Analiza policističnog sindroma jajnika logističkom regresijom i diskriminantnom analizom

Peček, Teodora

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Teodora Peček

ANALIZA POLICISTIČNOG SINDROMA JAJNIKA LOGISTIČKOM REGRESIJOM I DISKRIMINANTNOM ANALIZOM

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, veljača 2024.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Hvala mami, tati i seki na beskrajnoj ljubavi i potpori.
Hvala mentorici prof. dr. sc. Anamariji Jazbec na razumijevanju, strpljenju i korisnim savjetima.

Sadržaj

Uvod

Policistični sindrom jajnika (PCOS, eng. *policystic ovary syndrome*) je čest poremećaj endokrinog sustava kojeg karakteriziraju jajnici s folikularnim (vodenim) cistama, nepravilni menstrualni ciklusi, pretilost, poremećena tolerancija glukoze, inzulinska rezistencija i hiperandrogenizam (povećana razina androgenih hormona). Neki od simptoma hiperandrogenizma su akne, gubitak kose te pojačan rast dlaka lica i tijela.

Postoji nekoliko različitih kriterija za dijagnozu ovog sindroma. Prema kriteriju Europskog društva za humanu reprodukciju i embriologiju (ESHRE, eng. *European Society of Human Reproduction and Embryology*) PCOS se može dijagnosticirati ako su prisutna barem dva od sljedećih stanja: hiperandrogenizam, poremećaj ovulacije i policistični jajnici, uz uvjet da je isključena prisutnost drugih poremećaja koji uzrokuju ova stanja. Procjenjuje se da po ovom kriteriju PCOS ima 10 - 13% žena. [10]

U svrhu dijagnoze određuju se razine folikulostimulirajućeg hormona (FSH), luteinizirajućeg hormona (LH), testosterona, prolaktina, tiroidnog stimulirajućeg hormona (TSH), anti-Mullerovog hormona (AMH). Ultrazvukom se određuje broj cista na jajnicima te se prate moguće promjene i nepravilnosti na endometriju.

U ovom radu uvode se osnovni pojmovi i ideje logističke regresije i diskriminantne analize. Korištenjem programa SAS OnDemand analizira se baza podataka koja prati razne podatke o pacijenticama, među kojima je i prisutnost PCOS-a. Određuje se koji parametri značajno utječu na izglednost prisutnosti PCOS-a te se predviđa ima li pacijentica PCOS.

Poglavlje 1

Logistička regresija

1.1 Logistički model

Regresijska analiza je statistička metoda kojom se određuje veza između skupa nezavisnih varijabli ili varijabli poticaja i zavisne varijable ili varijable odziva. Rezultat regresijske analize je regresijski model - jednadžba koja matematički opisuje tu vezu. Najpoznatija vrsta regresijske analize je linearna regresija u kojoj je zavisna varijabla neprekidna, a ovisnost o nezavisnim varijablama je opisana linearnom jednadžbom.

Logistička regresija je vrsta regresijske analize čija je zavisna varijabla diskretna, odnosno poprima dvije ili više vrijednosti. Nezavisne varijable mogu biti neprekidne i diskretne. Logistička regresija često se primjenjuje u medicini, ekonomiji, *credit-scoringu*. U ovom radu koristi se logistička regresija s dihotomnom zavisnom varijablom.

Neka je Y zavisna varijabla koja poprima vrijednosti 0 i 1, gdje 1 označava da se događaj od interesa dogodio (npr. prisutna je bolest) i neka je X nezavisna varijabla. Cilj je procijeniti očekivanu vrijednost zavisne varijable ako je poznata vrijednost nezavisne varijable, tj. traži se $\mathbb{E}(Y|X = x)$. U slučaju linearne regresije tražio bi se model oblika

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x, \tag{1.1}$$

no u slučaju logističke regresije to nije moguće jer je

$$\mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x), \tag{1.2}$$

što je ograničena funkcija, dok linearna funkcija u (1.1) nije.

Kako bi se riješio taj problem, umjesto vjerojatnosti $\mathbb{P}(Y = 1|X = x)$, označene s $p(x)$, koristi se njezina transformacija *logit* funkcijom. Neka je prvo **izgled** ili **šansa** (eng. ***odd***)

$$odd(p(x)) = \frac{p(x)}{1 - p(x)}. \tag{1.3}$$

Logit ili ***log odds*** funkcija je tada $logit : (0, 1) \to (-\infty, +\infty)$

$$logit(p(x)) = \ln(odd(p(x))) = \ln\left(\frac{p(x)}{1 - p(x)}\right) = \ln(p(x)) - \ln(1 - p(x)). \tag{1.4}$$

Uvrštavanjem *logit* transformacije umjesto $p(x) = \mathbb{E}(Y|X = x)$ u (1.1) dobiva se **univarijatni logistički model** [2, 5]

$$logit(p(x)) = \ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x, \tag{1.5}$$

odnosno

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{1.6}$$

i

$$odd(p(x)) = e^{\beta_0+\beta_1 x}. \tag{1.7}$$

Na isti način dobiva se **multivarijatni logistički model** s n nezavisnih varijabli $X_1, X_2, ..., X_n$

$$logit(p(x)) = \ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x + \ldots + \beta_n x, \tag{1.8}$$

$$p(x) = \frac{e^{\beta_0+\beta_1 x} + \ldots + \beta_n x}{1 + e^{\beta_0+\beta_1 x} + \ldots + \beta_n x}, \tag{1.9}$$

$$odd(p(x)) = e^{\beta_0+\beta_1 x+\ldots+\beta_n x}. \tag{1.10}$$

1.2 Procjena parametara modela

Neka je dan uzorak od n nezavisnih opažanja parova (x_i, y_i), $i = 1, 2, \ldots, n$, gdje je x_i vrijednost nezavisne varijable u i-tom opažanju, a y_i vrijednost zavisne varijable u i-tom opažanju. Procjena parametara β_0 i β_1 logističkog modela dobiva se **metodom maksimalne vjerodostojnosti** (eng. ***maximum likelihood***). Traže se parametri β_0 i β_1 koji maksimiziraju funkciju vjerodostojnosti

$$l(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{y_i}, \tag{1.11}$$

gdje je kao i ranije $p(x) = \mathbb{P}(Y = 1|X = x)$.

Radi jednostavnosti funkcija vjerodostojnosti se logaritmira i traži se maksimum log-vjerodostojnosti.

$$L(\beta_0, \beta_1) = \ln(l(\beta_0, \beta_1)) = \sum_{i=1}^{n} \left[y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i)) \right]. \tag{1.12}$$

Parcijalnom derivacijom po β_0 i β_1 te izjednačavanjem tih derivacija s nulom dobivaju se **jednadžbe vjerodostojnosti** (eng. ***likelihood equations***)

$$\sum_{i=1}^{n} (y_i - p(x_i)) = 0 \tag{1.13}$$

i

$$\sum_{i=1}^{n} x_i(y_i - p(x_i)) = 0. \tag{1.14}$$

Jednadžbe vjerodostojnosti riješavaju se iterativnim metodama zbog nelinearne ovisnosti o parametrima β_0 i β_1 te se njihova rješenja nazivaju **procjeniteljima maksimalne vjerodostojnosti** (eng. ***maximum likelihood estimators***) za β_0 i β_1. [2]

1.3 Testiranje adekvatnosti modela

Nakon procjene parametara modela metodom maksimalne vjerodostojnosti testira se adekvatnost modela (eng. ***goodness of fit***) za procjenu zavisne varijable. Kako bi se testirala adekvatnost dobivenog modela, koristi se **devijanca** (eng. ***deviance***)

$$D = -2\ln\left(\frac{\textit{vjerodostojnost trenutnog modela}}{\textit{vjerodostojnost saturiranog modela}}\right), \tag{1.15}$$

gdje je saturirani model onaj čiji je broj parametara jednak broju opservacija.

Kako saturirani model ima najveću moguću vjerodostojnost [7], želimo da je **omjer vjerodostojnosti** (eng. ***likelihood ratio***), oznaka *LR*,

$$LR = \frac{\textit{vjerodostojnost trenutnog modela}}{\textit{vjerodostojnost saturiranog modela}} \tag{1.16}$$

najveći mogući, odnosno da je devijanca najmanja moguća.

Devijanca se najčešće ne koristi kao apsolutna mjera adekvatnosti modela, već za usporedbu ugniježđenih modela. [7] Testira se je li model s uključenim nezavisnim varijablama značajno bolji od modela koji se sastoji samo od slobodnog člana (eng. *intercept*) pomoću statistike *G*

$$G = D(\textit{model bez varijabli}) - D(\textit{model s varijablama}) \tag{1.17}$$

$$G = -2\ln\left(\frac{\textit{vjerodostojnost modela bez varijabli}}{\textit{vjerodostojnost saturiranog modela}}\right) - \left(-2\ln\left(\frac{\textit{vjerodostojnost modela s varijablama}}{\textit{vjerodostojnost saturiranog modela}}\right)\right)$$

$$G = -2\ln\left(\frac{\textit{vjerodostojnost modela bez varijabli}}{\textit{vjerodostojnost modela s varijablama}}\right). \tag{1.18}$$

Statistika *G* ima približno χ^2- distribuciju čiji je broj stupnjeva slobode jednak broju nezavisnih varijabli uključenih u model. [5] Ako je g realizacija statistike G, tada se računa $\mathbb{P}(G > g)$ i ako je ta vjerojatnost manja od unaprijed definirane razine značajnosti α, zaključuje se da je model s uključenim varijablama značajno bolji.

Na isti način može se testirati dolazi li do poboljšanja modela dodavanjem jedne od nezavisnih varijabli. Tada *G* ima χ^2 - distribuciju s jednim stupnjem slobode. [2]

1.4 Interpretacija parametara modela

Treba odrediti kako promjena vrijednosti nezavisne varijable utječe na zavisnu varijablu. Neka je dan univarijatni logistički model kao u (1.5). Tada je

$$logit(p(x+1)) - logit(p(x)) = \beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x) \tag{1.19}$$

$$\ln\left(\frac{p(x+1)}{1-p(x+1)}\right) - \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_1(x+1) - \beta_1 x \tag{1.20}$$

$$\ln\left(\frac{\frac{p(x+1)}{1-p(x+1)}}{\frac{p(x)}{1-p(x)}}\right) = \beta_1 \tag{1.21}$$

$$\ln\left(\frac{odds(p(x+1))}{odds(p(x))}\right) = \beta_1. \tag{1.22}$$

Dakle, parametar β_1 je logaritam **omjera šansi** (eng. ***odds ratio***), oznaka OR, definiranih s (1.3), odnosno vrijedi

$$OR = e^{\beta_1}. \tag{1.23}$$

[5]

Ako je nezavisna varijabla X dihotomna i poprima vrijednosti 0 i 1, tada omjer šansi označava koliko je izgledniji (ili manje izgledan) ishod $y = 1$ (od ishoda $y = 0$) u grupi u kojoj je $x = 1$, nego u grupi u kojoj je $x = 0$. [2]

Ako je nezavisna varijabla X kategorijska i poprima vrijednosti $x_1, x_2, \ldots, x_k$, $k > 2$, tada se bira jedna od vrijednosti kao referentna, npr. x_1 i uvodi se $k-1$ pomoćnih (eng. *dummy*) varijabli $D_2, D_3, \ldots, D_k$. *Dummy* varijabla D_i, $\forall i = 2, 3, \ldots, k$, poprima vrijednost 1 ako je $X = x_i$, a inače poprima vrijednost 0. Kako su *dummy* varijable dihotomne, omjeri šansi interpretiraju se na gore opisan način. Za svaku *dummy* varijablu D_i, $i = 2, 3, \ldots, k$, pripadni omjer šansi označava koliko je izgledniji ishod $y = 1$ u grupi u kojoj je $x = x_i$, nego u grupi u kojoj x poprima referentnu vrijednost x_1.

Ako je nezavisna varijabla X neprekidna, omjer šansi (1.23) označava koliko je izgledniji ishod $y = 1$ ako nezavisnu varijablu povećamo za 1. Često je od interesa povećanje nezavisne varijable za neku drugu konstantu c. Na isti način kao gore dobiva se da je tada omjer šansi $OR = e^{c\beta_1}$, odnos ishod $y = 1$ je $e^{c\beta_1}$ puta izgledniji od ishoda $y = 0$. [2]

1.5 ROC krivulja

Neka je, kao ranije, Y dihotomna varijabla koja poprima vrijednosti 0 i 1 te neka $Y = 1$ označava prisutnost bolesti. Logističkim modelom dobivaju se procijenjene vrijednosti $\hat{y} \in [0, 1]$. Ako se traže klasifikacije opservacija u grupe *bolesni* i *zdravi*, uvodi se granica odluke d i ako je $\hat{y} \geq d$, opservacija se svrstava u grupu *bolesni*, a inače u grupu *zdravi*. Tada su moguća četiri ishoda:

	$\hat{y} \geq d$	$\hat{y} < d$
$Y = 1$	stvarno pozitivni (*true positive, TP*)	lažno negativni (*false negative, FN*)
$Y = 0$	lažno pozitivni (*false positive, FP*)	stvarno negativni (*true negative, TN*)

Tablica 1.1: Klasifikacijska tablica

Kako bi se odredilo koliko dobro model klasificira opservacije, koriste se pojmovi osjetljivosti i specifičnosti te ROC krivulja. **Osjetljivost** (eng. ***sensitivity***) je omjer broja opservacija kojima je $Y = 1$ i klasificirane su u grupu *bolesni* i ukupnog broja opservacija kojima je $Y = 1$, tj.

$$osjetljivost = \frac{TP}{TP + FN}. \tag{1.24}$$

Specifičnost (eng. ***specificity***) je omjer broja opservacija kojima je $Y = 0$ i klasificirane su u grupu *zdravi* i ukupnog broja opservacija kojima je $Y = 0$, tj.

$$specifičnost = \frac{TN}{FP + TN}. \tag{1.25}$$

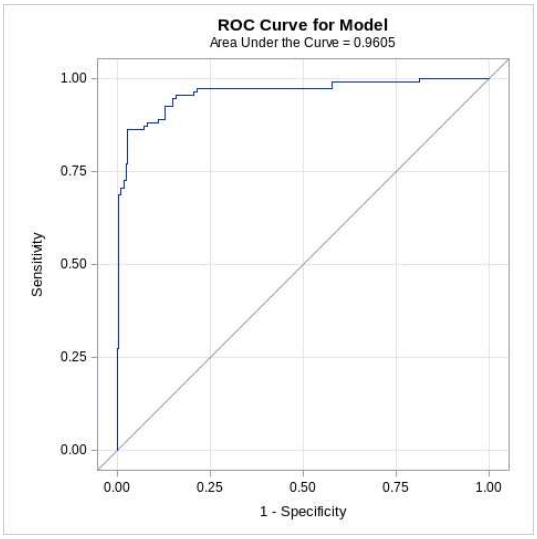
ROC krivulju (eng. ***receiver operating characteristic curve***) prikazuje kako se mijenja *osjetljivost* u odnosu na 1− *specifičnost*, tj. kako se mijenja **stopa stvarno pozitivnih** (eng. ***true positive rate***) $\frac{TP}{TP+FN}$ u odnosu na **stopu lažno pozitivnih** (eng. ***false positive rate***) $\frac{FP}{FP+TN}$. Površina ispod ROC krivulje, oznaka c ili AUC (eng. *area under curve*), predstavlja prediktivnu snagu modela. Kako bi se odredila ta površina, predviđene vrijednosti $\hat{y}$ dijele se u dva skupa $D_0 = \{\hat{y}_{01}, \hat{y}_{02}, \ldots, \hat{y}_{0n_0}\}$ i $D_1 = \{\hat{y}_{11}, \hat{y}_{12}, \ldots, \hat{y}_{1n_1}\}$ koji sadrže procijenjene vrijednosti $\hat{y}$ za opservacije kojima je stvarna vrijednost $Y = 0$ i $Y = 1$, redom. Tada prema [8] vrijedi

$$c = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \psi(\hat{y}_{0i}, \hat{y}_{1j}) \tag{1.26}$$

gdje je $n_0 = |D_0|$, $n_1 = |D_1|$, te

$$\psi(y_0, y_1)\begin{cases} 1 & y_0 < y_1 \\ \frac{1}{2} & y_0 = y_1 \\ 0 & y_0 > y_1 \end{cases}. \tag{1.27}$$

Statistika c poprima vrijednosti od 0 do 1. Ako je vrijednost manja od 0.5, smatra se da model nije dobar klasifikator.



Slika 1.1: Primjer ROC krivulje (ispis iz SAS-a)

1.6 Odabir nezavisnih varijabli u modelu

Postoji nekoliko opcija za odabir nezavisnih varijabli koje su uključene u model - proizvoljni odabir, puni model (sve nezavisne varijable su uključene u model), *forward* procedura, *backward* procedura, *stepwise* procedura.

Forward procedura počinje modelom bez nezavisnih varijabli, tj. modelom koji ima samo slobodni član. Traži se nezavisna varijabla koja je statistički značajna na unaprijed definiranoj razini značajnosti i dodaje se u model. Zatim se postupak ponavlja na preostalim neuključenim nezavisnim varijablama. Postupak završava kada nema dostupnih nezavisnih varijabli koje su statistički značajne ili kada se dosegne maksimalni poželjni broj varijabli u modelu. Jednom dodana varijabla više se ne izbacuje iz modela.

Backward procedura počinje punim modelom. Iz modela se izbacuje varijabla koja je najmanje značajna za model. Postupak se ponavlja do kad nisu isključene sve varijable koje nisu značajne na unaprijed definiranoj razini značajnosti.

Stepwise procedura je kombinacija *forward* i *backward* procedura. Počinje s modelom bez nezavisnih varijabli koje se dodaju jedna po jedna kao kod *forward* procedure, ali u *stepwise* proceduri nakon dodavanja k-te varijable moguće je izbaciti neku od ranije dodanih $k-1$ varijabli. [8, Poglavlje 78]

Poglavlje 2

Diskriminantna analiza

2.1 Opis metode

Diskriminantna analiza jedna je od klasifikacijskih metoda, tj. metoda kojima je cilj predviđanje pripadnosti grupi ili kategoriji na temelju danih prediktorskih varijabli. Za razliku od regresijske analize koja daje jednu jednadžbu kojom se predviđaju vrijednosti varijable odziva (eng. *response variable*), diskriminantna analiza daje po jednu jednadžbu za svaku od grupa. Neka su dane prediktorske varijable $X_1, X_2, \ldots, X_p$, n opservacija prediktorskih varijabli i neka je za svaku opservaciju poznato kojoj od J mogućih grupa pripada. Tada se iz danih opservacija određuje J klasifikacijskih jednadžbi temeljem nekog klasifikacijskog pravila (eng. *classification rule*). Opservacija se klasificira u grupu čija klasifikacijska jednadžba daje najbolju klasifikacijsku vrijednost (eng. *classification score*). [3] [9]

Postoje različite vrste diskriminantne analize. Ako se pretpostavi da $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ ima multivarijatnu normalnu distribuciju u svakoj grupi, dobivaju se linearna i kvadratna diskriminantna analiza. Ako nije poznato kakve su distribucije od $\mathbf{X}$ po grupama, tada se koriste neparametarske metode za procjenu funkcija gustoća. [8, Poglavlje 41] Ovaj rad bavi se linearnom diskriminantnom analizom.

2.2 Klasifikacijska pravila

Neka je dana opservacija $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ prediktorskih varijabli $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ koja pripada jednoj od J grupa. Traži se grupa kojoj pripada dana opservacija. To se može učiniti koristeći sljedeća **klasifikacijska pravila** (eng. **classification rules**): [4]

1. Neka je poznato matematičko očekivanje $\boldsymbol{\mu}_j$ i kovarijanca $\boldsymbol{\Sigma}_j$ od $\mathbf{X}$ za svaki $j = 1, 2, \ldots, J$. Tada se za svaki $j = 1, 2, \ldots, J$ računa **Mahalanobisovu udaljenost** $D_j(\mathbf{x})$:

$$D_j(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)}. \tag{2.1}$$

   Opservacija $\mathbf{x}$ klasificira se u grupu j_0 tako da vrijedi

$$D_{j_0}(\mathbf{x}) = \min_{1 \le j \le J} D_j(\mathbf{x}). \tag{2.2}$$

2. Neka je poznata funkcija gustoće $f(\cdot|j)$ od $\mathbf{X}$ u grupi $j = 1, 2, \ldots, J$. Tada je $L(j|\mathbf{x}) = f(\mathbf{x}|j)$ vjerodostnojnost grupe $j = 1, 2, \ldots, J$. Opservacija $\mathbf{x}$ klasificira se u grupu j_0 tako da vrijedi

$$L(j_0|\mathbf{x}) = \max_{1 \le j \le J} L(j|\mathbf{x}) \tag{2.3}$$

3. Neka je poznata funkcija gustoće $f(\cdot|j)$ od $\mathbf{X}$ u grupi $j = 1, 2, \ldots, J$ i apriorne vjerojatnosti π_j , $j = 1, 2, \ldots, J$. Bayesovskom metodom dobiva se da je aposteriorna gustoća

$$p(j|\mathbf{x}) \propto \pi_j f(j|\mathbf{x}). \tag{2.4}$$

   Opservacija $\mathbf{x}$ klasificira se u grupu j_0 tako da vrijedi

$$p(j_0|\mathbf{x}) = \max_{1 \le j \le J} p(j|\mathbf{x}). \tag{2.5}$$

2.3 Linearna diskriminantna analiza

Neka $\mathbf{X}$ ima multivarijatnu normalnu distribuciju u svakoj grupi j i neka su pripadne kovarijance jednake, tj. $\mathbf{X}|j \sim \mathbf{N}(\boldsymbol{\mu}_j, \Sigma)$, $j = 1, 2, \ldots, J$. Tada su funkcije gustoće $f(\cdot|j)$ od $\mathbf{X}|j$, $j = 1, 2, \ldots, J$, dane s

$$f(\mathbf{x}|j) = \frac{1}{\sqrt{(2\pi)^p}\sqrt{\det\Sigma}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right]. \tag{2.6}$$

Klasifikacijsko pravilo 2.3 koje koristi metodu maksimalne vjerodostojnosti sada glasi: opservacija $\mathbf{x}$ klasificira se u grupu j_0 za koju vrijedi

$$f(\mathbf{x}|j_0) = \max_{1\le j\le J} \frac{1}{\sqrt{(2\pi)^p}\sqrt{\det\Sigma}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right]. \tag{2.7}$$

Ekvivalentno je

$$\begin{aligned} &\max_{1\le j\le J} \frac{1}{\sqrt{(2\pi)^p}\sqrt{\det\Sigma}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right] \Longleftrightarrow \\ &\max_{1\le j\le J}\left(-(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right) \Longleftrightarrow \min_{1\le j\le J}(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j), \end{aligned} \tag{2.8}$$

tj. klasifikacijska pravila koja koriste metodu maksimalne vjerodostojnosti i Mahalanobisovu udaljenost su ekvivalentna.

Neka su dodatno sve apriorne vjerojatnosti π_j jednake. Tada iz 2.3 - 2.5 slijedi da su klasifikacijska pravila koja koriste metodu maksimalne vjerodostojnosti i bayesovsku metodu ekvivalentna. Dakle, u ovom posebnom slučaju sve metode su međusobno ekvivalentne. [3] [4]

Raspisivanjem izraza iz 2.8

$$(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) = \mathbf{x}'\Sigma^{-1}\mathbf{x} - 2\boldsymbol{\mu}_j'\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_j'\Sigma^{-1}\boldsymbol{\mu}_j, \tag{2.9}$$

vidi se da izraz $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ ne ovisi o j. Oduzimanjem te konstante i dijeljenjem s -2 dobiva se klasifikacijsko pravilo: opservacija $\mathbf{x}$ klasificira se u grupu j_0 ako je

$$\boldsymbol{\mu}_{j_0}'\Sigma^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_{j_0}'\Sigma^{-1}\boldsymbol{\mu}_{j_0} = \max_{1\le j\le J}\left(\boldsymbol{\mu}_j'\Sigma^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j'\Sigma^{-1}\boldsymbol{\mu}_j\right). \tag{2.10}$$

Kako **klasifikacijske jednadžbe**

$$L_j(\mathbf{x}) = \boldsymbol{\mu}_j'\Sigma^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j'\Sigma^{-1}\boldsymbol{\mu}_j,\ j = 1, 2, \ldots, J, \tag{2.11}$$

linearno ovise o $\mathbf{x}$, ovakva diskriminantna analiza naziva se **linearnom diskriminantnom analizom**. [1]

2.4 Unakrsna validacija

Nakon određivanja klasifikacijskih jednadžbi određuje se koliko je dobivena klasifikacija dobra. Točnost klasifikacije može se odrediti pomoću **stope pogreške** (eng. ***error rate***), tj. udjela netočno klasificiranih podataka. Budući da su klasifikacijske funkcije dobivene procjenom iz danih podataka, često daju visoku stopu točno klasificiranih opservacija (odnosno malu stopu pogreške) na tim podacima. Iz tog razloga poželjno je testirati točnost klasifikacije na novim opservacijama. Takvo testiranje naziva se **unakrsna validacija** (eng. ***cross-validation***). [9]

Jedan od načina na koji se provodi unakrsna validacija je eng. ***leave-one-out*** unakrsna validacija. Provodi se tako da se izostavi jedna opservacija, procijene klasifikacijske jednadžbe na temelju preostalih podataka i zatim klasificira izostavljena opservacija. Ovaj postupak provodi se za svaku opservaciju. Stopa pogreške je udio netočno klasificiranih opservacija. Ovakav način unakrsne validacije može precijeniti stopu pogreške. [1]

Za veće skupove podataka često se koristi i eng. ***K-fold*** unakrsna validacija. Provodi se tako da se opservacije na slučajan način podijele u K grupa, odrede se klasifikacijske jednadžbe na temelju opservacija iz $K - 1$ grupa i pomoću njih klasificiraju opservacije iz preostale grupe. Postupak se provodi K puta tako da je svaka od K grupa jednom izostavljena u procijeni klasifikacijskih jednadžbi. U svakom od K koraka računa se stopa pogreške koja je udio netočno klasificiranih podataka, a ukupna stopa pogreške je njihova aritmetička sredina. *Leave-one-out* unakrsna validacija je poseban slučaj *K-fold* unakrsne validacije u kojem je $K = n$. [1]

2.5 Odabir prediktorskih varijabli u diskriminantnoj analizi

Postoji nekoliko načina na koje se biraju prediktorske varijable uključujene u analizu. Jedan od načina je standardna (direktna) diskriminantna analiza u kojoj su uključujene sve prediktorske varijable. Preostali načini su *forward*, *backward* i *stepwise* procedure.

Forward procedura počinje bez prediktorskih varijabli, a zatim traži varijablu koja najviše pridonosi diskriminantnoj snazi modela. Zatim se postupak ponavlja na preostalim neuključenim prediktorskim varijablama. Postupak završava kada nema dostupnih prediktorskih varijabli koje statistički značajno doprinose diskriminatnoj snazi.

Backward procedura počinje s uključenim svim prediktorskim varijablama. Iz modela se izbacuje varijabla koja je najmanje značajna za diskriminatornu snagu modela. Postupak se ponavlja do kad nisu isključene sve varijable koje nisu značajne na unaprijed definiranoj razini značajnosti.

Stepwise procedura je kombinacija *forward* i *backward* procedura. Počinje modelom bez prediktorskih varijabli koje se zatim dodaju jedna po jedna kao kod *forward* procedure, ali u *stepwise* proceduri nakon dodavanja k-te varijable moguće je izbaciti neku od ranije dodanih $k-1$ varijabli. [8, Poglavlje 93]

Diskriminatorna snaga modela mjeri se pomoću **Wilksove lambde**. Računa se po formuli

$$\Lambda = \frac{\det(S_w)}{\det(S_t)} = \frac{\det(S_w)}{\det(S_b) + \det(S_w)} \tag{2.12}$$

gdje su S_t, S_b, S_w kovarijancije koje predstavljaju, redom, ukupnu varijancu prediktorskih varijabli, varijancu prediktorskih varijabli između grupa i varijancu prediktorskih varijabli unutar grupa. Ako je vrijednost Λ blizu nule, tada prediktorske varijable dobro razdvajaju grupe. [8, Poglavlje 93] [9]

Poglavlje 3

Primjer

U ovom poglavlju koriste se logistička regresija i diskriminantna analiza kako bi se analizirala baza podataka [6]. Baza sadrži podatke o pacijenticama zabilježne u nekoliko bolnica u Indiji. Prati razne atribute kao što su godine, BMI, razine anti-Mullerovog i luteinizirajućeg hormona, trajanje menstruacije i sl. koji će biti nezavisne, tj. prediktorske varijable, te prisutnost sindroma policističnih jajnika, što će biti zavisna varijabla, odnosno varijabla za klasifikaciju u grupe. Opisuju se varijable korištene za analizu, prikazuje ih se grafički i daje se njihova deskriptivna statistika, a zatim provodi logistička regresija i diskriminantna analiza. Sve statističke analize rade se u *online* programu SAS OnDemand (SAS Institut Inc.). Logistička regresija dobiva se korištenjem LOGISTIC procedure, a diskriminantna analiza korištenjem DISCRIM i STEPDISC procedura.

3.1 Opis varijabli i deskriptivna statistika

Koristi se 25 nezavisnih varijabli te jedna zavisna. Nakon izbacivanja opservacija s nepotpunim podacima (jedna opservacija) i netipičnih vrijednosti (eng. *outlier*) (deset opservacija) baza sadrži 310 opservacija. Opisuje se svaka varijabla, zatim se neprekidne varijable prikazuju grafički histogramom te se navode njihove aritmetičke sredine, standardne devijacije, minimumi, medijani i maksimumi, a za kategorijske varijable prikazuju se njihove frekvencijske tablice.

- **PCOS** - poprima vrijednosti 0 = *pacijentica nema sindrom policističnih jajnika*, 1 = *pacijentica ima sindrom policističnih jajnika*
- **Age** - starost izražena u godinama, neprekidna varijabla
- **BMI** - indeks tjelesne težine (eng. *body mass index*), neprekidna varijabla
- **Hb** - razina hemoglobina u krvi izražena u g/dL, neprekidna varijabla

- **Period_length** - trajanje menstruacije, diskretna varijabla
- **FSH** - razina folikulostimulirajućeg hormona izražena u IU/L, neprekidna varijabla
- **LH** - razina luteinizirajućeg hormona izražena u IU/L, neprekidna varijabla
- **LH_FSH_ratio** - omjer razina luteinizirajućeg i folikulostimulirajućeg hormona, neprekidna varijabla
- **TSH** - razina tiroidnog stimulirajućeg hormona izražena u mIU/L, neprekidna varijabla
- **AMH** - razina anti-Mullerovog hormona izražena u ng/mL, neprekidna varijabla
- **PRG** - razina progesterona izražena u ng/mL, neprekidna varijabla
- **PRL** - razina prolaktina izražena u ng/mL, neprekidna varijabla
- **Vit_D3** - razina vitamina D3 izražena u ng/mL, neprekidna varijabla
- **RBS** - razina glukoze u krvi izražena u mg/dL, neprekidna varijabla
- **Follicle_No_L** - broj folikularnih cista na lijevom jajniku, diskretna varijabla
- **Follicle_No_R** - broj folikularnih cista na desnom jajniku, diskretna varijabla
- **Avg_F_size_L** - prosječna veličina folikularnih cista na lijevom jajniku izražena u mm, neprekidna varijabla
- **Avg_F_size_R** - prosječna veličina folikularnih cista na desnom jajniku izražena u mm, neprekidna varijabla
- **Endometrium** - debljina endometrija maternice izražena u mm, neprekidna varijabla
- **Cycle** - poprima vrijednosti 0 = *pacijentica ima redovite menstrualne cikluse*, 1 = *pacijentica ima neredovite menstrualne cikluse*
- **Weight_gain** - poprima vrijednosti 0 = *nema povećanja tjelesne težine*, 1 = *povećanje tjelesne težine*
- **Hair_growth** - poprima vrijednosti 0 = *nema povećane dlakavosti*, 1 = *povećana dlakavost*
- **Skin_darkening** - poprima vrijednosti 0 = *nema hiperpigmentacije kože*, 1 = *hiperpigmentacija kože*

- **Hair_loss** - poprima vrijednosti 0 = *nema ispadanja kose*, 1 = *ispadanje kose*
- **Acne** - poprima vrijednosti 0 = *nema akni*, 1 = *pojava akni*
- **Reg_Exercise** - poprima vrijednosti 0 = *ne vježba redovito*, 1 = *vježba redovito*

PCOS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	205	66.13	205	66.13
1	105	33.87	310	100.00

Tablica 3.1: PCOS - frekvencijska tablica (ispis iz SAS-a)

Cycle	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	215	69.35	215	69.35
1	95	30.65	310	100.00

Tablica 3.2: Cycle - frekvencijska tablica (ispis iz SAS-a)

Weight_gain	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	188	60.65	188	60.65
1	122	39.35	310	100.00

Tablica 3.3: Weight_gain - frekvencijska tablica (ispis iz SAS-a)

Hair_growth	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	219	70.65	219	70.65
1	91	29.35	310	100.00

Tablica 3.4: Hair_growth - frekvencijska tablica (ispis iz SAS-a)

Skin_darkening	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	218	70.32	218	70.32
1	92	29.68	310	100.00

Tablica 3.5: Skin_darkening - frekvencijska tablica (ispis iz SAS-a)

Hair_loss	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	171	55.16	171	55.16
1	139	44.84	310	100.00

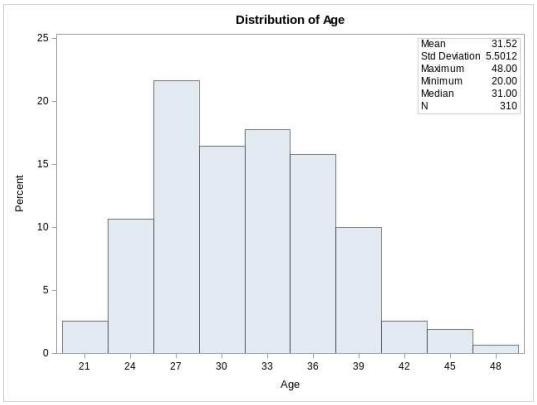
Tablica 3.6: Hair_loss - frekvencijska tablica (ispis iz SAS-a)

Acne	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	169	54.52	169	54.52
1	141	45.48	310	100.00

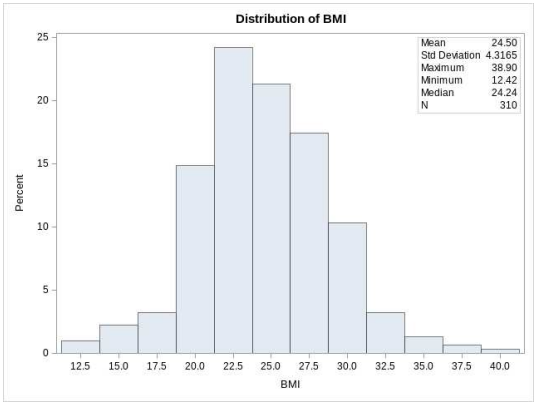
Tablica 3.7: Acne - frekvencijska tablica (ispis iz SAS-a)

Reg_Exercise	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	228	73.55	228	73.55
1	82	26.45	310	100.00

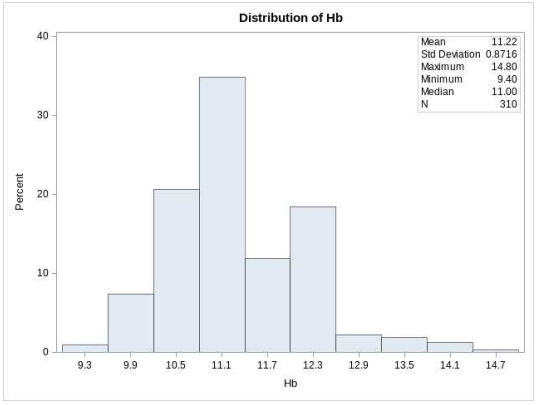
Tablica 3.8: Reg_exercise - frekvencijska tablica (ispis iz SAS-a)



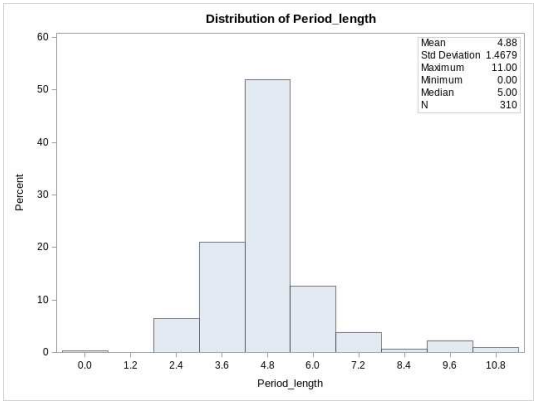
Slika 3.1: Age - histogram (ispis iz SAS-a)



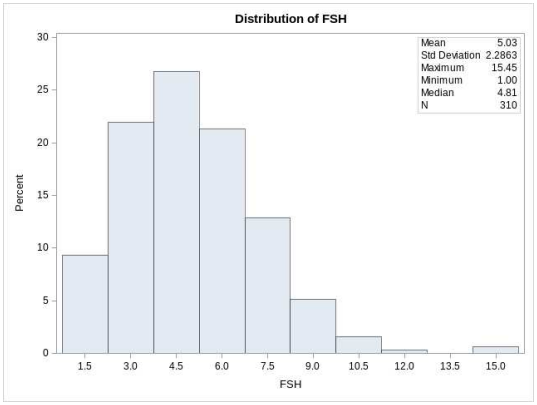
Slika 3.2: BMI - histogram (ispis iz SAS-a)



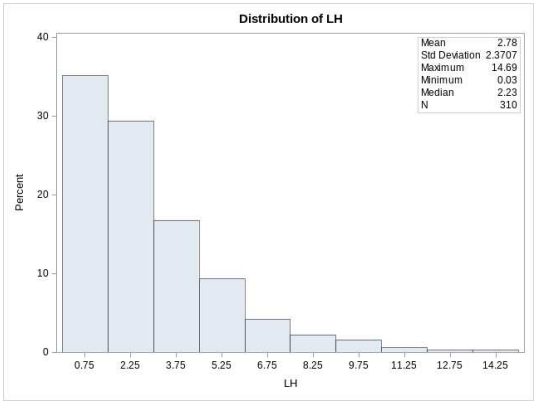
Slika 3.3: Hb - histogram (ispis iz SAS-a)



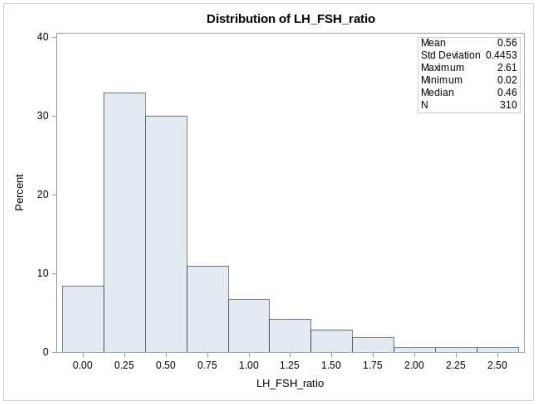
Slika 3.4: Period length - histogram (ispis iz SAS-a)



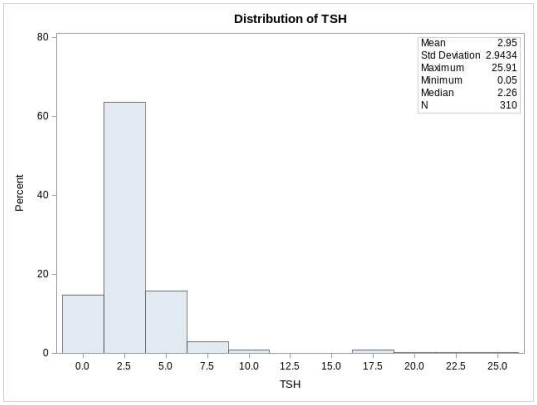
Slika 3.5: FSH - histogram (ispis iz SAS-a)



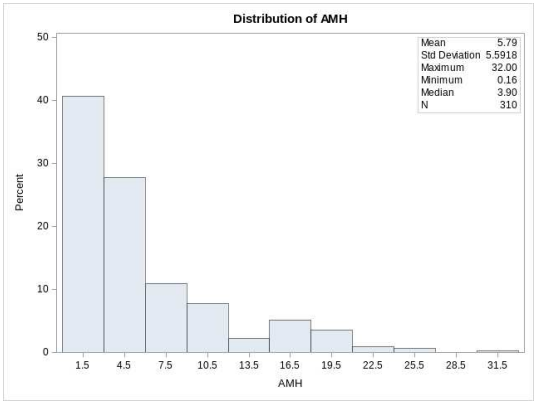
Slika 3.6: LH - histogram (ispis iz SAS-a)



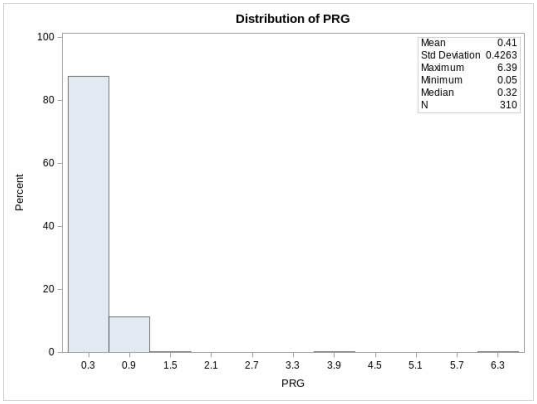
Slika 3.7: LH_FSH_ratio - histogram (ispis iz SAS-a)



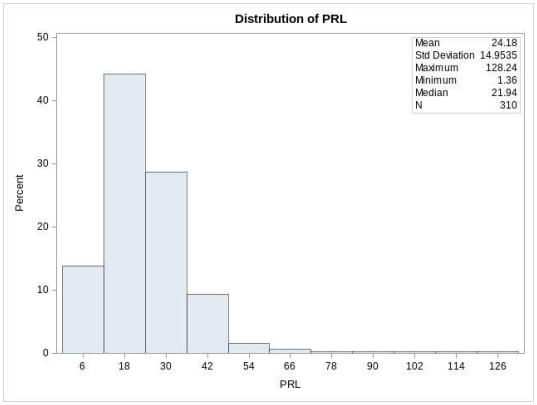
Slika 3.8: TSH - histogram (ispis iz SAS-a)



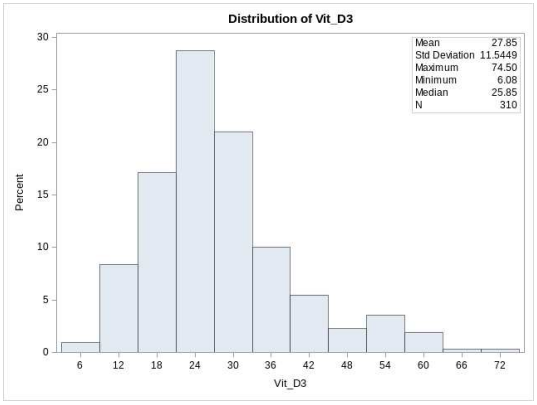
Slika 3.9: AMH - histogram (ispis iz SAS-a)



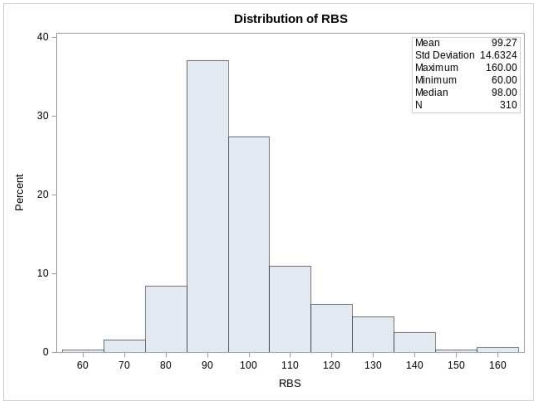
Slika 3.10: PRG - histogram (ispis iz SAS-a)



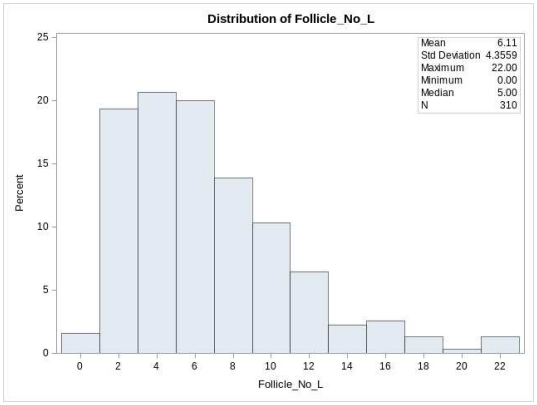
Slika 3.11: PRL - histogram (ispis iz SAS-a)



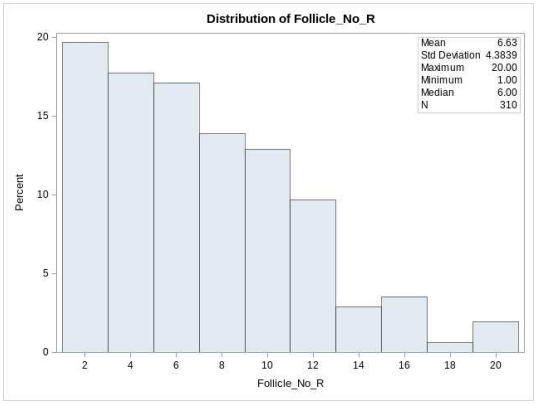
Slika 3.12: Vit_D3 - histogram (ispis iz SAS-a)



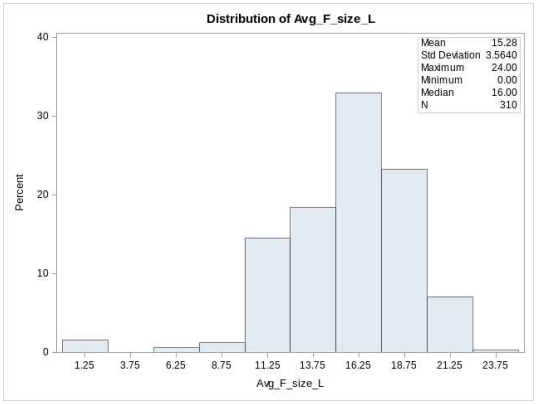
Slika 3.13: RBS - histogram (ispis iz SAS-a)



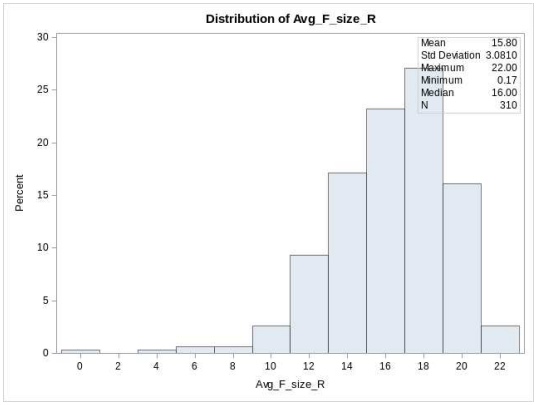
Slika 3.14: Follicle_No_L - histogram (ispis iz SAS-a)



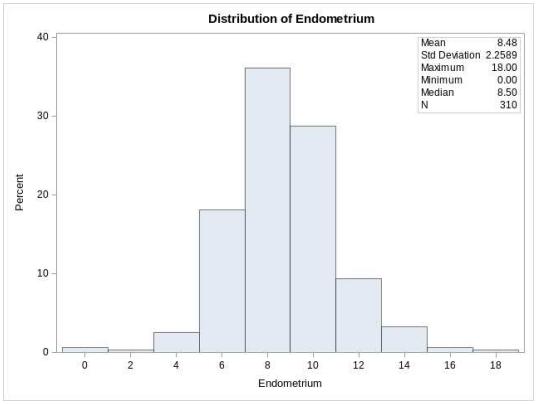
Slika 3.15: Follicle_No_R - histogram (ispis iz SAS-a)



Slika 3.16: Avg_F_size_L - histogram (ispis iz SAS-a)



Slika 3.17: Avg_F_size_R - histogram (ispis iz SAS-a)



Slika 3.18: Endometrium - histogram (ispis iz SAS-a)

3.2 Logistička regresija

U ovom poglavlju provodi se logistička regresija, najprije univarijatna za sve navedene nezavisne varijable, a zatim multivarijatna. U svakom modelu zavisna varijabla je PCOS.

Univarijatna logistička regresija

Rezultati testova adekvatnosti univarijatnih logističkih modela, procjene parametara i omjera šansi prikazani su tablično.

Variable	Likelihood ratio	p-value
Age	14.1224	0.0002
BMI	16.3327	< 0.0001
Hb	1.1749	0.2784
Period_length	4.6170	0.0317
FSH	16.8006	< 0.0001
LH	0.9410	0.3320
LH_FSH_ratio	7.7901	0.0053
TSH	0.5068	0.4765
AMH	17.3940	< 0.0001
PRG	1.1814	0.2771
PRL	0.0186	0.8914
Vit_D3	0.6296	0.4275
RBS	1.6289	0.2019
Follicle_No_L	121.8881	< 0.0001
Follicle_No_R	138.3385	< 0.0001
Avg_F_size_L	2.1582	< 0.0001
Avg_F_size_R	0.0157	0.9004
Endometrium	2.4535	0.1173
Cycle	59.0336	< 0.0001
Weight_gain	69.2147	< 0.0001
Hair_growth	66.0606	< 0.0001
Skin_darkening	64.2349	< 0.0001
Hair_loss	18.760	< 0.0001
Acne	23.9808	< 0.0001
Reg_exercise	1.3038	0.2535

Tablica 3.9: Rezultati testiranja adekvatnosti univarijatnih logističkih modela

Varijable čiji su univarijatni modeli statistički značajno bolji od modela bez nezavisnih varijabli na razini značajnosti $\alpha = 0.05$ su Age, BMI, Period_length, FSH, LH_FSH_ratio, AMH, Follicle_No_L, Follicle_No_R, Avg_F_size_L, Cycle, Weight_gain, Hair_growth, Skin_darkening, Hair_loss, Acne.

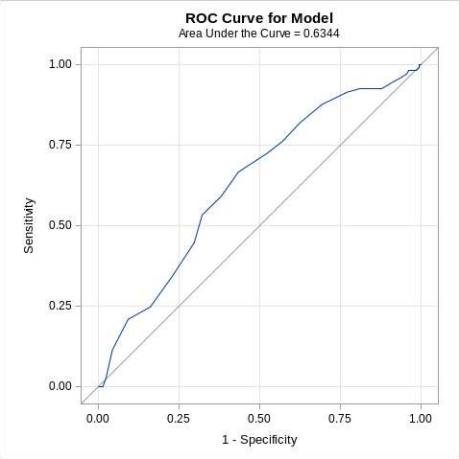
Variable	Parameter estimate	Odds ratio estimate	Odds ratio 95% CL
Age	-0.0861	0.918	(0.876, 0.961)
BMI	0.1168	1.124	(1.059, 1.192)
Hb	0.1483	1.160	(0.887, 1.516)
Period_length	-0.1830	0.833	(0.701, 0.989)
FSH	-0.2374	0.789	(0.699, 0.890)
LH	0.0485	1.050	(0.952, 1.157)
LH_FSH_ratio	0.7397	2.095	(1.238, 3.546)
TSH	0.0283	1.029	(0.952, 1.111)
AMH	0.0890	1.093	(1.047, 1.141)
PRG	-0.4190	0.658	(0.262, 1.651)
PRL	-0.00109	0.999	(0.983, 1.015)
Vit_D3	-0.00837	0.992	(0.971, 1.013)
RBS	0.0104	1.010	(0.995, 1.027)
Follicle_No_L	0.4106	1.508	(1.368, 1.662)
Follicle_No_R	0.4333	1.542	(1.400, 1.700)
Avg_F_size_L	0.0513	1.053	(0.981, 1.129)
Avg_F_size_R	0.00488	1.005	(0.931, 1.085)
Endometrium	0.0838	1.087	(0.978, 1.208)
Cycle	2.0169	7.515	(4.386, 12.877)
Weight_gain	2.1287	8.404	(4.926, 14.337)
Hair_growth	2.1691	8.751	(5.035, 15.209)
Skin_darkening	2.1295	8.411	(4.858, 14.562)
Hair_loss	1.0566	2.877	(1.769, 4.678)
Acne	1.2009	3.323	(2.032, 5.435)
Reg_exercise	0.3067	1.359	(0.805, 2.294)

Tablica 3.10: Rezultati procjene parametara i omjera šansi univarijatnih logističkih modela

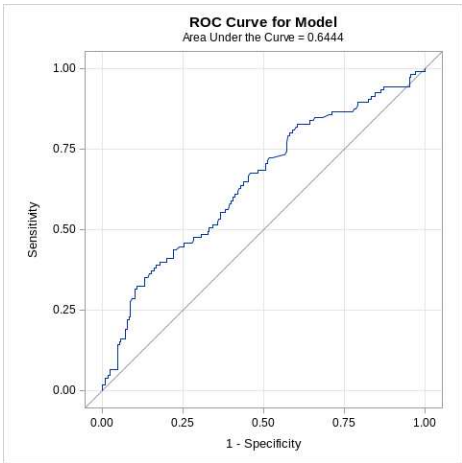
U Tablici 3.10 dane su procjene parametara, procjene omjera šansi za prelazak iz grupe $PCOS = 0$ u grupu $PCOS = 1$ i 95% pouzdani intervali za omjere šansi. Intervali pouzdanosti su jedan od načina testiranja značajnosti parametra u modelu - ako interval ne sadrži 1, parametar je značajan. U ovom primjeru varijable koje su značajne na razini $\alpha = 0.05$ i interpretacija pripadnih parametara je sljedeća:

- Age - povećanjem starosti pacijentice za jednu godinu izglednost prisutnosti PCOS-a povećava se 0.918 puta, odnosno smanjuje za 8.2%
- BMI - povećanjem BMI-a za jedan izglednost prisutnosti PCOS-a povećava se 1.124 puta, odnosno za 12.4%
- Period_length - povećanjem trajanja menstruacije za jedan dan izglednost prisutnosti PCOS-a povećava se 0.833 puta, odnosno smanjuje za 16.7%
- FSH - povećanjem razine folikulostimulirajućeg hormona za 1 IU/L povećava se izglednost prisutnosti PCOS-a povećava se 0.789, odnosno smanjuje za 21.1%
- LH_FSH_ratio - povećanjem omjera luteinizirajućeg hormona i folikulostimulirajućeg hormona za jedan povećava se izglednost prisutnosti PCOS-a 2.095 puta, odnosno za 109.5%
- AMH - povećanjem razine anti-Mullerovog hormona za jedan povećava se izglednost prisutnosti PCOS-a 1.093 puta, odnosno za 9.3%
- Follicle_No_L - povećanjem broja folikularnih cista na lijevom jajniku za jedan povećava se izglednost prisutnosti PCOS-a 1.508 puta, odnosno za 50.8%
- Follicle_No_R - povećanjem broja folikularnih cista na desnom jajniku za jedan povećava se izglednost prisutnosti PCOS-a 1.542 puta, odnosno za 54.2%
- Cycle - kod pacijentica s neredovitim ciklusima prisutnost PCOS-a izglednija je 7.515 puta nego kod pacijentica s redovitim ciklusima
- Weight_gain - kod pacijentica kojima se povećala tjelesna težina prisutnost PCOS-a je 8.404 puta izglednija nego kod pacijentica kojima se tjelesna težina nije povećala
- Hair_growth - kod pacijentica s pojačanim rastom dlaka prisutnost PCOS-a je 8.751 puta izglednija nego kod pacijentica koje nemaju pojačan rast dlaka
- Skin_darkening - kod pacijentica s hiperpigmentacijom kože prisutnost PCOS-a je 8.411 puta izglednija nego kod pacijentica koje nemaju hiperpigmentaciju kože
- Hair_loss - kod pacijentica kojima ispada kosa prisutnost PCOS-a je 2.877 puta izglednija nego kod pacijentica kojima ne ispada kosa
- Acne - kod pacijentica s aknama prisutnost PCOS-a je 1.359 puta izglednija nego kod pacijentica koje nemaju akne

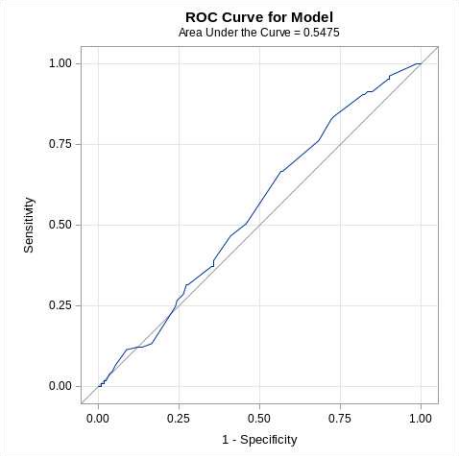
Prediktivna snaga logističkog modela mjeri se pomoću površine ispod ROC krivulje. Na sljedećim slikama prikazane su ROC krivulje i površine ispod njih za univarijatne logističke modele.



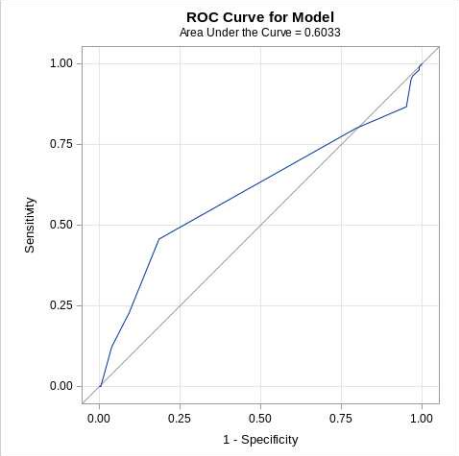
Slika 3.19: Age - ROC krivulja (ispis iz SAS-a)



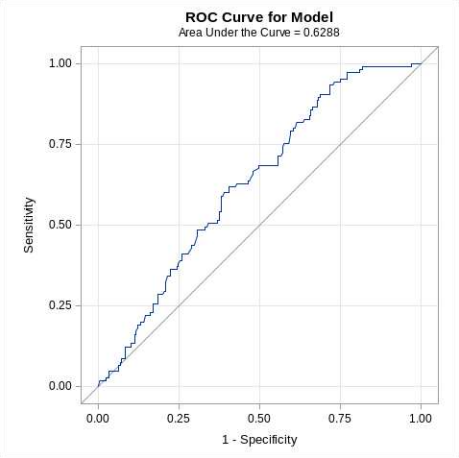
Slika 3.20: BMI - ROC krivulja (ispis iz SAS-a)



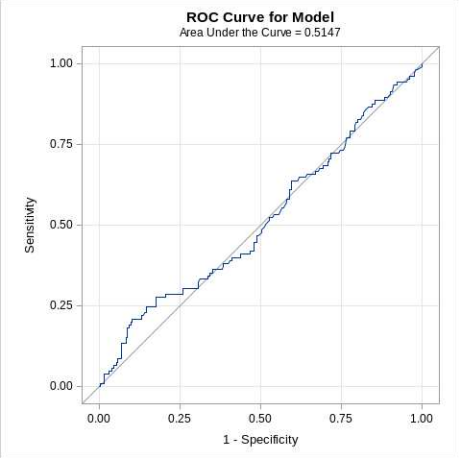
Slika 3.21: Hb - ROC krivulja (ispis iz SAS-a)



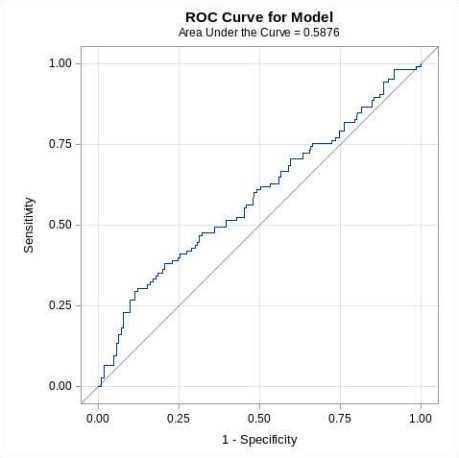
Slika 3.22: Period_length - ROC krivulja (ispis iz SAS-a)



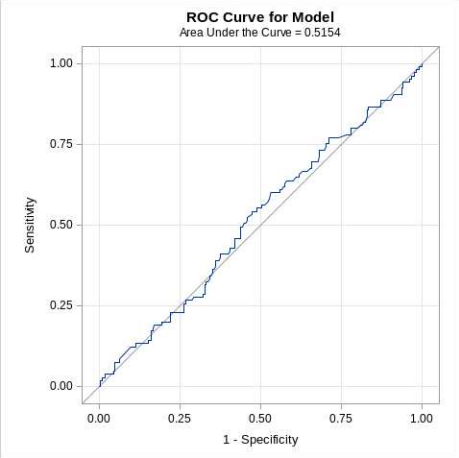
Slika 3.23: FSH - ROC krivulja (ispis iz SAS-a)



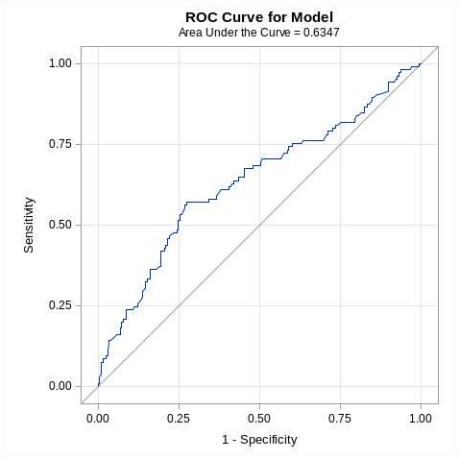
Slika 3.24: LH - ROC krivulja (ispis iz SAS-a)



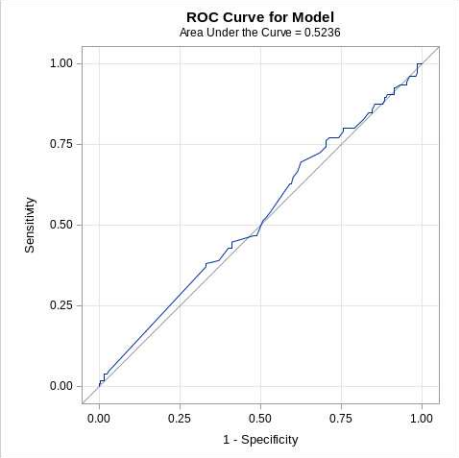
Slika 3.25: LH_FSH_ratio - ROC krivulja (ispis iz SAS-a)



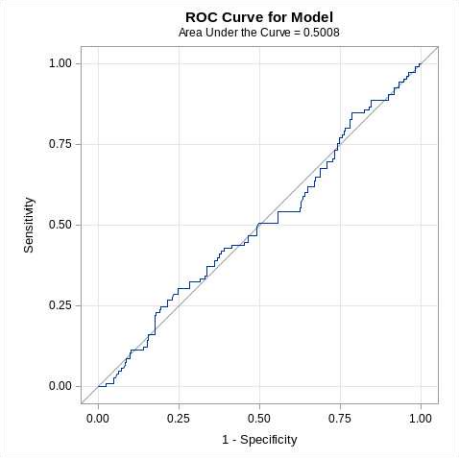
Slika 3.26: TSH - ROC krivulja (ispis iz SAS-a)



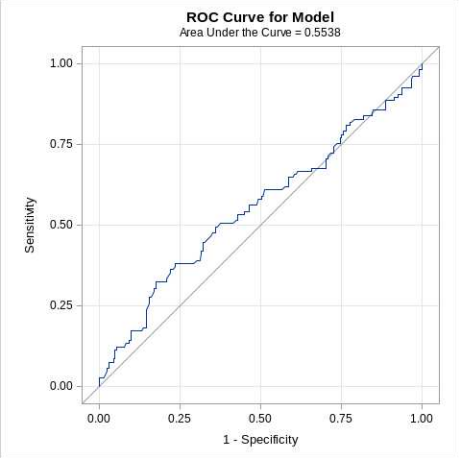
Slika 3.27: AMH - ROC krivulja (ispis iz SAS-a)



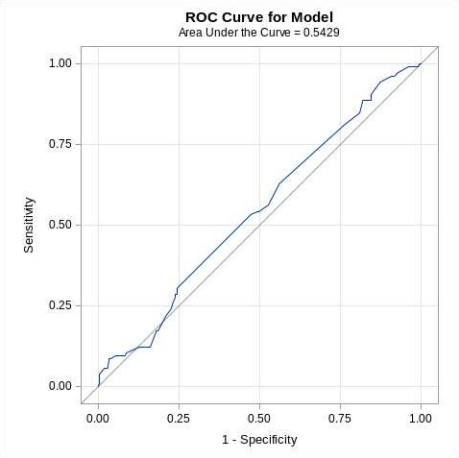
Slika 3.28: PRG - ROC krivulja (ispis iz SAS-a)



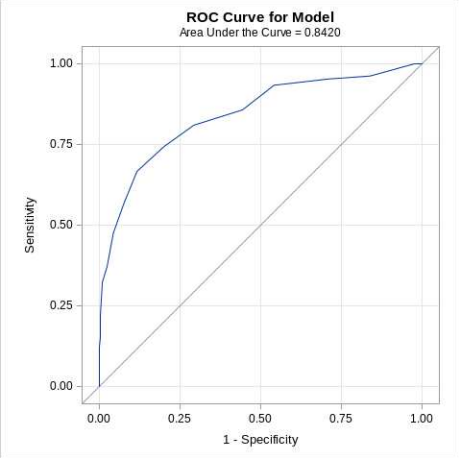
Slika 3.29: PRL - ROC krivulja (ispis iz SAS-a)



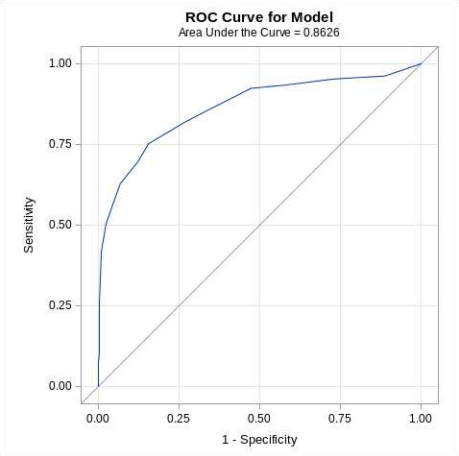
Slika 3.30: Vit_D3 - ROC krivulja (ispis iz SAS-a)



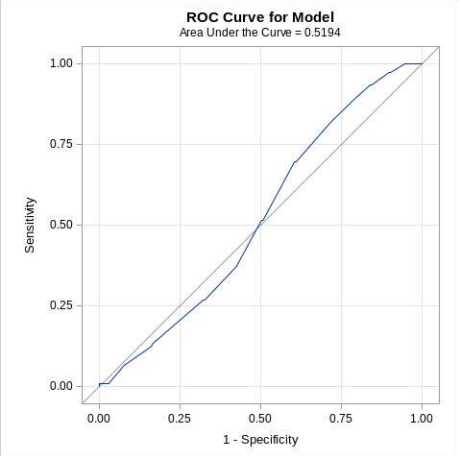
Slika 3.31: RBS - ROC krivulja (ispis iz SAS-a)



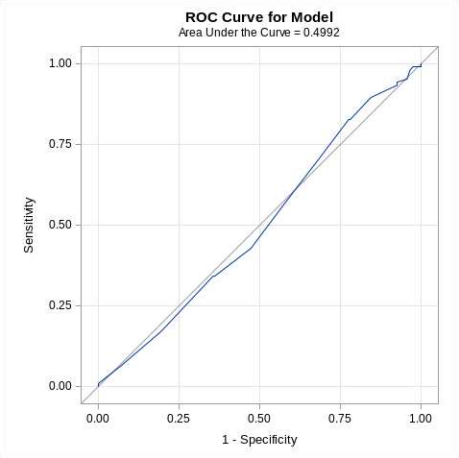
Slika 3.32: Follicle_No_L - ROC krivulja (ispis iz SAS-a)



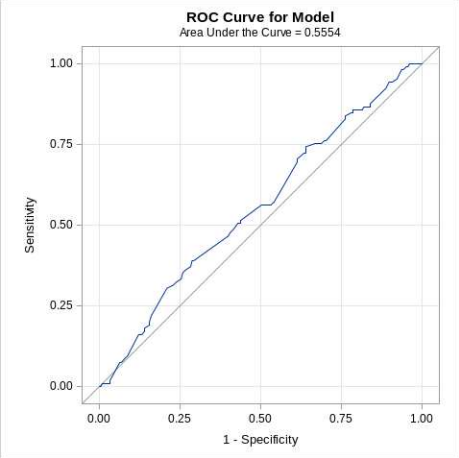
Slika 3.33: Follicle_No_R - ROC krivulja (ispis iz SAS-a)



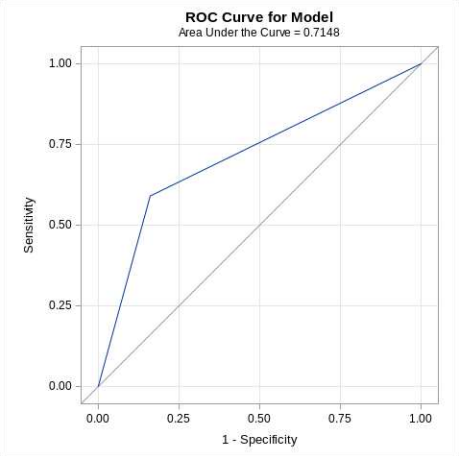
Slika 3.34: Avg_F_size_L - ROC krivulja (ispis iz SAS-a)



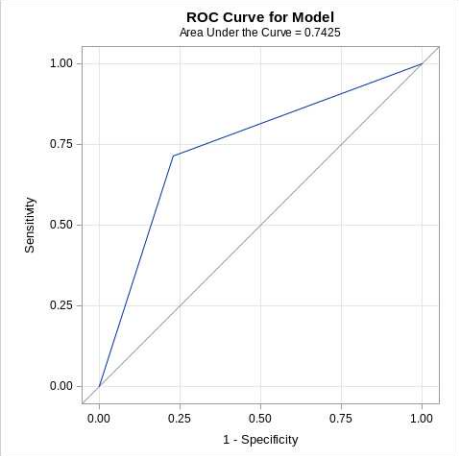
Slika 3.35: Avg_F_size_R - ROC krivulja (ispis iz SAS-a)



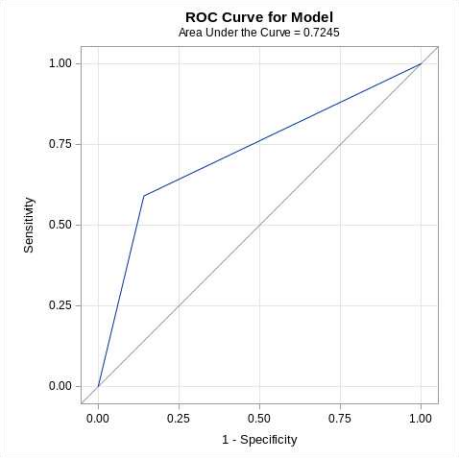
Slika 3.36: Endometrium - ROC krivulja (ispis iz SAS-a)



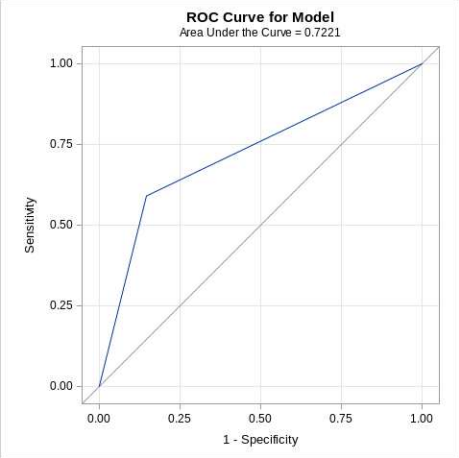
Slika 3.37: Cycle - ROC krivulja (ispis iz SAS-a)



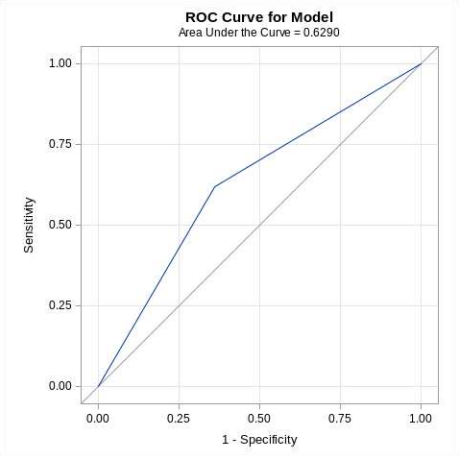
Slika 3.38: Weight_gain - ROC krivulja (ispis iz SAS-a)



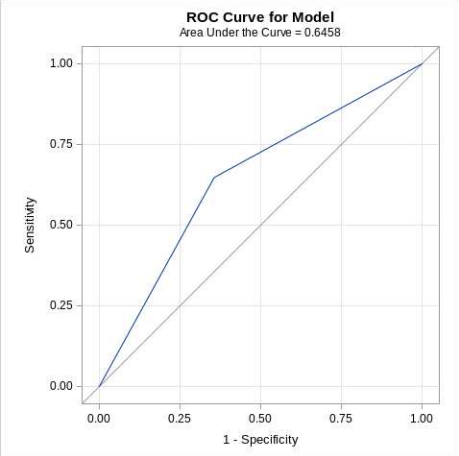
Slika 3.39: Hair_growth - ROC krivulja (ispis iz SAS-a)



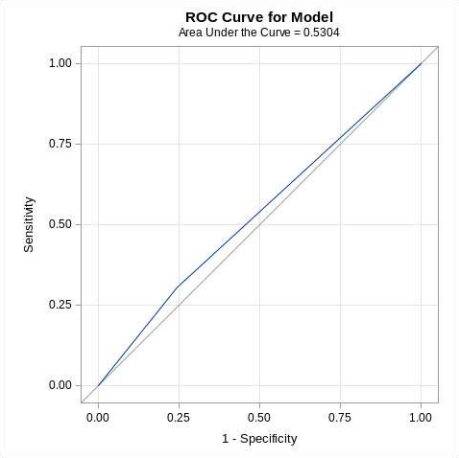
Slika 3.40: Skin_darkening - ROC krivulja (ispis iz SAS-a)



Slika 3.41: Hair_loss - ROC krivulja (ispis iz SAS-a)



Slika 3.42: Acne - ROC krivulja (ispis iz SAS-a)



Slika 3.43: Reg_exercise - ROC krivulja (ispis iz SAS-a)

Multivarijatna logistička regresija

Određen je multivarijatni logistički model koji uključuje sve nezavisne varijable čiji su se univarijatni modeli pokazali statistički značajnima na razini značajnosti $\alpha = 0.05$. Rezultati su prikazani tablično kao i ranije.

Likelihood ratio	df	p-value
244.8293	15	< 0.0001

Tablica 3.11: Rezultati testiranja adekvatnosti multivarijatnog logističkog modela

P-vrijednost testa omjera vjerodostojnosti je manja od 0.0001 pa je model statistički značajno bolji od modela bez nezavisnih varijabli na razini značajnosti $\alpha = 0.05$.

Variable	Parameter estimate	p-value
Intercept	-7.6866	0.003
Age	-0.0588	0.1316
BMI	0.0647	0.2851
Period_length	0.1008	0.4977
FSH	-0.1004	0.3797
LH_FSH_ratio	0.4645	0.3373
AMH	0.0498	0.2224
Follicle_No_L	0.1375	0.127
Follicle_No_R	0.3613	< 0.0001
Avg_F_size_L	0.04	0.6078
Cycle	1.1148	0.0175
Weight_gain	1.0757	0.0397
Hair_growth	1.3007	0.0101
Skin_darkening	1.0707	0.0271
Hair_loss	0.5797	0.2188
Acne	1.0464	0.0208

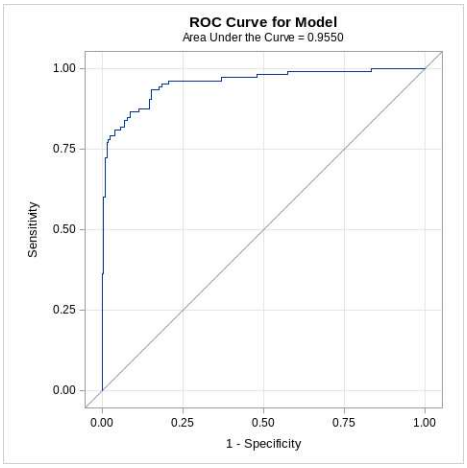
Tablica 3.12: Rezultati procjene parametara i pripadne p-vrijednosti za multivarijatni logistički model

Statistički značajne varijable na razini značajnosti $\alpha = 0.05$ su Follicle_No_R, Cycle, Weight_gain, Hair_growth, Skin_darkening, Acne.

Variable	Odds ratio estimate	Odds ratio 95% CL
Age	0.943	(0.874, 1.018)
BMI	1.067	(0.948, 1.201)
Period_length	1.106	(0.826, 1.48)
FSH	0.904	(0.723, 1.132)
LH_FSH_ratio	1.591	(0.616, 4.11)
AMH	1.051	(0.97, 1.139)
Follicle_No_L	1.147	(0.962, 1.369)
Follicle_No_R	1.435	(1.209, 1.703)
Avg_F_size_L	1.041	(0.893, 1.213)
Cycle	3.049	(1.215, 7.648)
Weight_gain	2.932	(1.052, 8.172)
Hair_growth	3.672	(1.363, 9.893)
Skin_darkening	2.917	(1.129, 7.539)
Hair_loss	1.785	(0.709, 4.498)
Acne	2.847	(1.172, 6.915)

Tablica 3.13: Rezultati procjene omjera šansi i pripadnih 95% pouzdanih intervala za multivarijatni logistički model

Interpretacija omjera šansi jednaka je kao kod univarijatnog modela. Površina ispod ROC krivulje je $c = 0.955$, što znači da je prediktivna snaga modela velika.



Slika 3.44: ROC krivulja multivarijatnog logističkog modela (ispis iz SAS-a)

Stepwise procedura logističke regresije

Provedena je *stepwise* procedura logističke regresije na razini značajnosti $\alpha = 0.15$ za dodavanje i izbacivanje varijabli iz modela.

Likelihood ratio	df	p-value
242.7550	9	< 0.0001

Tablica 3.14: Rezultati testiranja adekvatnosti logističkog modela dobivenog *stepwise* procedurom

Step	Effect entered	Effect removed	p-value
1	Follicle_No_R		< 0.0001
2	Weight_gain		< 0.0001
3	Hair_growth		< 0.0001
4	Cycle		0.0003
5	Skin_darkening		0.001
6	Follicle_No_L		0.0303
7	Acne		0.1
8	Age		0.0621
9	RBS		0.0413

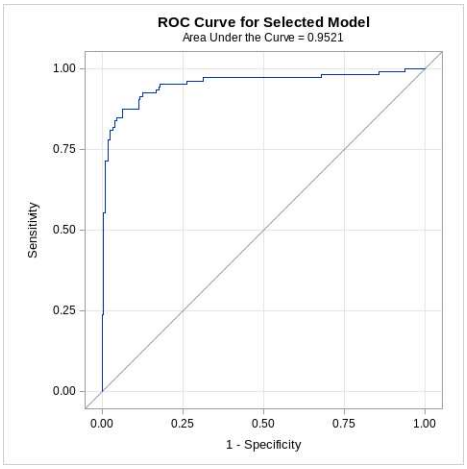
Tablica 3.15: Koraci *stepwise* procedure logističke regresije

Variable	Odds ratio estimate	Odds ratio 95% CL
Age	0.926	(0.86, 0.996)
RBS	1.032	(1.001, 1.064)
Follicle_No_L	1.17	(0.995, 1.376)
Follicle_No_R	1.419	(1.199, 1.68)
Cycle	3.894	(1.639, 9.252)
Weight_gain	4.231	(1.763, 10.158)
Hair_growth	3.817	(1.463, 9.959)
Skin_darkening	3.455	(1.384, 8.627)
Acne	2.541	(1.078, 5.988)

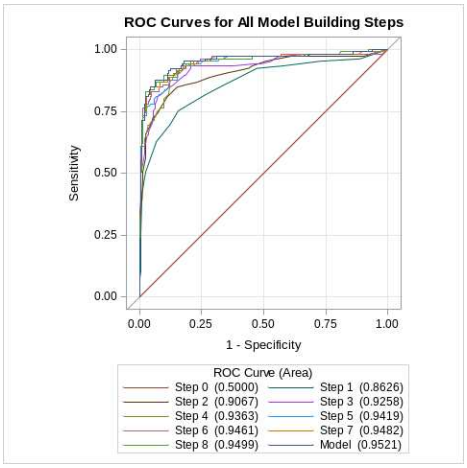
Tablica 3.16: Rezultati procjene omjera šansi i pripadnih 95% pouzdanih intervala za logistički model dobiven *stepwise* procedurom

U model su redom dodane varijable Follicle_No_R, Weight_gain, Hair_growth, Cycle, Skin_darkening, Follicle_No_L, Acne, Age i RBS. Ni jedna varijabla nije izbačena iz modela.

Površina ispod ROC krivulje je $c = 0.952$ pa model ima veliku prediktivnu snagu.



Slika 3.45: ROC krivulja logističkog modela dobivenog *stepwise* procedurom (ispis iz SAS-a)



Slika 3.46: ROC krivulja po koracima *stepwise* procedure logističke regresije (ispis iz SAS-a)

3.3 Diskriminantna analiza

Jedna od pretpostavki linearne diskriminantne analize bila je multivarijatna normalna distribucija prediktorskih varijabli po grupama. Budući da u ovom primjeru postoje dihotomne varijable, ta pretpostavka očito nije zadovoljena. No, prema [1] linearna diskriminantna analiza je robusna na odstupanja od normalnosti čak i u slučaju dihotomnih varijabli pa se zanemaruju pretpostavke i provodi linearna diskriminantna analiza.

Direktna diskriminantna analiza

Provedena je direktna diskriminantna analiza koristeći sve prediktorske varijable. Koristi se pretpostavka da su kovarijance prediktorskih varijabli jednake po grupama $PCOS = 0$ i $PCOS = 1$ te da su apriorne vjerojatnosti 0.5 za obje grupe. Dobivene su sljedeće klasifikacijske funkcije:

Linear Discriminant Function for PCOS		
Variable	0	1
Constant	-221.33948	-229.30662
Age	1.44930	1.38621
BMI	1.73496	1.77302
Hb	17.56242	17.62331
Period_length	3.14576	3.17389
FSH	3.86007	3.63421
LH	-4.81544	-4.47204
LH_FSH_ratio	30.12620	28.93847
TSH	0.35332	0.39344
AMH	0.22700	0.27749
PRG	5.44785	5.17610
PRL	-0.01287	-0.01526
Vit_D3	0.30477	0.28348
RBS	0.38696	0.40583
Follicle_No_L	-1.05723	-0.93986
Follicle_No_R	1.25065	1.65845
Avg_F_size_L	0.97769	0.99366
Avg_F_size_R	1.92606	1.93698
Endometrium	2.49725	2.49469
Cycle	0.85255	2.24027
Weight_gain	-2.87077	-1.45548
Hair_growth	-7.28249	-5.34432
Skin_darkening	7.66765	9.21102
Hair_loss	2.22631	2.57093
Acne	2.34265	3.18717
Reg_Exercise	-0.45372	-0.21230

Tablica 3.17: Klasifikacijske funkcije direktne diskriminantne analize (ispis iz SAS-a)

Uvrštavanjem opservacija u dobivene klasifikacijske funkcije dobivaju se aposteriorne vjerojatnosti, npr. za prvu opservaciju, koja ima vrijednosti $PCOS = 0, Age = 28, BMI = 19.30, Hb = 10.48, Period\_length = 5, FSH = 7.95, LH = 3.68, LH\_FSH\_ratio = 0.4629, TSH = 0.68, AMH = 2.07, PRG = 0.57, PRL = 45.16, Vit\_D3 = 17.1, RBS = 92, Follicle\_No\_L = 3, Follicle\_No\_R = 3, Avg\_F\_size\_L = 18, Avg\_F\_size\_R = 18,$ $Endometrium = 8.5, Cycle = 0, Weight\_gain = 0, Hair\_growth = 0, Skin\_darkening = 0, Hair\_loss = 0, Acne = 0, Reg\_Exercise = 0$, dobiva se

$$\begin{aligned} L_0 = -221.33948 &+ 1.4493Age + 1.73496BMI + 17.56242Hb + 3.14576Period\_length \\ &+3.86007FSH - 4.81544LH + 30.1262LH\_FSH\_ratio + 0.35332TSH + 0.227AMH \\ &+5.44785PRG - 0.01287PRL + 0.30477Vit\_D3 + 0.38696RBS - 1.05723Follicle\_No\_L \\ &+1.25065Follicle\_No\_R + 0.97769Avg\_F\_size\_L + 1.92606Avg\_F\_size\_R \\ &+2.49725Endometrium + 0.85255Cycle - 2.87077\ Weight\_gain - 7.28249Hair\_growth \\ &+7.66765Skin\_darkening + 2.22631Hair\_loss + 2.34265Acne - 0.45372Reg\_Exercise \end{aligned} \tag{3.1}$$

$$\begin{aligned} L_0 = -221.33948 &+ 1.4493 * 28 + 1.73496 * 19.30 + 17.56242 * 10.48 + 3.14576 * 5 \\ &+3.86007 * 7.95 - 4.81544 * 3.68 + 30.1262 * 0.4629 + 0.35332 * 0.68 + 0.227 * 2.07 \\ &+5.44785 * 0.57 - 0.01287 * 45.16 + 0.30477 * 17.1 + 0.38696 * 92 - 1.05723 * 3 \\ &+1.25065 * 3 + 0.97769 * 18 + 1.92606 * 18 \\ &+2.49725 * 8.5 + 0.85255 * 0 - 2.87077 * 0 - 7.28249 * 0 \\ &+7.66765 * 0 + 2.22631 * 0 + 2.34265 * 0 - 0.45372 * 0 \end{aligned} \tag{3.2}$$

$$L_0 = 0.9976 \tag{3.3}$$

$$\begin{aligned} L_1 = -229.30662 &+ 1.38621Age + 1.77302BMI + 17.62331Hb + 3.17389Period\_length \\ &+3.63421FSH - 4.47204LH + 28.93847LH\_FSH\_ratio + 0.39344TSH + 0.27749AMH \\ &+5.1761PRG - 0.01526PRL + 0.28348Vit\_D3 + 0.40583RBS - 0.93986Follicle\_No\_L \\ &+1.65845Follicle\_No\_R + 0.99366Avg\_F\_size\_L + 1.93698Avg\_F\_size\_R \\ &+2.49469Endometrium + 2.24027Cycle - 1.45548Weight\_gain - 5.34432Hair\_growth \\ &+9.21102Skin\_darkening + 2.57093Hair\_loss + 3.18717Acne - 0.2123Reg\_Exercise \end{aligned} \tag{3.4}$$

$$L_1 = 0.0024 \tag{3.5}$$

Prva opservacija klasificira se u grupu $PCOS = 0$ jer je $L_0 > L_1$.

Kako bi se ocijenilo koliko je dobra klasifikacija dobivena temeljem ovih klasifikacijskih funkcija, gledaju se stope pogreške klasifikacije i stope pogreške *leave-one-out* unakrsne validacije.

Number of Observations and Percent Classified into PCOS

From PCOS	0	1	Total
0	192 93.66	13 6.34	205 100.00
1	13 12.38	92 87.62	105 100.00
Total	205 66.13	105 33.87	310 100.00
Priors	0.5	0.5	

Error Count Estimates for PCOS

	0	1	Total
Rate	0.0634	0.1238	0.0936
Priors	0.5000	0.5000	

Tablica 3.18: Stope pogreške direktne diskriminantne analize (ispis iz SAS-a)

Stopa lažno pozitivnih (odnosno opservacija koje pripadaju grupi *PCOS = 0*, a klasificirane su u grupu *PCOS = 1*) je 6.34%. Stopa lažno negativnih (odnosno opservacija koje pripadaju grupi *PCOS = 1*, a klasificirane su u grupu *PCOS = 0*) je 12.38%. Ukupna stopa pogreške je 9.36% (26 od 310 opservacija).

Number of Observations and Percent Classified into PCOS			
From PCOS	0	1	Total
0	180 87.80	25 12.20	205 100.00
1	18 17.14	87 82.86	105 100.00
Total	198 63.87	112 36.13	310 100.00
Priors	0.5	0.5	

Error Count Estimates for PCOS			
	0	1	Total
Rate	0.1220	0.1714	0.1467
Priors	0.5000	0.5000	

Tablica 3.19: Stope pogreške unakrsne validacije direktne diskriminantne analize (ispis iz SAS-a)

Stopa lažno pozitivnih je 12.2%, a stopa lažno negativnih 17.14%. Ukupna stopa pogreške je 14.67% (43 od 310 opservacija).

Stepwise procedura diskriminantne analize

Određene su klasifikacijske funkcije korištenjem *stepwise* metode izbora varijabli s razinom značajnosti $\alpha = 0.15$ za dodavanje i izbacivanje varijabli.

Step	Entered	Removed	Wilks' Lambda	p-value
1	Follicle_No_R		0.61141582	< 0.0001
2	Weight_gain		0.5142122	< 0.0001
3	Hair_growth		0.46966565	< 0.0001
4	Cycle		0.44594899	< 0.0001
5	Skin_darkening		0.42657696	< 0.0001
6	AMH		0.41871046	< 0.0001
7	Acne		0.41317319	< 0.0001
8	Age		0.41013322	< 0.0001
9	RBS		0.40690821	< 0.0001
10	Follicle_No_L		0.4035817	< 0.0001

Tablica 3.20: Koraci *stepwise* procedure diskriminantne analize

Redom su dodane varijable Follicle_No_R, Weight_gain, Hair_growth, Cycle, Skin_darkening, AMH, Acne, Age, RBS, Follicle_No_L te ni jedna varijabla nije izbačena. Dobivene su sljedeće klasifikacijske funkcije:

Linear Discriminant Function for PCOS		
Variable	0	1
Constant	-39.85340	-47.06743
Age	1.01643	0.94680
AMH	0.36709	0.43117
RBS	0.42057	0.44190
Follicle_No_L	-0.07904	0.04784
Follicle_No_R	0.67580	1.05518
Cycle	-0.06958	1.45686
Weight_gain	0.08858	1.71751
Hair_growth	-1.04011	0.76951
Skin_darkening	1.92595	3.61137
Acne	1.87065	2.82471

Tablica 3.21: Klasifikacijske funkcije diskriminantne analize dobivene *stepwise* procedurom (ispis iz SAS-a)

Uvrštavanjem opservacija u dobivene klasifikacijske funkcije dobivaju se aposteriorne vjerojatnosti, npr. za prvu opservaciju, koja ima vrijednosti $PCOS = 0, Age = 28, AMH = 2.07, RBS = 92, Follicle\_No\_L = 3, Follicle\_No\_R = 3, Cycle = 0, Weight\_gain = 0, Hair\_growth = 0, Skin\_darkening = 0, Acne = 0$, dobiva se

$$\begin{aligned} L_0 = -39.85340 + 1.01643Age + 0.36709AMH + 0.42057RBS - 0.07904Follicle\_No\_L \\ +0.67580Follicle\_No\_R - 0.06958Cycle + 0.08858Weight\_gain - 1.04011Hair\_growth \\ +1.92595Skin\_darkening + 1.87065Acne \end{aligned} \tag{3.6}$$

$$L_0 = 0.9961 \tag{3.7}$$

$$\begin{aligned} L_1 = -47.06743 + 0.94680Age + 0.43117AMH + 0.44190RBS + 0.04784Follicle\_No\_L \\ +1.05518Follicle\_No\_R + 1.45686Cycle + 1.71751Weight\_gain + 0.76951Hair\_growth \\ +3.61137Skin\_darkening + 2.8247Acne \end{aligned} \tag{3.8}$$

$$L_1 = 0.0039 \tag{3.9}$$

Kako je $L_0 > L_1$, opservacija se klasificira u grupu $PCOS = 0$.

Stopa lažno pozitivnih diskriminantne analize dobivene *stepwise* procedurom je 6.83%, a stopa lažno negativnih 13.3%. Ukupna stopa pogreške je 10.08% (28 od 310 opservacija).

Stopa lažno pozitivnih unakrsne validacije diskriminantne analize dobivene *stepwise* procedurom je 9.27%, a stopa lažno negativnih 17.14%. Ukupna stopa pogreške je 13.21% (37 od 310 opservacija).

Number of Observations and Percent Classified into PCOS			
From PCOS	0	1	Total
0	191 93.17	14 6.83	205 100.00
1	14 13.33	91 86.67	105 100.00
Total	205 66.13	105 33.87	310 100.00
Priors	0.5	0.5	

Error Count Estimates for PCOS			
	0	1	Total
Rate	0.0683	0.1333	0.1008
Priors	0.5000	0.5000	

Tablica 3.22: Stope pogreške diskriminantne analize dobivene *stepwise* procedurom (ispis iz SAS-a)

Number of Observations and Percent Classified into PCOS			
From PCOS	0	1	Total
0	186 90.73	19 9.27	205 100.00
1	18 17.14	87 82.86	105 100.00
Total	204 65.81	106 34.19	310 100.00
Priors	0.5	0.5	

Error Count Estimates for PCOS			
	0	1	Total
Rate	0.0927	0.1714	0.1321
Priors	0.5000	0.5000	

Tablica 3.23: Stope pogreške unakrsne validacije diskriminantne analize dobivene *stepwise* procedurom (ispis iz SAS-a)

3.4 Usporedba modela

Univarijatnom logističkom regresijom dobiveno je da su na razini značajnosti $\alpha = 0.05$ značajne varijable Age, BMI, Period_length, FSH, LH_FSH_ratio, AMH, Follicle_No_L, Follicle_No_R, Avg_F_size_L, Cycle, Weight_gain, Hair_growth, Skin_darkening, Hair_loss, Acne.

Multivarijatnom logističkom regresijom dobiveno je da su na razini značajnosti $\alpha = 0.05$ značajne varijable Follicle_No_L, Follicle_No_R, Cycle, Weight_gain, Hair_growth, Skin_darkening, Acne. Površina ispod ROC krivulje je $c = 0.955$.

Stepwise procedurom logističke regresije s razinom značajnosti $\alpha = 0.15$ dobiveno je da su značajne varijable Follicle_No_R, Weight_gain, Hair_growth, Cycle, Skin_darkening, Follicle_No_L, Acne, Age, RBS. Površina ispod ROC krivulje je $c = 0.952$.

U obje verzije multivarijatne regresije nisu uključene sve varijable koje su se pokazale značajnima tijekom univarijatne analize, ali oba modela imaju visoku prediktivnu snagu ($c > 0.9$).

Direktna diskriminantna analiza daje stopu pogreške klasifikacije 9.36% i stopu pogreške unakrsne validacije 14.67%. *Stepwise* procedurom diskriminantne analize s razinom značajnosti $\alpha = 0.15$ dobivene su klasifikacijske funkcije koje uključuju varijable Follicle_No_R, Weight_gain, Hair_growth, Cycle, Skin_darkening, AMH, Acne, Age, RBS i Follicle_No_L te je dobivena stopa pogreške klasifikacije 10.08% i stopa pogreške unakrsne validacije 13.21%.

Varijable koje su izabrane *stepwise* procedurom za obje metode su Age, RBS, Follicle_No_L, Follicle_No_R, Cycle, Weight_gain, Hair_growth, Skin_darkening i Acne.

Poglavlje 4

Dodatak

4.1 SAS kod

```
/* histogrami*/
ods noproctitle;
ods graphics / imagemap=on;

proc means data=WORK.PCOS chartype mean std min max median
        n vardef=df qmethod=os;
var Age BMI Hb Period_length FSH LH TSH AMH PRL Vit_D3 PRG RBS
        Follicle_No_L Follicle_No_R Avg_F_size_L Avg_F_size_R
        Endometrium LH_FSH_ratio;
run;

proc univariate data=WORK.PCOS vardef=df noprint;
var Age BMI Hb Period_length FSH LH TSH AMH PRL Vit_D3 PRG RBS
        Follicle_No_L Follicle_No_R Avg_F_size_L Avg_F_size_R
        Endometrium LH_FSH_ratio;
histogram Age BMI Hb Period_length FSH LH TSH AMH PRL Vit_D3 PRG RBS
Follicle_No_L Follicle_No_R Avg_F_size_L Avg_F_size_R
        Endometrium LH_FSH_ratio;
inset mean (9.2) std (10.4) max (9.2) min (9.2) median (9.2)
        n / position=ne;
run;

/*frekvencijske tablice */
proc freq data=WORK.PCOS;
```

```
tables PCOS Cycle Weight_gain Hair_growth Skin_darkening Hair_loss
        Acne Reg_Exercise / plots=(freqplot);
run;

/*univarijatne logističke regresije*/
/*Cycle*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
class Cycle / param=ref ref=first;
model PCOS(event='1')=Cycle / link=logit technique=fisher;
run;

/*weight gain*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
class Weight_gain / param=ref ref=first;
model PCOS(event='1')=Weight_gain / link=logit technique=fisher;
run;

/*Hair growth*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
class Cycle Hair_growth / param=ref ref=first;
model PCOS(event='1')=Hair_growth / link=logit technique=fisher;
run;

/*Skin darkening*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
class Skin_darkening / param=ref ref=first;
model PCOS(event='1')=Skin_darkening / link=logit technique=fisher;
```

```
run;

/*Hair loss*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
class Hair_loss / param=ref ref=first;
model PCOS(event='1')=Hair_loss / link=logit technique=fisher;
run;

/*Acne*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
class Acne / param=ref ref=first;
model PCOS(event='1')=Acne / link=logit technique=fisher;
run;

/*Regular exercise*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
class Reg_Exercise / param=ref ref=first;
model PCOS(event='1')=Reg_Exercise / link=logit technique=fisher;
run;

/*Age*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=Age / link=logit technique=fisher;
run;

/*Hb*/
ods noproctitle;
```

```
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=Hb / link=logit technique=fisher;
run;

/*Period_length*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=Period_length / link=logit technique=fisher;
run;

/*FSH*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=FSH / link=logit technique=fisher;
run;

/*LH*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=LH / link=logit technique=fisher;
run;

/*TSH*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=TSH/ link=logit technique=fisher;
run;

/*AMH*/
```

```
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=AMH / link=logit technique=fisher;
run;

/*PRL*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=PRL/ link=logit technique=fisher;
run;

/*Vitamin D3*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=Vit_D3/ link=logit technique=fisher;
run;

/*PRG*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=PRG/ link=logit technique=fisher;
run;

/*RBS*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=RBS / link=logit technique=fisher;
run;
```

```
/*Follicle No L*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=Follicle_No_L / link=logit technique=fisher;
run;

/*Follicle No R*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=Follicle_No_R / link=logit technique=fisher;
run;

/*Avg F Size L*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=Avg_F_size_L / link=logit technique=fisher;
run;

/*Avg F Size R*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=Avg_F_size_R / link=logit technique=fisher;
run;

/*Endometrium*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=Endometrium/ link=logit technique=fisher;
run;
```

```
/*BMI*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=BMI/ link=logit technique=fisher;
run;

/*LH:FSH*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
model PCOS(event='1')=LH_FSH_ratio / link=logit technique=fisher;
run;

/*model s uključenim varijablama koje su značajne u univarijatnoj
regresiji za alpha = 0.05*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=(roc);
class Cycle Weight_gain Hair_growth Skin_darkening Hair_loss
        Acne / param=ref ref=first;
model PCOS(event='1')=Age BMI Period_length FSH LH_FSH_ratio
        AMH Follicle_No_L Follicle_No_R Avg_F_size_L Cycle
        Weight_gain Hair_growth Skin_darkening Hair_loss
        Acne/ link=logit technique=fisher;
run;

/*stepwise logistička regresija - alpha = 0.15*/
ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.PCOS plots=roc;
class Cycle Weight_gain Hair_growth Skin_darkening Hair_loss
        Acne Reg_Exercise / param=ref ref=first;
model PCOS(event='1')=Cycle Weight_gain Hair_growth Skin_darkening
```

```
        Hair_loss Acne Reg_Exercise Age Hb Period_length FSH LH
        TSH AMH PRL Vit_D3 PRG RBS Follicle_No_L Follicle_No_R
        Avg_F_size_L Avg_F_size_R Endometrium BMI LH_FSH_ratio
        / link=logit rsquare selection=stepwise slentry=0.15
        slstay=0.15 hierarchy=none technique=fisher;
run;

/*diskriminantna analiza*/
/*sve varijable*/
ods noproctitle;

proc discrim data=WORK.PCOS pool=yes list crossvalidate listerr;
class PCOS;
var Age BMI Hb Period_length FSH LH LH_FSH_ratio TSH AMH PRG PRL
        Vit_D3 RBS Follicle_No_L Follicle_No_R Avg_F_size_L
        Avg_F_size_R Endometrium Cycle Weight_gain Hair_growth
        Skin_darkening Hair_loss Acne Reg_Exercise;
run;


/*stepwise diskriminantna analiza - alpha = 0.15*/
ods noproctitle;

proc stepdisc data=WORK.PCOS method=sw slentry=0.15
        slstay=0.15 short;
class PCOS;
var Age BMI Hb Period_length FSH LH LH_FSH_ratio TSH AMH PRG PRL
        Vit_D3 RBS Follicle_No_L Follicle_No_R Avg_F_size_L
        Avg_F_size_R Endometrium Cycle Weight_gain Hair_growth
        Skin_darkening Hair_loss Acne Reg_Exercise;
run;

proc discrim data=WORK.PCOS pool=yes crossvalidate listerr;
class PCOS;
var &_stdvar;
run;
```

Bibliografija

[1] R. Christensen, *Advanced Linear Modeling*, Springer, 2001.

[2] D. W. Hosmer i S. Lemeshow, *Applied Logistic Regression*, Wiley & Sons, 1989.

[3] C.J. Huberty i S. Olejnik, *Applied MANOVA and Discriminant Analysis*, Wiley & Sons, 2006.

[4] M. Huzak, *Primijenjena statistika*, PMF-MO, nastavni materijali, 2022.

[5] A. Jazbec, *Odabrane statističke metode u biomedicini*, PMF-MO, nastavni materijali, 2022.

[6] Kaggle, *Polycystic ovary syndrome (PCOS)*, `https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos/data`.

[7] P. McCullagh i J. A. Nelder, *Generalized Linear Models*, Chapman and Hall, 1989.

[8] SAS Institute Inc., *SAS/STAT® 15.2 User's Guide*, 2020.

[9] B. G. Tabachnick i L. S. Fidell, *Using Multivariate Statistics*, Allyn and Bacon, 2001.

[10] H. J. Teede et al., *Recommendations from the 2023 International Evidence-based Guideline for the Assessment and Management of Polycystic Ovary Syndrome*, (2023), `https://doi.org/10.1093/humrep/dead156`.

Sažetak

U ovom radu uvedeni su osnovni pojmovi logističke regresije - logistički model, izgled, omjer izgleda. Opisano je kako se testira adekvatnost logističkog modela, interpretiraju dobiveni parametri i biraju nezavisne varijable modela. Uveden je pojam diskriminantne analize i poseban slučaj linearne diskriminantne analize. Opisano je kako se provodi validacija dobivenih rezultata i kako se biraju prediktorske varijable.

Korištenjem programa SAS OnDemand analizirana je baza podataka pacijentica iz nekoliko bolnica u Indiji kako bi se odredilo koje varijable utječu na pojavnost sindroma policističnih jajnika. Dobiveni su univarijatni logistički modeli, multivarijatni logistički model i logistički model *stepwise* procedure. Varijable koje su značajne u svim modelima su Follicle_No_L (broj folikularnih cista na lijevom jajniku), Follicle_No_R (broj folikularnih cista na desnom jajniku), Cycle (redovitost menstrualnog ciklusa), Weight_gain (povećanje tjelesne težine), Hair_growth (pojačan rasta dlaka), Skin_darkening (hiperpigmentacija kože) i Acne (akne). Korištena je direktna linearna diskriminantna analiza i *stepwise* procedura linearne diskriminantne analize za klasifikaciju podataka. *Stepwise* procedurom diskriminantne analize izabrane su varijable Follicle_No_R, Weight_gain, Hair_growth, Cycle, Skin_darkening, AMH (razina anti-Mullerovog hormona), Acne, Age (starost), RBS (razina glukoze u krvi), Follicle_No_L.

Summary

In this thesis basic terms of logistic regression are introduced - logistic model, odds, odds ratio. It is described how to test the adequacy of the logistic model, interpret the obtained parameters and select the independent variables of the model. The concept of discriminant analysis and the special case of linear discriminant analysis are introduced. It is described how the validation of the obtained results is carried out and how the predictor variables are selected.

Using SAS OnDemand, a database of patients from several hospitals in India was analyzed to determine which variables influence the incidence of polycystic ovary syndrome. Univariate logistic models, multivariate logistic model and logistic model using *stepwise* procedure were obtained. Variables that are significant in all models are Follicle_No_L (number of follicular cysts on the left ovary), Follicle_No_R (number of follicular cysts on the right ovary), Cycle (regularity of the menstrual cycle), Weight_gain , Hair_growth, Skin_darkening and Acne. Direct linear discriminant analysis and linear discriminant analysis using *stepwise* procedure were used for data classification. *Stepwise* discriminant analysis selected the variables Follicle_No_R, Weight_gain, Hair_growth, Cycle, Skin_darkening, AMH (anti-Mullerian hormone level), Acne, Age, RBS (glucose level in blood), Follicle_No_L.

Životopis

Rođena sam 23.4.1999. u Varaždinu. Pohađala sam Osnovnu školu Ivana Kukuljevića Sakcinskog u Ivancu te Srednju školu Ivanec, smjer opća gimnazija, koju sam završila 2018. Iste godine upisala sam preddiplomski studij matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Nakon završetka preddiplomskog studija 2018. godine upisala sam diplomski studij Matematičke statistike na istom fakultetu. Od prosinca 2022. zaposlena sam u PLIVI u odjelu Istraživanja i razvoja.