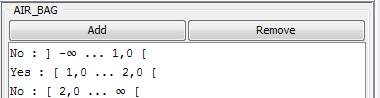
**BDYTY\_IM** değişkeninde **Door** ve **Sedan** kasa tipli araçların karıştığı kazalardaki yaralanmaların daha fazla olduğunu fark ettik ve bu sebepten aşağıdaki gibi bir gruplama yaptık. Kullandığımız diğer değişkenlerin gruplaması da aşağıda gösterilecektir.



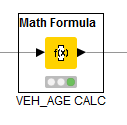
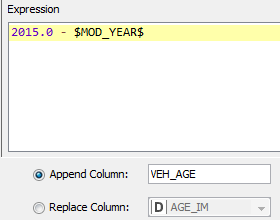




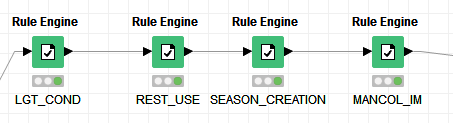


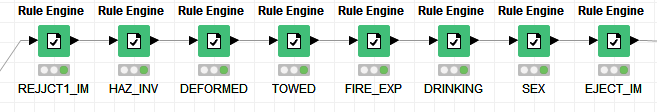


Araç yaşlarının kazadaki yaralanmaya etkisinin oldukça fazla olduğunu düşündüğümüzden ve verimizde bu değişken olmadığından **VEHICLE\_AGE** değişkenini ekledik. Bu değişkeni **math formula** kullanarak oluşturduk. VEHICLE\_AGE ’i “**2015 – MOD\_YEAR”**  formülünü kullanarak hesapladık.

Kullanılan değişkenlerin bir kısmını **rule engine** ile grupladık.





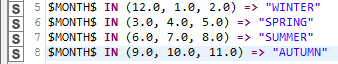
**LGT\_COND** değişkenini **light** ve **dark** olarak grupladık.



**REST\_USE** değişkenini **YES** ve **NO** olarak grupladık.



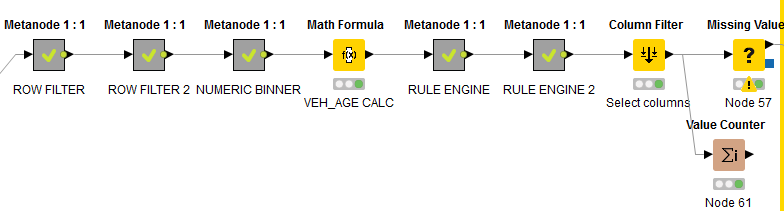
**SEASON** değişkeninde, yılın aylarını mevsimlere göre; **AUTUMN**, **WINTER, SPRING** ve **SUMMER** olarak grupladık



**MANCOL\_IM** değişkenini **front** ve **angle** olarak grupladık



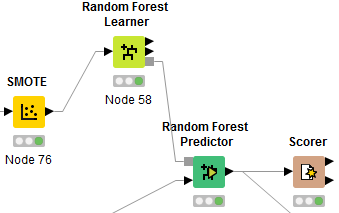
**Data Preparation** kısmını, aşağıda gösterildiği gibi tamamlamş bulunmaktayız.



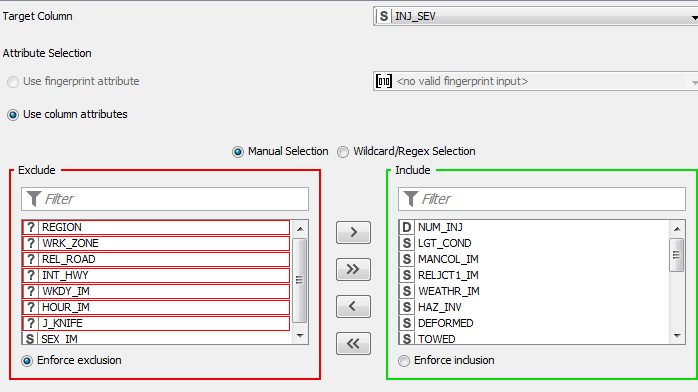
**4-) Modeling**

Bu bölümde dört ayrı model kuruldu, bunlar; **Random Forest, Gradient Boosted, Naive Bayes ve Artificial Neural Networks**

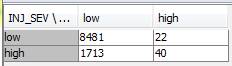
**a-) Random Forest**



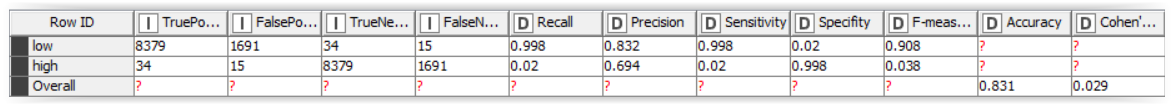
**Random Forest Learner’** da bazı değişkenleri, modele etkisinin daha az olacagını düşünerek model dışında bıraktık. Örneğin **J\_KNIFE** pozisyonunda yapılan kazalarda yaralanmanın cok az oldugunu gördük, bu sebeple, değişkeni model dışında bıraktık.



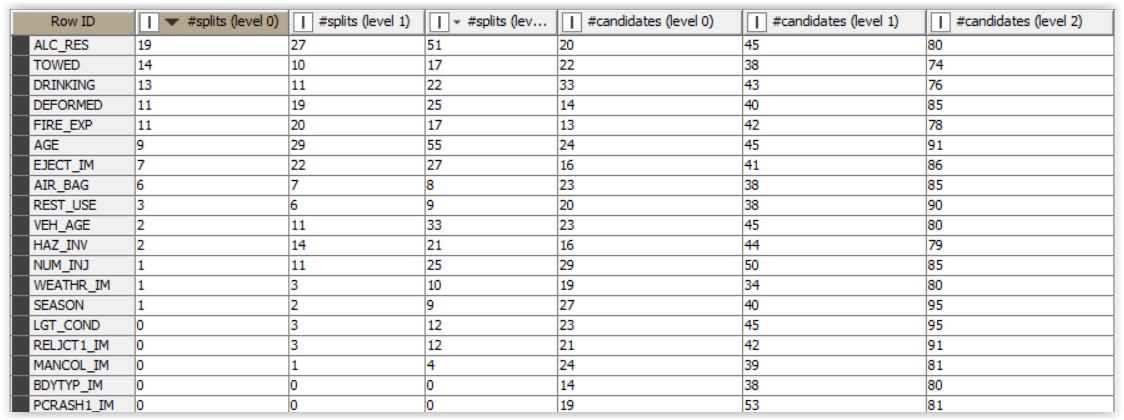
Kullandığımız bu modeldeki **confusion matrix** aşağıdaki gibi oluştu.



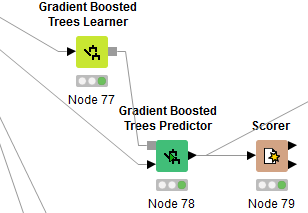
Kullandığımız bu modeldeki **accuracy** değerleri aşağıdaki gibi oluştu.



Bu model için; değişkenlerin model oluşumundaki önemleri ve ağaç oluşturulurken değişkenlerin hangi seviyelerde ve hangi sıklıkla kullanıldıklarını gösteren istatistik verileri aşağıdaki gibi oluştu. Görüleceği üzere, ilk seviye için en sık seçilen değişkenler: alkol test sonucu, araç çekildi mi, hasarın boyutu gibi yaralanma seviyesine doğrudan etki edecek olanlar.



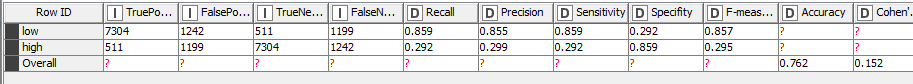
**b-) Gradient Boosted**



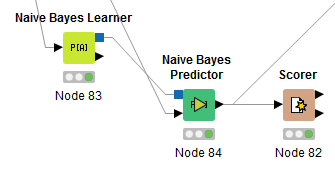
Kullandığımız bu modeldeki **confusion marix** aşağıdaki gibi oluştu.



Kullandığımız bu modeldeki **accuracy** değerleri aşağıdaki gibi oluştu.



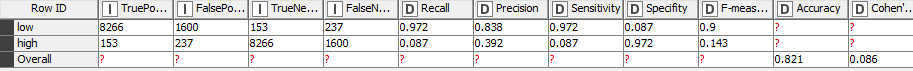
**c-) Naive Bayes**



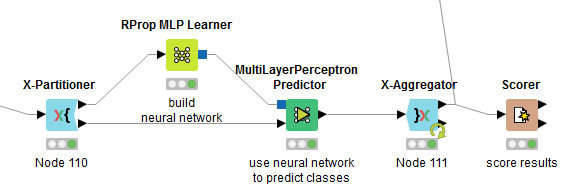
Kullandığımız bu modeldeki **confusion marix** aşağıdaki gibi oluştu.



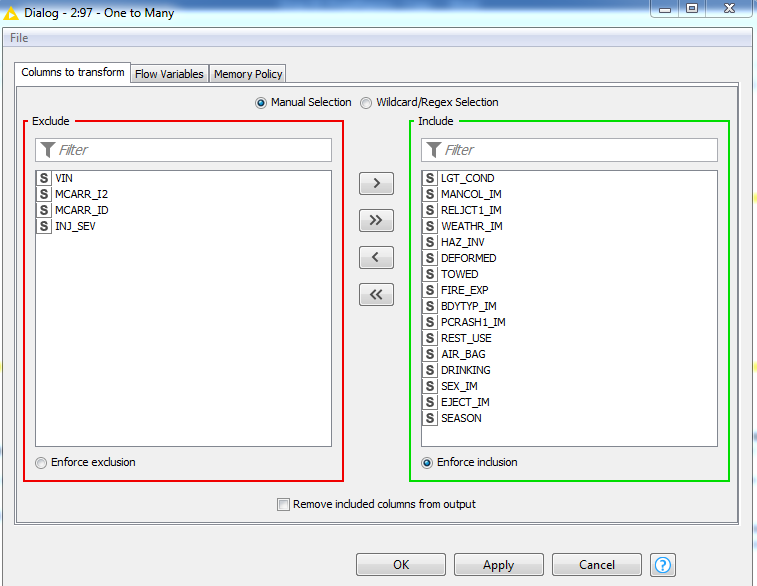
Kullandığımız bu modeldeki **accuracy** değerleri aşağıdaki gibi oluştu.

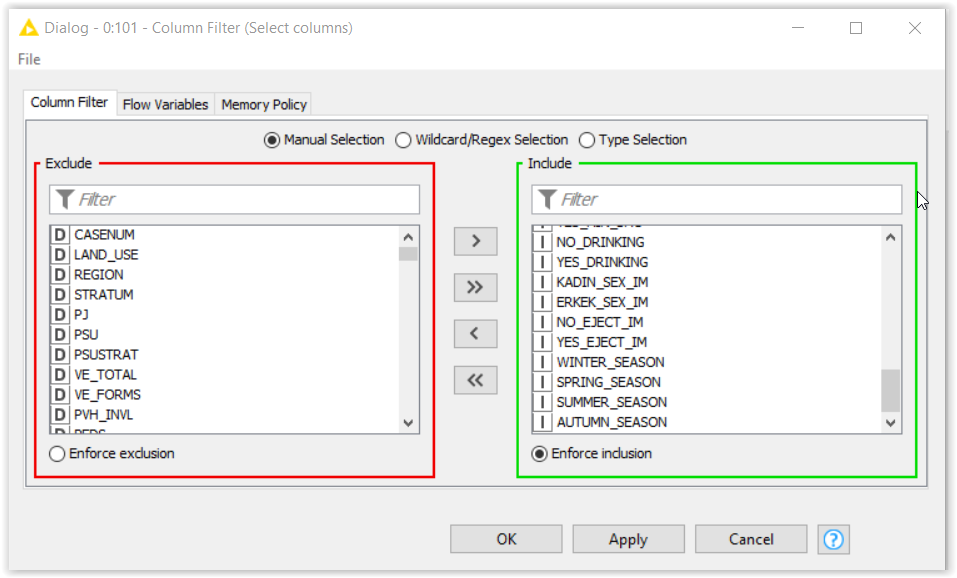


**d-) Artificial Neural Networks**

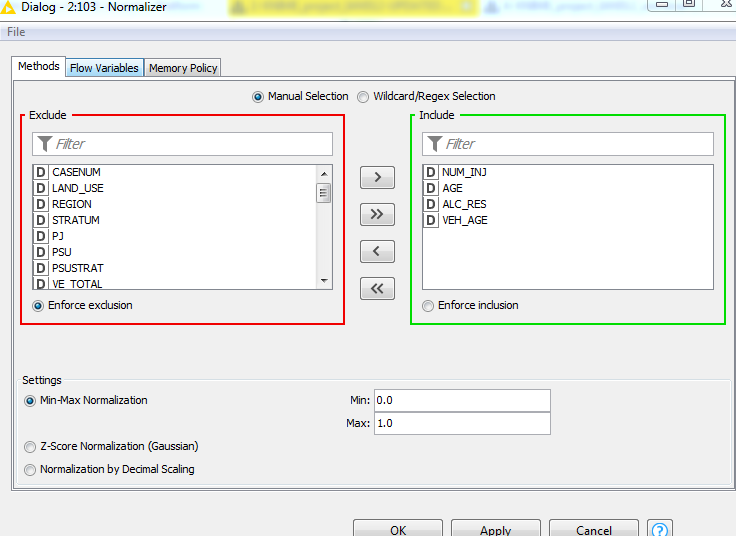


Oluşturduğumuz yapay sinir ağları modelinde, diğer modellerde kullandığımız string değerleri kullanamadığımız için buradaki değerleri numeric’te bıraktık, ayrıca **one to many** ile modelimize birçok yeni değişken eklemiş olduk. Sonrasında column filter işlemi ile, oluşan bu yeni değişkenlerin modele dahil edilmesi sağlandı. Ekran görüntüsü aşağıda bulunuyor.

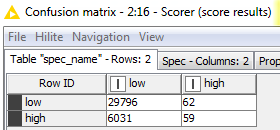




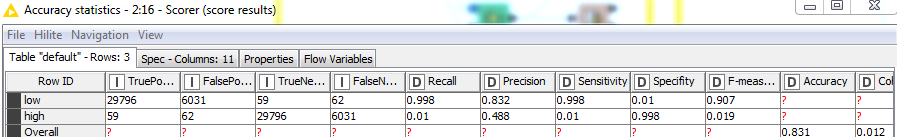
**Normalize** ile **NUM\_INJ, AGE, ALC\_RES, VEH AGE** değişkenleri **0** ile **1** arasında normalize edilmiştir.



Kullandığımız bu modeldeki **confusion marix** aşağıdaki gibi oluştu.



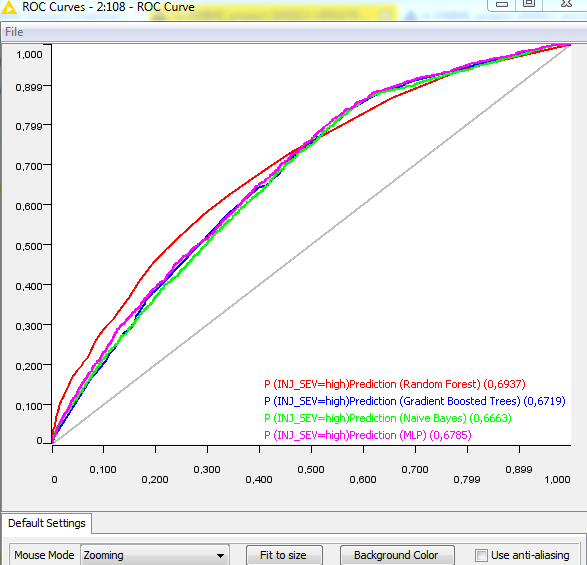
Kullandığımız bu modeldeki **accuracy** değerleri aşağıdaki gibi oluştu.



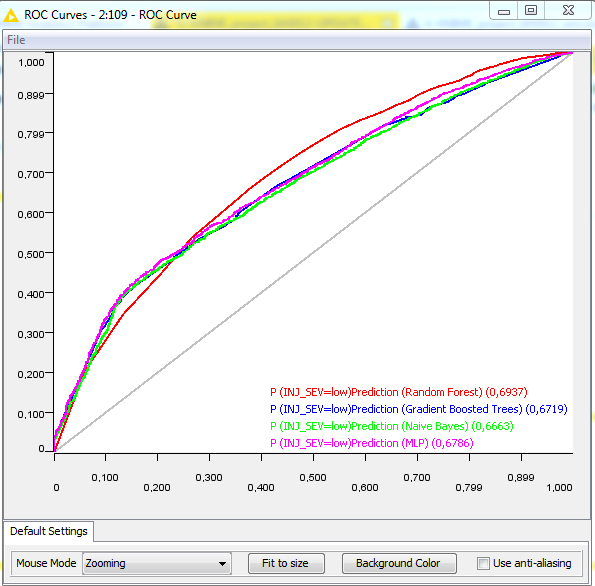
**5 -) Evaluation**

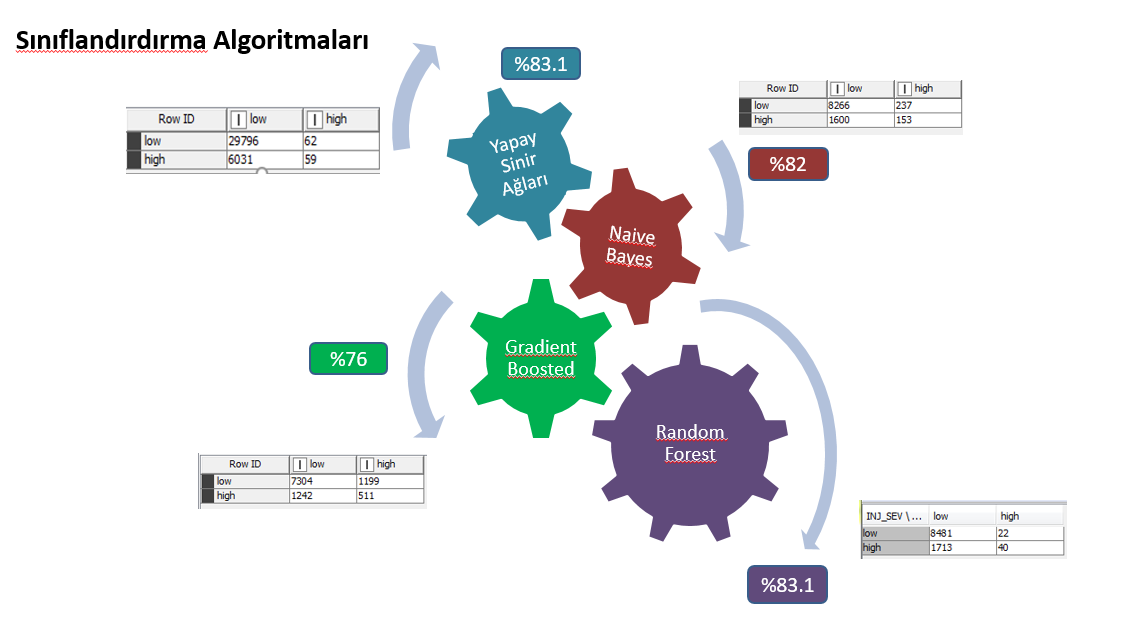
Kullandığımız dört modelde de birbirine yakın accuracy değerleri elde edilirken, en iyi tahmin sonuçlarını **Random Forest** modelinin verdiğini gördük.

**Roc Curve High**

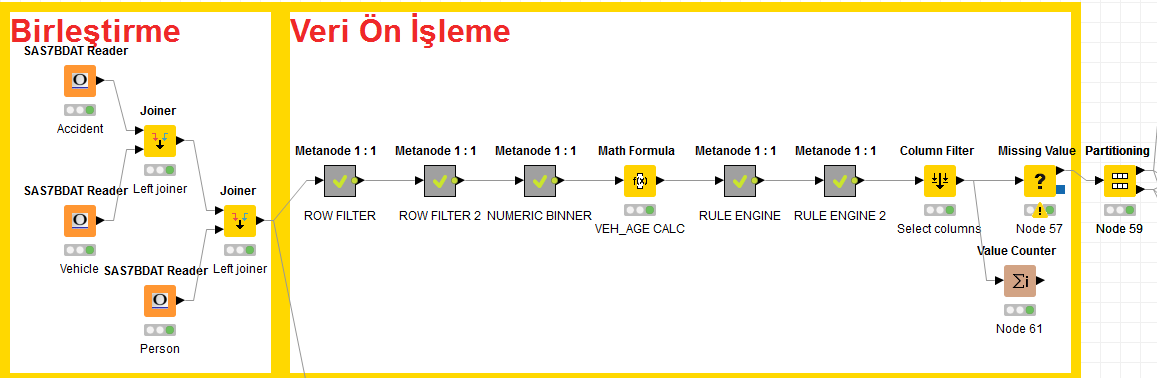


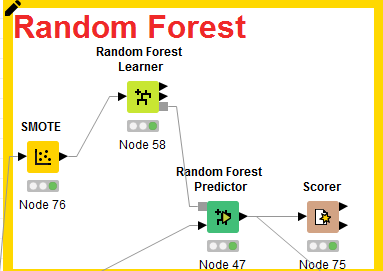
**Roc Curve Low**

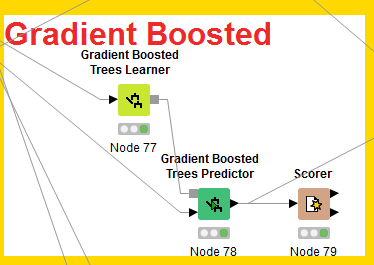


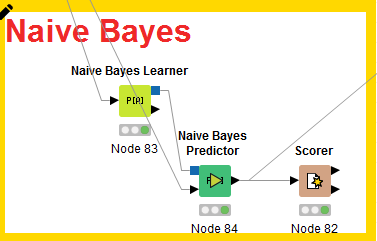


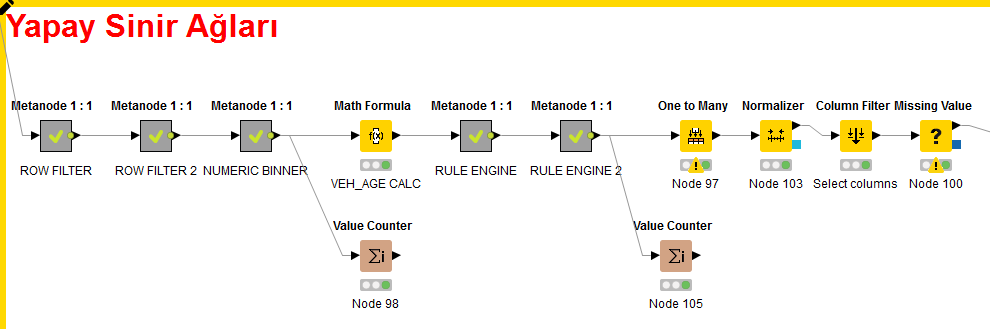
**Knime Workflow Çalışması**

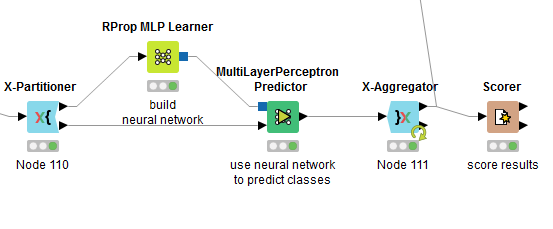


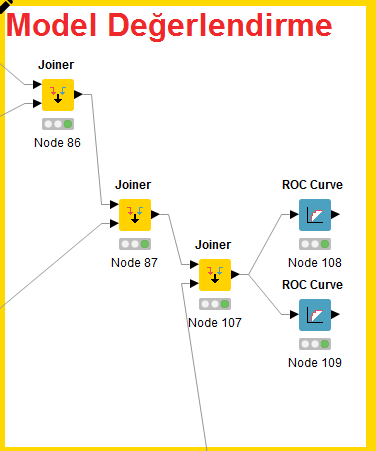












**GANTT CHART**



**PERT DIAGRAM**

