Astım, solunum yollarının kronik ve tekrarlayıcı bir şekilde iltihaplanmasıyla karakterize edilen bir solunum hastalığıdır. Bu durum, hava yollarının daralmasına, öksürük, hırıltı ve nefes darlığı gibi semptomlara neden olabilir. Astım genellikle çocukluk döneminde başlar, ancak her yaşta ortaya çıkabilir ve yaşam boyu devam edebilir.

Astımın belirtileri ve şiddeti kişiden kişiye değişebilir. Bazı insanlar arasında hafif semptomlarla sınırlı kalırken, diğerleri daha ciddi semptomlar yaşayabilir ve günlük yaşam aktivitelerini etkileyebilir. Astım semptomları genellikle tetikleyiciler adı verilen çeşitli faktörlerle ilişkilidir. Bu tetikleyiciler arasında viral enfeksiyonlar, alerjenler (polen, evcil hayvan tüyleri, küf), egzersiz, soğuk hava, hava kirliliği ve irritan maddeler (sigara dumanı, parfüm kokuları) yer alabilir.

Bizlerin çalışmasında da ana konumuz astımın tetikleyicisi olan “Cinsiyet, Yaş, Yaşadığın Bölge, Eğitim, Kentleşme, Alerji, Hareketlilik, Sigara” gibi durumların astım üzerinde olasılıksal olarak ne kadar etkisi var bunun analizini yapmak ve sonuçları sizlerle paylaşmak olacaktır.

Çalışmamızda Kaggle üzerinde herkese açık olarak paylaşılmış “ASHTMA” veri seti (“<https://www.kaggle.com/code/marcotucci/asthma-proj-bayesian-network>”) üzerinde analiz yapılmıştır. Veri setimiz 9 öznitelikten ve kategorik verilerden oluşmaktadır. Kategorik özniteliklerimiz kendi altında 2 veya 3 alt veriye ayrılmıştır. Bunlardan bazıların; cinsiyet özniteliğinin “female” “male” olması gibi.

Astım Veri Seti Tanıtımı

Bu veri seti, astımın risk faktörlerini belirlemek ve hastalığın yaygınlığını anlamak amacıyla toplanmıştır. Veri seti, farklı demografik özellikleri, yaşam tarzı faktörlerini ve astım durumunu içermektedir. Bayes ağları gibi olasılıksal yöntemlerle kullanılmak üzere tasarlanmıştır.

Değişkenler:

Cinsiyet (**SEX**): Katılımcının cinsiyetini belirten bir kategorik değişken. Değerler "male" ve "female" şeklinde kodlanır.

Yaş (**AGE**): Katılımcının yaşını belirten bir sürekli değişken. Yaş, katılımcının astım gelişimi üzerinde önemli bir etkiye sahip olabilir. Değerler "young", “adult” ve "old" şeklinde kodlanır.

Kentleşme (**URB**): Katılımcının yaşadığı yerin kentsel veya bölgesel bir alan olup olmadığını gösteren bir kategorik değişken. Bu, çevresel faktörlerin astım riski üzerindeki etkisini değerlendirmek için önemlidir. Değerler "low",”medium” ve "high" şeklinde kodlanır.

Eğitim Seviyesi (**EDU**): Katılımcının eğitim seviyesini belirten bir kategorik değişken. Eğitim seviyesi, yaşam tarzı ve sağlık davranışları üzerinde etkili olabilir. Değerler "low" ve "high" şeklinde kodlanır.

Coğrafi Bölge (**GEO**): Katılımcının yaşadığı coğrafi bölgeyi belirten bir kategorik değişken. Coğrafi bölge, çevresel faktörlerin ve hava kalitesinin astım üzerindeki etkisini değerlendirmek için önemlidir. Değerler "north", ”south/islands” ve "centre" şeklinde kodlanır.

Alerji (**ALG**): Katılımcının alerji durumunu belirten bir kategorik değişken. Alerjik reaksiyonlar, astım semptomlarını tetikleyebilir. Değerler "yes" ve "no" şeklinde kodlanır.

Sigara Kullanımı (**SMK**): Katılımcının sigara kullanım alışkanlığını belirten bir kategorik değişken. Sigara içmek, astım semptomlarını şiddetlendirebilir. Değerler "no” ve "yes" şeklinde kodlanır.

Hareketsiz Yaşam Tarzı (**SED**): Katılımcının günlük yaşamında ne kadar hareketsiz olduğunu belirten bir kategorik değişken. Hareketsizlik, astım semptomlarını etkileyebilir. Değerler "yes" ve "no" şeklinde kodlanır.

Astım (**ASTHMA**): Katılımcının astım durumunu belirten bir kategorik değişken. Bu değişken, diğer değişkenlerle ilişkili olarak astımın nedenlerini ve etkilerini anlamak için kullanılacaktır. Değerler "yes" ve "no" şeklinde kodlanır.

|  |  |
| --- | --- |
| **Öznitelik** | **Kategoriler** |
| Cinsiyet | Male – Female, 2 – 1 |
| Yaş | Young – Adult – Old, 3 – 1 – 2 |
| Kentleşme | Low – Medium – High, 2 – 3 – 1 |
| Eğitim Seviyesi | Low – High, 2 – 1 |
| Coğrafi Bölge | North – South/Islands – Centre, 2 – 3 – 1 |
| Alerji | No – Yes, 1 – 2 |
| Sigara Kullanımı | No – Yes, 1 – 2 |
| Hareketsiz Yaşam Tarzı | No – Yes, 1 – 2 |
| Astım | No – Yes, 1 – 2 |

Bu veri seti, astımın karmaşık ilişkilerini ve risk faktörlerini değerlendirmek için Bayes ağları gibi olasılıksal modellerin kullanılmasını sağlar. Veri setinin analizi, astımın nedenlerini ve önleyici tedbirlerini belirlemek için önemli bir adım olabilir.

Çalışmamızda Bayesian Network’ün Expert System ve Hill Climbing tabanlı kullanımının karşılaştırılacaktır. Ve bu iki farklı modelin sonuçları BIC ve AIC skorlarını kriter olarak belirledik. En son olarakta Bayesin Network’ün olasılık gösterimini gerçekleştirip analizimizi bitirmiş olacağız. Analizimizi R programlama dili üzerinde yapacağız çünkü bayes ağları gibi olasılıksal grafik modelleri, değişkenler arasındaki ilişkileri ve belirsizlikleri temsil etmek için güçlü bir araçtır. R, istatistiksel analiz ve veri bilimi alanlarında yaygın olarak kullanılan bir programlama dili ve çevreler bütünüdür. Bayes ağlarının oluşturulması, tahmin yapılması ve modelin doğrulanması için bir dizi paket ve fonksiyon içeren R, bu tür modellerin etkili bir şekilde uygulanmasını sağlar. R, aynı zamanda grafiklerin oluşturulması ve sonuçların görselleştirilmesi için geniş bir grafik ve görselleştirme araçlarına da sahiptir. Bu nedenlerden dolayı, Bayes ağları üzerinde çalışırken R kullanmak, modelleme sürecini kolaylaştırabilir ve sonuçların analizini daha erişilebilir hale getirebilir.

metin, elektronik donanım, ekran görüntüsü, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazılım, ekran, görüntüleme içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bu veri setinin yapısı hakkında bilgi veren str(data) çıktısını inceledikten sonra şu sonuçlara ulaşıyoruz:

Veri seti bir data frame yapısına sahiptir.

Toplamda 2755 gözlem (observation) ve 9 değişken (variable) bulunmaktadır.

Değişkenlerin tümü karakter (chr) türündedir.

Değişkenler şunlardır:

* sex: Cinsiyet (erkek veya kadın)
* age: Yaş grubu (yetişkin veya yaşlı)
* urbanization: Şehirleşme düzeyi (düşük, orta veya yüksek)
* education: Eğitim seviyesi (düşük veya yüksek)
* geographic\_area: Coğrafi bölge (güney/adalar, merkez veya kuzey)
* allergy: Alerji durumu (var veya yok)
* smoke: Sigara içme durumu (evet veya hayır)
* sedentary: Seyrek aktivite durumu (evet veya hayır)
* asthma: Astım durumu (var veya yok)

Bu bilgiler, veri setinizdeki değişkenlerin ve veri türlerinin anlaşılmasına yardımcı olur ve veri setinizin yapısını daha iyi anlamanıza yardımcı olur.

Burada kategorik verilerin ”chr” veri tipinde olduğunu görüyoruz. Analiz yapabilmemiz için bu veri tipini “as.factor()” metodu ile numerik hale çevirmemiz lazım.

metin, ekran görüntüsü, yazı tipi, menü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Veri setimizdeki bütün öznitelikleri “lapply()” metodu kullanarak “as.factor()” ile factor veri tipi dönüşümü yaptık. Veri setimizin geri kalan yapısı aynı kalmak sureti ile kategorik verileri artık analiz yapabileceğimiz uygun formata çevirmiş olduk.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Özniteliklerimizin ismi çok uzun olduğu bunları names(data) <- new\_names kodu ile yeni adlandırmaları ile değiştiriyoruz.

Veri setimiz eksiksiz olduğu için veri ön işleme adımlarını kullanmıyoruz eğer ki verilerimizde eksik, gürültü, hatalı type, aykırı veri olsaydı veri önişleme adımlarını gerçekleştirecektik.

Veri Ön İşleme Nedir?

Veri ön işleme, veri analizi ve modelleme sürecinde veri setini hazırlama ve düzenleme işlemidir. Veri setindeki eksik veya anlamsız verilerin temizlenmesi, gereksiz değişkenlerin kaldırılması ve veri setinin modele uygun hale getirilmesi veri ön işleme adımlarına örnektir.

Veri Ön İşleme Adımları:

1. **Veri Yükleme:** İlk adım veri setinin R ortamına yüklenmesidir. read.csv(), read.table() gibi fonksiyonlarla veri seti okunur ve bir değişkene atanır.
2. **Eksik Verilerin İncelenmesi ve Temizlenmes**i: Veri setindeki eksik değerler is.na() fonksiyonu ile kontrol edilir ve bu değerler silinebilir (na.omit()), ortalama değerlerle doldurulabilir (mean()), veya başka bir yöntemle yerine konabilir.
3. **Anlamsız Değişkenlerin Kaldırılması:** Modelleme sürecinde gereksiz veya anlamsız olduğu düşünülen değişkenler veri setinden çıkarılabilir. subset() veya $ operatörü ile değişkenlerin seçimi yapılabilir.
4. **Kategorik Değişkenlerin Dönüştürülmesi:** Kategorik değişkenler faktör veri tipine dönüştürülerek modelleme sürecinde kullanılabilir hale getirilebilir. factor() fonksiyonu kullanılabilir.
5. **Veri Normalizasyonu ve Standartizasyonu:** Veri setindeki sayısal değerlerin farklı ölçeklerde olması durumunda, veri normalizasyonu veya standartizasyonu yapılabilir. scale() veya benzeri fonksiyonlar kullanılabilir.
6. **Özellik Mühendisliği:** Var olan değişkenlerin üzerinde işlem yaparak yeni özellikler türetme işlemidir. Örneğin, iki değişkenin toplamı ya da çarpımı ile yeni bir değişken oluşturulabilir.
7. **Veri Setinin Bölünmesi:** Genellikle veri seti model eğitimi ve testi için bölünür. sample() veya createDataPartition() gibi fonksiyonlarla veri seti rastgele ikiye bölünebilir.

metin, elektronik donanım, ekran görüntüsü, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

Grafımızı oluşturuyoruz, yukarıda düğüm ve kenarlarımızı değişkelere atıyoruz.

metin, elektronik donanım, ekran görüntüsü, bilgisayar içeren bir resim

Açıklama otomatik olarak oluşturuldu

Ağ yapısını graf'a ekle: arcs(graf) <- kenar komutu, oluşturduğumuz graf adlı boş ağa, önceden belirlediğimiz kenarları (kenar) ekler. Bu kenarlar, düğümler arasındaki ilişkileri temsil eder.

graf'ın özelliklerini göster: graf adlı ağın özelliklerini göstermek için bu komutu kullanırız. Özellikler arasında düğüm ve kenar sayısı gibi bilgiler yer alır.

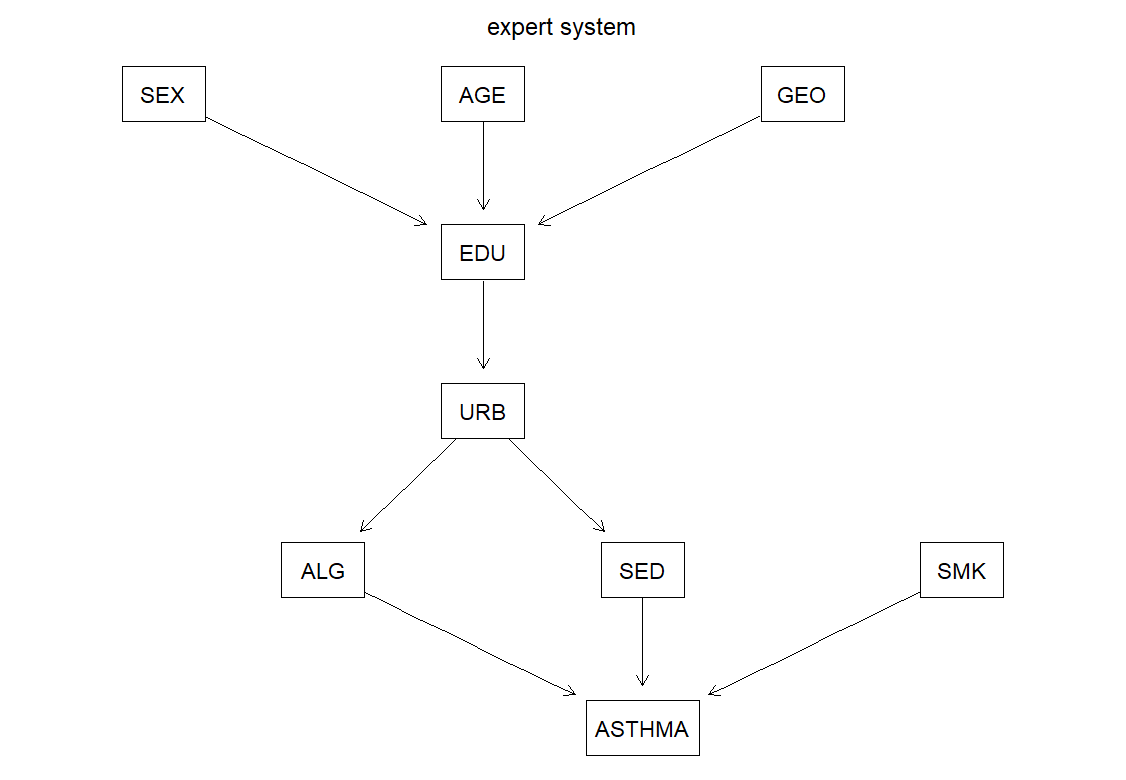
Ağdaki düğümler: graf adlı ağdaki düğümleri (nodes) listeler. Yani, ağdaki değişkenleri gösterir.

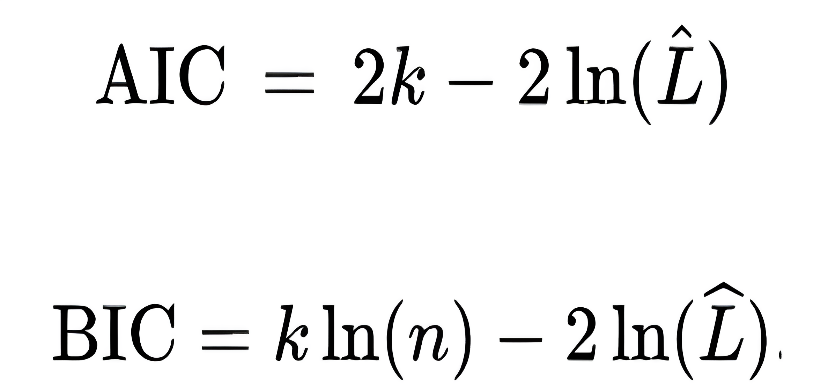
Ağdaki bağlantılar: graf adlı ağdaki bağlantıları (arcs) listeler. Yani, düğümler arasındaki ilişkileri gösterir.

graf'ın sınıfını göster: graf adlı ağın sınıfını (class) gösterir. Ağın Bayesian network olduğunu belirtir.

graf'ın model dizesini göster: graf adlı ağın model dizesini (modelstring) gösterir. Model dizesi, ağdaki düğümler ve aralarındaki ilişkileri açıklayan bir metinsel ifadedir.

graf'ı görselleştir: Son olarak, graf adlı ağı görselleştiririz. Bu komut, ağı bir grafik olarak görüntüler ve belirtilen başlıkla (expert system) görselleştirmeyi sağlar.





Bayesian Information Criterion (BIC) ve Akaike Information Criterion (AIC), istatistiksel model seçiminde kullanılan iki önemli bilgi kriteridir. Her ikisi de bir modelin karmaşıklığı ve uyumunu dengeler ve farklı modeller arasında karşılaştırma yapmamıza yardımcı olur.

1. BIC (Bayesian Information Criterion):

BIC, Schwarz Kriteri olarak da bilinir. Amaç, bir modelin karmaşıklığına ve veriye uyumunu dengeleyerek en iyi modeli seçmektir. BIC skoru, aşağıdaki formülle hesaplanır:

BIC = −2 × log-likelihood + k × ln(n)

Log-likelihood, bir olasılık dağılımının parametreleri kullanılarak gözlemlenen veri setinin olasılığını hesaplamak için kullanılan bir terimdir. Daha spesifik olarak, log-likelihood, veri setinin belirli bir olasılık dağılımı altında ne kadar "iyi" uyum sağladığını ölçmek için kullanılır.

Log-likelihood, logaritmik bir fonksiyondur ve genellikle olasılık hesaplamalarında kullanılan küçük sayıları büyük sayılara dönüştürerek hesaplama kolaylığı sağlar. Ayrıca, olasılıkların çarpımı yerine toplamını kullanarak hesaplama sürecini basitleştirir.

Burada:

* log-likelihood, modelin veri setine uyumunu ölçer ve maksimum olabilirlik tahminini ifade eder.
* k, modelin serbest parametre sayısını ifade eder. Daha karmaşık modeller daha fazla serbest parametreye sahiptir, bu da BIC skorunu etkiler.
* n, veri setindeki gözlem sayısını ifade eder.

BIC skoru düşük olmalıdır, yani daha düşük BIC skorları daha iyi bir model performansını temsil eder.

2. AIC (Akaike Information Criterion):

AIC, bir modelin karmaşıklığını ve veri uyumunu dengelerken, daha geniş bir model yelpazesinde esnekliği artırmak için kullanılır. AIC skoru aşağıdaki formülle hesaplanır:

AIC = −2 × log-likelihood + 2 × k

Burada:

log-likelihood, modelin veri setine uyumunu ölçer.

k, modelin serbest parametre sayısını ifade eder.

AIC skoru da düşük olmalıdır. Yani, daha düşük AIC skorları daha iyi bir model performansını temsil eder.

Farklar:

BIC, model seçimi için daha kısıtlayıcıdır ve daha fazla parametre sayısına ceza uygular. Bu nedenle, BIC daha basit modelleri tercih eder.

AIC ise, model seçimi için daha esnek bir yaklaşım sunar ve daha geniş bir model yelpazesine izin verir.

Genel olarak, her iki kriter de model seçimi için önemlidir ve doğru modelin seçilmesine yardımcı olabilir.

metin, ekran görüntüsü, yazılım, bilgisayar içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bu R kodu, bir Hill Climbing algoritması tabanlı bir model oluşturmak için kullanılabilir. Kod, öncelikle bir siyah liste (black\_list) ve bir beyaz liste (white\_list) oluşturur. Siyah liste, hangi değişkenlerin birbirine bağlı olamayacağını belirtirken, beyaz liste hangi değişkenlerin birbirine bağlı olması gerektiğini belirtir.

Hill Climbing algoritması, bu siyah ve beyaz listeleri kullanarak Bayes ağı modelini oluşturur. Başlangıçta, algoritma rastgele bir Bayes ağı oluşturur ve ardından mevcut modele bazı değişiklikler yaparak (örneğin, düğüm ekleme, düğüm çıkarma veya kenar ekleme/çıkarma) modelin skorunu artırmaya çalışır. Siyah liste ve beyaz liste kriterlerine uygun olarak değişiklikler yapılır. Algoritma, belirli bir süre boyunca modelin skorunu artırmaya çalışır veya bir belirleme noktasına ulaşana kadar devam eder.

Bu şekilde, Hill Climbing algoritması, veri setinden çıkarılan kurallar doğrultusunda Bayes ağı oluşturur ve modelin doğruluğunu artırmaya çalışır. Bu model daha sonra çeşitli istatistiksel analizlerde veya tahminlerde kullanılabilir.

Kod, siyah\_liste ve beyaz\_liste olmak üzere iki matris tanımlar. Siyah liste, hangi değişkenlerin birbirine bağlı olamayacağını belirtirken, beyaz liste hangi değişkenlerin birbirine bağlı olması gerektiğini belirtir. Bu listeler, Bayes ağının yapısını oluşturmak için kullanılacaktır. metin, ekran görüntüsü, yazılım, multimedya yazılımı içeren bir resim

Açıklama otomatik olarak oluşturuldudiyagram, çizgi, metin, teknik çizim içeren bir resim

Açıklama otomatik olarak oluşturuldu

Biz, bu R kodu parçacığını kullanarak bir Bayes ağı modeli öğretiyoruz.

Bu modeli öğretmek için Hill Climbing algoritmasını kullanıyoruz. Bu algoritma, veri setimizdeki ilişkileri göz önünde bulundurarak en iyi modeli bulmaya çalışır.

Modeli oluştururken BIC (Bayesian Information Criterion) skorunu kullanıyoruz. BIC, modelin karmaşıklığını ve veriye uygunluğunu dengeleyerek en iyi modeli seçmemize yardımcı olur.

Beyaz ve siyah listeler kullanarak, oluşturulan modelin yapısını belirliyoruz. Beyaz liste, belirli değişkenler arasında olması gereken bağlantıları belirtirken, siyah liste belirli bağlantıların olmaması gerektiğini belirtir.

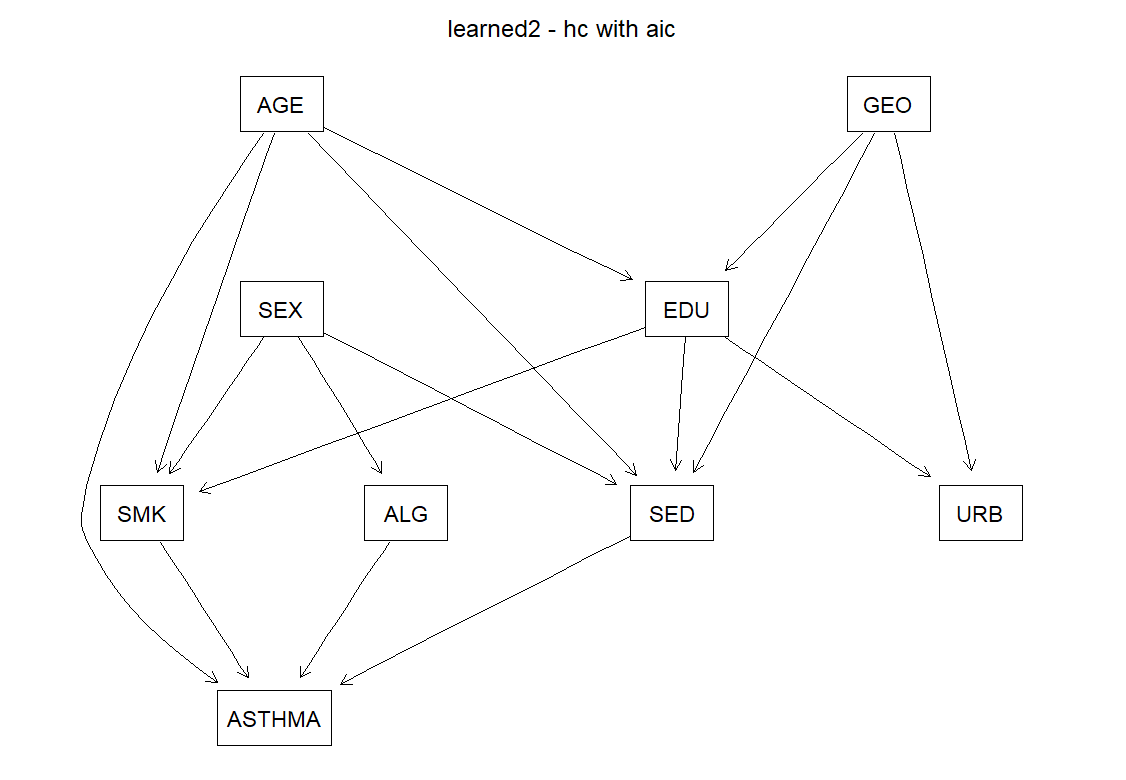
Modeli görselleştirmek için graphviz.plot işlevini kullanıyoruz. Bu, Bayes ağının yapısını görsel olarak sunar.

score işleviyle modelin BIC skorunu hesaplayarak, modelin veriye uygunluğunu ve karmaşıklığını değerlendiriyoruz.

nodes işleviyle, oluşturulan modeldeki düğümleri görüntülüyoruz.

arcs işleviyle, oluşturulan modeldeki kenarları listeliyoruz.

Bu kod parçacığı, veri setimizden çıkarılan kurallar doğrultusunda bir Bayes ağı modeli oluşturur ve modelin yapısını değerlendirmek için çeşitli işlevleri kullanır. Bu, veri setimizdeki değişkenler arasındaki ilişkileri ve astımın potansiyel risk faktörlerini anlamamıza yardımcı olur.metin, ekran görüntüsü, yazılım, ekran, görüntüleme içeren bir resim

Açıklama otomatik olarak oluşturuldu

Oluşturulan Model: Hill Climbing algoritması kullanılarak veri setiniz üzerinde AIC skoru kullanılarak bir Bayes ağı modeli oluşturdunuz. Model, 9 düğüm (node) ve 16 kenar (arc) içeriyor. Bu kenarlar, değişkenler arasındaki ilişkileri temsil eder.

Düğümler ve Kenarlar: Oluşturulan modeldeki düğümler şunlardır: SEX, AGE, GEO, EDU, ALG, URB, SMK, SED ve ASTHMA. ASTHMA düğümü, diğer değişkenlerden etkilenebilecek astım durumunu temsil eder. ASTHMA düğümüne gelen kenarlar, AGE, ALG, SMK ve SED değişkenlerine bağlıdır.

Ortalama Değerler: Modelde ortalama markov örtüsü boyutu 5.33, ortalama komşuluk boyutu 3.56 ve ortalama dallanma faktörü 1.78'dir.

Öğrenme Algoritması ve Skor: Model, Hill-Climbing öğrenme algoritması kullanılarak oluşturuldu. Skor olarak AIC (düşüklikle düzeltilmiş Akaike Bilgi Kriteri) kullanıldı ve bu modelin AIC skoru -17850.03 olarak hesaplandı.

Bu sonuçlar, veri setinizdeki değişkenler arasındaki ilişkileri ve astımın potansiyel risk faktörlerini anlamak için oluşturulan Bayes ağı modelinin yapısını gösterir. Bu model, astımın nedenlerini ve etkilerini anlamak için önemli bir araç olabilir.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bayesian Information Criterion (BIC) Karşılaştırması:

Uzman sistem ve öğrenilmiş model için BIC skorları hesaplandı.

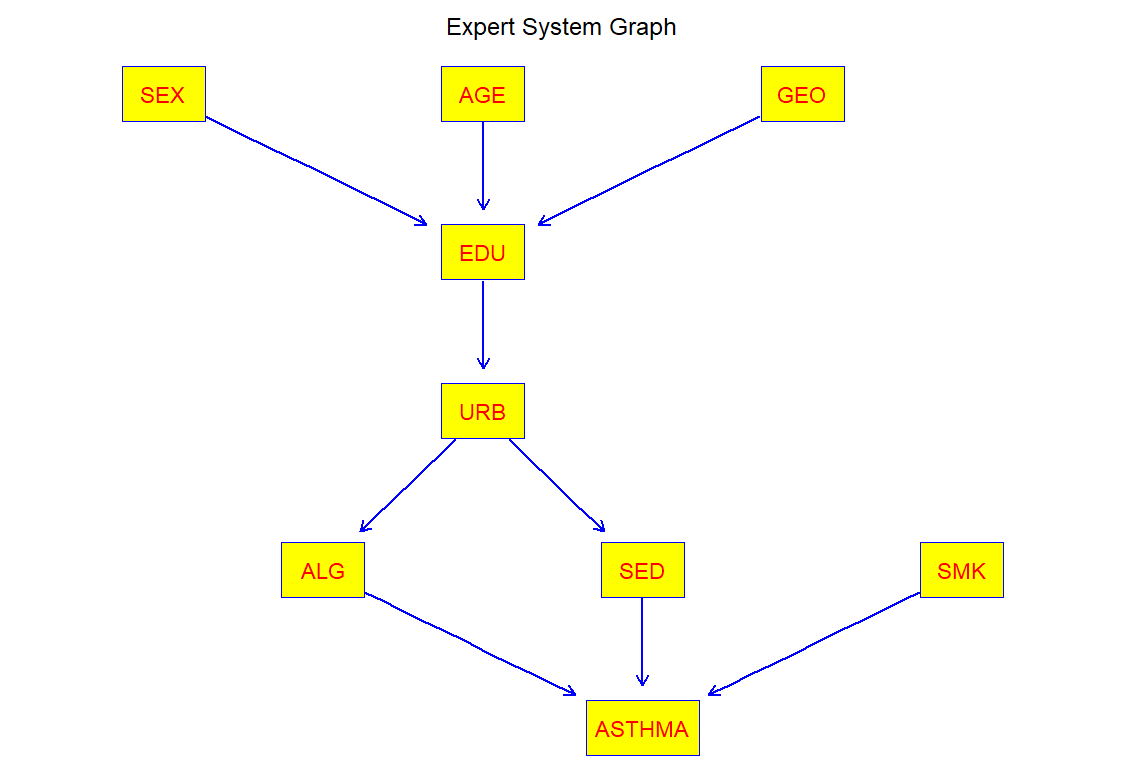
Uzman sistem için BIC skoru -18230.42, öğrenilmiş model için ise -18019.94 olarak belirlendi.

BIC skorlarına göre, öğrenilmiş modelin uzman sistemden daha düşük bir skora sahip olduğu görülmektedir.

Akaike Information Criterion (AIC) Karşılaştırması:

Uzman sistem ve öğrenilmiş model için AIC skorları hesaplandı.

Uzman sistem için AIC skoru -18106.08, öğrenilmiş model için ise -17928.16 olarak belirlendi.

AIC skorlarına göre de, öğrenilmiş modelin uzman sistemden daha düşük bir skora sahip olduğu görülmektedir.

Görselleştirme:

Uzman sistem grafiği görselleştirildi ve düğümler ile kenarlar mavi renkle vurgulandı.

Bu grafiğin üzerinde, değişkenler arasındaki ilişkiler açıkça görülebilir. Bu karşılaştırmalar ve görselleştirme, uzman sistem ve öğrenilmiş model arasındaki farkları ve öğrenilmiş modelin uzman sistemden daha iyi bir performansa sahip olduğunu göstermektedir.