T.C.

MERSİN ÜNİVERSİTESİ

ERDEMLİ UYGULAMALI TEKNOLOJİ VE İŞLETMECİLİK YÜKSEK OKULU

BİTİRME PROJESİ

Proje Danışmanı:

Öğretim Görevlisi Vedia Bennu GİLAN

19701801 Mikail DOĞRUER

Mersin, 2021

İçindekiler

İNME GEÇİRME ANALİZ ve TAHMİNLEME ÇALIŞMASI

1. Problem Tanımı ve Veri Seti Hikayesi

Dünya Sağlık Örgütü'ne göre felç, ölümlerin yaklaşık %11'ini oluşturan ve önde gelen 2. ölüm nedenidir. Erken teşhis sayesinde ilaçlar, beslenme şeklini değiştirme veya fiziksel aktiviteler ile bu oranın azaltılmasını amaçlayan çalışmadır.

2. Felç Nedir?

Felç (inme), beyin ya da kalbin belli bir bölgesinde kan akışının azalması ya da kesilmesi sonucunda gerçekleşen ani krizlerdir. Dünya genelinde ölüme en çok neden olan rahatsızlıklardan biri olan felcin kalıcı sakatlanmalara da neden olduğu bilinmektedir. Bu nedenle felç riskinin önceden belirlenmesi ölüm ya da kalıcı sakatlık riskinin azaltılması için oldukça önemlidir.

3. Veri Seti Öznitelikleri

Bu çalışmada Kaggle veri deposunda yaş, cinsiyet, sigara içme durumları gibi özniteliklere göre 5110 adet bireyin kayıtlarını içeren "Stroke Prediction Dataset" veri seti kullanılmıştır. Veri seti felç geçirme risk sınıflandırması ve tahmini için kullanılan 10 adet girdi ve felç geçirme durum etiketlerini belirten 1 adet çıktı olmak üzere toplam 11 adet öznitelikten oluşmaktadır. Bu öznitelikler ve bu özniteliklere ait açıklamalara aşağıda yer verilmiştir.

3.1**Yaş:** Bu öznitelik bir bireyin yaş bilgisinin içermektedir. Sayısal türde verilerden oluşmaktadır.

3.2 **Cinsiyet:** Bu öznitelik bir bireyin cinsiyeti bilgisini içermektedir. Kategorik türde verilerden oluşmaktadır.

3.3 **Hipertansiyon:** Bu öznitelik bir bireyin hipertansiyon rahatsızlığı olup olmadığı bilgisini içermektedir. Sayısal türde verilerden oluşmaktadır (1 – Hipertansiyon var, 0 – Hipertansiyon yok).

3.4 **İş Türü:** Bu öznitelik bireyin meslek bilgilerini içermektedir (Devlette çalışıyor, çalışmıyor, serbest meslek vb.). Kategorik türde verilerden oluşmaktadır.

3.5 **Kalp Hastalığı:** Bu öznitelik bir bireyin kalp hastalığı olup olmadığı ile ilgili bilgisini içermektedir (0 – Kalp hastalığı yok, 1 – kalp hastalığı var). Sayısal türde verilerden oluşmaktadır.

3.6 **Evlilik Durumu:** Bu öznitelik bir bireyin evlilik durumu bilgisini içermektedir (Evet-Hayır). Kategorik türde verilerden oluşmaktadır.

3.7 **Yaşanılan Bölge:** Bu öznitelik bir bireyin yaşadığı bölge bilgisini içermektedir (Kent-Kırsal). Kategorik türde verilerden oluşmaktadır.

3.8 **BMI:** Bu öznitelik bir bireyin vücut kitle indeksi bilgisini içermektedir. Sayısal türde verilerden oluşmaktadır.

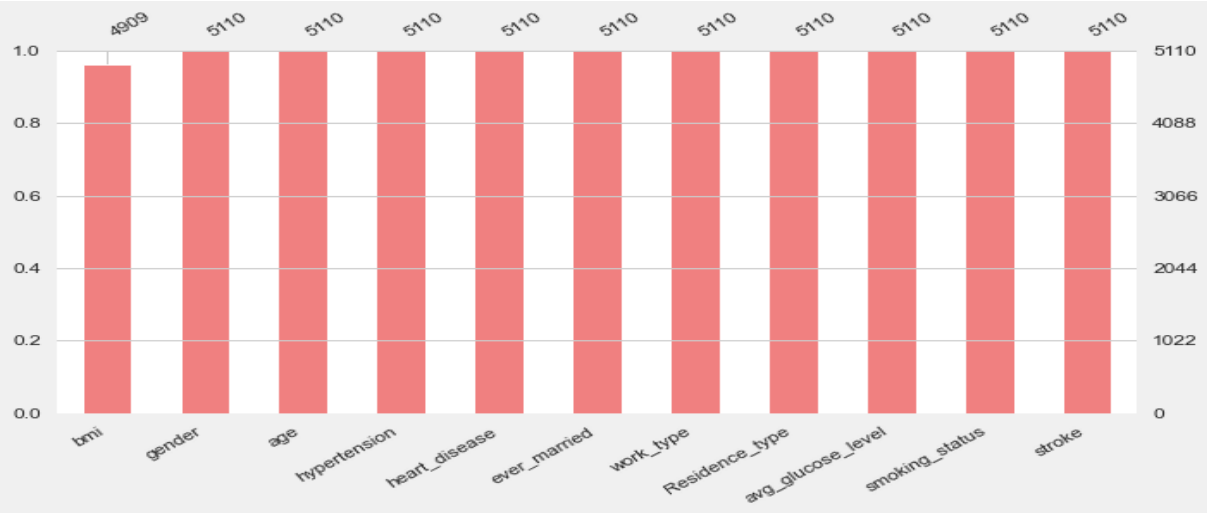
3.9 **Ortalama Glikoz Seviyesi:** Bu öznitelik bir bireyin kanındaki ortalama glikoz seviyesi bilgisini içermektedir. Sayısal türde verilerden oluşmaktadır.

3.10 **Sigara Kullanma Durumu:** Bireyin sigara kullanıp kullanmadığı bilgisini içeren verilerden oluşmaktadır (Daha önce sigara kullandı-Hala kullanıyor-Hiç kullanmamış). Kategorik türde verilerden oluşmaktadır.

3.11 **Felç Durumu:** Bu öznitelik bir bireyin daha önce felç geçirip geçirmediği bilgisini içermektedir (1- Felç geçirdi, 0-Felç geçirmedi). Sayısal türde verilerden oluşmaktadır.

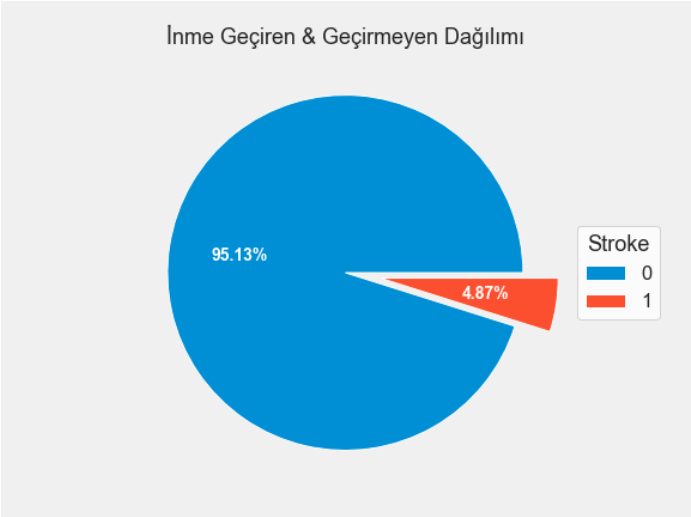
4. Özniteliklerin Eksik Değerleri ve Doldurulması

Yukarıda bahsedildiği üzere 11 öznitelik ve toplamda 5110 gözlem değerine sahip olan veri setinde "bmi" özniteliğinde 201 adet kayıp değerimiz bulunmaktadır. Özniteliklerin dağılımına ilişkin yapılan Shapiro Wilk Testine göre vücut kitle indeksi: Normal (Gaussian) dağılımdan gelmemektedir. Bu nedenle eksik olan değerleri medyan(ortanca) değeriyle doldurma işlemi yapılmıştır.



5. Hedef Öznitelik Dağılımı

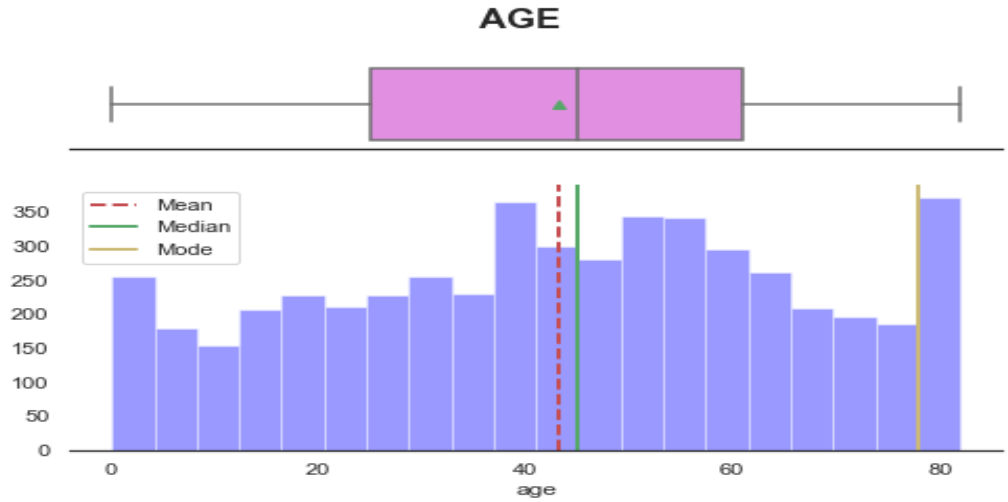
Analiz için belirlenen veri setinde ele alınan gözlem değerlerinde inme geçirenlerin sayısı 249 toplam veri setine oranla 4,87% iken inme geçirmeyenlerin sayısı 4861 yani veri setinin 95,13% oluşturmaktadır. Söz konusu veri setinde tahmin işlemi için dengesizlik (Unbalanced) söz konusudur. İleri ki aşamalarda değineceğimiz SMOTE (Synthetic Minority Over-Sampling Technique) ile hedef değişkenimizi dengeleyici sentetik veri üretimi yapılacaktır.



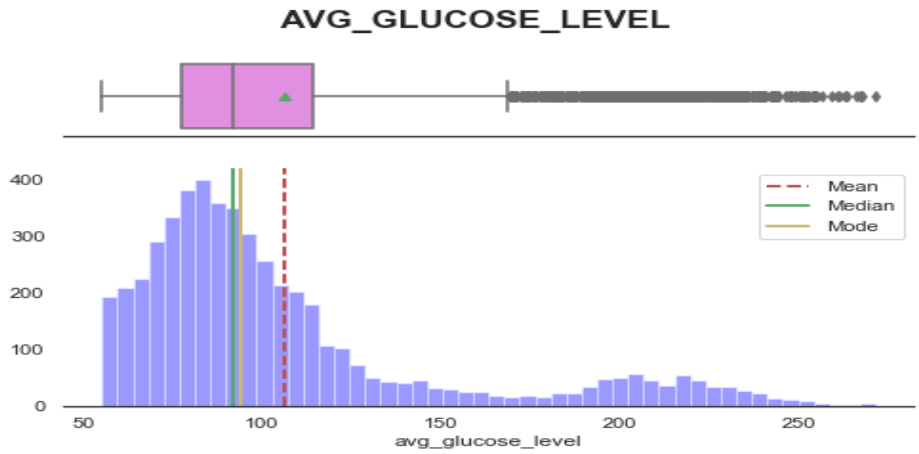
6. Sürekli Değere Sahip Özniteliklerin Dağılımları

Veri seti içerisinde bahsedildiği üzere 3 öznitelik sürekli değerlere sahiptir. Bu öznitelikler: **Yaş**, **Ortalama Glikoz Seviyesi** ve **Vücut Kitle İndeksi** değerleridir. Bu özniteliklerin dağılımına bakılarak veri seti hakkında fikir edinilecektir.

6.2 Yaş**:** Öznitelik değerlerine bakıldığında; Veri seti ortalama yaşı 43, medyan değeri 45 ve en çok tekrar eden yaş ise 78 olarak karşımıza çıkmaktadır



6.3 **Ortalama Glikoz Seviyesi:** Öznitelik değerlerine bakıldığında: Ortalama değeri 106, medyan değeri 91 ve en çok tekrar eden, görülme sıklığı çok olan değer ise 93 glikoz seviyesidir



6.4 **Vücut Kitle İndeksi:** Dünya Sağlık Teşkilatının VKİ oranlarına göre:

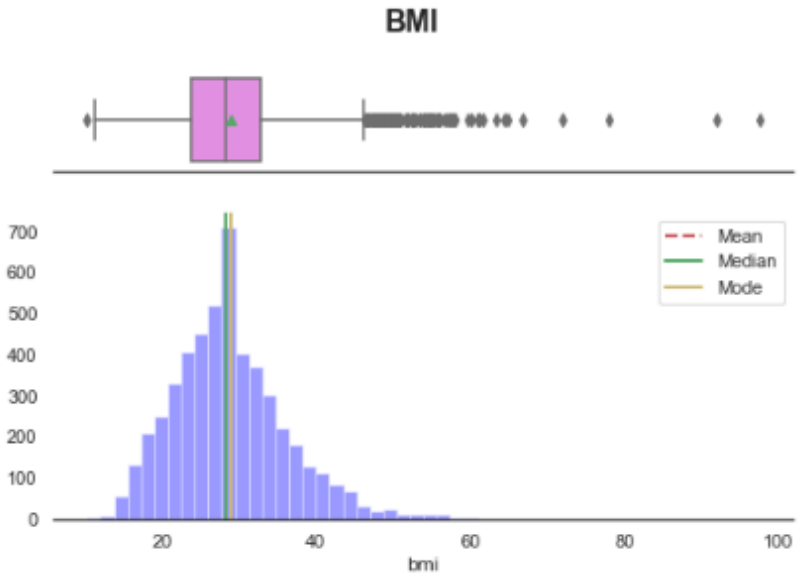
18, 5 kg/m2 altında olanlar: Zayıf

18, 5-24, 9 kg/m2: Normal kilolu

25-30 kg/m2: Fazla kilolu

30-35 kg/m2: Tip 1 Obez

35-40 kg/m2: Tip 2 Obez

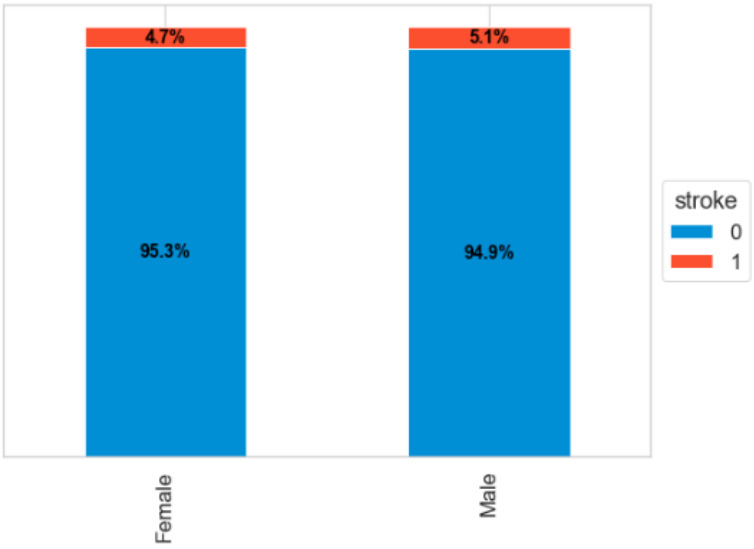


Bu öznitelik değerlerine bakıldığında, rassallığa bağlı olarak ortalama, medyan ve en çok tekrar eden değer 28 olarak görülmektedir. Dünya Sağlık Teşkilatına göre fazla kilolu sınıflamasına girmektedir.

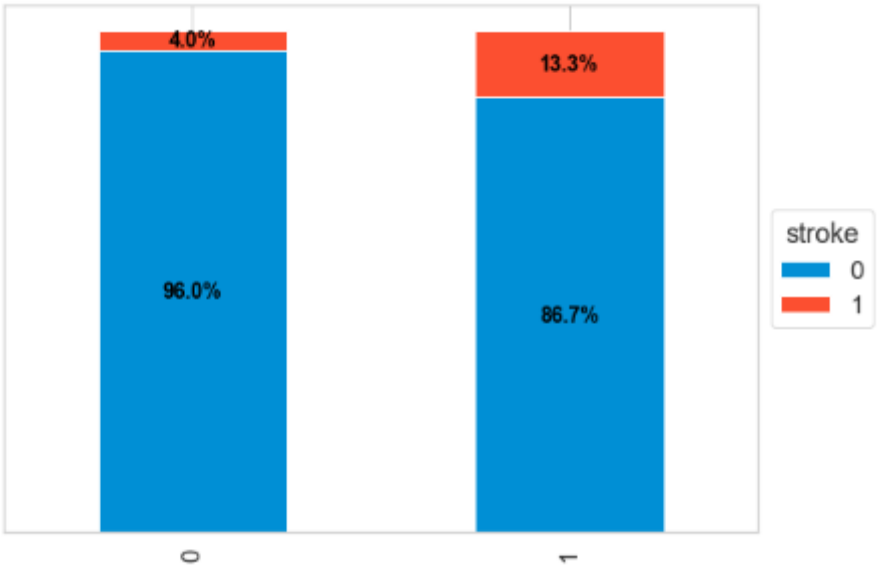
7. Kategorik Değere Sahip Özniteliklerin Dağılımları

Veri seti çeşitliliğe bağlı olarak kategorik ve nominal özniteliklerden oluşmaktadır. Bunlardan kategorik olanları: **Cinsiyet**, **Hipertansiyon, Evlilik Durumu, Yaşanılan Bölge, Sigara, İş Türü Kullanma Durumu.** Nominal Öznitelikler ise**: Hipertansiyon, Kalp Hastalığı** özniteliklerdir. Veri seti içerisinde kategorik ve nominal özniteliklerin bütün veri setine oranına ve bilgilerine değinilecektir.

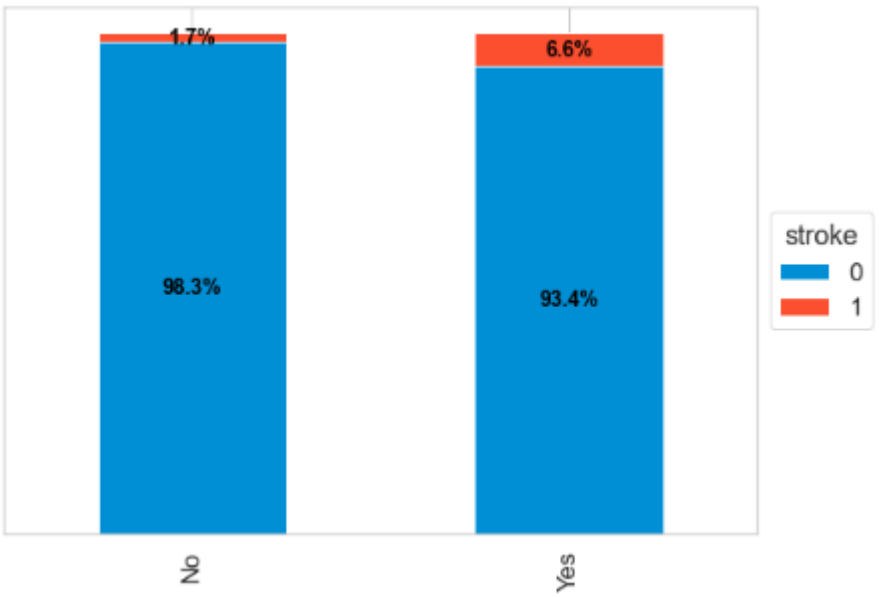
7.1 **Cinsiyet:** Cinsiyet özniteliğine bakıldığına erkeklerin kadınlara oranla felç geçirme dağılımı daha fazladır.



7.2 **Hipertansiyon:** Hipertansiyon hem koroner ateroskleroz yani damar sertliği gelişimini hızlandırır hem de artan kas dokusu koroner dolaşımda değişikliklere yol açar. Koroner damarlarda direnç artar ve rezerv azalır. Bütün bu nedenlere bağlı olarak hipertansiyonu olan kişilerde kalbin beslenmesi bozulabilir. Bu öngörüyü destekler nitelikte veri setimizin dağılımında hipertansiyona sahip kişilerin inme geçirme dağılımı daha fazladır.



7.3 **Evlilik Durumu:** Evlilik öznitelik değerlerimizde evli olanların bu veri seti özelinde inme geçirme dağılımı fazladır**.**



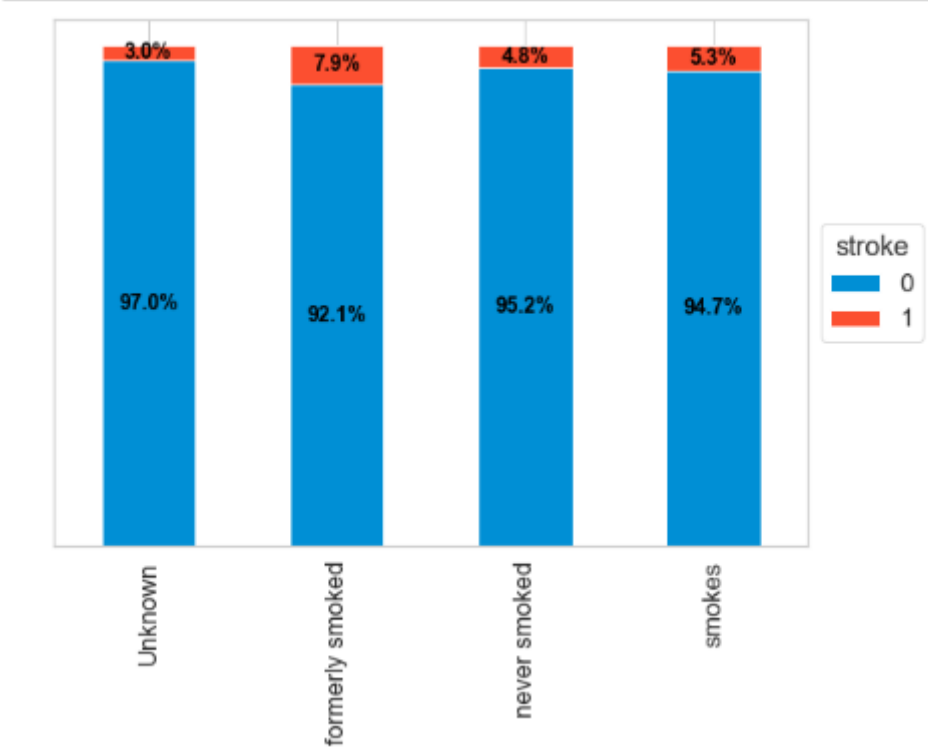
7.4 **Sigara İçme Durumu:** Sigara başta akciğer kanseri olmak üzere hemen hemen vücuttaki her organda çoğu ölümcül birçok hastalığa yol açıyor. Sigara kullanımı veya sadece dumanına maruz kalınmasının bile inme başta olmak üzere ciddi damar hastalıklarına yol açtığı bilinmektedir. Veri bütünlüğü açısından aşağıda ki işlemler model eğitim sırasında yapılacaktır

Hiç sigara içmeyen bir hastanın inme geçirmeme olasılığı en yüksektir

- Bilinmeyen ve daha önce inme geçirmemiş kişileri, daha önce sigara kullanmamış kategorisine alınarak veri bütünlüğü sağlanacaktır

Daha önce sigara içen bir hastanın inme geçirme olasılığı en yüksektir

- Bilinmeyen ve daha önce inme geçirmiş kişileri, daha önce sigara kullanmış kategorisine alınarak veri bütünlüğü sağlanacaktır



Kategorik ve sürekli özniteliklerin dağılımı ve veri setine oranları verilmiştir. Veri seti hakkında her bir öznitelik için keşifçi analiz ve görselleştirmeler ile içerik zenginleştirilmeye çalışılmış model ve istatistik testler için zemin hazırlanmıştır

8. Hipotez Testlerine İlişkin Uygulamalar ve Çıktılar

Örnekleme ile test edilmeye çalışılan bir popülasyonun ilgili parametresi hakkında ortaya sunulan iddiadır. Örneğin;

– A dersi için vize ortalaması 50'nin altındadır

– A ve B lastik firmalarının ürettikleri lastiklerin kaliteleri aynıdır.

Görüldüğü gibi bir konu hakkında öne sürülen ve doğruluğu henüz ispatlanmamış görüşler hipotezlerdir. Hipotezler üzerinde çeşitli işlemler yapılarak ifadenin "doğruluğu/yanlışlığı" araştırılır.

8.1 **Shapiro Wilk Testi**: İstatistikte uygulanan birçok analiz ve tahmin yöntemleri X değişkeninin Normal dağılış göstermesi halinde geçerlidir. Bu nedenle dağılışın Normal olamadığı hallerde yapılan analiz geçerliliğini kaybeder. Dolayısıyla bu dağılış istatistikte önemli bir yere sahiptir. Verilere parametrik testlerin uygulanabilmesi için verilerin dağılımının normal olması gerekmektedir.

Öznitelik	İstatistik	P-value	Durum
Yaş	0.967	1.37e-32	H0 Hipotezi Ret
Ortalama Glikoz Seviyesi	0.806	0.00	H0 Hipotezi Ret
Vücut Kitle İndeksi	0.952	4.28e-38	H0 Hipotezi Ret

Sürekli olan özniteliklerimiz değerlerine Shapiro Wilk Testi uygulaması sonucu verilerimizin normal dağılımından gelmediği H_0 hipotezimiz olan veri setinin dağılımı normaldir hipotezini reddeder, alternatif hipotezimiz olan H_1 hipotezini reddedemeyiz. Dağılımın normal olmadığı veri setlerinde parametrik olmayan (Non-parametric) testler uygulanması gerekmektedir.

8.2 **Levene Testi:** Tek yönlü gruplar arası varyans analizine (ANOVA) eşdeğerdir.
İki veya daha fazla grup için hesaplanan bir değişken için varyansların eşitliğini değerlendirmek için kullanılan bir hipotez testidir. **Levene testi** (veya homojen varyans testi ile) gruplar'daki ölçülen değişken varyanslarının eşit olup olmadığına bakılır. Homogeneity of variance test (varyansların homojenliği testi) tablosundaki anlamlılık(p) değeri incelenir, p>0,05 ise **varyanslar homojen** dağılmıştır, p<0,05 ise **varyanslar homojen** dağılmamıştır.

Öznitelik	Test İstatistiği	P-Value	Durum
Yaş	12299.40	0.0	Örneklem için varyans homojen değildir
Hipertansiyon	90.318	2.48E-21	Örneklem için varyans homojen değildir
Kalp Hastalığı	1,46	0.22	Örneklem için varyans homojendir
Ortalama Glikoz Seviyesi	3489.71	0.0	Örneklem için varyans homojen değildir
Vücut Kitle İndeksi	6155.96	0.0	Örneklem için varyans homojen değildir

Uygulanan test sonucunda hedef değişkenimiz (bağımlı değişken) olan felç durumu ile tek yönlü varyans analizi yapılmıştır. H_0 Hipotezimiz olan varyanslar arasında fark yoktur hipotezini; Özniteliğimiz olan **Kalp Hastalığında** reddederek H_1 Hipotezini reddedemeyiz. Yani, Kalp Hastalığı ve Felç Durumu arasında homojen bir dağılım olduğunu kabul ederiz. Diğer özniteliklerimiz olan Yaş, Hipertansiyon, Ortalama Glikoz Seviyesi ve Vücut Kitle İndeksi değerlerinde H_0 reddedemeyiz. Yani, bu öznitelikler arasında bir homojenlik söz konusu yok kanısına varabiliriz.

9. Kategorik Değişkenlerde Testler:

Kategorik değişkenler arasındaki ilişkileri test etmek veya kategorik bağımlı değişkenli bir regresyon modeli kurmak için **kategorik veri analizi** tekniklerine başvurulmaktadır. Analizlerde, kategorik veriler için genellikle sayılabilir özet tablolar kullanılmaktadır. Bu tip verilerde, standart normal dağılım verilerinde kullanılan istatistiksel analiz yöntemleri uygulanmamaktadır. Kategorik verilerin analizi için çeşitli istatistik metotları ile test edilebilmektedir. Bunlara örnek olarak;

- **Ki-kare bağımsızlık testleri**
- **Lojistik regresyon analizleri**
- **Log-lineer modeller**

9.1 **Ki-Kare Testi**

Eğer iki kategorik değişken arasında ilişki olup olmadığını merak ediyorsak kullanacağımız istatistik yöntemi Pearson Ki-Kare testi olacaktır. Kategorik değişkenlerin analizleri genelde frekanslar üzerinden yapılır**.** Ki-Kare testi her bir kategori çiftine düşen frekans sayısı ile bu durumlara şansla düşebilecek frekans sayılarının karşılaştırılmasına dayanır. Gözlenen frekans ile beklenen frekans karşılaştırması diyebiliriz.

9.2 **Örnek Uygulama**

Örneğin inme geçiren ve geçirmeyen kişilerin Cinsiyetlerine göre dağılımını merak ettiğimiz bir araştırma sorusunda 4 farklı durum ortaya çıkabilir:

(0-Kadın, 0-Erkek, 1-Kadın ve 1-Erkek). Bu durumların hepsini aşağıdaki çaprazlık tablosu ile gösterebiliriz.

İnme	**Geçirmemiş**	**Geçirmiş**	**Toplam**
Kadın	**2853**	**141**	**2994**
Erkek	**2008**	**108**	**2116**
Toplam	**4861**	**249**	**5110**

Bu değerin anlamlı bir fark doğurup doğurmadığını test edebilmemiz için serbestlik değerine ihtiyacımız vardır. Değişkenlerimizin kategorisi ikişer kategoridir yani:

(2-1) x (2-1) = 1 =serbestlik derecesi

Daha sonra bu serbestlik derecesini ve ki-kare değerlerini alarak istatistik tablolarından bulabileceğimiz kritik değer ile karşılaştırdığımızda ki-kare sonucunun anlamlı bulunup bulunmadığını test edebiliriz. Eğer p-değeri 0.05'ten küçük bulunursa cinsiyet ile inme geçirme arasında bir ilişki vardır şeklinde belirtebiliriz.

9.3 **Örnek Hipotez**

H_0: İnme geçirme durumunun kadın ve erkek arasında fark yoktur.

H_1: İnme geçirme durumunun kadın ve erkek arasında fark vardır.

Yapılan Ki-Kare testi sonucunda Ki-Kare Test İstatistiği 0.416, P-value değeri 0.518, Serbestlik Derecesi 1, Kritik Değer 3.84 bulunmuştur. Yani H_0 hipotezini reddedemiyoruz. Bu test sonucunda diyebiliriz ki inme geçirme ile cinsiyet arasında bir bağ yoktur.

9.4 **Ki-Kare Hipotez Örneği**

H_0: Kentte yaşayan ve yaşamayanların inme geçirme oranları arasında fark yoktur.

H_1: Kentte yaşayan ve yaşamayanların inme geçirme oranları arasında fark vardır.

Yapılan Ki-Kare testi sonucunda Ki-Kare Test İstatistiği1,22, P-value değeri 0.269, Serbestlik Derecesi 1, Kritik Değer 3.84 bulunmuştur. Yani H_0 hipotezini reddedemiyoruz. Bu test sonucunda diyebiliriz ki inme geçirme ile kentte veya kırsalda yaşama arasında bir bağ yoktur.

10. MANN WHITNEY-U Testi

Normal dağılım özelliği göstermeyen bir dağılımda iki bağımsız grup ortalamalarını karşılaştırmak amacıyla kullanılan non-parametrik bir yöntemdir. Parametrik test varsayımları yerine getirilmeden, iki ortalama arasındaki farkın önemlilik testinin uygulanması varılan kararın hatalı olmasına neden olabilir. Parametrik test varsayımları yerine gelmediğinde kullanılabilecek en güçlü testtir.

10.1 **MANN WHITNEY-U Testi Hipotez Örneği**

H_0: Kentte yaşayan kadın ile, Kırsal kesimde yaşayan kadının VKİ oranında fark yoktur.

H_1: Kentte yaşayan kadın ile, Kırsal kesimde yaşayan kadının VKİ oranında fark vardır.

Yapılan U Testi sonucunda Test İstatistiği 1121833.5, P-value değeri 0.93. Yani H_0 hipotezin i reddedemiyoruz. Bu test sonucunda diyebiliriz ki kentte veya kırsal alanda yaşayan kadınları n **Vücut Kitle İndeksi** oranları arasında anlamlı bir farklılık yoktur.

11. Makine Öğrenmesi Kullanarak Felç Geçirme Tahminleme Çalışması

11.1 Tahmin: Kavram olarak daha genel bir anlam taşır. **Öngörü** (*forecasting*) ise geleceğin ta hmini olacaktır Makine öğrenmesi ile hastalıkların erken evre teşhisinde, tanı koyma ve karar a lma süreçlerinde, sağlıktan sosyal bilimlere birçok alanda kullandığı bilinmektedir. Veri setimi ze uygulayacağımız algoritmalar istatistik tabanlı ve anlaşılması kolay olan OLS ve Binary Re gression (İkili Sınıflandırma) modellerini kullanarak hem özniteliklerin arasında ki bağı(patter n) yakalanmaya çalışacak hem de doğruluk skorları çıkarılarak tahmin işlemi yapılacaktır.

12. Sıradan En Küçük Kareler Yöntemi (Ordinary Least Squares -OLS)

En küçük kareler yöntemi, birbirine bağlı olarak değişen iki fiziksel büyüklük arasındaki ilişki yi mümkün olduğunca gerçeğe uygun bir denklem olarak yazmak için kullanılan, standart bir regresyon yöntemidir. Bir başka deyişle bu yöntem, ölçüm sonucu elde edilmiş veri noktaların a "mümkün olduğu kadar yakın" geçecek bir işlev eğrisi bulmaya yarar. Gauss-Markov Teore mi'nin en küçük kareler yöntemi, regresyon için optimal yöntemdir. İstatistikte sıradan en küç ük kareler (OLS) veya doğrusal en küçük kareler, gözlemlenen yanıtlar arasındaki değişkenler in karelerinin toplamını en aza indirgemek amacı ile doğrusal regresyon modelinde bilinmeye n parametrelerin tahmin edilmesi için kullanılan bir yöntemdir.

- Sıradan en küçük kareler yöntemidir. Doğrusal regresyon modeli kurulurken verilerin ortasından geçen eğime (çizgiye) en az kare farkı ile yaklaşmaya çalışan bir yöntemdir

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{SPxy}{SSx}$$

and $\tilde{\beta}_0 = \bar{Y} - \tilde{\beta}_1 \bar{X}.$

13. OLS Özet Tablosu ve Yorumlanması

OLS Regression Results

Dep. Variable:	stroke	R-squared:	0.097
Model:	OLS	Adj. R-squared:	0.095
Method:	Least Squares	F-statistic:	39.25
Date:	Fri, 10 Jun 2022	Prob (F-statistic):	1.56e-102
Time:	12:43:49	Log-Likelihood:	858.52
No. Observations:	5110	AIC:	-1687.
Df Residuals:	5095	BIC:	-1589.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0315	0.020	-1.595	0.111	-0.070	0.007
age	0.0029	0.000	13.270	0.000	0.002	0.003
hypertension	0.0393	0.010	3.851	0.000	0.019	0.059
heart_disease	0.0484	0.013	3.617	0.000	0.022	0.075
avg_glucose_level	0.0003	6.66e-05	4.409	0.000	0.000	0.000
bmi	-0.0008	0.000	-1.882	0.060	-0.002	3.32e-05
gender_Male	-0.0042	0.006	-0.705	0.481	-0.016	0.007
ever_married_Yes	-0.0353	0.009	-4.153	0.000	-0.052	-0.019
work_type_Never_worked	0.0443	0.045	0.983	0.326	-0.044	0.133
work_type_Private	0.0149	0.009	1.669	0.095	-0.003	0.032
work_type_Self-employed	-0.0051	0.011	-0.466	0.641	-0.026	0.016
work_type_children	0.0694	0.014	4.789	0.000	0.041	0.098
Residence_type_Urban	0.0045	0.006	0.783	0.434	-0.007	0.016
smoking_status_never smoked	-0.0678	0.008	-8.568	0.000	-0.083	-0.052
smoking_status_smokes	-0.0487	0.010	-4.854	0.000	-0.068	-0.029

Omnibus:	3731.788	Durbin-Watson:	0.195
Prob(Omnibus):	0.000	Jarque-Bera (JB):	44958.528
Skew:	3.563	Prob(JB):	0.00
Kurtosis:	15.663	Cond. No.	2.00e+03

13.1 **R-Squared (R2):** Verilerin yerleştirilmiş regresyon çizgisine ne kadar yakın olduğunun istatistiksel bir ölçüsüdür**.** Bizim tablomuzda 0.97 çok iyi bir sonuç diyebiliriz.

13.2 **Adjusted (Düzenlenmiş) R-Squared**: Bu değer genellikle çoklu regresyon analizinde kullanılır. Modele değişken eklendikçe R2 değeri değişkeni etkili varsayıp varyansını açıkladığını sanabilir. Ancak düzenlenmiş R2 mantığında her bir değişkenin bağımlı değişken üzerindeki etkisi ölçülerek daha optimum sonuçlar vermektedir. Tabloyu yorumlarken bu iki değerin birbirine yakın çıkması önemli bir konu olup bizim tablomuzda her iki değerinde (0.97-0.95) birbirine yakın olduğu görülmektedir.

13.3 **Coefficients (Katsayılar)**: Regresyon katsayısı, iki veya daha fazla değişken arasındaki ortalama fonksiyonel ilişkinin istatistiksel bir ölçüsüdür.

13.4 **P-Value**: Bu değer bize bağımsız değişkenden çıkan katsayının istatistiksel olarak anlamlı olup olmadığını vermektedir. Bilim dünyasında genel olarak kabul edilen anlamlılık değeri x<0.05'dir. Eğer bu değerin üzerinde bir sonuç çıkıyorsa ilgili bağımsız özniteliğin modele etkisi anlamlı değildir. Etkisiz olan öznitelikleri eğitilecek olan modele dahil etmeyerek daha yansız sonuçlar ile tahminleme çalışması yapılabilir.

13.5 **Ortalama Kare Hata:** Bir regresyon eğrisinin bir dizi noktaya ne kadar yakın olduğunu söyler. Modelinin, tahminleyicinin performansını ölçmemizi sağlar. Bu model için hata oranı 0.41 bulunmaktadır. Ortalama Kare Hata değeri 0'a yakın oldukça model başarılıdır diyebiliriz.

$$MSE = \frac{1}{n}\sum_{j=1}^{n} e_j^2$$

14. Sckit-learn Logistic Regression Model

Sckit-learn kütüphanesi çeşitli makine öğrenmesi modelleri içerisinde bulunduran açık kaynak kodlu bir pakettir. Python programlama dili dahil olmak üzere birçok dil ile kullanılabilir. Sckit-learn kütüphanesi içerisinde bulunan **Logistic Regression** Modelini kullanarak tahminleme yapılacaktır.

Veri setimizi eğitime hazır gele getirmek için test ve eğitim seti olarak ayırma işlemi yapıyoruz. Yani verimizin %80'i eğitim için %20'si test olacak şekilde ayırıp modelimizde eğitiyoruz. Eğitilen modelin performansını değerlendirmek için farklı ölçüt tipleri vardır. Bu tiplerden bazıları şunlardır:

14.1 **Ortalama Mutlak Hata:** İki sürekli değişken arasındaki farkın ölçüsüdür. MAE, her gerçek değer ile veriye en iyi uyan çizgi arasındaki ortalama dikey mesafedir. MAE aynı zamanda her veri noktası ile en iyi uyan çizgi arasındaki ortalama yatay mesafedir. MAE değeri kolay yorumlanabilir olduğu için regresyon ve zaman serisi problemlerinde sıkça kullanılmaktadır.

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |e_j|$$

TP (True Positive — Doğru Pozitif): Hastaya hasta demek.

FP (False Positive — Yanlış Pozitif): Hasta olmayana hasta demek.

TN (True Negative — Doğru Negatif): Hasta olmayana hasta değil demek.

FN (False Negative — Yanlış Negatif): Hasta olana hasta değil demek.

14.2 Doğruluk Skoru (Accuracy): Anlaşılması ve yorumlanması en basit ölçütlerden birisidir. Makine öğrenmesi sınıflandırma algoritmalarının testlerinde sıklıkla kullanılır. Accuracy skoru aşağıdaki gibi hesaplanır. Accuracy skoru 0 ve 1 arasında olup 1'e yaklaşan skorlarda model başarılı kabul edilir

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

14.3 **Kesinlik (Precision):** Positive olarak tahminlediğimiz değerlerin gerçekten kaç adedinin Positive olduğunu göstermektedir.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

14.4 **Duyarlılık (Recall):** Positive olarak tahmin etmemiz gereken işlemlerin ne kadarını Positive olarak tahmin ettiğimizi gösteren bir metriktir.

$$\text{Recall} = \frac{TP}{TP + FN}$$

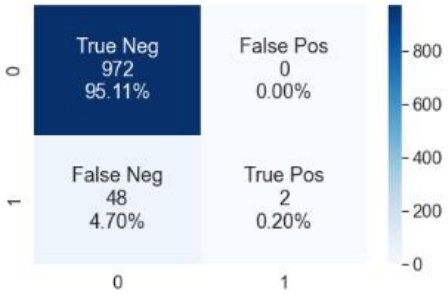
14.5 **F1 Score:** Bize Kesinlik (Precision) ve Duyarlılık (Recall) değerlerinin harmonik ortalamasını göstermektedir. Basit bir ortalama yerine harmonik ortalama olmasının sebebi ise uç durumları da gözardı etmememiz gerektiğidir.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Doğruluk (Accuracy) yerine F1 Score değerinin kullanılmasının en temel sebebi eşit dağılmayan veri kümelerinde hatalı bir model seçimi yapmamaktır. Ayrıca sadece False Negative ya da False Positive değil tüm hata maliyetlerini de içerecek bir ölçme metriğine ihtiyaç duyulduğu içinde F1 Score bizim için çok önemlidir.

15. Eğitilen Modelin Sonuçları

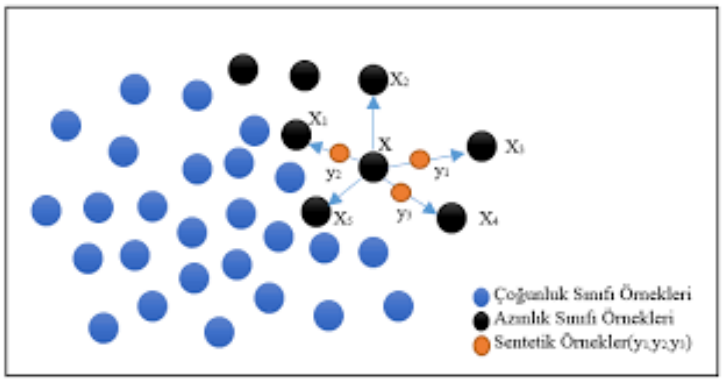
	precision	recall	f1-score	support
0	0.95	1.00	0.98	972
1	1.00	0.02	0.04	50
accuracy			0.95	1022
macro avg	0.98	0.51	0.51	1022
weighted avg	0.95	0.95	0.93	1022



Model değerlendirme sonuçlarına
baktığımızda doğruluk skorumuz 0,95 oranında harika bir başarı elde ederken, aslında yukarıda bahsetmiş olduğumuz f1-score değerimiz bizim için performans ölçütü olarak baz alınacak metriktir. F1-Score değerine baktığımızda inme geçirmemiş kişileri tahmin ederken çok iyi bir performans sergilerken geçirmiş kişilerin tahmininde kabul edilebilir bir başarı elde edememiştir. Veri setimiz zaten dengesiz bir veri olduğu için bu model ile eğitilen makine inme geçirmiş kişileri tahmin etmekte başarısızdır.

Model Geliştirme ve SMOTE (Synthetic Minority Over-Sampling Technique)
Bir önceki modelimiz inme geçirmeyen kişileri tahmin ederken %95 oranında bir doğruluk payı ile tahminleme yapmaktaydı. Performans değerlendirme ölçütlerinde değindiğimiz konuda doğruluk skorunun tek başına yeterli olmadığı, özellikle tahminleme aşamasında dengesiz dağılan veri setleri için F1-Score metriğinin çok önemli olduğunu belirtmiştik. Bu öngörüye dayanarak: Yeniden eğitilecek olan modelimizin hiperparametre (hyperparameter) değerlerini optimize ederek aynı zamanda bağımlı özniteliğimizi de SMOTE veri artırıma yöntemi ile dağılımı eşitlemeye çalışılmıştır.

SMOTE (Synthetic Minority Over-Sampling Technique): Sentetik veri üretilmesini sağlayan bir aşırı örnekleme sürecidir. Veri bilimi projelerinden en sık kullanılan yöntemlerden biridir. Yöntemin ana fikri, azınlık sınıfının örnekleri arasında belirli işlemler yaparak yeni azınlık sınıfı örnekleri yaratmaktır



16. Modelin Yeniden Eğitilmesi ve Sonuçlar

Bir önceki modelden edindiğimiz dengesiz dağılım ve performans ölçütümüz olan f1-skor değerinin yetersiz olması sebebiyle yeni modelimiz SMOTE kullanarak inme geçirmemiş özniteliğimize sentetik veri üreterek eşitlik sağlanmıştır. Model hiperparametre olarak: K katlı çapraz doğrulama yöntemi ile 40 defa Ceza katsayıları ve L1, L2 düzeltme normları ile modelimiz tekrar eğitilmiştir. K katlı çapraz doğrulama ile verdiğimiz hiperparametre değerlerinden en iyi optime olanı C:0.9, Düzeltme Norm: L2 ve Maksimum iterasyon(yineleme) sayısı da 100 olduğu saptanmıştır. Bu hiperparametreler eğitilen model sonuçları aşağıdadır:

	precision	recall	f1-score	support
0	0.98	0.87	0.92	972
1	0.18	0.58	0.28	50
accuracy			0.85	1022
macro avg	0.58	0.72	0.60	1022
weighted avg	0.94	0.85	0.89	1022

Doğruluk skoruna baktığımızda 0,85 oranında doğruluk oranı olduğunu görüyoruz modelimizin. Bir önce ki modelimize göre düşük doğruluk oranına sahip ama performans ölçütü olarak belirlediğimiz f1-skoru değerini SMOTE ile yükseltmiş olduk. Artık tahmin işlemimiz bir önceki modele göre daha tutarlı ve güvenilirdir.

17. Sonuç:

Elimizde bulunan veri setine önce keşifçi veri analizi yaparak her bir öznitelik hakkında derinlemesine bilgi sahibi olduk. Sürekli öznitelik değerlerimizin dağılımına ve kategorik öznitelik değerlerimizin ise genel veri setine oranla olasılıklarına ve frekans değerlerini inceledik. Veri setimize hipotez testlerin doğru ve yansız sonuçlar vermesi için normallik varsayımı ve homojenlik testleri uyguladık. Uyguladığımız testler sonucunda normal dağılmadığını ve non-parametrik testlerin bu veri seti için daha doğru sonuçlar vereceğini test etmiş olduk. Basit doğrusal Regression modeli ile veri setimizin istatistiki anlamda birbiri üzerine dağılımı, ilişkiyi ve inme geçiren veya geçirmeyen kişilere etkisi üzerine istatistik tabanlı bir model inşa edildi. Sckit-learn kütüphanesini ve python programlama dilini kullanarak Logistic Regression modeli ile bir makine öğrenmesi algoritması tahminleme için kuruldu. Kurulan modelde doğruluk skorunun yetersiz olduğu ve yanlı sonuçlar ile tahminleme yapacağı için model geliştirilmesi yapıldı. SMOTE tekniğini kullanarak sentetik veri üreterek dengesiz olan bağımlı özniteliğimiz dengelenmeye çalışıldı. SMOTE ile geliştirilen model bir önceki modele göre performans ölçütü olarak belirlediğimiz f1-skor daha yüksek ve tahminlerimizin daha doğru olmasına olanak sağladı.