

PIMA INDIANS DIABETES VERİ SETİ'NİN VERİ MADENCİLİĞİ YÖNTEMLERİYLE ANALİZ EDİLMESİ

Veri madenciliği daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veritabanlarından elde edilmesi ve bu bilgilerin işletme kararları verirken kullanılmasıdır. Veri madenciliği günümüzde pazarlama, perakendecilik, biyoloji, tıp, sağlık, sigortacılık, sanayi gibi farklı alanlarda uygulanmaktadır. Verinin büyüyen hızı her geçen gün daha da yakalanamaz olmaktadır. Büyük hacimde olan, hızlı büyüyen, çok çeşitli kaynaklardan çok çeşitli şekillerde gelen (yazı, ses, blog, xlm..) verilerin işlenmesi rolünde veri madenciliğinin önemi büyüktür.

Çalışmamız için kullanacağımız “diabetes” veri seti NIDDK'den (National Institute of Diabetes and Digestive and Kidney) alınmış gerçek bir veri seti olup 9 değişken ve 768 gözlemden oluşmaktadır.

Veri kümesinin amacı, veri kümesine dahil edilen belirli tanı ölçümlerine dayanarak bir hastanın diyabet olup olmadığını teşhis amaçlı olarak tahmin etmektir.

1. VERİ TANIMLAMA

Pregnancies: Hamilelik sayısı

Glucose: Glikoz konsantrasyonu

BloodPressure: Diyastolik kan basıncı (mm Hg)

SkinThickness: Triceps cilt katlama kalınlığı (mm)

Insulin: 2 saatlik serum insülini (mu U / ml)

BMI: Vücut kitle indeksi (kg olarak ağırlık / (m olarak yükseklik) ^ 2)

DiabetesPedigreeFunction: Diyabet soyağacı işlevi

Age: Yaş (yıl)

Outcome: Sınıf değişkeni (0 veya 1)

1.1 Betimsel İstatistikler

Veri kümesinin farklı alanlarının merkezi eğilimini ve tanımlayıcı istatistiklerini elde etmek için SPSS ile betimleyici istatistiklere bakılır. Outcome değişkenimiz target olarak belirlenmiştir.

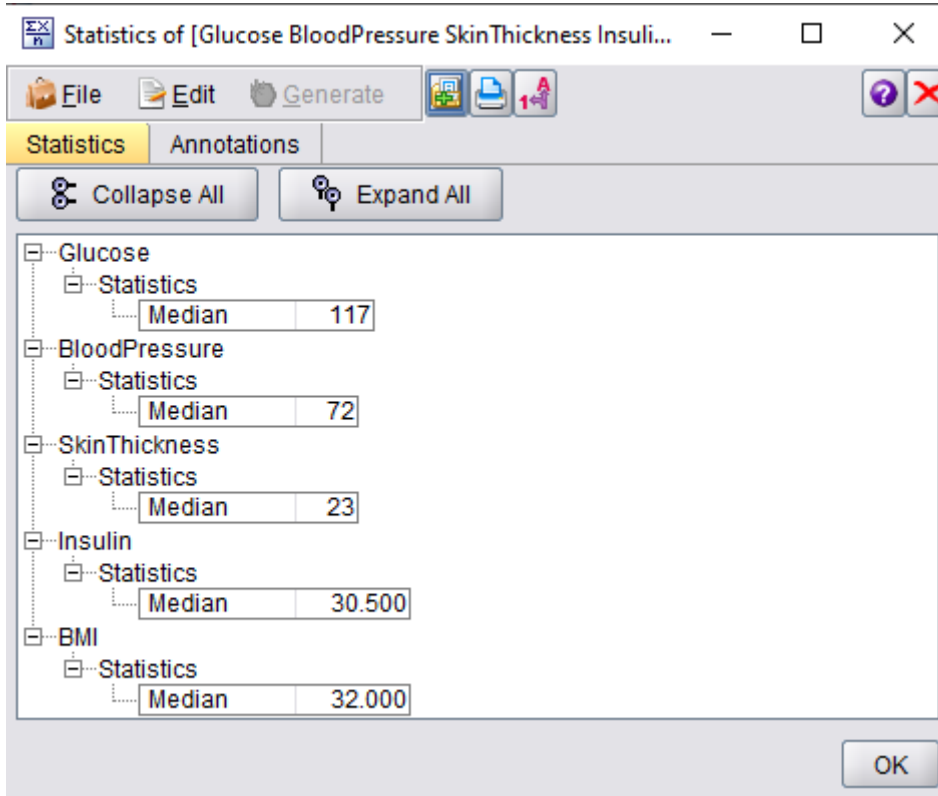
Diabetes Veri Kümesi'nin merkezi eğilim ölçülerine bakacak olursak ortalama, medyan, modu içerir, tanımlayıcı istatistikler ise standart sapma, minimum değer, ilk çeyrek (Q1), medyan (Q2), üçüncü çeyrek (Q3), maksimum değeri içerir ve veri setimiz için sonuçlar şu şekildedir:

Statistics

		Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
N	Valid	768	768	768	768	768	768	768	768	768
	Missing	0	0	0	0	0	0	0	0	0
Mean		3,85	120,89	69,11	20,54	79,80	31,9926	,47188	33,24	,35
Median		3,00	117,00	72,00	23,00	30,50	32,0000	,37250	29,00	,00
Mode		1	99 ^a	70	0	0	32,00	,254 ^a	22	0
Std. Deviation		3,370	31,973	19,356	15,952	115,244	7,88416	,331329	11,760	,477
Variance		11,354	1022,248	374,647	254,473	13281,180	62,160	,110	138,303	,227
Minimum		0	0	0	0	0	,00	,078	21	0
Maximum		17	199	122	99	846	67,10	2,420	81	1
Percentiles	25	1,00	99,00	62,00	,00	,00	27,3000	,24325	24,00	,00
	50	3,00	117,00	72,00	23,00	30,50	32,0000	,37250	29,00	,00
	75	6,00	140,75	80,00	32,00	127,75	36,6000	,62675	41,00	1,00

a. Multiple modes exist. The smallest value is shown

Fakat BMI ve BloodPressure gibi bazı değişkenlerin 0 değerlerini içermesi sorunu söz konusudur. Vücut kitle indeksi ve kan basıncının 0 olması söz konusu olmadığından eksik veri durumu ortaya çıkar dolayısıyla algoritmalar çok iyi sonuç vermeyebilir. Bu verileri tamamen silmemiz ise veri kaybına yol açacaktır. Araştırmalarım sonucunda belirli bir sütun için medyan değerini hesaplayarak ve bu değeri o sütundaki sıfır olan yere ikame edebiliriz.



Yukarıdaki değişkenlerimiz için 0 olan değerleri medyan değerleri ile değiştiriyoruz. Ve BMI modellerde okuttuğumuzda ise oluşan düzenlenmiş veri setimiz aşağıdaki gibidir.

Var. File

Preview Refresh

C:\Users\User\Desktop\VERİMADEN\diabetes.csv

File Data Filter Types Annotations

File: C:\Users\User\Desktop\VERİMADEN\diabetes.csv

```
Pregnancies;Glucose;BloodPressure;SkinThickness;Insulin;BMI;DiabetesPedigreeF
6;148;72;35;30;5;33;6;0;627;50;1
1;85;66;29;30;5;26;6;0;351;31;0
8;183;64;23;30;5;23;3;0;672;32;1
```

☒ Read field names from file ☐ Specify number of fields 1

Skip header characters: 0 EOL comment characters:

Strip lead and trail spaces: ☒ None ☐ Left ☐ Right ☐ Both

Invalid characters: ☒ Discard ☐ Replace with

Encoding: Stream default Decimal symbol: Comma (,)

☒ Line delimiter is newline character Lines to scan for column and type: 50

Field delimiters

☐ Space ☐ Comma ☐ Tab

☐ Newline ☒ Other ;

☐ Non-printing characters

☐ Allow multiple blank delimiters

☒ Automatically recognize dates and times

☐ Treat square brackets as lists

Quotes

Single quotes: Discard

Double quotes: Discard

OK Cancel Apply Reset

diabetes.csv

Preview Refresh

C:\Users\User\Desktop\VERİMADEN\diabetes.csv

File Data Filter **Types** Annotations

Read Values Clear Values Clear All Values

Field	Measurement	Values	Missing	Check	Role
Pregnancies	Ordinal	0,1,2,3,4,5...		None	Input
Glucose	Continuous	[44,199]		None	Input
BloodPressu...	Continuous	[24,122]		None	Input
SkinThickness	Continuous	[7,123]		None	Input
Insulin	Continuous	[14.0,846.0]		None	Input
BMI	Continuous	[2.0,67.1]		None	Input
DiabetesPed...	Continuous	[0.001,2.3...		None	Input
Age	Continuous	[21,81]		None	Input
Outcome	Flag	1/0		None	Target

Table (9 fields, 768 records) #1

File Edit Generate

Table	Annotations									
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	
1	6	148	72	35	30.500	33.600	0.627	50	1	
2	1	85	66	29	30.500	26.600	0.351	31	0	
3	8	183	64	23	30.500	23.300	0.672	32	1	
4	1	89	66	23	94.000	28.100	0.167	21	0	
5	0	137	40	35	168.0...	43.100	2.288	33	1	
6	5	116	74	23	30.500	25.600	0.201	30	0	
7	3	78	50	32	88.000	3.100	0.248	26	1	
8	10	115	72	23	30.500	35.300	0.134	29	0	
9	2	197	70	45	543.0...	30.500	0.158	53	1	
10	8	125	96	23	30.500	32.000	0.232	54	1	
11	4	110	92	23	30.500	37.600	0.191	30	0	
12	10	168	74	23	30.500	3.800	0.537	34	1	
13	10	139	80	23	30.500	27.100	1.441	57	0	
14	1	189	60	23	846.0...	30.100	0.398	59	1	
15	5	166	72	19	175.0...	25.800	0.587	51	1	
16	7	100	72	23	30.500	3.000	0.484	32	1	
17	0	118	84	47	230.0...	45.800	0.551	31	1	
18	7	107	74	23	30.500	29.600	0.254	31	1	
19	1	103	30	38	83.000	43.300	0.183	33	0	
20	1	115	70	30	96.000	34.600	0.529	32	1	
21	3	126	88	41	235.0...	39.300	0.704	27	0	
22	8	99	84	23	30.500	35.400	0.388	50	0	
23	7	196	90	23	30.500	39.800	0.451	41	1	
24	9	119	80	35	30.500	2.900	0.263	29	1	
25	11	143	94	33	146.0...	36.600	0.254	51	1	
26	10	125	70	26	115.0...	31.100	0.205	41	1	
27	7	147	76	23	30.500	39.400	0.257	43	1	
28	1	97	66	15	140.0...	23.200	0.487	22	0	
29	13	145	82	19	110.0...	22.200	0.245	57	0	
30	5	117	92	23	30.500	34.100	0.337	38	0	
31	5	109	75	26	30.500	3.600	0.546	60	0	
32	3	158	76	36	245.0...	31.600	0.851	28	1	
33	3	88	58	11	54.000	24.800	0.267	22	0	

1.2 Korelasyon Analizi

Değişkenlerimiz için aralarındaki ilişki açısından korelasyonlarına bakacak olursak Glucose değişkeni ile Outcome değişkeni arasındaki korelasyon 0,47 iken Age ile Pregnancies değişkenleri arasındaki korelasyon 0,54'tür. Bu değişkenler arasındaki ilişkiler göz önünde bulundurulabilir.

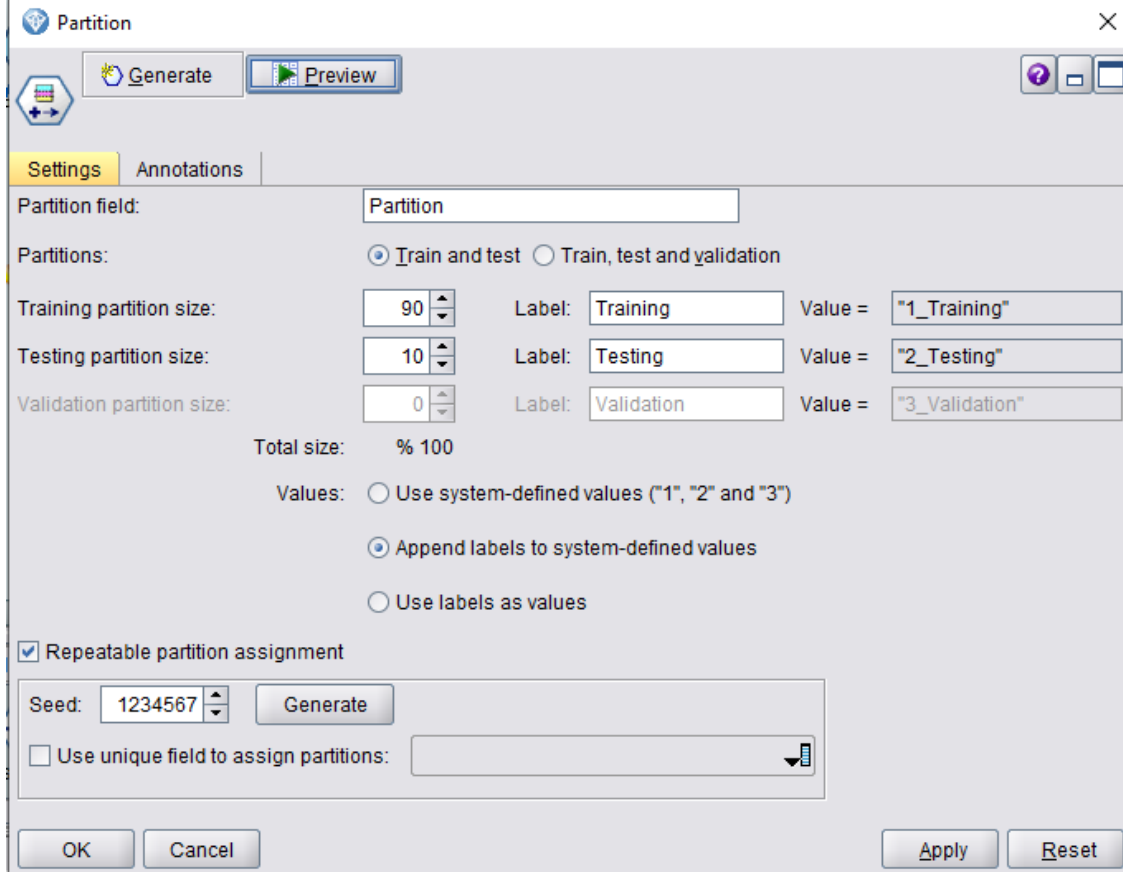
Correlations

		Pregnancies	Glucose	Age	Outcome
Pregnancies	Pearson Correlation	1	,129**	,544**	,222**
	Sig. (2-tailed)		,000	,000	,000
	N	768	768	768	768
Glucose	Pearson Correlation	,129**	1	,264**	,467**
	Sig. (2-tailed)	,000		,000	,000
	N	768	768	768	768
Age	Pearson Correlation	,544**	,264**	1	,238**
	Sig. (2-tailed)	,000	,000		,000
	N	768	768	768	768
Outcome	Pearson Correlation	,222**	,467**	,238**	1
	Sig. (2-tailed)	,000	,000	,000	
	N	768	768	768	768

** . Correlation is significant at the 0.01 level (2-tailed).

2. ANALİZ

Şimdi verileri dönüştürdüğümüze göre, veri kümesini iki bölüme ayırmalıyız: bir eğitim veri kümesi ve bir test veri kümesi. Veri kümesini bölmek, denetimli makine öğrenimi modelleri için çok önemli bir adımdır. Partition nodu ile bunu IBM modellerda yapabiliriz.



The image shows the 'Partition' dialog box in IBM SPSS. The 'Settings' tab is active. The 'Partition field' is set to 'Partition'. The 'Partitions' section has 'Train and test' selected. The 'Training partition size' is 90, 'Testing partition size' is 10, and 'Validation partition size' is 0. The 'Total size' is % 100. The 'Values' section has 'Append labels to system-defined values' selected. The 'Repeatabile partition assignment' checkbox is checked. The 'Seed' is 1234567. The 'Generate' button is visible. The 'Use unique field to assign partitions' checkbox is unchecked. The 'OK', 'Cancel', 'Apply', and 'Reset' buttons are at the bottom.

Partition

Generate Preview

Settings Annotations

Partition field: Partition

Partitions: ☒ Train and test ☐ Train, test and validation

Training partition size: 90 Label: Training Value = "1_Training"

Testing partition size: 10 Label: Testing Value = "2_Testing"

Validation partition size: 0 Label: Validation Value = "3_Validation"

Total size: % 100

Values: ☐ Use system-defined values ("1", "2" and "3")
☒ Append labels to system-defined values
☐ Use labels as values

☒ Repeatabile partition assignment

Seed: 1234567 Generate

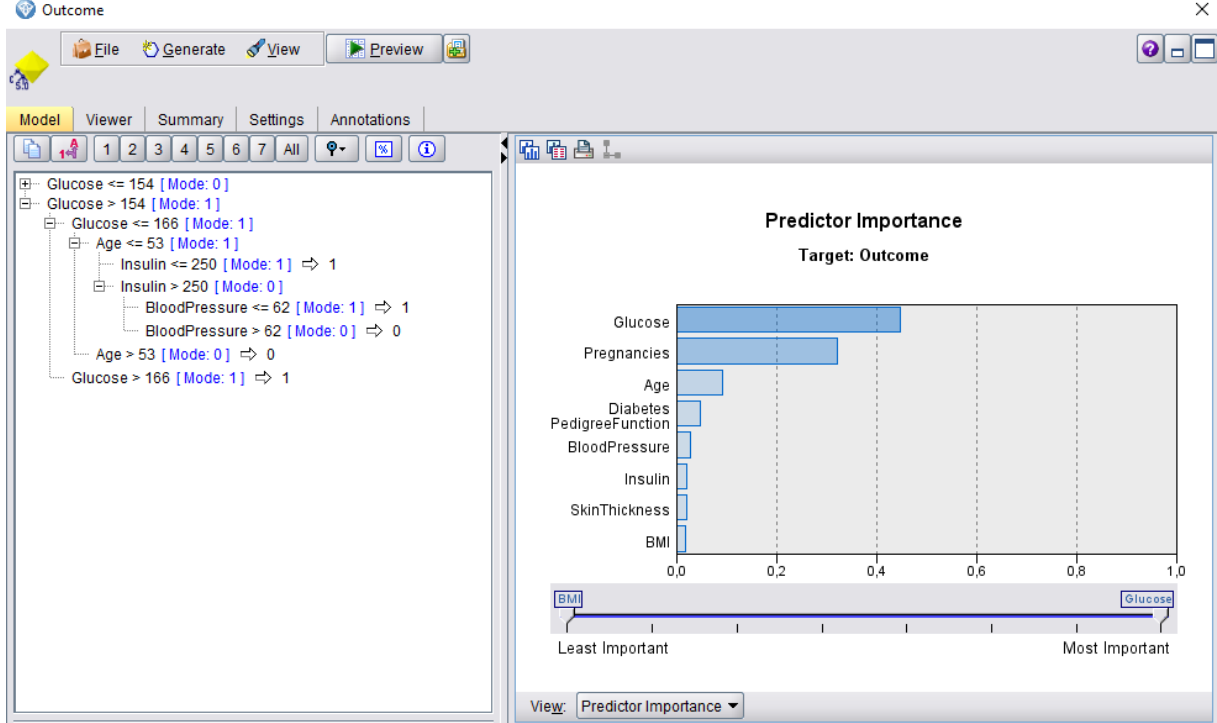
☐ Use unique field to assign partitions:

OK Cancel Apply Reset

2.1 IBM SPSS Modeler ile Analiz

2.1.1 Karar Ağacı Yöntemi ile Tahmin

1. C5.0 Algoritması



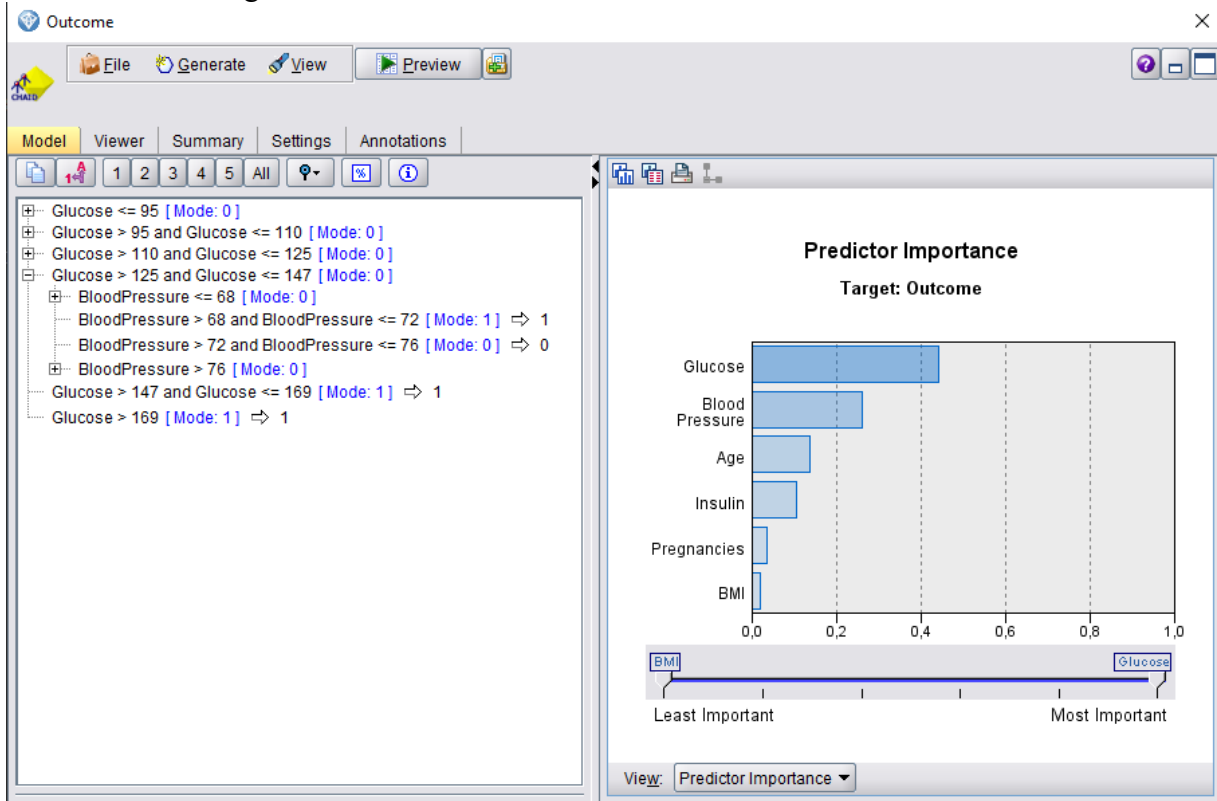
C5 algoritmasına göre glikoz önemli bir değişken iken BMI daha az önemlidir.

Bu karar ağacına bakılarak hasta olup olmama durumu bakımından; glikoz seviyesi 154'ten büyük ve 166'dan küçük eşit olan, yaşı 53'den küçük veya eşit, insülini 250 ml'den fazla olup kan basıncı da 62'den küçük veya eşit olanlar diyabet hastasıdır çıkarımında bulunabiliriz.

2. C&RT Algoritması

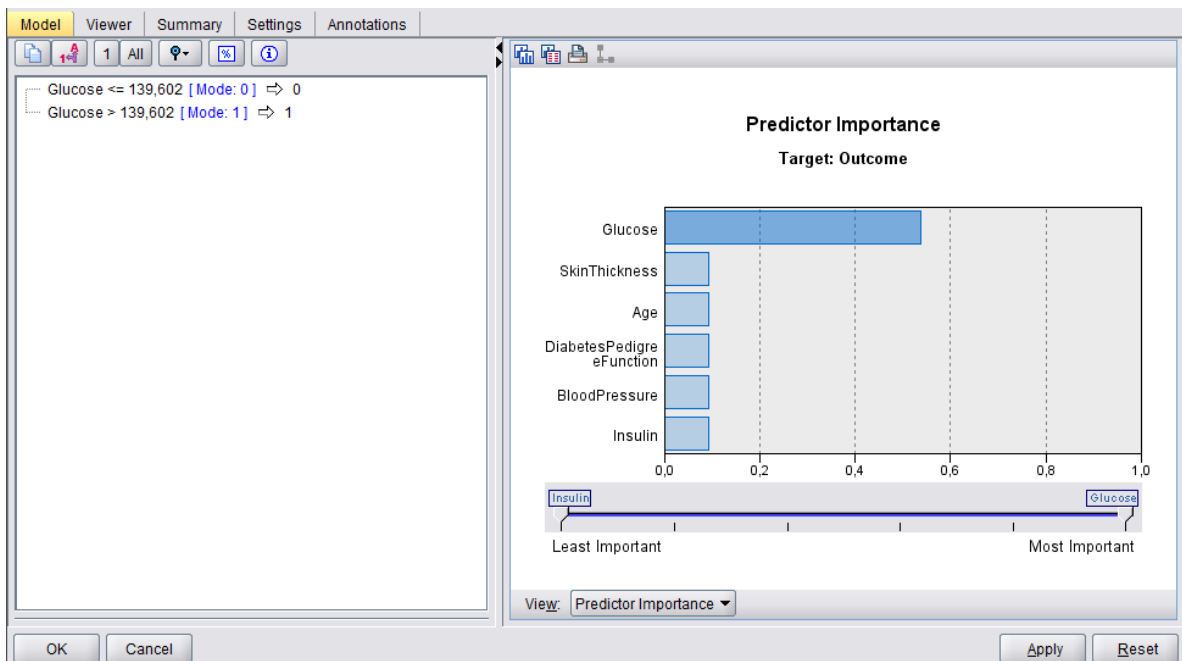
Sürekli hedef değişkenleri ile çalışan bir algoritmadır. Hedef değişkenimiz olan outcome sürekli değişken olmadığından önemli bir tahmin edici de bulamadı.

3. CHAID Algoritması

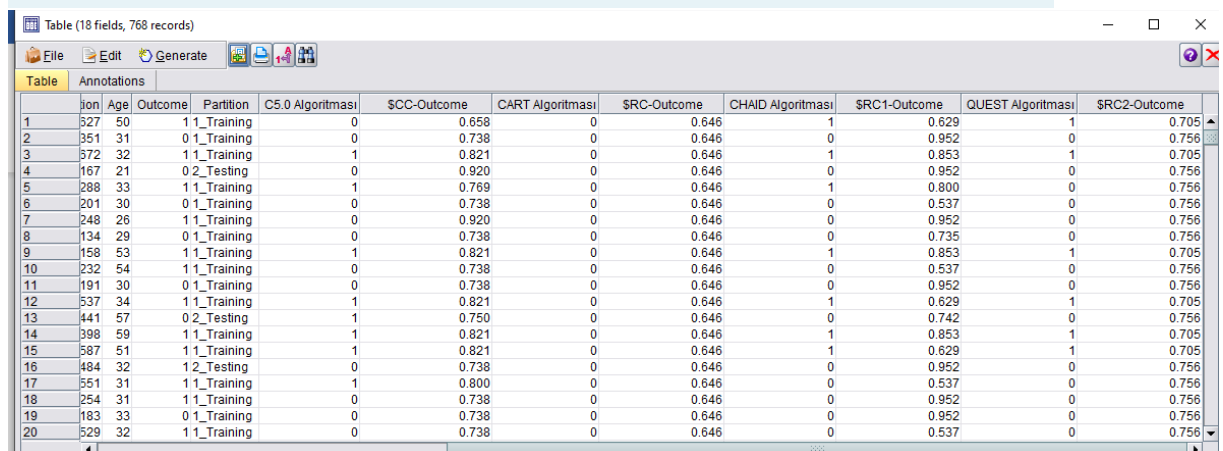


CHAID algoritmasına göre glikoz önemli bir değişken iken vücut kitle endeksi(BMI) daha az önemlidir.

4. QUEST Algoritması



Quest algoritmasında glikoz önemli değişken bulunurken diğer değişkenlerin önem dereceleri aynıdır.



8

Matrix of Outcome by C5.0 Algoritması		
C5.0 Algoritması		
Outcome	0	1
0	447	53
1	88	180

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 264,143, df = 1, probability = 0

OK

Buradaki confusion matrixi satırlarda gerçek değer, sütunlarda tahmin değeri yer almaktadır. Sınıflandırma tahmin sonuçları şekildeki matrixte gösterilmiştir.

Results for output field Outcome
Individual Models
Comparing cart with Outcome
Correct 597 % 77,73
Wrong 171 % 22,27
Total 768
Comparing chaid with Outcome
Correct 596 % 77,6
Wrong 172 % 22,4
Total 768
Comparing Quest with Outcome
Correct 574 % 74,74
Wrong 194 % 25,26
Total 768
Comparing c5 with Outcome
Correct 645 % 83,98
Wrong 123 % 16,02
Total 768
Agreement between cart chaid Quest c5
Agree 539 % 70,18
Disagree 229 % 29,82
Total 768
Comparing Agreement with Outcome
Correct 480 % 89,05
Wrong 59 % 10,95
Total 539

Burada doğru sınıflandırma başarı yüzdesi en yüksek olan C5.0 algoritmasıdır ve en iyi tahmini yaptığı söylenebilir bu sebeple de yeni hasta kaydının diyabet olup olmadığını karar

ağaçları algoritmasından C5.0 algoritması ile tahmin edebiliriz. Ayrıca her dört algoritma da 768 veriden 539 tanesine ortak atama yapmıştır. Bu atamalardan %89'u doğru tahmin edilmiştir.

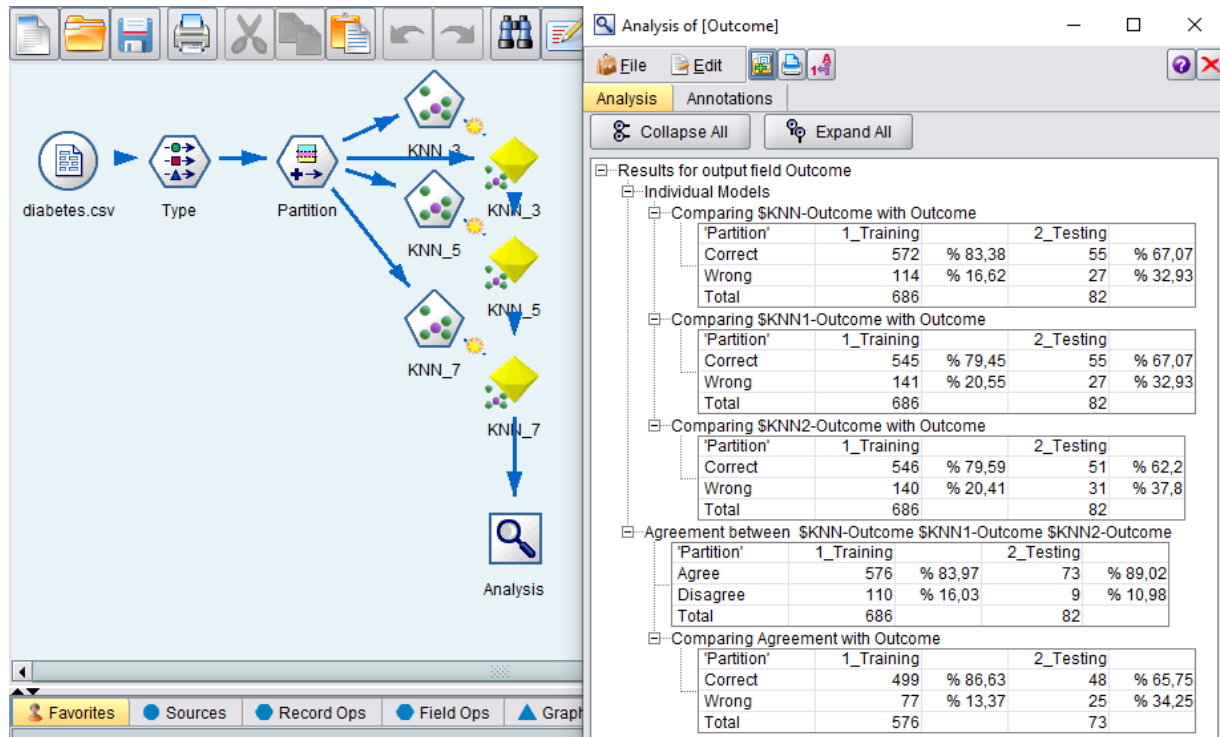
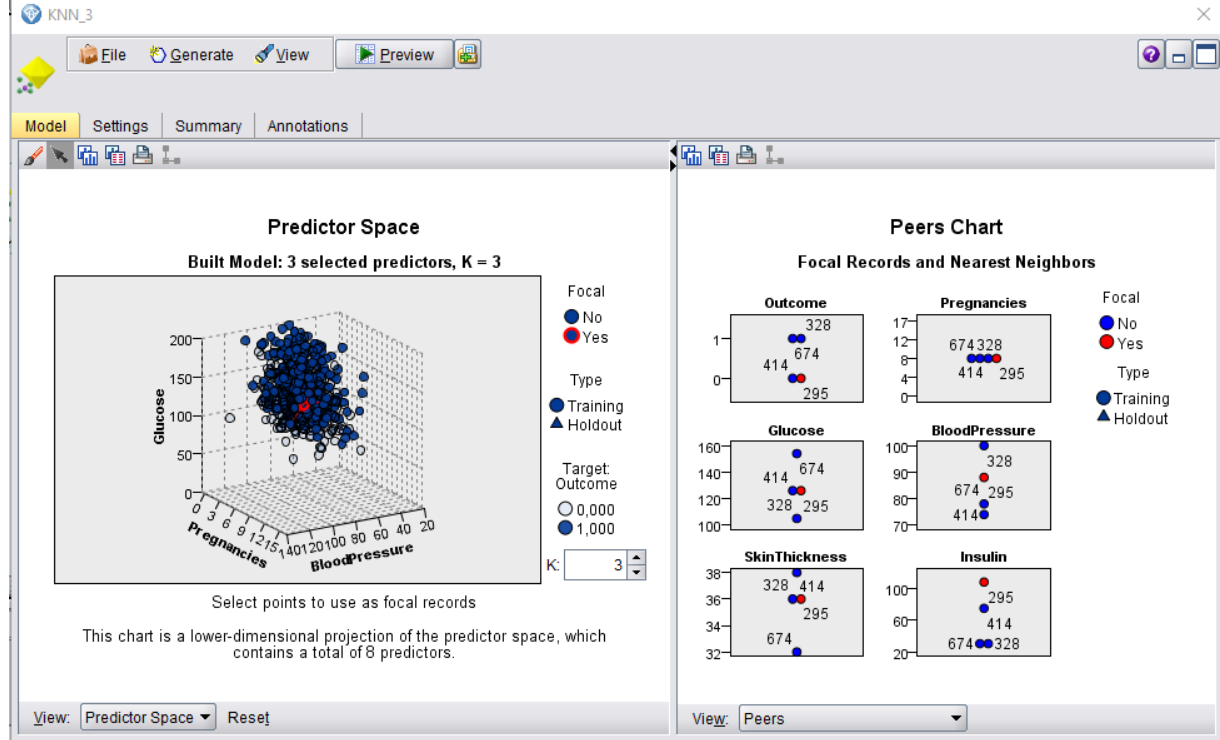
768. hasta için Outcome değeri olan 0'ı silerek bu hastanın diyabet olup olmadığını tahmin etmek istersek C5.0 algoritması bize tahmin sonucunu aşağıda yer aldığı gibi verecektir. Burada da görüldüğü gibi 768. denegin diyabet hastası olmadığı şeklinde tahmin etmiştir.

Table (18 fields, 768 records) #2

File Edit Generate

Table	Annotations	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	Partition	C5.0 Algoritması	\$CC-Outcome
738		8	65	72	23	30.500	32		6	42	0 1_Training	0	0.881
739		2	99	60	17	160.0...	366		453	21	0 1_Training	0	0.924
740		1	102	74	23	30.500	395		293	42	1 1_Training	0	0.658
741		11	120	80	37	150.0...	423		785	48	1 1_Training	1	0.696
742		3	102	44	20	94.000	308		4	26	0 1_Training	0	0.924
743		1	109	58	18	116.0...	285		219	22	0 1_Training	0	0.924
744		9	140	94	23	30.500	327		734	45	1 1_Training	1	0.696
745		13	153	88	37	140.0...	406		1174	39	0 1_Training	1	0.696
746		12	100	84	33	105.0...	30		488	46	0 1_Training	0	0.857
747		1	147	94	41	30.500	493		358	27	1 2_Testing	0	0.913
748		1	81	74	41	57.000	463		1096	32	0 1_Training	0	0.881
749		3	187	70	22	200.0...	364		408	36	1 1_Training	1	0.827
750		6	162	62	23	30.500	243		178	50	1 1_Training	1	0.827
751		4	136	70	23	30.500	312		1182	22	1 1_Training	1	0.833
752		1	121	78	39	74.000	39		261	28	0 1_Training	0	0.924
753		3	108	62	24	30.500	26		223	25	0 1_Training	0	0.924
754		0	181	88	44	510.0...	433		222	26	1 1_Training	1	0.827
755		8	154	78	32	30.500	324		443	45	1 1_Training	1	0.696
756		1	128	88	39	110.0...	365		1057	37	1 1_Training	1	0.696
757		7	137	90	41	30.500	32		391	39	0 1_Training	1	0.714
758		0	123	72	23	30.500	363		258	52	1 1_Training	0	0.658
759		1	106	76	23	30.500	375		197	26	0 1_Training	0	0.924
760		6	190	92	23	30.500	355		278	66	1 1_Training	0	0.750
761		2	88	58	26	16.000	284		766	22	0 1_Training	0	0.924
762		9	170	74	31	30.500	44		403	43	1 1_Training	1	0.827
763		9	89	62	23	30.500	225		142	33	0 1_Training	0	0.881
764		10	101	76	48	180.0...	329		171	63	0 2_Testing	0	0.658
765		2	122	70	27	30.500	368		34	27	0 1_Training	0	0.924
766		5	121	72	23	112.0...	262		245	30	0 1_Training	0	0.857
767		1	126	60	23	30.500	301		349	47	1 1_Training	1	0.696
768		1	93	70	31	30.500	304		315	23	\$null\$ 1_Training	0	0.924

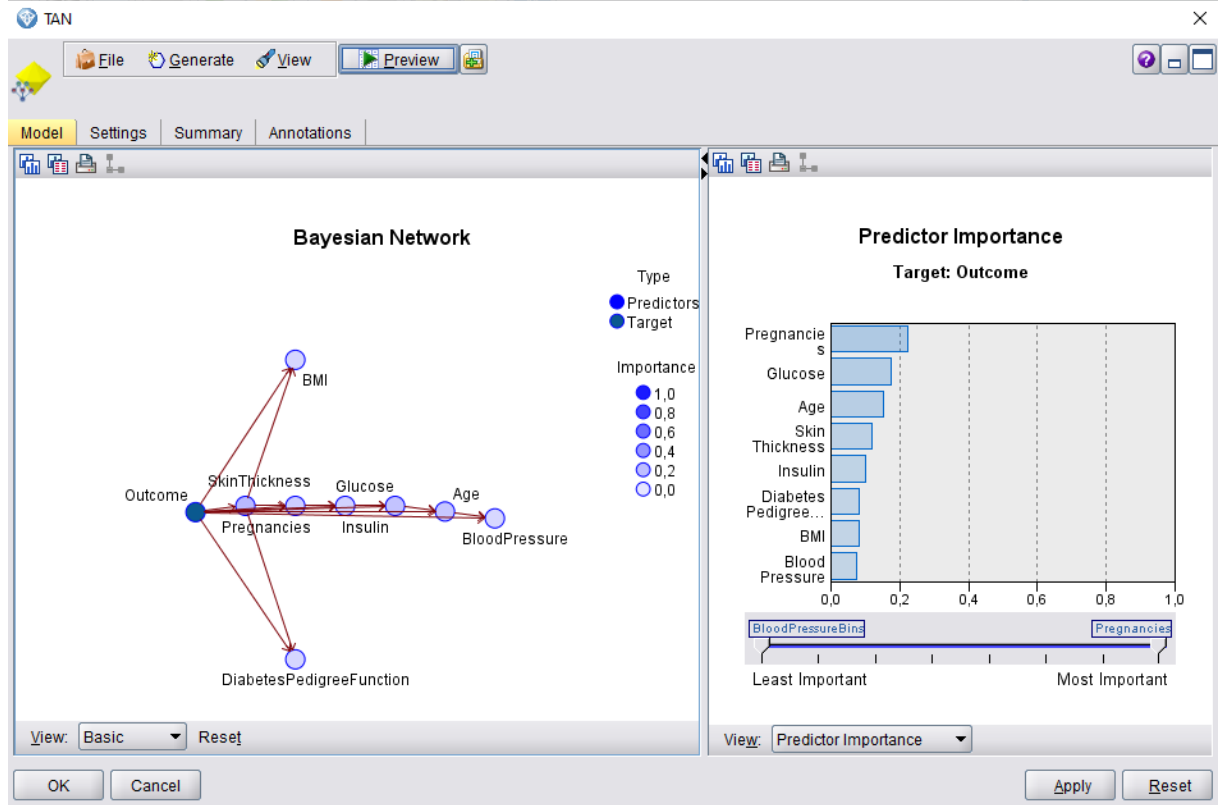
2.1.2_K En Yakın Komşu ile Tahmin



Buradaki 3, 5 ve 7'lik komşu algoritmaları içerisinde k=3 ve k=5 için oluşturulan model en iyi tahmini yapan modellerdir. (%67,07). Üç farklı k değeri için çalıştırılan KNN algoritmaları 82 test verisi içerisinde 73 tanesine ortak atama yapmıştır. Bu 73 ortak atamadan da 478tanesi (%65,75) doğru atanmıştır.

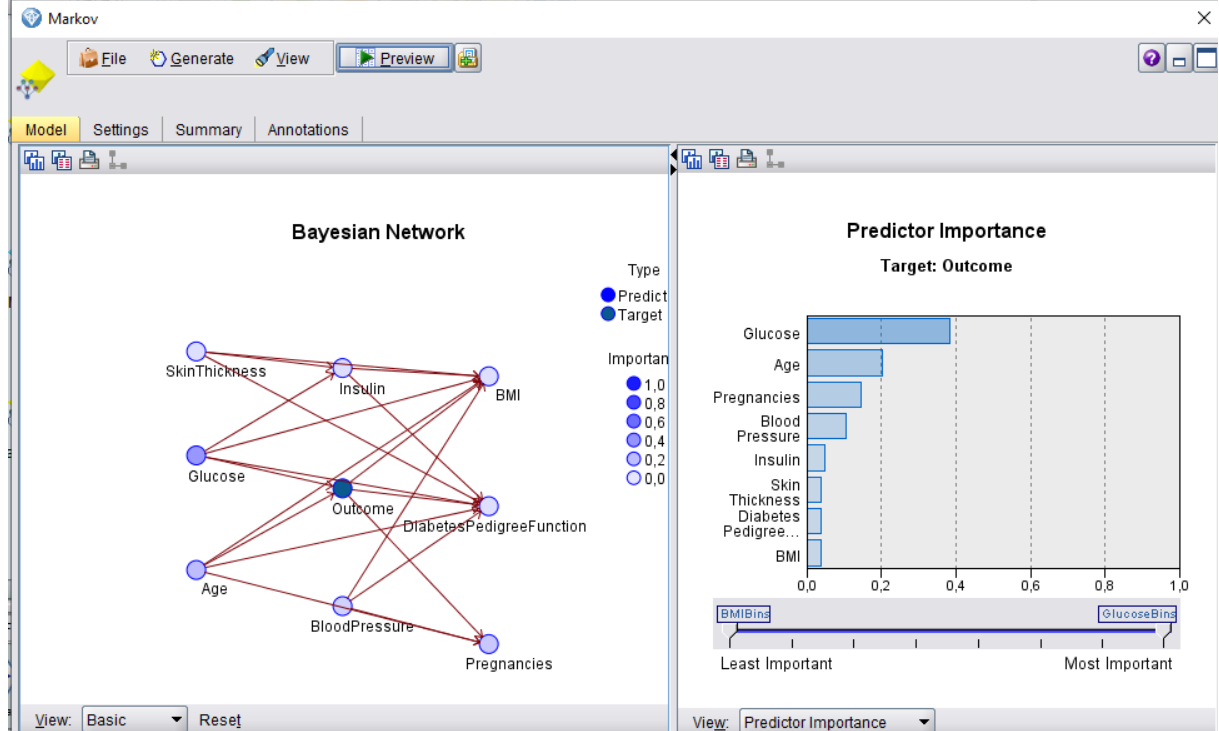
2.1.3 Bayes ile Tahmin

■ TAN Modeli



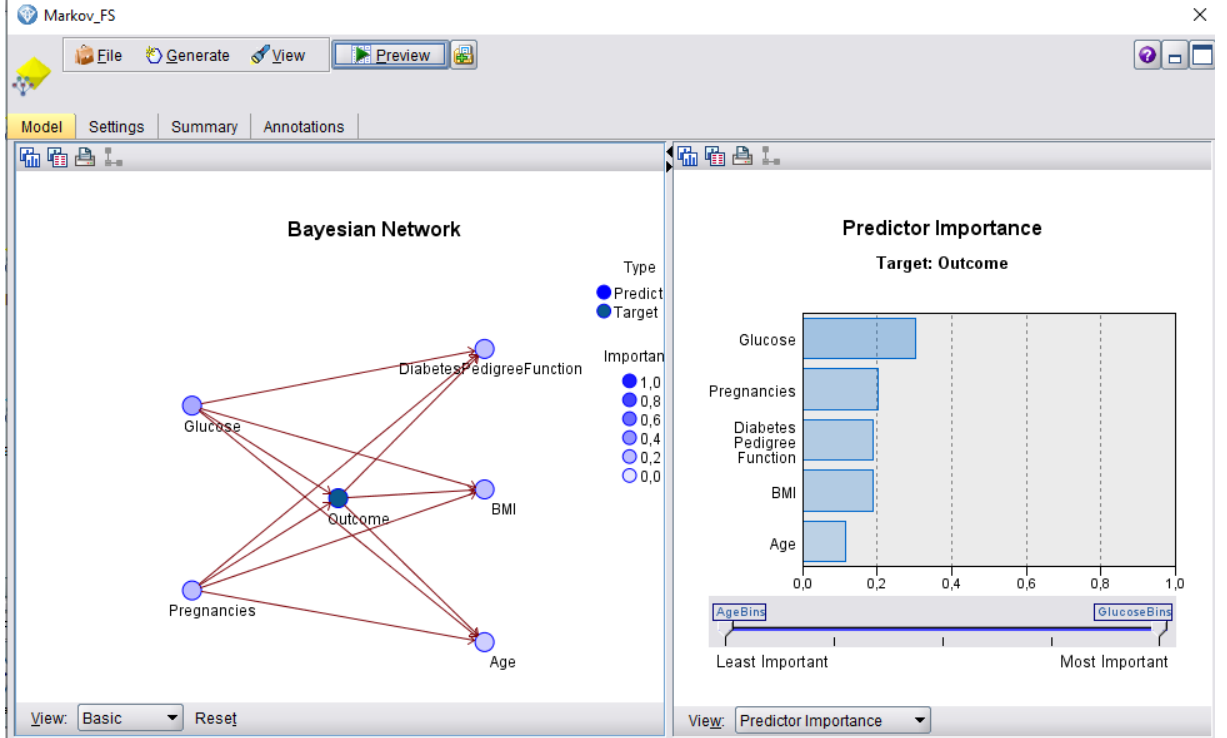
Sade bayes modeli için en önemli tahmin edici hamilelik değişkeni olmuştur.

■ MARKOV Modeli

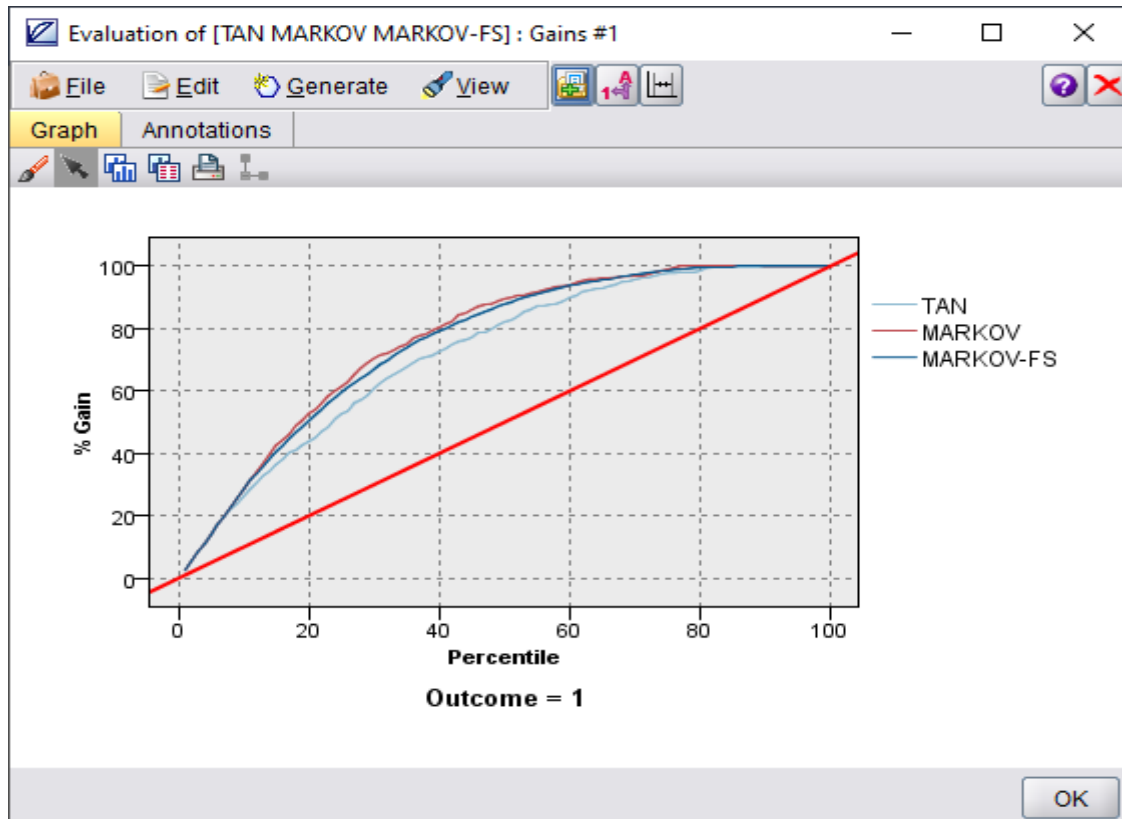


Markov modeli için en önemli tahmin edici glikozdur.

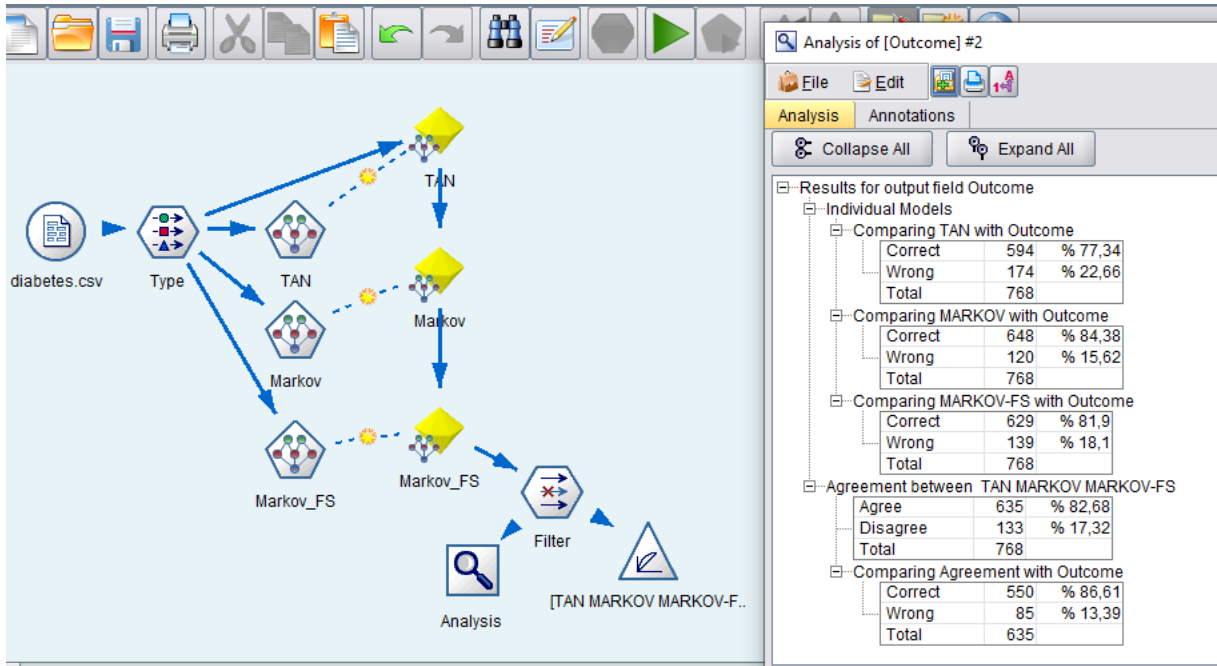
MARKOV-FS Modeli



Markov-FS modeli için en önemli tahmin edici glikozdur.

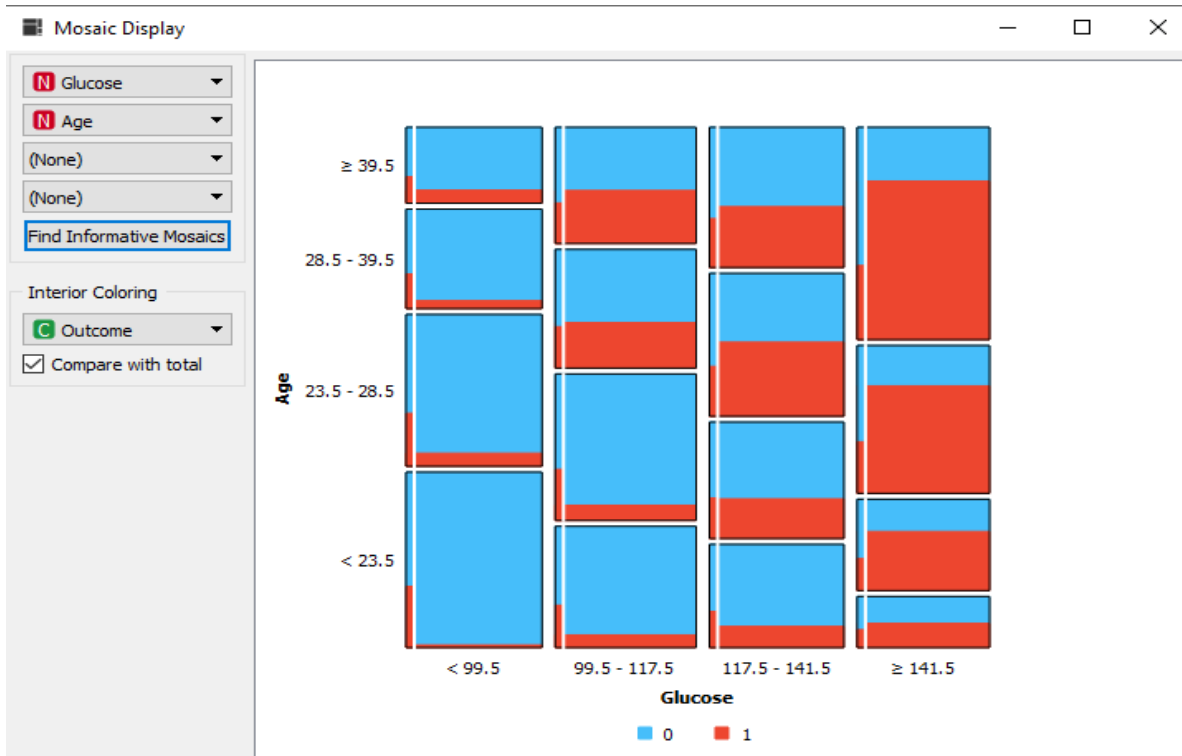


Burada TAN, MARKOV ve MARKOV-FS ile yapılan analizler sonucunda MARKOV'un kazanç yüzdesinin daha fazla olduğu bu sebeple de tahmin için MARKOV modelinin seçilebileceğini söyleyebiliriz.

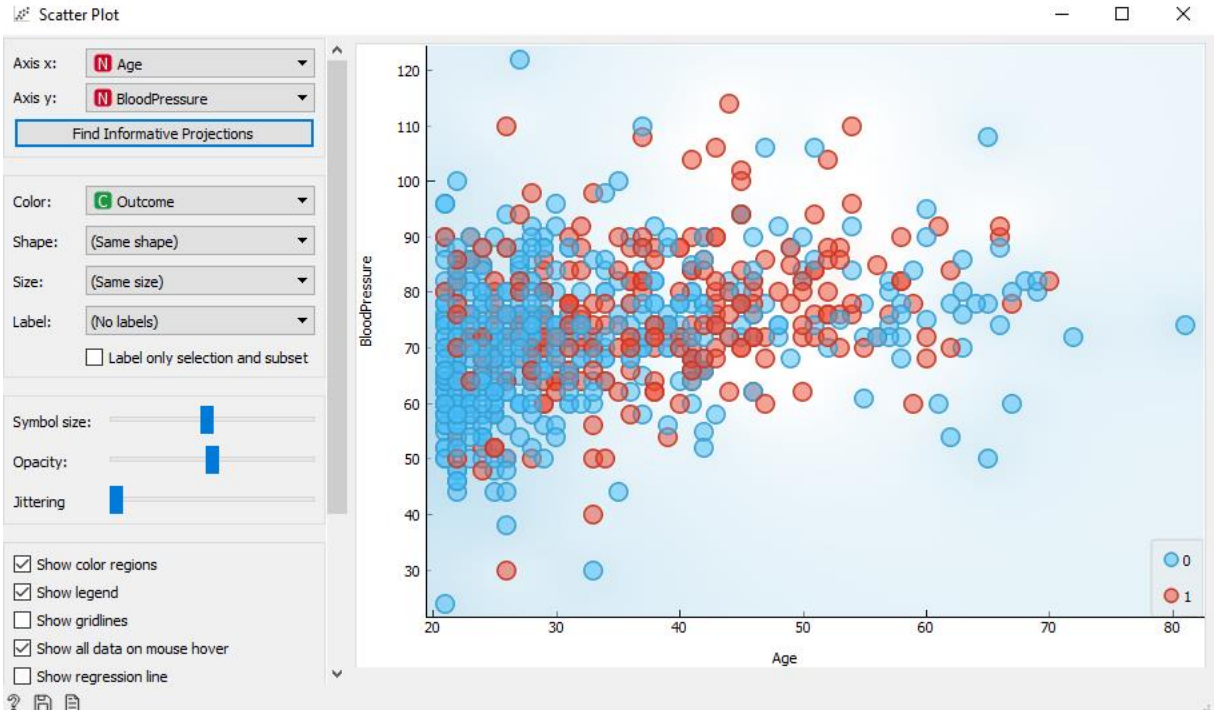


En iyi modelin %84,38'lik doğru tahmin oranıyla Markov modeli olduğu gözlenmiştir. Ayrıca 3 bayes ağının 768 gözlemden 635'i ortak atanmış olup bu 635 atamadan da her üç bayesci model de 550'sine doğru atama yapmıştır.

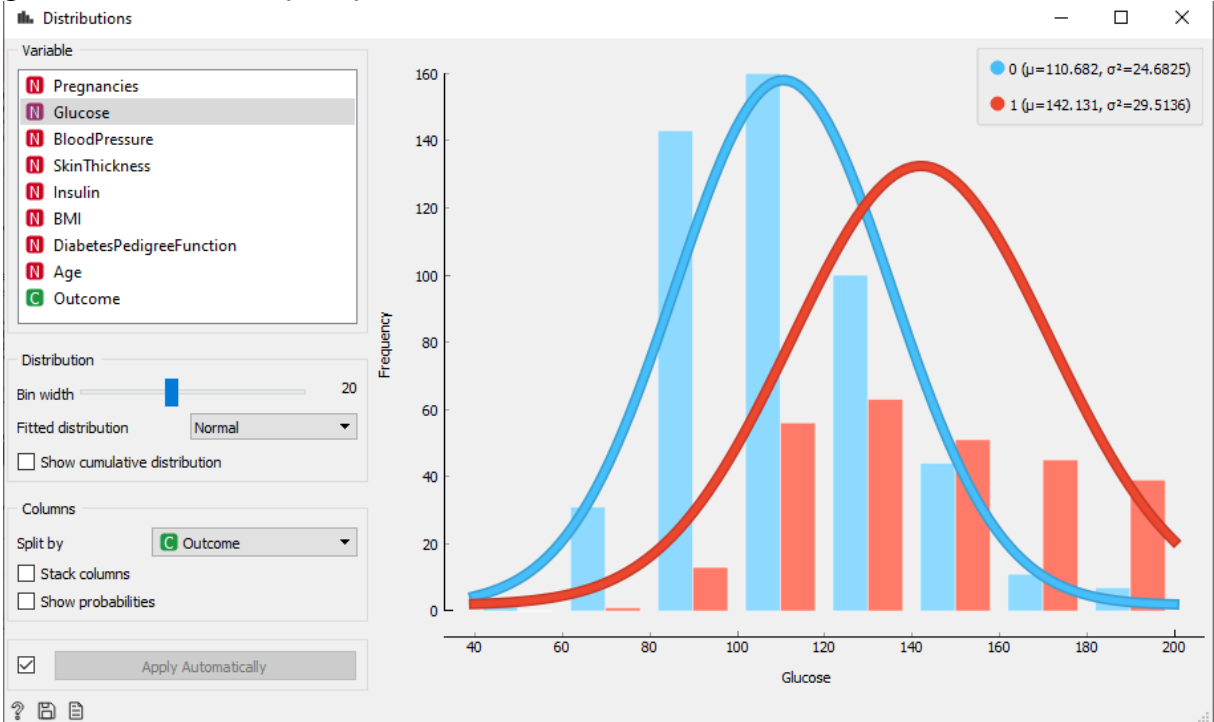
2.2 Orange İle Analiz



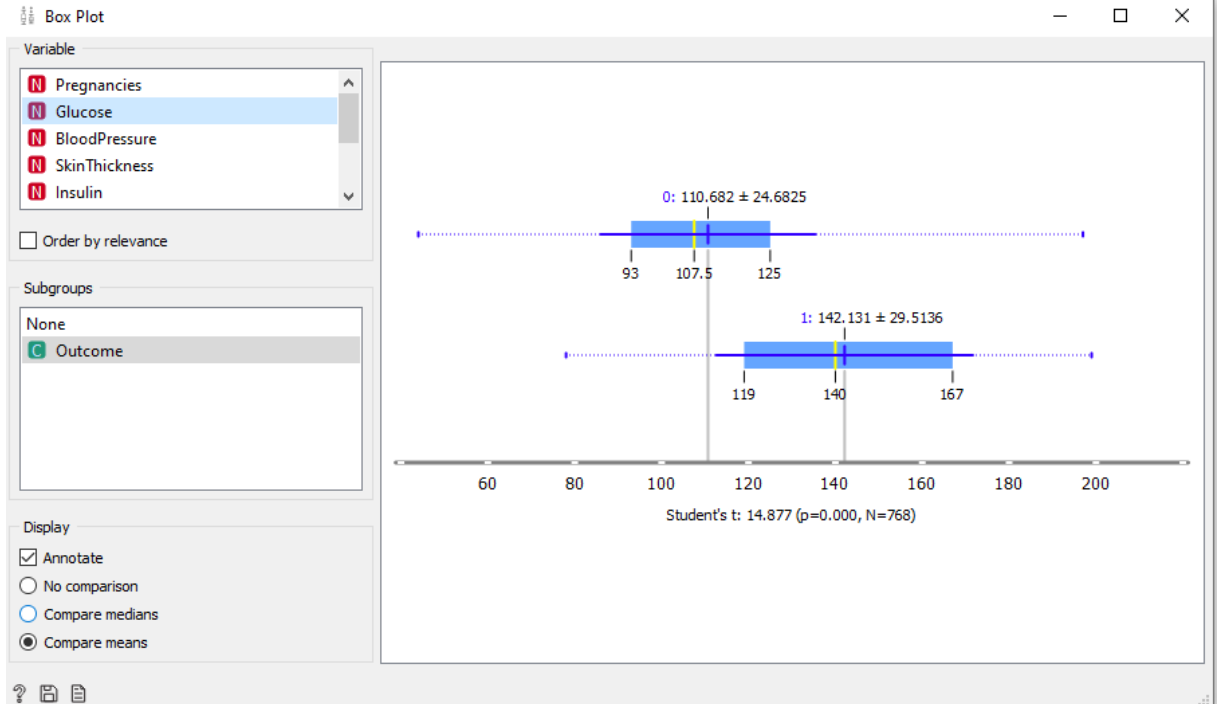
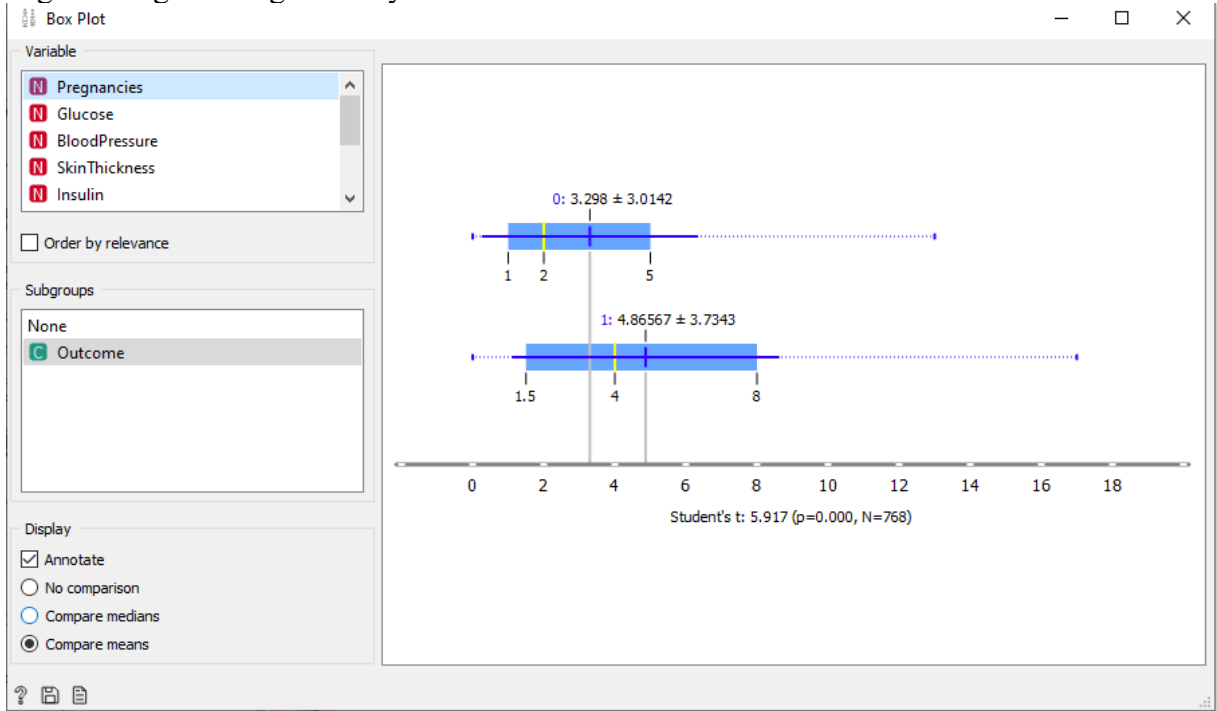
Bu mozaik grafikte yaş ve glikoz değişkenleri arttıkça diyabet hastalığı gözlenme oranı da artmıştır. Ayrıca yaşı genç olup glikoz miktarı 99,5'in altında olanlarda diyabet hastalığı düşük oranda gözlenmiştir.

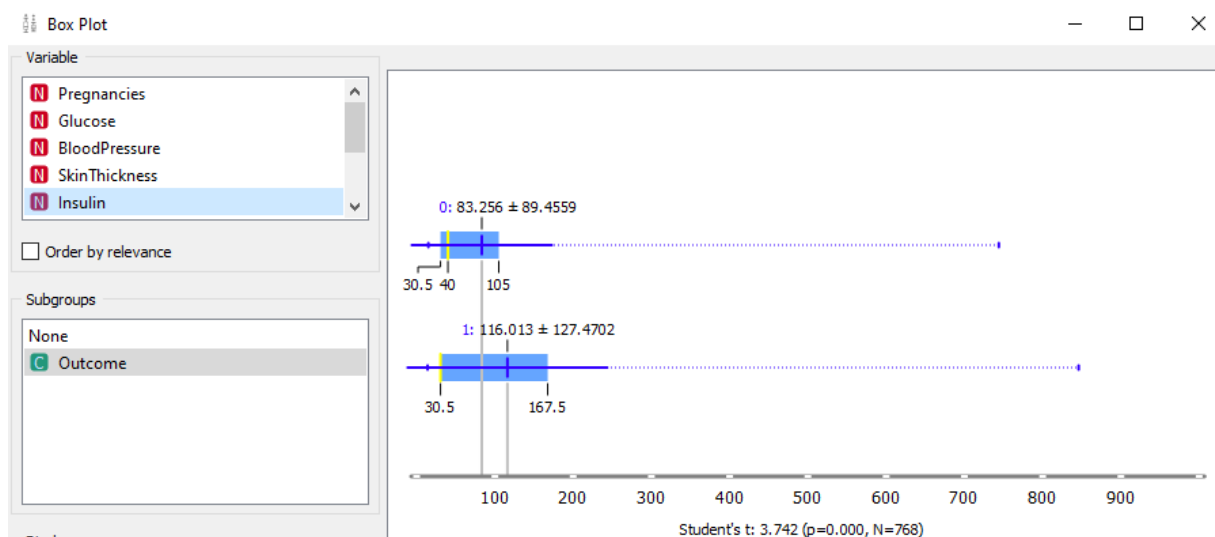
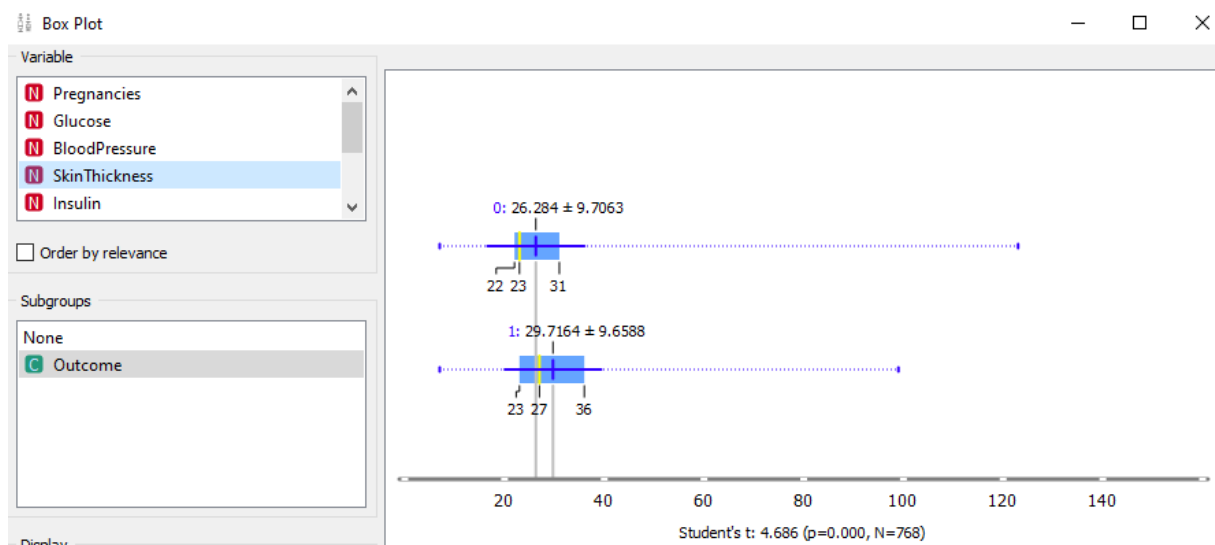
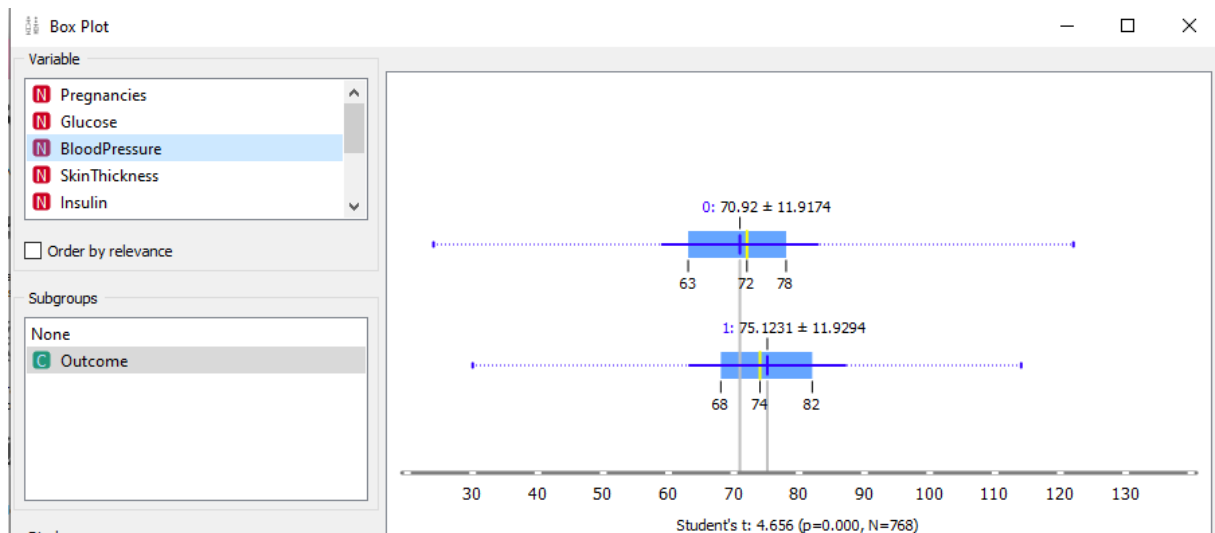


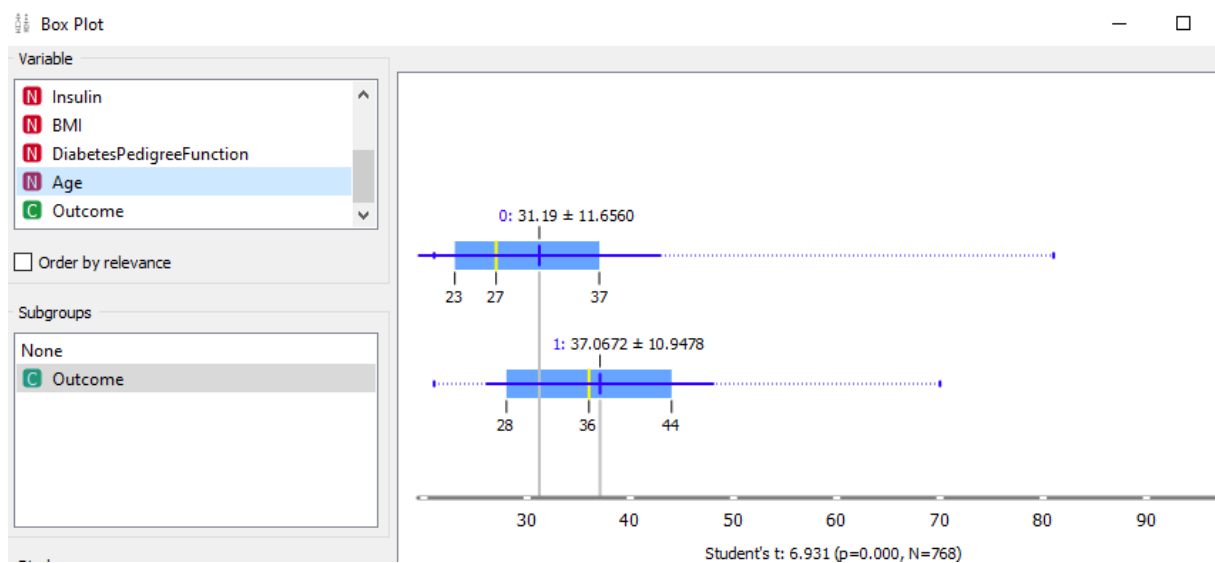
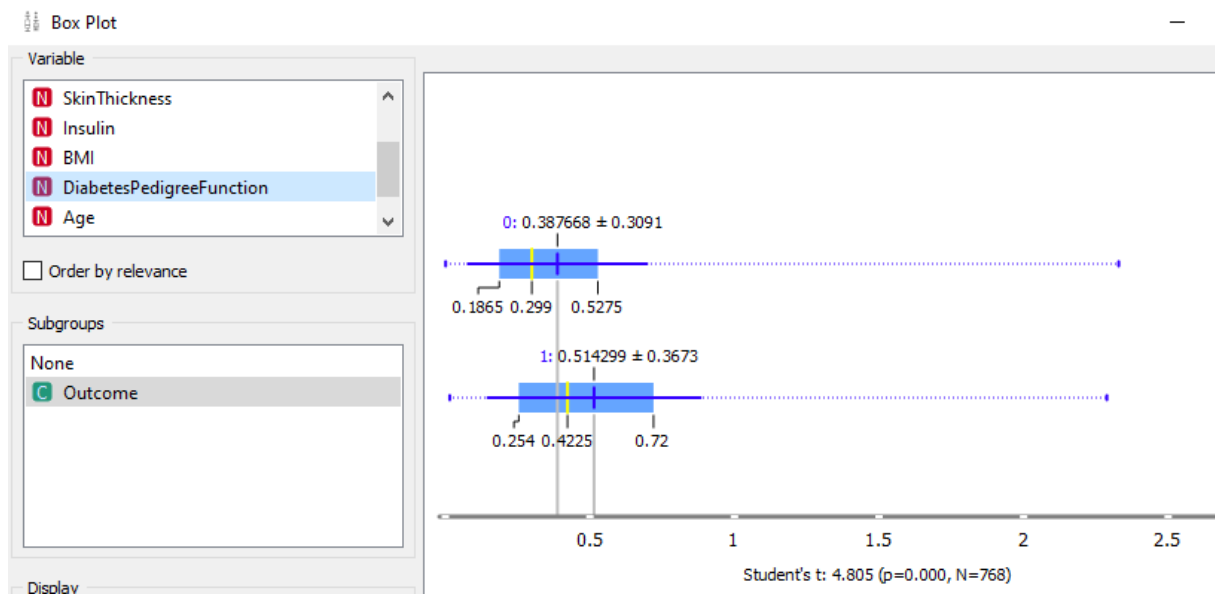
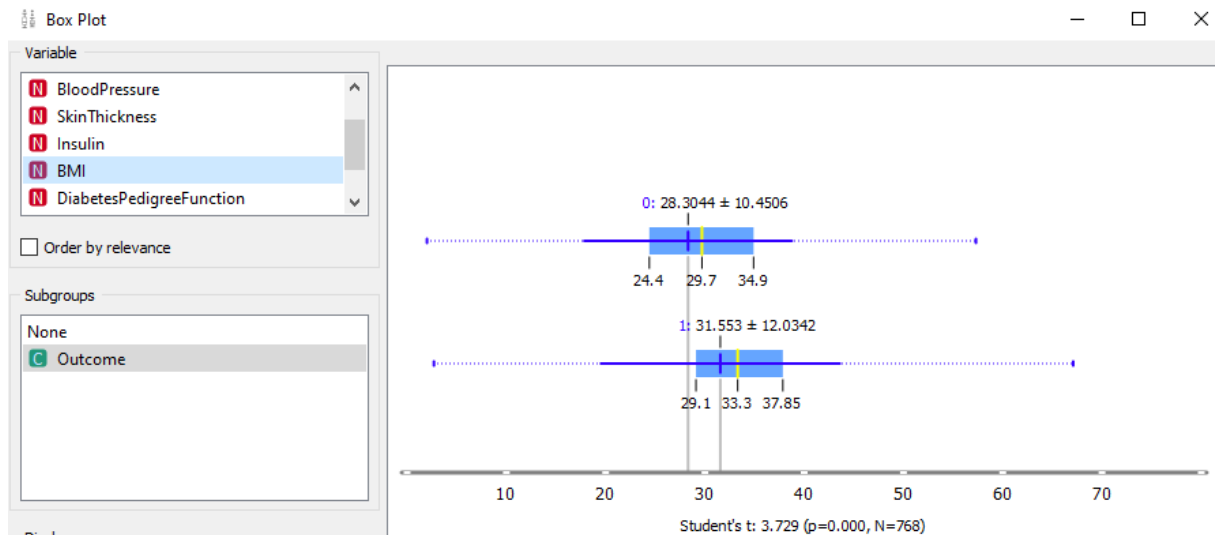
Yaşı 20-30 arasında olup kan basıncı 50-90 arasında değişen gözlemlerde diyabet hastalığı gözlenme oranı oldukça düşüktür.

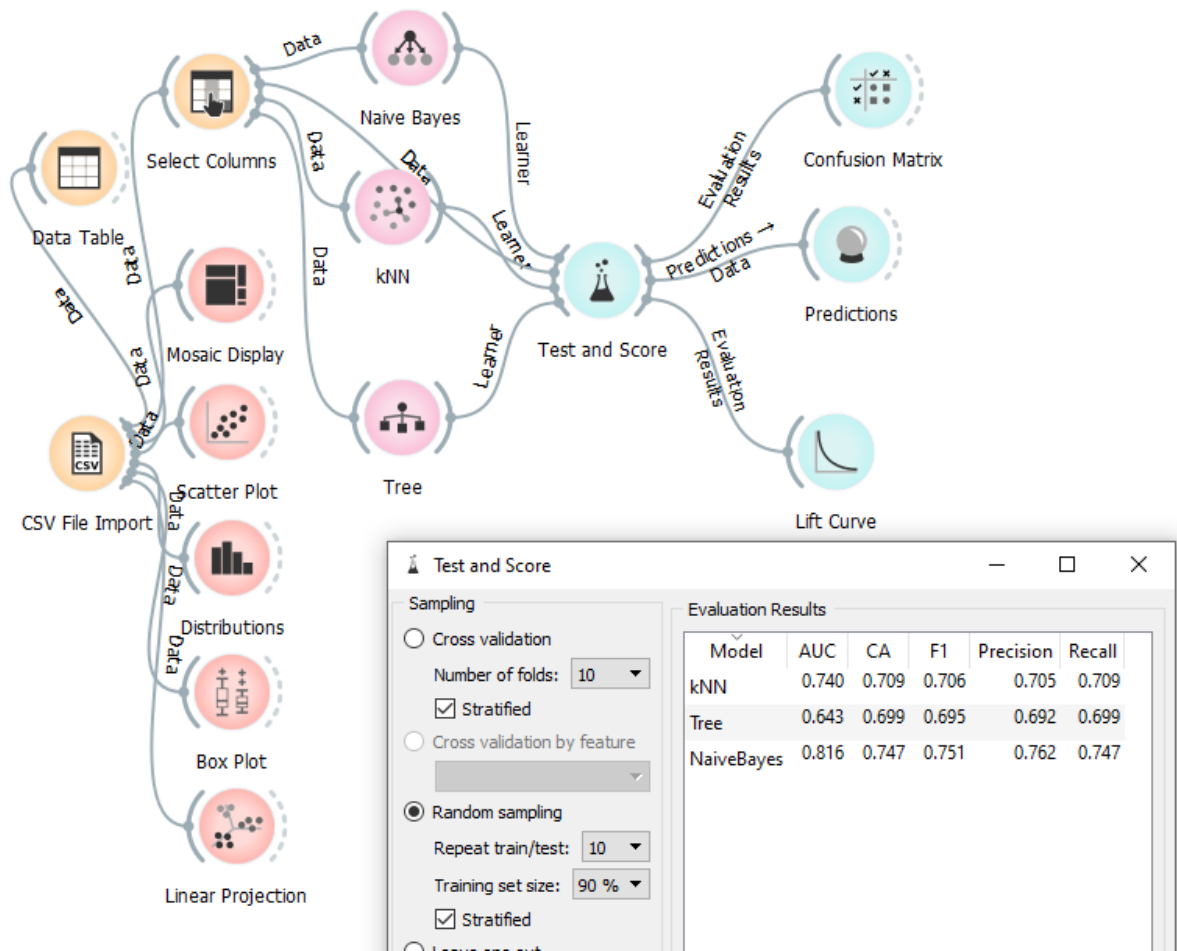


Aşağıdaki box-plot grafiklerinde her bir değişken için ortalama, çeyreklikler, min ve max değerlerini gösteren grafikler yer almaktadır.









Orange programı ile karar ağacı algoritması, naive bayes algoritması ve k en yakın komşu algoritmasının tahmin sonuçlarını karşılaştıracak olursak NaiveBayes algoritmasının 0.747'lik oranla en iyi tahmini yaptığını söyleyebiliriz.

Confusion Matrix

Model: kNN, Tree, NaiveBayes

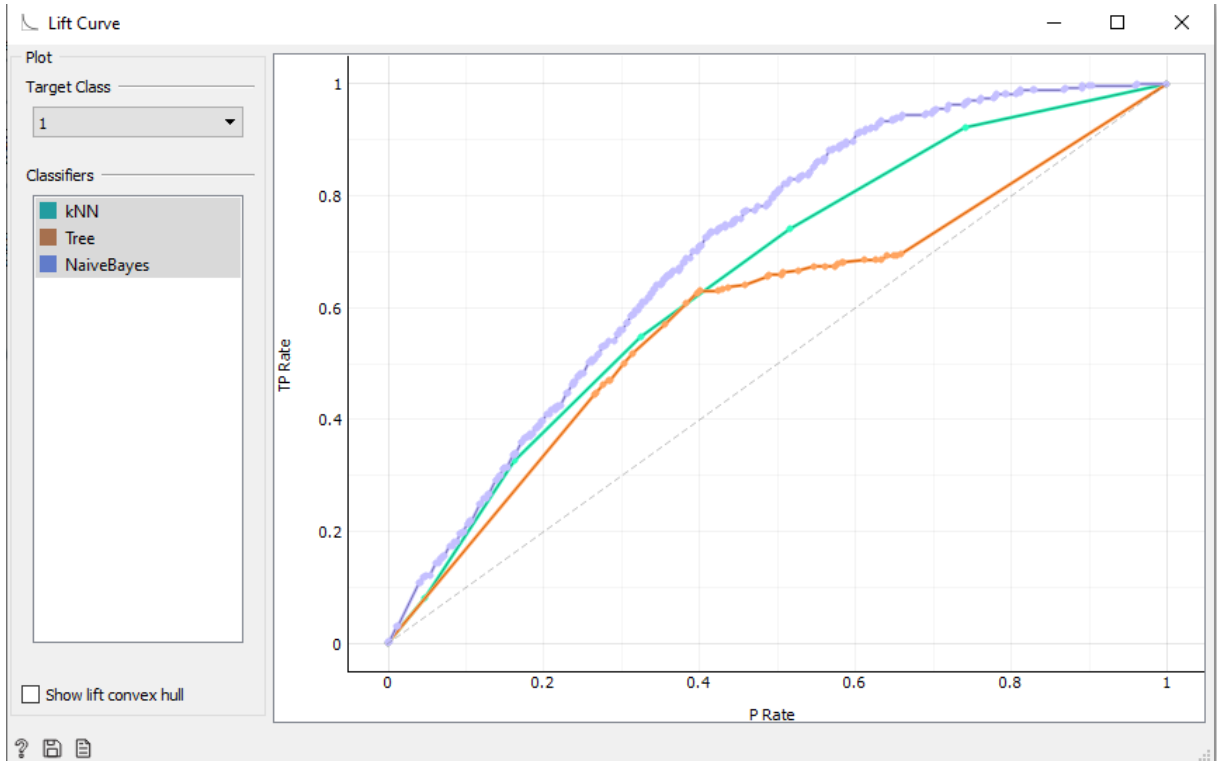
Output: ☒ Predictions ☐ Probabilities

☒ Send Automatically

Show: Number of instances

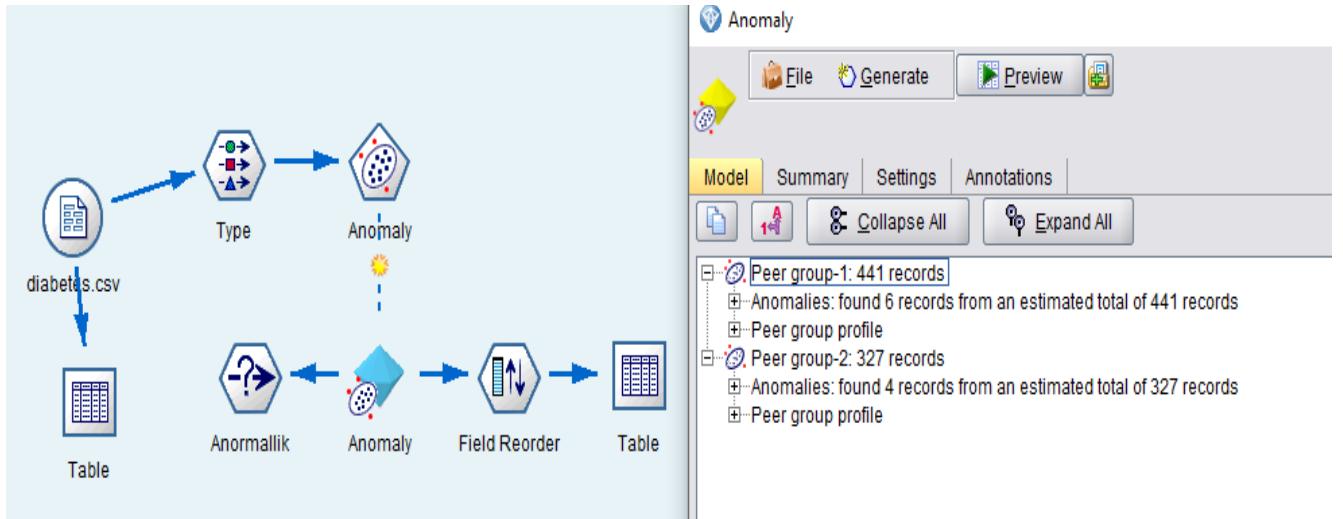
		Predicted		Σ
		0	1	
Actual	0	376	124	500
	1	71	199	270
Σ		447	323	770


Select Correct Select Misclassified Clear Selection







Grafiğe bakacak olursak naive bayes algoritması sol köşeye en yakın grafik çizgisini oluşturmuş bu durumda da diğer iki algoritmadan daha iyi sonuç verdiği gözlenmiştir.





2.3 Yapay Sinir Ağları ile Tahmin






 Anomaly

 File  Generate  Preview 

Model Summary Settings Annotations


   Collapse All  Expand All



  Peer group-1: 441 records


 Anomalies: found 6 records from an estimated total of 441 records

Contribution	Count	Average index
BloodPressure	3	0,189
DiabetesPedigreeFunction	2	0,559
Glucose	3	0,177
SkinThickness	4	0,43
Age	1	0,365
BMI	5	0,172

Residual of the unreported reasons: % 14,02


 Peer group profile

  Peer group-2: 327 records

 Anomalies: found 4 records from an estimated total of 327 records

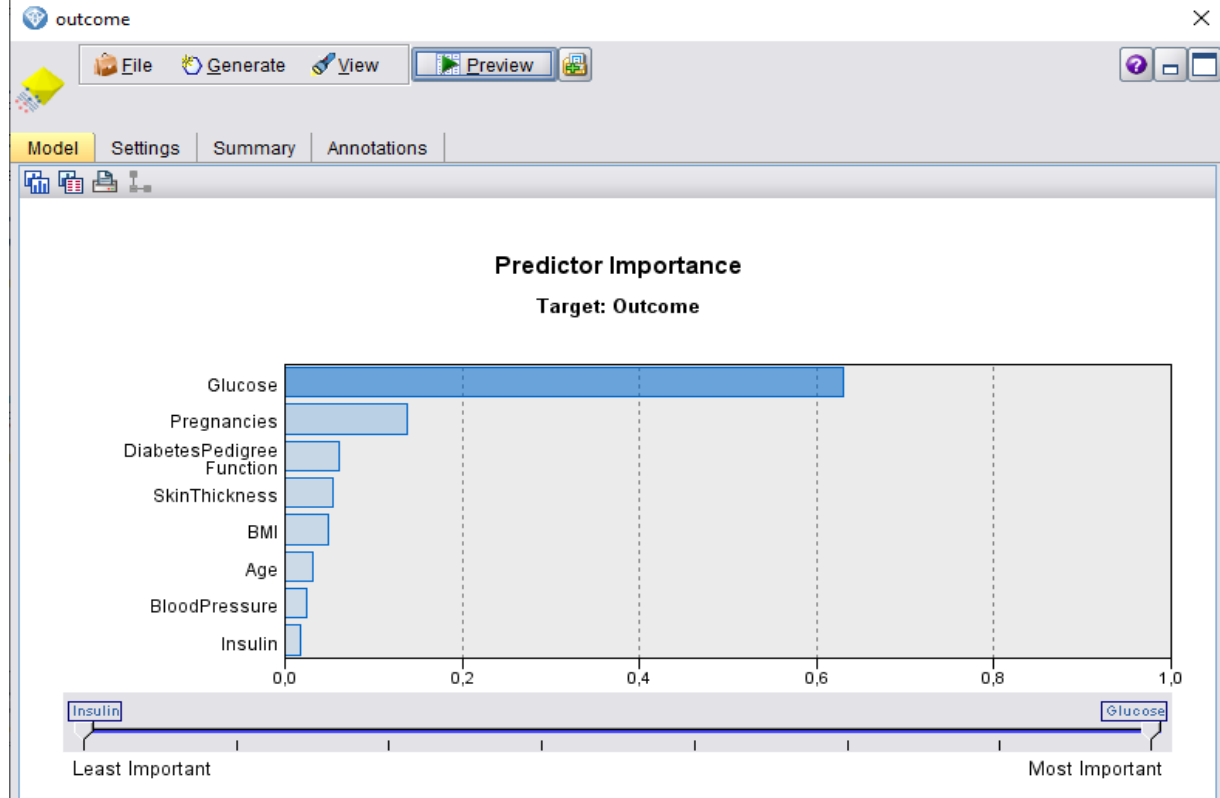
Contribution	Count	Average index
DiabetesPedigreeFunction	2	0,45
Glucose	2	0,073
SkinThickness	1	0,131
Insulin	4	0,433
Pregnancies	2	0,091
BMI	1	0,119

Residual of the unreported reasons: % 19,71

 Peer group profile

2.4 Destek Vektör Analizi

2.4.1 Radyal Tabanlı SVM



Deneğin diyabet hastası olup olmadığına dair en önemli özellik glikoz seviyesidir.

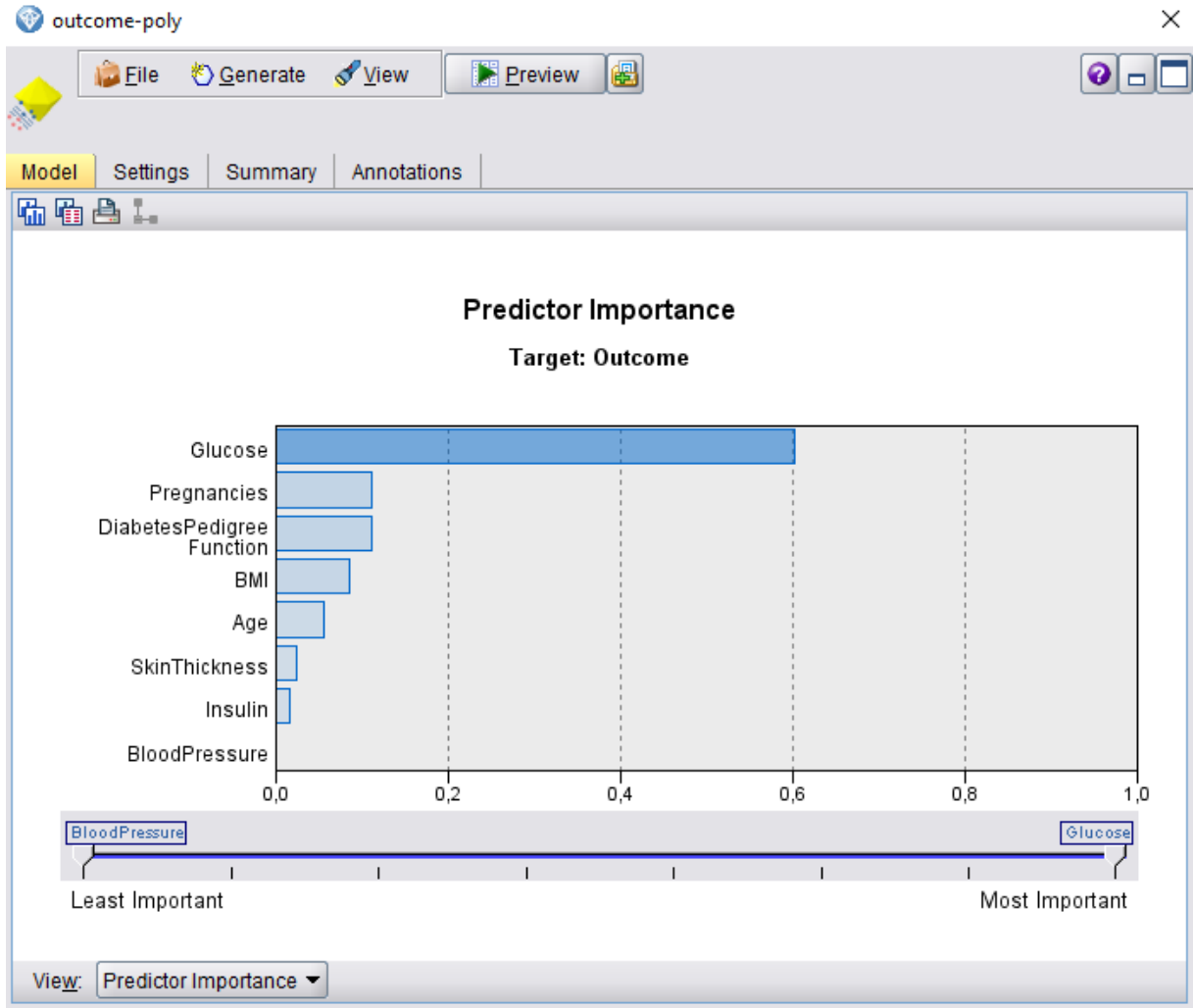
Table (11 fields, 768 records)

File Edit Generate

Table	Annotations								
	ure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	\$S-Outcome	
1	72	35	30.500	33....	0.627	50	1	1	
2	66	29	30.500	26....	0.351	31	0	0	
3	64	23	30.500	23....	0.672	32	1	1	
4	66	23	94.000	28....	0.167	21	0	0	
5	40	35	168.0...	43....	2.288	33	1	1	
6	74	23	30.500	25....	0.201	30	0	0	
7	50	32	88.000	3.1...	0.248	26	1	0	
8	72	23	30.500	35....	0.134	29	0	0	
9	70	45	543.0...	30....	0.158	53	1	1	
10	96	23	30.500	32....	0.232	54	1	0	
11	92	23	30.500	37....	0.191	30	0	0	
12	74	23	30.500	3.8...	0.537	34	1	1	
13	80	23	30.500	27....	1.441	57	0	1	
14	60	23	846.0...	30....	0.398	59	1	1	
15	72	19	175.0...	25....	0.587	51	1	1	
16	72	23	30.500	3.0...	0.484	32	1	0	
17	84	47	230.0...	45....	0.551	31	1	0	
18	74	23	30.500	29....	0.254	31	1	0	
19	30	38	83.000	43....	0.183	33	0	0	
20	70	30	96.000	34....	0.529	32	1	0	

Outcome değerlerimiz gerçek değerlerimiz iken \$S-outcome tahmin değerlerini gösterir. Örneğin 7. denek için gerçekte diyabet hastalığı var sonucu elde edilmiş fakat tahmin değeri diyabet hastası olmadığı yönündedir.

2.4.2 Polynomial Kernel SVM



Polinomial destek vektör modelinde de deneğin diyabet hastası olup olmadığına dair en önemli özellik glikoz seviyesidir.

Results for output field Outcome

Individual Models

Comparing \$S-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	536	% 78,13	57	% 69,51
Wrong	150	% 21,87	25	% 30,49
Total	686		82	

Comparing \$S1-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	609	% 88,78	54	% 65,85
Wrong	77	% 11,22	28	% 34,15
Total	686		82	

Comparing \$S2-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	495	% 72,16	56	% 68,29
Wrong	191	% 27,84	26	% 31,71
Total	686		82	

Comparing \$S3-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	530	% 77,26	56	% 68,29
Wrong	156	% 22,74	26	% 31,71
Total	686		82	

Agreement between \$S-Outcome \$S1-Outcome \$S2-Outcome \$S3-Outcome

'Partition'	1_Training		2_Testing	
Agree	510	% 74,34	61	% 74,39
Disagree	176	% 25,66	21	% 25,61
Total	686		82	

Comparing Agreement with Outcome

'Partition'	1_Training		2_Testing	
Correct	455	% 89,22	45	% 73,77
Wrong	55	% 10,78	16	% 26,23
Total	510		61	

\$S-Outcome radyal tabanlı modeldir ve doğru tahmin etme yüzdesi %69,51 olup tahminde daha başarılı olduğu gözlenmiştir. Yeni bir gözlem geldiğinde %69,51 olasılıkla radyal tabanlı sınıflandırma doğru tahmin yapacaktır. \$S1-Outcome ise polinomial, \$S2-Outcome sigmoid, \$S3-Outcome lineer modeldir. Destek vektör makinesi algoritması için dört model de %74,39 ortak atama yapmış olup bu ortak atamanın da %73,77'si doğru atanmıştır.

3. SPSS ile Analiz

3.1. ROC Eğrileri

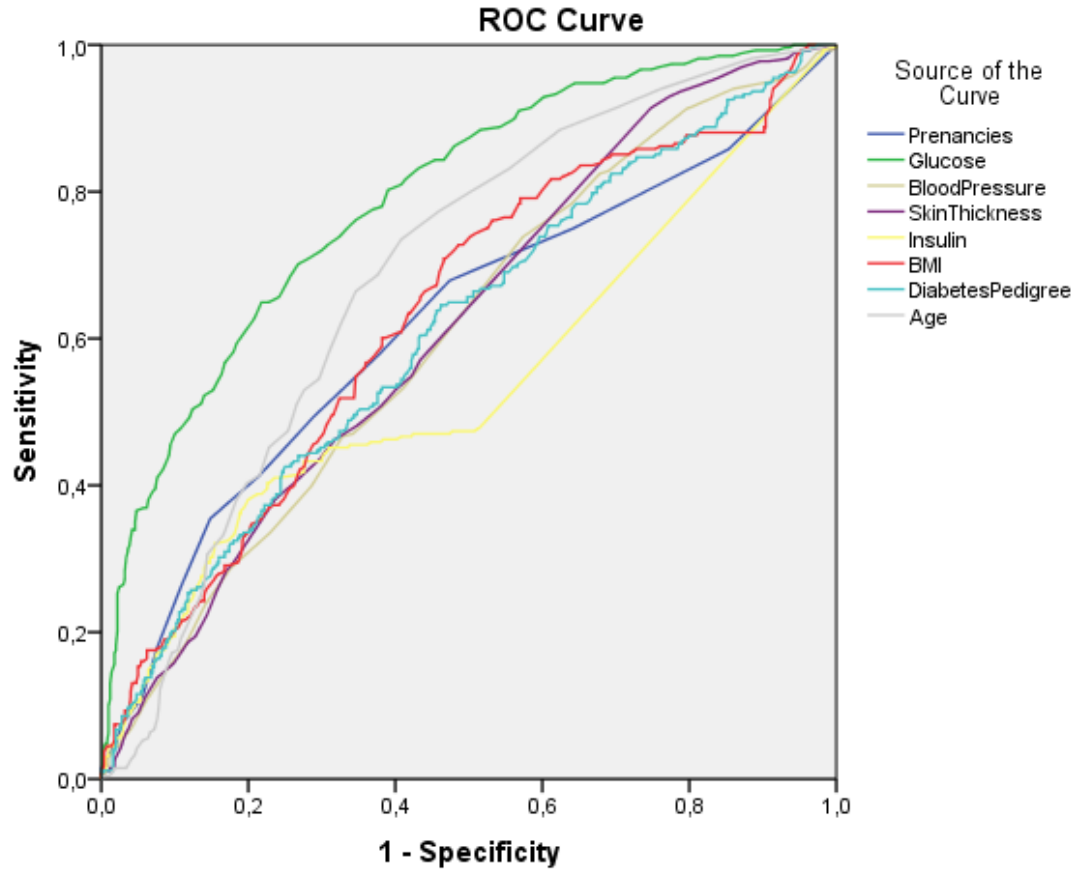
Case Processing Summary

Outcome ^a	Valid N (listwise)
Positive ^b	268
Negative	500

Larger values of the test result variable(s) indicate stronger evidence for a positive actual state.

- a. The test result variable(s):
BloodPressure has at least one tie between the positive actual state group and the negative actual state group.
- b. The positive actual state is 1.

Diyabet hastalığı pozitif olan 268 denek, diyabet hastalığı negatif olansa 500 denek vardır.

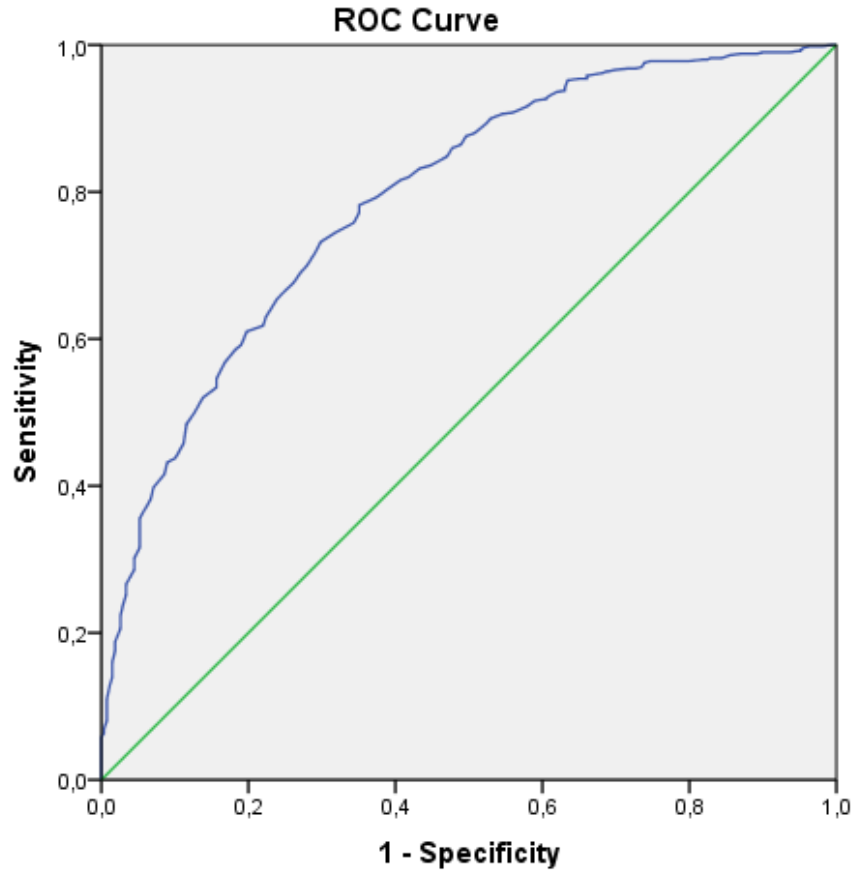


ROC eğrisine baktığımızda sol köşeye en yakın değişkenin glikoz değişkeni olduğunu görüyoruz.

Area Under the Curve	
Test Result Variable(s)	Area
Prenancies	,620
Glucose	,792
BloodPressure	,603
SkinThickness	,615
Insulin	,542
BMI	,632
DiabetesPedigree	,611
Age	,687

The test result variable(s): Prenancies,
Glucose, BloodPressure,
SkinThickness, Insulin, BMI,
DiabetesPedigree, Age has at least
one tie between the positive actual
state group and the negative actual
state group. Statistics may be biased.

Diyabet hastalığını araştırırken en etkili değer 0,792 ile glikozdur.
Glikoz için outcome=0 gözlemlere bakacak olursak grafiğimiz şöyledir;

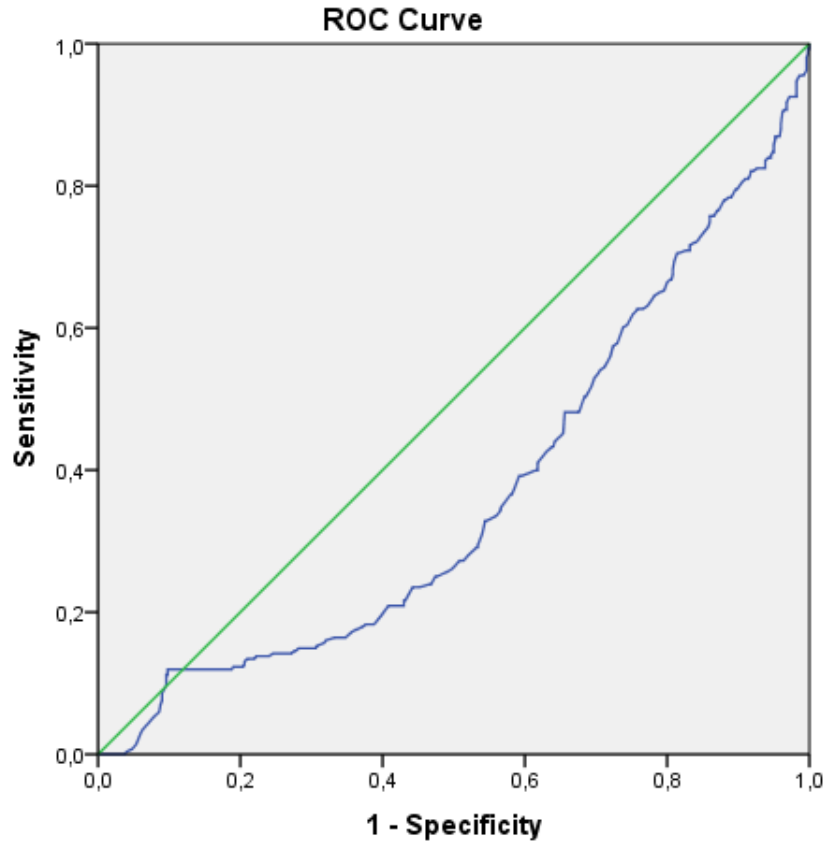


Diagonal segments are produced by ties.

Glikoz değişkeni için roc eğrisi sol köşeye yakındır testin ayırt etme gücü iyidir.



Diyabet hastası olan bir kişinin vücut kitle endeksi ne olmalıdır?
Bunun cevabı için BMI değişkeni ve Outcome=1 olan ROC eğrisine bakabiliriz.



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): BMI

Area
,368

The test result variable(s): BMI has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

BMI için 0,368 olan ROC puanını 0,5'ten küçük olduğundan bu değişken hastalığı test etmede yararlıdır diyemeyiz.

Area Under the Curve

Test Result Variable(s): BMI

Area	Std. Error ^a	Asymptotic Sig. ^b	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,368	,021	,000	,327	,410

The test result variable(s): BMI has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Burada BMI (vücut kitle endeksi) için anlamlılığa bakacak olursak sig. değeri 0,5'ten küçük olduğundan anlamlıdır diyebiliriz. Ayrıca 0,5 değerinin güven aralığında yer almadığını görürüz.

❖ CUT-OFF DEĞERİ

Excelde 1-(1-specificity) işlemiyle specificity sütunu oluşturulur ardından da sensitivity+specificity işlemiyle duyarlılığın ve seçiciliğin max. noktası bulunup buna karşılık gelen değer olan cut-off değeri bulunmuş olur.

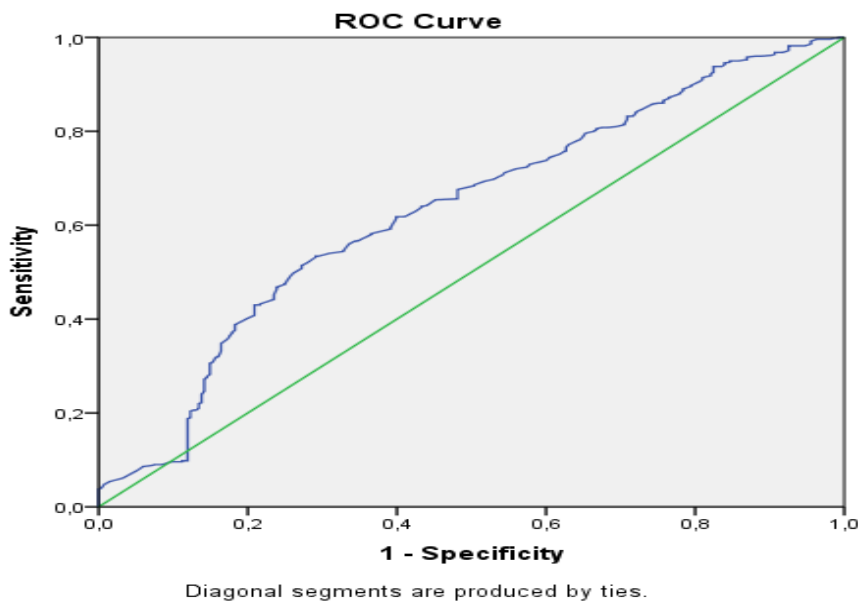
1	Coordinates of the Curve				
2	Test Result Variable(s): BMI				
3	Positive if Less Than or Equal To ^a	sensitivity	1 - Specificity	SPECIFICITY	sensitivity + specificity
4	1,0000	0,000	0,000	1,000	1,000
5	2,0500	0,000	,002	0,998	0,998
6	2,2000	0,000	,006	0,994	0,994
7	2,3500	0,000	,010	0,990	0,990
8	2,4500	0,000	,018	0,982	0,982
9	2,5500	0,000	,030	0,970	0,970
10	2,6500	0,000	,038	0,962	0,962
11	2,7500	,004	,040	0,960	0,964
12	2,8500	,007	,048	0,952	0,959
13	2,9500	,015	,054	0,946	0,961
14	3,0500	,030	,060	0,940	0,970
15	3,1500	,034	,062	0,938	0,972
16	3,3000	,052	,078	0,922	0,974
17	3,4500	,060	,086	0,914	0,974
18	3,5500	,071	,088	0,912	0,983

	A	B	C	D	E
11	2,7500	,004	,040	0,960	0,964
12	2,8500	,007	,048	0,952	0,959
13	2,9500	,015	,054	0,946	0,961
14	3,0500	,030	,060	0,940	0,970
15	3,1500	,034	,062	0,938	0,972
16	3,3000	,052	,078	0,922	0,974
17	3,4500	,060	,086	0,914	0,974
18	3,5500	,071	,088	0,912	0,983
19	3,6500	,075	,090	0,910	0,985
20	3,7500	,078	,090	0,910	0,988
21	3,8500	,086	,090	0,910	0,996
22	3,9500	,093	,094	0,906	0,999
23	4,0500	,097	,096	0,904	1,001
24	4,1500	,101	,096	0,904	1,005
25	4,3000	,104	,096	0,904	1,008
26	4,4500	,112	,096	0,904	1,016
27	4,7500	,112	,098	0,902	1,014
28	5,2500	,116	,098	0,902	1,018
29	11,8500	,119	,098	0,902	1,021
30	18,3000	,119	,104	0,896	1,015
31	18,7500	,119	,106	0,894	1,012

Diyabet hastası olan bir kişinin BMI (vücut kitle endeksi) değeri 11,85 olmalıdır.



Diyabet hastası olmayan bir kişinin vücut kitle endeksi ne olmalıdır?
Bunun cevabı için BMI değişkeni ve Outcome=0 olan ROC eğrisine bakabiliriz.



Area Under the Curve

Test Result Variable(s): BMI

Area	Std. Error ^a	Asymptotic Sig. ^b	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,632	,021	,000	,590	,673

The test result variable(s): BMI has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Burada BMI (vücut kitle endeksi) için anlamlılığa bakacak olursak sig. değeri 0,5'ten küçük olduğundan anlamlıdır diyebiliriz. Ayrıca 0,5 değerinin güven aralığında yer almadığını görürüz.

❖ CUT-OFF DEĞERİ


	A	B	C	D	E
1	Coordinates of the Curve				
2	Test Result Variable(s): BMI				
3	Positive if Less Than or Equal To ^a	Sensitivity	1 - Specificity	specificity	sensitivity +specificity
4	1,0000	0,000	0,000	1,000	1,000
5	2,0500	,002	0,000	1,000	1,002
6	2,2000	,006	0,000	1,000	1,006
7	2,3500	,010	0,000	1,000	1,010
8	2,4500	,018	0,000	1,000	1,018
9	2,5500	,030	0,000	1,000	1,030
10	2,6500	,038	0,000	1,000	1,038
11	2,7500	,040	,004	0,996	1,036
12	2,8500	,048	,007	0,993	1,041
13	2,9500	,054	,015	0,985	1,039
14	3,0500	,060	,030	0,970	1,030
15	3,1500	,062	,034	0,966	1,028
16	3,3000	,078	,052	0,948	1,026
17	3,4500	,086	,060	0,940	1,026
18	3,5500	,088	,071	0,929	1,017
19	3,6500	,090	,075	0,925	1,015
20	3,7500	,090	,078	0,922	1,012

	A	B	C	D	E
105	28,8500	,468	,239	0,761	1,229
106	29,0500	,474	,250	0,750	1,224
107	29,2500	,476	,250	0,750	1,226
108	29,4000	,484	,254	0,746	1,230
109	29,5500	,492	,257	0,743	1,235
110	29,6500	,498	,261	0,739	1,237
111	29,7500	,508	,272	0,728	1,236
112	29,8500	,514	,272	0,728	1,242
113	30,0000	,520	,280	0,720	1,240
114	30,1500	,532	,291	0,709	1,241
115	30,2500	,534	,291	0,709	1,243
116	30,3500	,534	,295	0,705	1,239
117	30,4500	,540	,310	0,690	1,230
118	30,6000	,544	,328	0,672	1,216
119	30,7500	,546	,328	0,672	1,218
120	30,8500	,560	,336	0,664	1,224
121	31,0000	,566	,343	0,657	1,223
122	31,1500	,566	,347	0,653	1,219
123	31,2500	,580	,366	0,634	1,214
124	31,4500	,582	,366	0,634	1,216
125	31,7500	,592	,392	0,608	1,200

Diyabet hastası olmayan bir kişinin BMI (vücut kitle endeksi) değeri 30,25 olmalıdır.

3.2 Kümeleme

❖ K-means

 K-Means Cluster Analysis

Variables:

- Zscore(Glucose) [ZGlucose]
- Zscore(Insulin) [ZInsulin]
- Zscore(BMI) [ZBMI]

Label Cases by:

- Outcome

Number of Clusters: 3

Method:

- ☒ Iterate and classify
- ☐ Classify only

Cluster Centers:

☐ **Read initial:**

- ☐ Open dataset
- ☐ External data file

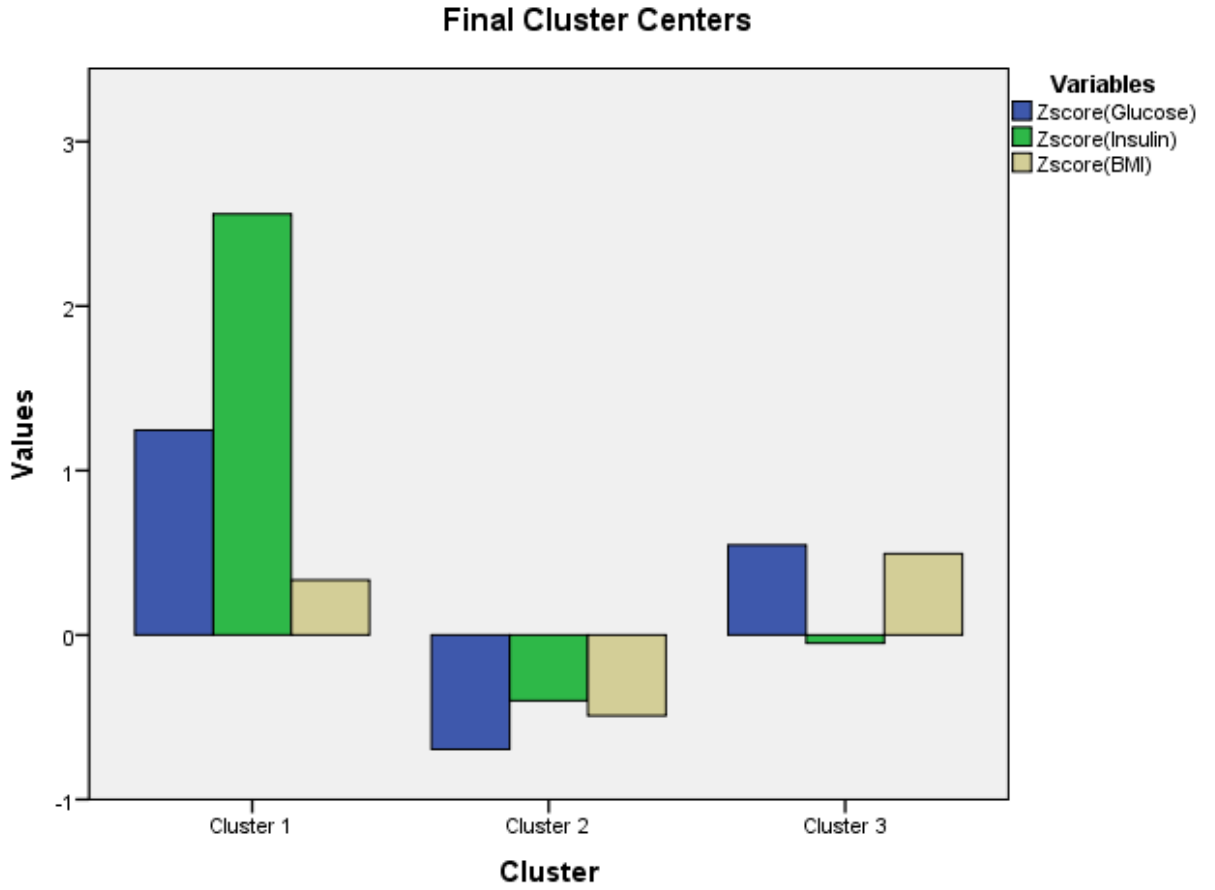
☐ **Write final:**

- ☐ New dataset

Buttons: Iterate..., Save..., Options...

Spss ile kümeleme analizinde ilk olarak k-means yöntemi ile 3 kümeye gruplama işlemi yapılır. Bunun için glikoz, insülin ve BMI değişkenleri için betimleyici istatistikler bulunup k-means sekmesinde bunların değerleri variablesa atılır.

Final Cluster Centers			
	Cluster		
	1	2	3
Zscore(Glucose)	1,24506	-,69431	,54706
Zscore(Insulin)	2,55925	-,39992	-,04994
Zscore(BMI)	,33384	-,49023	,49432



Grafiğe bakacak olursak 1. kümedeki insülin, BMI ve glikozdan yüksek çıkmıştır.

2. kümede ise bütün değerler negatif çıkmıştır.

3. kümede insülin negatif iken glikoz ve BMI pozitif olup birbirine çok yakındır.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore(Glucose)	189,850	2	,506	765	374,994	,000
Zscore(Insulin)	243,264	2	,367	765	663,515	,000
Zscore(BMI)	88,757	2	,771	765	115,183	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Glikoz, insülin ve BMI için sig. değerleri anlamlı olduğundan 3 kümeye gruplama işlemi yapılabilir.

Number of Cases in each

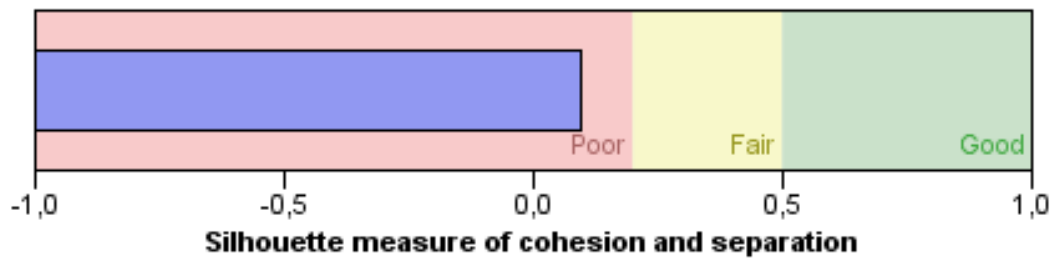
Cluster		
	1	65,000
Cluster	2	375,000
	3	328,000
Valid		768,000
Missing		,000

❖ Two-Step

Model Summary

Algorithm	TwoStep
Inputs	8
Clusters	3

Cluster Quality



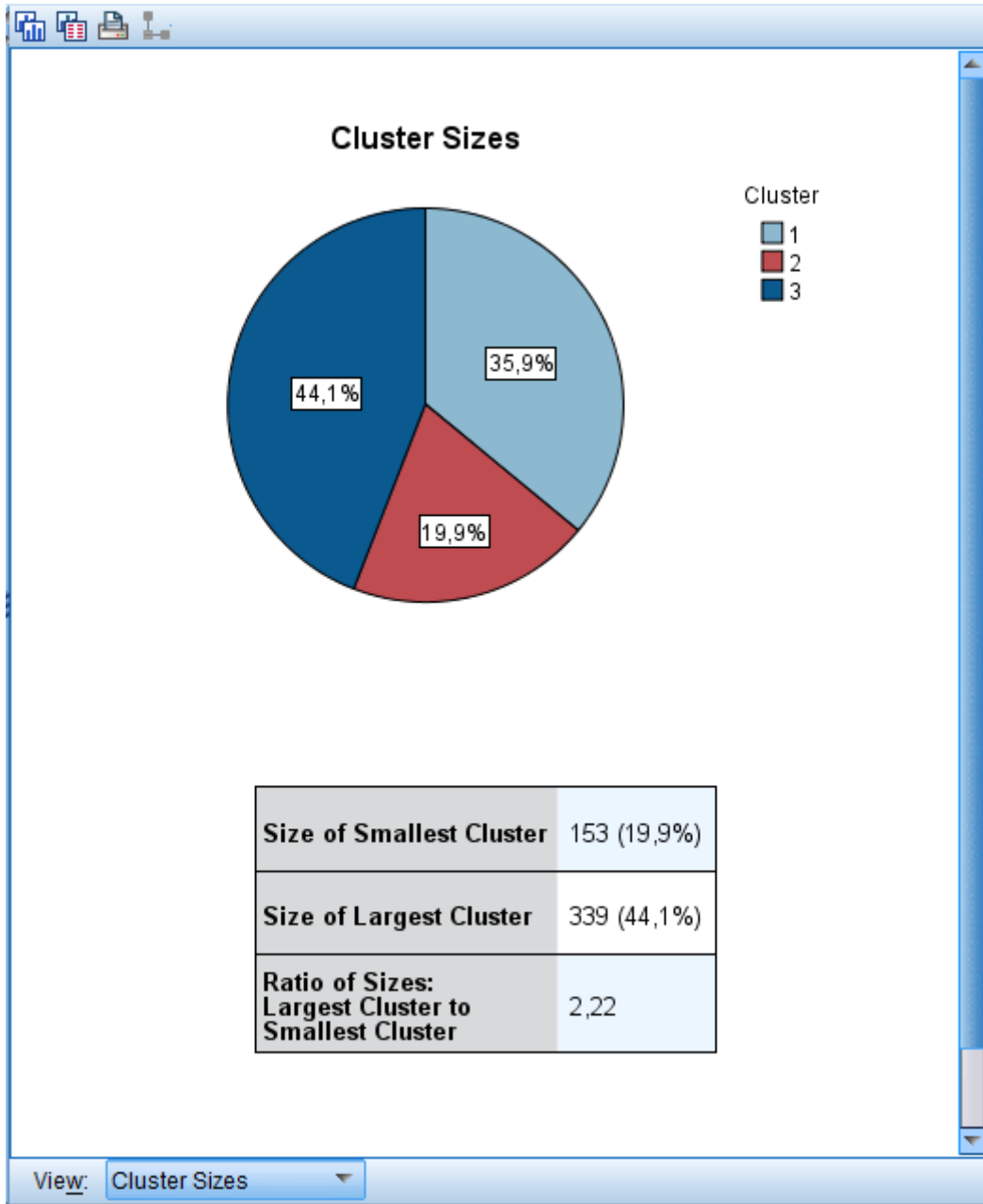
8 değişkeni kullanarak 3 kümeye ayrılmıştır.

Burada cluster quality yani algoritmanın kümelemeyi ne kadar kaliteli yaptığına dair bilgi yer alır kümelememiz poor çıkmıştır, biz fair ve gooda daha yakın olmasını isteriz çünkü ne kadar sağa yakın olursa o kadar küme içi benzerlik yani homojenlik fazla olur.



Kayıtlı hastalar kendi aralarında gruplanabilir mi?

Elimizdeki verinin olabildiğince homojen gruplara ayrılıp ayrılmadığını görmek için kümeleme analizinden yararlanabiliriz.



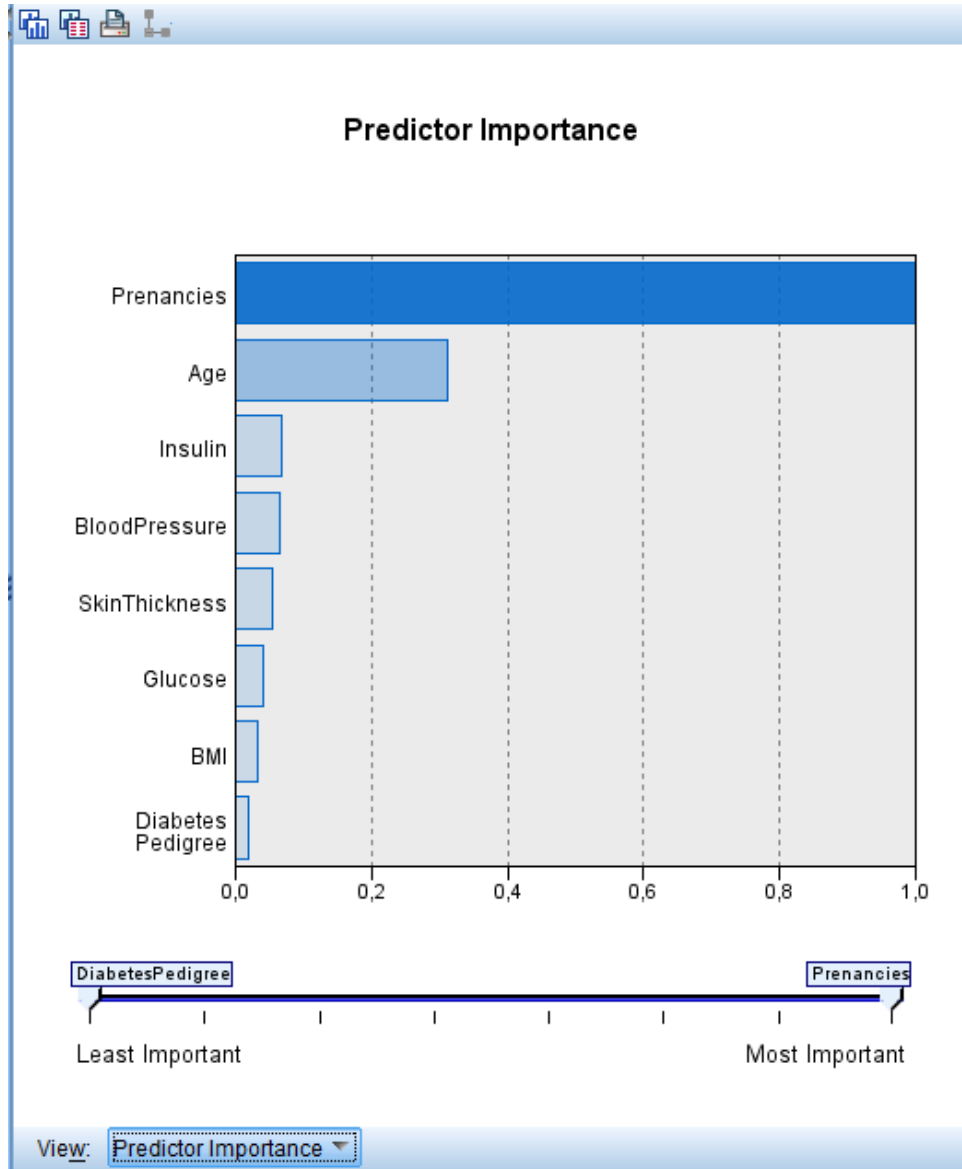
Burada en küçük ve en büyük kümenin kaç örneklemden oluştuğunu yani boyutunu gösterir.

3. küme 339 nesne (%44,1) ile en büyük boyutlu kümedir.

1. küme 276 nesne (%35,9) ile orta boyutlu bir kümedir.

2. küme 153 nesne (%19,9) ile en küçük boyutlu kümedir.

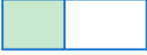
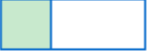

2,22 ise en büyük kümenin en küçük kümeye oranıdır.



Kümeleme yapılırken en önemli etkenin hamilelik sayısı olduğu çıkarsanmaktadır.

Clusters

Input (Predictor) Importance
 1,0 0,8 0,6 0,4 0,2 0,0

Cluster	3	1	2
Label			
Description			
Size	 44,1% (339)	 35,9% (276)	 19,9% (153)
Inputs	Prenancies 1 (32,2%)	Prenancies 5 (20,3%)	Prenancies 0 (72,5%)
	Age 27,46	Age 42,58	Age 29,21
	Insulin 83,26	Insulin 73,40	Insulin 158,40
	BloodPressure 68,50	BloodPressure 76,35	BloodPressure 73,84
	SkinThickness 25,14	SkinThickness 27,70	SkinThickness 32,27
	Glucose 113,68	Glucose 126,56	Glucose 130,50
	BMI 27,54	BMI 29,28	BMI 33,93
	DiabetesPedigree 0,40	DiabetesPedigree 0,42	DiabetesPedigree 0,53

Burada da her bir kümeye ilişkin her bir değişken açısından özellikler yer alır. Mesela hamilelik sayısını ele alırsak üçüncü kümede hamilelik sayısı 1 olanlar, birinci kümede hamilelik sayısı 5 olanlar ve ikinci kümede hamilelik sayısı 0 olanlar en yaygın olanlardır yorumunu yapabiliriz.

KAYNAKÇA

- <https://www.kaggle.com/avinash2203/pima-diabetes-dataset-exploratory-data-analysis>
- H. Yılmaz, O. 2014, “Random Forest Yönteminde Kayıp Veri Probleminin İncelenmesi Ve Sağlık Alanında Bir Uygulama”, Osmangazi Üniversitesi Biyoistatistik Anabilim Dalı Yüksek Lisans Tezi
- Larose, D. T. 2005. Discovering Knowledge in Data: An Introduction in Data Mining, Wiley, USA
- Savaş, S., Topaloğlu N., Yılmaz, M. 2012. “Veri Madenciliği ve Türkiye’deki Uygulama Örnekleri,” İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, sayı 21, s. 1-23
- T. Wendler, S. Gröttrup, Data Mining with SPSS Modeler: Theory, Exercises and Solutions
- Gökay, G. E. ve Taşkın, Ç., 2005, Veri madenciliğinde karar ağaçları ve bir satış analizi, Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, 6, 2, 221-239 s.
- Silahtaroglu, Gökhan. Veri Madenciliği (Kavram ve Algoritmaları) / Gökhan Silahtaroglu. - İstanbul: Papatya Yayıncılık Eğitim, 2013
- Spss ile Kümeleme (cluster) Analizi, <https://www.youtube.com/watch?v=haThRKBNpk0>
- M. Majnik, Z. Bosni’c, ROC Analysis of Classifiers in Machine Learning: A Survey, Technical report MM-1/2011