



Универзитет у Новом Саду
Природно-математички факултет
Департман за математику и информатику



ПРИМЕНА МУЛТИВАРИЈАНТНЕ ДИСКРИМИНАЦИОНЕ АНАЛИЗЕ У ПРОЦЕСУ РЕВИЗИЈЕ

Мастер рад

Ментор др Наташа Спахић

Студент Миљана Вукоје

Нови Сад, 2015.

САДРЖАЈ

УВОД.....	1-4
1 ЗНАЧЕЊЕ ПОЈЕДИНИХ ПОЈМОВА.....	1-7
2 ДИСКРИМИНАЦИОНА АНАЛИЗА.....	2-9
2.1 Претпоставке дискриминационе анализе.....	2-9
2.1.1 Бајесова теорема	2-11
2.1.2 Коришћење Бајесове теореме за класификацију.....	2-13
2.1.3 Средина и коваријациона матрица	2-15
2.1.4 Махаланобисово одстојање.....	2-17
2.2 Циљеви дискриминационе анализе	2-18
2.2.1. Дескриптивна дискриминациона анализа	2-20
2.3 Мултиваријантна дискриминациона анализа	2-23
2.4 Аналогија са регресијом.....	2-26
2.5 Фазе дискриминационе анализе	2-29
2.6 Проблеми у примени дискриминационе анализе	2-38
2.7 SPSS активност - дискриминациона анализа.....	2-40
3 ОСНОВЕ DATA MINING-А И ЊЕГОВИХ ОСНОВНИХ ТЕХНИКА.....	3-43
3.1 Стабла одлучивања.....	3-43
3.2. Правила асоцијације.....	3-45
3.3. Метода најближег суседа.....	3-45
3.4. Неуронске мреже	3-46
4 ПРИМЕНА МУЛТИВАРИЈАНТНЕ ДИСКРИМИНАЦИОНЕ АНАЛИЗЕ У ПРОЦЕСУ РЕВИЗИЈЕ.....	4-47
4.1 Процес ревизије	4-47
4.2 Финансијски извештаји предузећа дуванске индустрије у Републици Србији	4-49
4.3 Примена мултиваријантне дискриминационе анализе на примеру у SPSS-у	4-55
4.4 Примена мултиваријантне дискриминационе анализе у програмском алату STATISTICA ..	4-65
4.5 Примена стабла одлучивања.....	4-69

5	РЕЗУЛТАТИ ПРИМЕНЕ.....	5-74
5.1	Примена анализе са додатим варијаблама.....	5-79
6	ЗАКЉУЧАК.....	6-85
	ЛИТЕРАТУРА	6-87
	СПИСАК ТАБЕЛА, ГРАФИКОНА И СЛИКА.....	6-90
	БИОГРАФИЈА.....	6-95

Увод

Изузетна је способност проналажења одређених метода и процедура којима би могли експерти из области ревизије да на један изузетно прецизан начин предвиде будућа дешавања у оквирима једне фирме.

Из широког спектра постојећих процедура у овом раду је обрађена дискриминациона анализа тачније мултиваријациона, где се кроз објашњење са теоријског-практичног аспекта разматра њен значај за саму ревизорску праксу.

Циљ рада јесте да предочи да ли је и до које мере је ова анализа корисна у циљу налажења решења којим би се извршила предикција.

У раду је посебно обрађено пословање дуванске индустрије на простору Србије.

Кроз сам процес ревизије, анализирањем финансијских извештаја из претходних година, тачније одабиром одређених ставки из биланса стања и успеха, у раду је покушано да се баш мултиваријационом дискриминационом анализом предочи да ли је могуће да се предвиди мишљење ревизора.

У раду су коришћени подаци преузети са сајта Агенције за привредне регистре, која се бави вођењем законом утврђених регистара као јединствених, централизованих, јавних електронских база података свих привредних субјеката који послују на територији Републике Србије.

Нова времена изискују и нове мере којима би требао да се имплементира један шири методолошки принцип за проналажење скривених знања до сада занемариваних у финансијским извештајима. Сама предност ове методе огледа се у симулацији резонувања људског мозга.

„Истраживање података (енгл. *data mining*) као посебна област вештачке интелигенције, као таква обухвата примену тзв машинског учења и односи се на развој алгорита и техника који омогућавају да компјутери кроз њу уче.“[41] Методи којима се служи *data mining* заснивају се на индуктивном учењу, које треба разликовати од надзираног учења јер оно представља учење уз помоћ неког другог. Индуктивно учење заснива се на налажењу суштински нових теорија, на основу више пута понављаних сличних искустава приликом разматрања истих карактеристика. У истраживању података се најчешће користи метод индуктивног учења јер њим добијамо конкретне резултате за директну интерпретацију.

Осим што нам је познато да човек свакодневно учи, у свету технологије и компјутери уче, то је тзв. машинско учење.

Процедура учења може бити спроведена у два случаја: учење са надгледањем и учење без надгледања. Учење са надгледањем је техника машинског учења на основу скупа података који називамо *training set*, који се састоји од улазних објеката и жељених излаза. „Излазна функција може да предвиди класу улазног објекта.“ [41] Главни циљ учења са надгледањем јесте да се постигне предикција вредности излазне функције за сваку опсервацију на основу одређених већ виђених примера тог типа. Како би се што ефикасније имплементирало ово учење постоје наредни кораци:

- Утврдити тип података (у овом случају су у питању квантитативни подаци добијени из финансијских извештаја).
- Сакупити податаке.
- Утврдити број улазних карактеристика које описују објекат.
- Изабрати одређени алгоритам учења (нпр. стабла одлучивања, неуронске мреже,...).
- Покренути изабрани алгоритам учења на датим, сакупљеним подацима.

Учење без надгледања не карактеришу предикције о излазним вредностима. У овом случају се прикупљен скуп улазних објеката посматра као скуп случајних (*random*) варијабли.

Примена мултиваријантне дискриминационе методе почиње тридесетих године прошлог века, као статистичка категорија. Временом је почела да се примењује и у економији, у анализи банкрота одређених корпорација, у одређивању кредитно-способних клијената, у класификацији могућих рачуноводствених грешака итсл. „Међутим, примена дискриминационе анализе у контексту проналажења имплицитних знања (*data mining*) је специфична пошто се анализа не своди на примену модела, већ се врши надзирано учење на узорку модела дискриминационе анализе у више наврата, да би решење било што тачније.“ [41]

На крају или пак на почетку захвалила бих се менторки др Наташи Спахић на пруженој подршци, саветима и незаборавној сарадњи током студија. Такође, захваљујем се и члановима комисије др Наташи Крејић и др Сањи Рапајић на пренетом знању, стручном усмеравању и на прилици да и са њима остварим сарадњу уз помоћ које сам пласирала своје знање.

Велику захвалност дугујем наравно мојој породици и мојим драгим људима који су несебично током студирања били уз мене!

Миљана Вукоје

1 Значење појединих појмова

Финансијски извештај

Скуп извештаја који се састављају на крају пословне године и који презентују резултате пословања клијента у претходном периоду. Најчешће их чине биланс стања, биланс успеха, извештај о променама на капиталу, биланс токова готовине, напомене уз финансијске извештаје и сл.

Ревизија

Поступак испитивања и оцене финансијских извештаја, где се на основу усклађености са одређеним правилима и процедурама, критеријумима или циљевима, даје стручно мишљење о реалности и објективности стања имовине, капитала, обавеза и резултата пословања.

Интерна контрола

Интерна контрола је „процес који је дизајниран, имплементиран и одржаван од стране управљача, менаџмента и другог особља са циљем да пружи разумно уверавање у остваривање циљева ентитета у вези са поузданошћу финансијског извештавања, ефективношћу и ефикасношћу активности и усагласношћу са одговарајућим законима и прописима“.[1]

Један интерни ревизор преиспитује и област пословања и финансијску област у организацији са циљем процене интерних контрола. Практично све процедуре интерне ревизије су усредсређене на неки облик евалуације интерних контрола.

Факторска анализа

Важан алат који омогућава да се утврде главни фактори који утичу на понашање потрошача.

Кластер анализа

Овај поступак омогућава издвајање испитаника чији су одговори изразито слични у исти кластер, па је кластер анализа одлична за опис и квантификацију сегмената потрошача. Дакле, то је мултиваријациона техника са циљем груписања случајева са сличним профилима у дефинисаном скупу карактеристика.

Регресиона анализа

Регресиона анализа је скуп статистичких метода којима се открива да ли постоје везе између посматраних варијабли, и од коликог су значаја постојеће везе.

Машинско учење	Машинско учење је подобласт вештачке интелигенције чији је циљ конструисање алгоритама и рачунарских система који су способни да се адаптирају на нове ситуације и уче на основу претходног „искуства“. Развијене су различите технике учења за извршавање различитих задатака.
Вештачка интелигенција	Вештачка интелигенција има за циљ истраживања развој програма (софтвера), који ће рачунарима омогућити да „уче“ и „стичу искуство“, те да на тај начин испољавају одређену врсту интелигенције.
Корелациона матрица	Табела која показује интеркорелације између свих варијабли.
Варијанса грешке	Представља варијансу варијабле, услед насталих грешака током прикупљања података или током мерења.
Грешка мерења	Односи се на настале нетачности у мерењу „истинских“ вредности варијабли, при чему треба узети у обзир погрешивост инструмената мерења, постојање грешака приликом уноса итд.
Мултиколинеарност	Мултиколинеарност заправо показује колика је међузависност између независних варијабли. Када се деси да су две или више независних променљивих у корелацији једна са другом, тешко је одредити која од ових променљивих појединачно утиче на функцију и колики је тај утицај. Код већине података постоји мултиколинеарност између варијабли. Уколико степен мултиколинеарности није висок, може да се занемари.
Кросвалидација (Cross-validation)	Ова теорија је донета од стране Симора Гајсера, и односи се на стварање партиција скупа података на подскупове, на којима се даље врши појединачна анализа.

2 Дискриминациона анализа

2.1 Претпоставке дискриминационе анализе

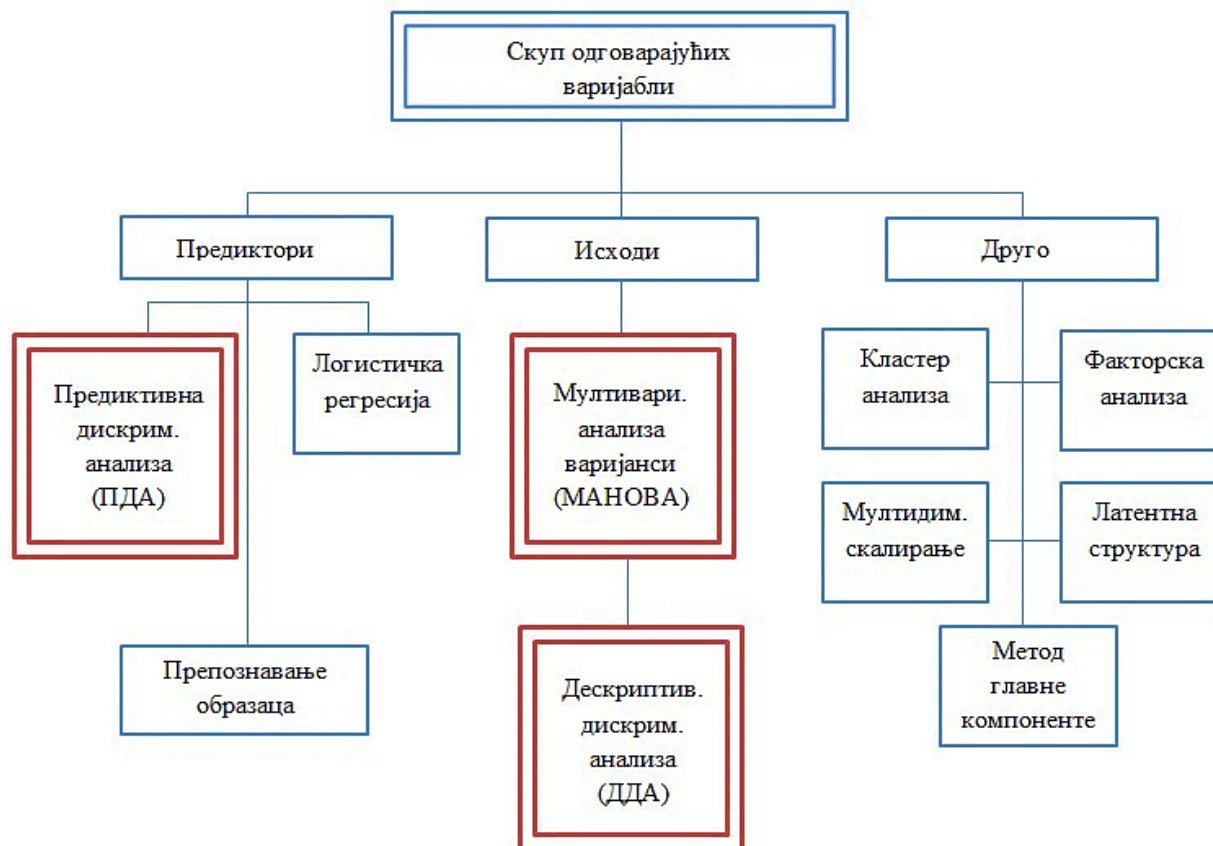
Мултиваријациона анализа представља низ мултиваријационих техника које симултано анализирају међусобни однос између великог броја варијабли. У самом називу ове анализе уочавамо појам „мултиваријациони“ који нам указује да је у питању велики број варијабли. Осим овог израза може се у литератури наићи и на термине „мултиваријантна“, „мултиваријатна“ анализа, али можда најисправнији термин би био „мултиваријабилна анализа“ чиме управо наглашавамо велики број варијабли.

Технике које чине ову анализу могу да се поделе у две групе: технике зависности и технике међузависности. Код техника зависности можемо уочити однос између варијабли где на једној страни имамо зависне варијабле на које утицај врше независне варијабле. У ову групу спадају следеће технике: вишеструка регресија и корелација, логит анализа, дискриминациона анализа, каноничка анализа, conjoint анализа.

Од свих наведених техника истичемо примену дискриминационе анализе, која је својом појавом омогућила идентификацију баш оне променљиве која највише доприноси раздвајању група, а на основу вредности скупа независних променљивих предвиђа вероватноћу да ће објект припасти једној од тих група. Дакле „метод мултиваријационе анализе који се бави раздвајањем различитих група и алокацијом опсервација у унапред дефинисане групе називамо дискриминациона анализа.“ [2]

Постоје три типа дискриминационе анализе: стандардна, секвенцијална и статистичка (постепена).

Сликом 1. је представљено како се дискриминациона анализа уклапа у општу шему анализе мултиваријационих података.



Слика 1. Шема анализе мултиваријационих података (извор: „Applied MANOVA and Discriminant Analysis”, [8])

Први пут ова техника се јавља негде 1936. године. Међутим, њена популарност је расла заједно са појавом моћних рачунара и развојем технологије током 60-их година.

Дискриминациона анализа је изузетно примењива у случајевима када су независне варијабле $X_1, X_2, \dots, X_n$ нумеричке, а зависна варијабла Z је категоричка (номинална, дескриптивна). Код ове технике карактеристично је то да се зависна варијабла састоји од већег броја група или категорија (у случају мултиваријационе дискриминационе анализе). Сама тежња ове анализе јесте управо класификација у групе, према одређеним критеријумима ($X_1, X_2, \dots, X_n$).

Постоје и основне претпоставке дискриминационе анализе које подразумевају да су варијабле $X_1, X_2, \dots, X_n$ су независне једна од друге, затим да се групе међусобно искључују и величине група се не разликују много, да број независних варијабли није већи од величине $(n - 2)$, да грешке (резидуали) имају насумичну расподелу и да за потребе тестирања значајности, независне варијабле имају нормалну расподелу.

Дакле, постоји више алгоритама за дискриминациону анализу, а неки од њих на темељу дискриминационих варијабли, условно речено предвиђају припадност објекта појединој групи. Због тога се оригиналне варијабле (које описују ентитете и из којих се изводе дискриминативне варијабле) називају још и предикторима, а варијабле (селекторске) којима се одређује припадност ентитета појединој групи називају критеријумске (*grouping*) варијабле.

2.1.1 Бајесова теорема

Пођимо прво од дефинисања простора вероватноћа $(\Omega, \mathcal{F}, P)$ и догађаја $Z, X \in \mathcal{F}$. Посматрајући случај да реализација догађаја Z , зависи од реализације догађаја X , заправо говоримо о условној вероватноћи.

Вероватноћа да је испуњен догађај Z , ако је испуњен догађај X се записује као $P(Z | X)$.

Дефиниција 1. [19] Нека је $(\Omega, \mathcal{F}, P)$ простор вероватноћа и $Z, X \in \mathcal{F}$, при чему је $P(X) > 0$. Условна вероватноћа $P(Z | X)$ (вероватноћа догађаја Z под условом да се реализовао догађај X) је

$$P(Z|X) = \frac{P(Z \cap X)}{P(X)}. \quad (2.1)$$

Ознаком $P(X)$ бележимо вероватноћу да се десио догађај X , а ознаком $P(Z \cap X)$ бележимо вероватноћу истовременог јављања догађаја Z и X (вероватноћа пресека догађаја Z и X).

Условне вероватноће догађаја из истог простора вероватноћа, у односу на неки догађај из тог простора, имају све особине вероватноће, тј задовољавају наредне аксиоме:

- 1° $P(Z|X) \geq 0$,
- 2° $P(\Omega|X) = 1$,
- 3° $P(\sum_j Z_j | X) = \sum_j P(Z_j | X)$.

Дефиниција 2. [20] Нека су Z и X два догађаја. Вероватноћа производа (или пресека) два догађаја може се добити помоћу условних вероватноћа :

$$P(Z \cap X) = P(ZX) = P(XZ) = P(X) \cdot P(Z|X) = P(Z) \cdot P(X|Z). \quad (2.2)$$

У претходној дефиницији уведена је релација (2.2) која је позната као *правило или закон множења вероватноћа* и може се проширити и на више од два догађаја.

Заменом релације (2.2) у релацију (2.1) добијамо

$$P(Z|X) = \frac{P(Z \cap X)}{P(X)} = \frac{P(Z) \cdot P(X|Z)}{P(X)}. \quad (2.3)$$

Претходно добијена формула (2.3) назива се *Бајесова формула*, односно Бајесова¹ теорема.

Бајесовом теоремом се врши преиспитивање условне вероватноће, применом расположивих информација, уз обезбеђење процедура за одлучивање да ли тврдње о вероватноћи требају да буду усклађене са датим додатним информацијама или не. [20]

Код коначно много дисјунктних догађаја, $Z_1, Z_2, Z_3, \dots, Z_n$, који припадају истом простору елементарних исхода важи:

$$P(Z_1 Z_2 Z_3 \dots Z_n) = P(Z_1) \cdot P(Z_2|Z_1) \cdot P(Z_3|Z_1 Z_2) \cdot \dots \cdot P(Z_n|Z_1 Z_2 \dots Z_{n-1}). \quad (2.4)$$

Дефиниција 3. Нека су $Z_1, Z_2, Z_3, \dots, Z_n$, међусобно искључиви догађаји и нека је X неки други догађај. Вероватноћу догађаја Z_i , када је дат догађај X , можемо израчунати применом Бајесове теореме:

$$P(Z_i|X) = \frac{P(Z_i) \cdot P(X|Z_i)}{P(X)}, i = 1, 2, \dots, n \quad (2.5)$$

Према формули потпуне вероватноће именилац можемо да искажемо на следећи начин:

$$P(X) = \sum_{k=1}^n P(Z_k) \cdot P(X|Z_k), \quad (2.6)$$

па сада заменом (2.6) у релацију (2.5) добија се:

$$P(Z_i|X) = \frac{P(Z_i) \cdot P(X|Z_i)}{\sum_{k=1}^n P(Z_k) \cdot P(X|Z_k)}, i = 1, 2, \dots, n \quad (2.7)$$

Такође, Бајесова формула се још назива и *формула вероватноћа хипотеза (или узорак)* јер $Z_1, Z_2, Z_3, \dots, Z_n$ можемо посматрати као различите узорке који могу довести до реализације догађаја X . У литератури се иначе *хипотезе* означавају са $H_1, H_2, H_3, \dots, H_n$, а са $\{H_1, H_2, H_3, \dots, H_n\}$ се означава *потпун систем догађаја*.

Дакле, ако се догађај X реализовао, интересује нас колика је вероватноћа да је у питању неки одређени узорак Z_i или пак нека постављена хипотеза H_i .

¹ Томас Бајес (1702-1761), енглески математичар и свештеник. Као математичар је био члан Лондонског краљевског друштва. Математичка област којој је највише посвећивао пажњу била је теорија вероватноће. Познат је по томе што је формулисао специјалан случај формуле тоталне вероватноће, који је познат под именом Бајесова формула. Међутим, Бајес никада није објавио рад у коме је формулисао ову теорему. Бајес је за време свог живота објавио само два рада, један теолошки рад и један математички.

Дефиниција 4. [19] Нека је $(\Omega, \mathcal{F}, P)$ простор вероватноћа и $H_1, H_2, H_3, \dots, H_n \in \mathcal{F}$ задовољавају следеће услове:

- 1° $P(H_i) > 0, i = 1, 2, \dots, n,$
- 2° $H_i \cap H_j = \emptyset, i, j = 1, 2, \dots, n, i \neq j,$
- 3° $\sum_{i=1}^n H_i = \Omega.$

Теорема 1. [19] (Формула тоталне вероватноће) Ако је $\{H_1, H_2, H_3, \dots, H_n\}$ потпун систем догађаја, онда

$$P(X) = \sum_{i=1}^n P(H_i) \cdot P(X|H_i), \forall X \in \mathcal{F}. \quad (2.8)$$

Коришћењем формуле тоталне вероватноће, и уношењем ње у (2.7) доказујемо наредну теорему.

Теорема 2. [19] (Бајесова формула) Ако је $\{H_1, H_2, H_3, \dots, H_n\}$ потпун систем догађаја, онда

$$P(H_m|X) = \frac{P(H_m) \cdot P(X|H_m)}{\sum_{i=1}^n P(H_i) \cdot P(X|H_i)}, m = 1, 2, \dots, n, X \in \mathcal{F}, P(X) > 0. \quad (2.9)$$

2.1.2 Коришћење Бајесове теореме за класификацију

Претпоставимо да желимо да класификујемо неку опсервацију у једну од K класа, где је $K \geq 2$. Другим речима, квалитативно одговарајућа променљива Z може преузети K могућих различитих и неуређених вредности.

Нека π_k представља вероватноћу да случајно изабрана опсервација потиче из k -те класе; тј. ово је вероватноћа да дата опсервација је повезана са k -том категоријом одговарајуће променљиве Z . Означавамо са $f_k(X) = P(Z = k|X = x)$ функцију густине од X за опсервацију која долази из k -те класе. Другим речима, $f_k(X)$ је релативно велика ако постоји велика вероватноћа да је опсервација у k -тој класи има $X \approx x$, и $f_k(X)$ је мала уколико то не важи. Онда Бајесова теорема стоји иза следеће формуле:

$$P(Z = k | X = x) = \frac{\pi_k f_k(X)}{\sum_{l=1}^K \pi_l f_l(X)}. \quad (2.10)$$

Означаваћемо убудуће $P(Z = k | X = x)$ са p_k .

У принципу, процена π_k је једноставна ако имамо случајан узорак Z -а из популације: ту се само израчуна разломак опсервација које припадају k -тој класи. Међутим, прорачун $f_k(X)$ је донекле захтевнији, осим ако претпоставимо неке једноставне форме за ове густине. Знамо да Бајесов класификатор има најнижу могућу стопу грешке од свих класификатора.

Дакле, налажењем начина да $f_k(X)$ буде процењено, онда се може усавршити класификатор који би био приближан Бајесовом класификатору, који додељује сваку опсервацију најпогоднијој класи.

Уколико посматрамо да је $n = 1$ тиме постављамо да имамо само један предиктор. Е сада како би била процењена вредност p_k , треба нам прво добра процена за $f_k(X)$ коју ћемо заменити у формулу (2.10).

У циљу процене $f_k(X)$, прво ћемо направити неке претпоставке о облику функције густине. Претпоставимо да $f_k(X)$ је дата за нормалну или Гаусову расподелу. У једнодимензионалној форми, густина код нормалне расподеле се рачуна на следећи начин:

$$f_k(X) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right), \quad (2.11)$$

где су дати μ_k и σ_k^2 средња вредност и варијанса за параметре k -те класе.

Претпоставимо додатно да важи хомоскедастичност (тачније да су варијансе те класе све међусобно једнаке), те ће заједничка варијанса бити означена са σ^2 . Убацивањем (2.11) у (2.10) добија се следећа једнакост:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}. \quad (2.12)$$

Нагласимо да π_k није везан за константу $\pi \approx 3,14159$. Бајесов класификатор подразумева доделу опсервације $X = x$ класи за коју је (2.12) највећи.

Сада проширује се дискриминациони класификатор за случај више предиктора. Да бисмо то урадили, претпоставићемо да је дато $X = (X_1, X_2, X_3, \dots, X_n)$ са мултиваријационом нормалном расподелом, са специфичним вектором очекивања и са заједничком одговарајућом коваријационом матрицом за класу.

Мултиваријантна нормална расподела претпоставља да сваком појединачном предиктору одговара једнодимензионална нормална расподела, као и формула (2.12), при чему постоји корелација између сваког пара предиктора.

Како би било указано да n -димензионална случајна променљива X има мултиваријациону нормалну расподелу, пише се $X \sim N(\mu, \Sigma)$. Овде је $E(X) = \mu$ средња вредност X -а (вектора са n компоненти), и $Cov(X) = \Sigma$ представља коваријациону матрицу димензије $n \times n$. Формално, функција густине је дефинисана на следећи начин:

$$f(X) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right). \quad (2.13)$$

У случају када постоји више предиктора тј. $n > 1$, дискриминациони класификатори претпоставља да узорци из k -те класе су узети из мултиваријационе нормалне расподеле $N(\mu_k, \Sigma)$, где је μ_k је вектор средњих вредности које одговарају специфичној класи, а Σ је коваријациона матрица која је заједничка за све k класе.

2.1.3 Средина и коваријациона матрица

У претходном одељку (2.1.2.) коришћени су појмови као што су средина и коваријациона матрица. Појаснимо ближе ове појмове.

Претпоставимо да је дат $X (n \times 1)$ случајан вектор и сваки елемент тог вектора представља једнодимензиону случајну променљиву. За сваки тај елемент, тј. једнодимензиону случајну променљиву може се израчунати средина $E(X_j) = \mu_j$ и варијанса која је дата са $\sigma_j^2 = E(X_j - \mu_j)^2, j = 1, 2, \dots, n$. У различитим литературама можете наићи и на следеће ознаке варијансе: $\sigma_{jj}, Var(X_j)$.

Средина случајног вектора X се означава са μ , и то је $(n \times 1)$ вектор чији су елементи $E(X_j) = \mu_j, j = 1, 2, \dots, n$.

$$\mu = E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}. \quad (2.14)$$

Коваријанса за било који пар случајних променљивих X_j и X_k се дефинише:

$$\sigma_{jk} = Cov(X_j, X_k) = E[(X_j - \mu_j)(X_k - \mu_k)]. \quad (2.15)$$

А при томе важи и да је $Cov(X_j, X_j) = Var(X_j)$ и $Cov(X_j, X_k) = Cov(X_k, X_j) = \sigma_{jk} = \sigma_{kj}$.

За случајан вектор X , дефинише се $(n \times n)$ симетрична матрица, којој се на дијагонали налазе $\sigma_{jj} = Var(X_j)$, а остале елементе чине $\sigma_{jk} = Cov(X_j, X_k), j \neq k$. [2]

Оваква матрица се назива *коваријациона матрица*. Означава се са: $Var(X)$, $Cov(X)$ или Σ .

$$\begin{aligned} Cov(X) = \Sigma = [\sigma_{jk}] &= \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix} \\ &= \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_n) \\ Cov(X_1, X_2) & Var(X_2) & \cdots & Cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_n) & Cov(X_2, X_n) & \cdots & Var(X_n) \end{bmatrix}. \end{aligned} \quad (2.16)$$

Коваријациона матрица је заправо очекивана вредност случајне матрице. Посматрамо симетричну случајну матрицу квадрата, односно производа одступања елемената случајног вектора од одговарајуће средине:

$$\begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \dots & (X_1 - \mu_1)(X_n - \mu_n) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \dots & (X_2 - \mu_2)(X_n - \mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (X_n - \mu_n)(X_1 - \mu_1) & (X_n - \mu_n)(X_2 - \mu_2) & \dots & (X_n - \mu_n)^2 \end{bmatrix}. \quad (2.17)$$

Очекивана вредност ове матрице може се записати:

$$E[(X - \mu)(X - \mu)'] = \Sigma. \quad (2.18)$$

Посматрамо линеарну комбинацију n случајних променљивих из случајног вектора X , и то записујемо на следећи начин: $Z = a_1X_1 + a_2X_2 + \dots + a_nX_n = a'X$, при чему са a' означавамо $[a_1, a_2, \dots, a_n]$. Линеарном комбинацијом дефинишемо нову случајну променљиву Z , чија функција густине $f_Z(z)$ зависи од функције густине X , тј. $f_X(x)$.

Средина ове случајне променљиве је дата са

$$\mu_Z = E(Z) = E(a'X) = a'\mu, \quad (2.19)$$

а варијанса од Z дата је као квадратна форма

$$\sigma_Z^2 = Var(Z) = Var(a'X) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \sigma_{ij} = a' \Sigma a. \quad (2.20)$$

Уколико бисмо посматрали случај са m линеарних комбинација и n случајних променљивих

$$\begin{aligned} Z_1 &= a_{11}X_1 + a_{12}X_2 + \dots + a_{1n}X_n \\ Z_2 &= a_{21}X_1 + a_{22}X_2 + \dots + a_{2n}X_n \\ &\vdots \\ Z_m &= a_{m1}X_1 + a_{m2}X_2 + \dots + a_{mn}X_n \end{aligned} \quad (2.21)$$

или записан у матричној нотацији $Z = AX$, онда би средина добијеног случајног вектора Z била

$$\mu_Z = E(Z) = E(AX) = A\mu_X, \quad (2.22)$$

где је Z вектор димензије $m \times 1$, а A матрица коефицијената линеарних комбинација димензије $m \times n$, а μ_X је средина случајног вектора X . А коваријациона матрица дата је са

$$\Sigma_Z = Cov(Z) = Cov(AX) = A\Sigma_X A', \quad (2.23)$$

где је Σ_X коваријациона матрица случајног вектора X .

2.1.4 Махаланобисово одстојање

Еуклидска раздаљина или Еуклидска метрика је уобичајена раздаљина између две тачке. На пример Еуклидска раздаљина између тачака $X = (x_1, x_2, \dots, x_n)$ и $\bar{X} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$, у Еуклидском n -простору, се дефинише као:

$$\sqrt{(x_1 - \bar{x}_1)^2 + (x_2 - \bar{x}_2)^2 + \dots + (x_n - \bar{x}_n)^2} = \sqrt{\sum_i^n (x_i - \bar{x}_i)^2}. \quad (2.24)$$

Заправо, код метода мултиваријационе анализе које се баве испитивањем међузависности посматраних објеката, одстојање је основна величина у анализи. Међутим, у статистичкој анализи Еуклидско одстојање није најоптималније, стога се користи Махаланобисово² одстојање које представља уопштење Еуклидског растојања.

„Махаланобис не само да приступа процесу стандардизације скалираних података у односу на стандардну девијацију, већ је интегрисан унутар групе варијансе-коваријансе која регулише интеркорелације између варијабли.“[24]

Махаланобисово растојање корисно је када постоји корелација неких атрибута објеката, уз евентуални додатак у да атрибути који су у корелацији да имају различите опсеге вредности, тј. различите варијансе. Такође ово одстојање се користи када је расподела података приближно нормална (Гаусова). Уколико бисмо тражили Махаланобисово растојање два објекта X и $\bar{X}$ ³ изразили бисмо то преко формуле:

$$d^2 = (X - \bar{X})\Sigma^{-1}(X - \bar{X})^T, \quad (2.25)$$

где је Σ узорачка коваријациона матрица за улазне податаке, а Σ^{-1} инверзна матрица матрици коваријанси података.

² Прасанта Чандра Махаланобис (29. јун 1893 - 28. јун 1972) био је индијски научник из области примењене статистике. Запамћен је управо по Махаланобисовој удаљености.

³ $\bar{X}$ је у улози узорачке средине.

2.2 Циљеви дискриминационе анализе

Постоје два циља дискриминационе анализе: дискриминација (раздвајање између група) и класификација (алокација опсервација).

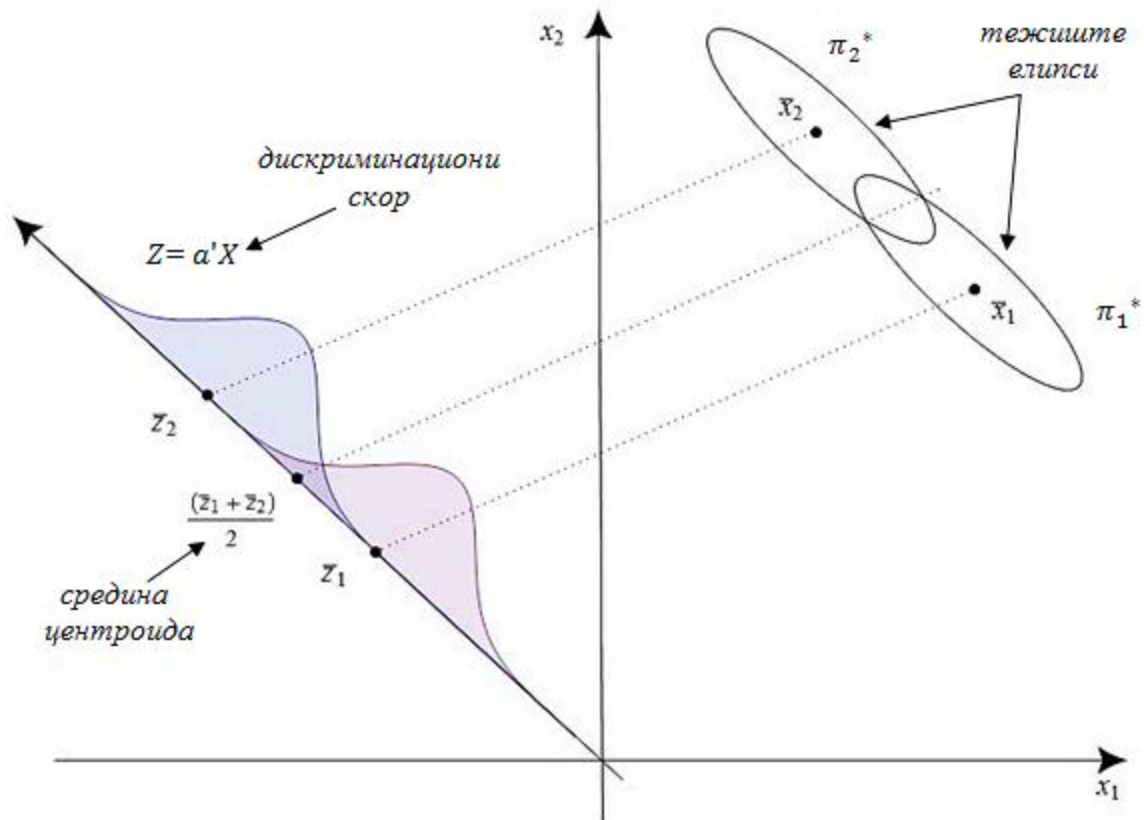
Дискриминација је један од циљева дискриминационе анализе којим се утврђује да ли постоји статистички значајна разлика у срединама две или више група, и одређује се променљива која највише доприноси датој разлици. „ Други циљ односи се на утврђивање поступка за класификацију опсервација на основу вредности неколико променљивих у две или више раздвојених, унапред дефинисаних група. У конкретном истраживању ова два циља често се међусобом преклапају, па средства анализе која користимо за раздвајање између група истовремено служе и за класификацију опсервација у те, унапред дефинисане групе.“ [2]

Осим ових назива за циљеве дискриминационе анализе, често се може наићи и на наслове дескриптивна дискриминациона анализа и метод класификације, где они редом одговарају раније наведеном првом и другом циљу.

Суштински, овом анализом се тежи ка постизању максимизације релативног односа варијанси између и унутар група, тј тежи се ка тежи се ка формирању линеарних комбинација независних променљивих којим ће се та дискриминација између унапред дефинисаних група извршити тако да се минимизира грешка која је настала погрешном класификацијом опсервација.

Један од основних појмова којим се служимо јесте дискриминациони скор који представља број добијен линеарном комбинацијом независних променљивих. Потом се тај скор трансформише у апостериорну вероватноћу којом проверавамо припадност посматраног објекта одређеној групи.

Посматрајмо упрошћен случај, где се презентује проблем дискриминације само између две групе (π_1^* и π_2^*) и уз помоћ две независне променљиве (X_1 и X_2). На *Слици 2.* презентован је графички циљ дискриминационе анализе, где су елипсама презентоване групе π_1^* и π_2^* .



Слика 2. Приказ дискриминационе анализе две групе (извор: Ковачић Ј. З. „Мултиваријациона анализа“, [2])

Као што је већ наглашено, дискриминациони скор се добија на основу линеарне комбинације вредности две независне променљиве, па уколико се искористе следеће ознаке: са Z је представљен дискриминациони скор, са a је представљен вектор коефицијената линеарне комбинације тј. дискриминационих коефицијената. И тако долазимо до формуле којом описујемо ту комбинацију:

$$Z = a'X \quad (2.26)$$

На претходно наведеној слици се може уочити пројекција тачака са дијаграма растурања на Z осу.

Том пројекцијом генерисана је једнодимензионална расподела дискриминационих скорова две популације π_1^* и π_2^* . Испрекиданим линијама су тежишта елипси $\bar{x}_1$ и $\bar{x}_2$ повезана са њиховим пројекцијама на Z оси (редом тачке $\bar{z}_1$ и $\bar{z}_2$). Са обзиром да је претходно већ истакнуто да је са Z је представљен дискриминациони скор, а појам којим описујемо средине дискриминационих скорова за ове две групе, назива се *центроид*. Поређењем центроида може се утврдити колико су групе међусобно удаљене. На Слици 2. може се уочити и средина центроида. Уколико посматрамо површину добијену преклапањем два једнодимензиона распореда, уочавамо да део површине лево од средине центроида

представља вероватноћу погрешне алокације објеката који припадају групи π_1^* , а алоциран је у групи π_2^* , и обрнуто за површину десно од средине центроида.

Овим путем се поједностављује проблем раздвајања популације у групе. Можемо уочити из претходно наведеног проблема да се уместо дводимензионих распореда променљивих X_1 и X_2 посматрамо једнодимензионе распореде дискриминационих скорова Z .

2.2.1. Дескриптивна дискриминациона анализа

Током 1936. године, Фишер⁴ је први дефинисао прецизније циљ дискриминационе анализе, представивши поступак дискриминације између две и више група.

Израчунавање Фишерове дискриминационе функције

Првобитно се посматра случај расподеле дискриминационих скорова једне популације, која се састоји од група π_1^* и π_2^* .

Сваку од ове две групе презентујемо вишедимензионалном функцијом густине, $f_1(x)$ и $f_2(x)$, међутим да би се димензија посматраног проблема смањила врши се трансформација у једнодимензионалну опсервацију у (дискриминациони скор). Циљ је да вредности које представљају средине дискриминационих скорова обе групе, означне са μ_1 и μ_2 буду максимално међусобно удаљене.

Уведимо одређене параметре којим описујемо формуле.

$\mu_k = E(X \pi_k^*)$	- очекивана вредност ⁵ случајног вектора X ако потиче из k -те популације
Σ	- коваријациона матрица
$\mu_{1Z} = E(Z \pi_1^*) = a'\mu_1$, односно $\mu_{2Z} = E(Z \pi_2^*) = a'\mu_2$	- очекивана вредност линеарне комбинације $Z = a'X$
σ_Z^2	- варијанса дискриминационих скорова

Затим се уводи количник који се назива *Фишеров дискриминациони критеријум*

$$\frac{(\mu_{1Z} - \mu_{2Z})^2}{\sigma_Z^2} = \frac{(a'\mu_1 - a'\mu_2)^2}{a'\Sigma a} = \frac{(a'\delta)^2}{a'\Sigma a}, \quad (2.27)$$

где је $\delta = \mu_1 - \mu_2$. Суштина је да се вектор коефицијената линеарне комбинације a одреди тако да максимизира наведени количник (2.27).

⁴ Ronald Aylmer Fisher (17. фебруар 1890 - 29. јул 1962) био је познати енглески статистичар, биолог, математичар, и генетичар.[42]

⁵ Очекивана вредност у овом случају представља такозвано *условно очекивање*, уз помоћ којег налазимо најбољу процену за X , уколико нам је познато из које групе потиче.

Максимизација се постиже уколико је

$$a = c\Sigma^{-1}(\mu_1 - \mu_2), c \neq 0. \quad (2.28)$$

Уколико се деси ситуација да је $c = 1$, онда на основу вектора a добијамо линеарну комбинацију

$$Z = a'X = (\mu_1 - \mu_2)' \Sigma^{-1}X, \quad (2.29)$$

која се назива **ФишEROVA линеарна дискриминациона функција**. Проблем оптимизације ове функције се решава уз ограничење да је $a'\Sigma a = 1$. „За овај избор коефицијената линеарне комбинације максимална вредност Фишеровог дискриминационог критеријума“ [2] дата је са

$$\max_a \frac{(a'\delta)^2}{a'\Sigma a} = \delta'\Sigma^{-1}\delta = (\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2). \quad (2.30)$$

Иначе, ФишEROVA линеарна дискриминациона функција се користи и за алокацију опсервација у једну од две групе. Одређује се тачка која се налази на средини између средина дискриминационих скорова из прве и друге групе и означена је са μ_Z , где је

$$\mu_Z = \frac{1}{2}(\mu_{1Z} + \mu_{2Z}) = \frac{1}{2}(a'\mu_1 + a'\mu_2) = \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2). \quad (2.31)$$

Уколико посматрамо нову опсервацију x_0 и њој одговарајући дискриминациони скор $z_0 = (\mu_1 - \mu_2)' \Sigma^{-1}x_0$, може се показати да важи правило класификације нове опсервације: x_0 се алоцира у π_1^* уколико је $z_0 = (\mu_1 - \mu_2)' \Sigma^{-1}x_0 \geq \mu_Z$, а алоцира се x_0 у π_2 ако је $z_0 = (\mu_1 - \mu_2)' \Sigma^{-1}x_0 < \mu_Z$.

Уколико нам нису познати параметри μ_1 , μ_2 и Σ , онда их оцењујемо на основу случајног узорка из прве и друге групе. При томе уводимо и узорачке показатеље $\bar{X}_1$, $\bar{X}_2$, S_1 и S_2 . Да би била формирана коваријациона матрица $\bar{S}$, корисимо наведене узорачке коваријационе матрице

$$\bar{S} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{(n_1 + n_2 - 2)}, \quad (2.32)$$

при чему су n_1, n_2 редом број опсервација из прве, односно друге групе.

Следеће што уочавамо јесте да је *максимална вредност Фишеровог дискриминационог критеријума* једнака квадрату *Махаланобисовог одстојања* између средина вектора $\bar{X}_1$ и $\bar{X}_2$. (2.33)

$$\max_a \frac{(\hat{a}'\hat{\delta})^2}{\hat{a}'\bar{S}\hat{a}} = \hat{\delta}'\bar{S}^{-1}\hat{\delta} = (\bar{X}_1 - \bar{X}_2)' \bar{S}^{-1}(\bar{X}_1 - \bar{X}_2) = d^2. \quad (2.33)$$

Како би се проверило да ли је раздвајање група статистички значајно уводи се нулта хипотеза $H_0 : \mu_1 = \mu_2$ наспрам алтернативне хипотезе $H_1 : \mu_1 \neq \mu_2$. Тест статистика која се користи јесте

$$\frac{n_1 n_2}{n_1 + n_2} \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} d^2 : F_{p, (n_1 + n_2 - p - 1)}. \quad (2.34)$$

Уколико одбацимо нулту хипотезу значи постоји статистичка значајност раздвајања група. На основу реализованих вредности случајног узорка добија се узорачка Фишера дискриминациона функција

$$Z = a'X = (\bar{x}_1 - \bar{x}_2)' \bar{S}^{-1}x. \quad (2.35)$$

И као што је већ било наведено тачка која се налази на средини у ознаци $\hat{\mu}_Y$, где је

$$\hat{\mu}_Z = \frac{1}{2}(\bar{z}_1 + \bar{z}_2) = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' \bar{S}^{-1}(\bar{x}_1 + \bar{x}_2). \quad (2.36)$$

Оно што још треба нагласити јесте да постоји ограничење типа $(n_1 + n_2 - 2) > p$, како би се избегао сингуларитет узорачке коваријационе матрице, где би последица била да њена инверзна матрица не постоји.

2.3 Мултиваријантна дискриминациона анализа

Дакле, применом Фишеровог приступа за две групе добија се једна линеарна комбинација независних променљивих. Том комбинацијом се постиже максимизација односа варијација између и унутар група.

Међутим уколико се посматра проблем са више група не може се очекивати да ће једна дискриминациона функција успешно да раздвоји те групе. Тиме се управо објашњава мултиваријациона дискриминациона анализа.

Мултиваријантна дискриминациона анализа (Multivariate discriminant analysis – MDA (или МДА)) или у одређеним литературама наведена као „мултиваријациона“, односно „мултиваријабилна“ дискриминациона анализа је статистички метод којим се дефинише природа и значајност евентуалних разлика које постоје између или унутар две или више унапред дефинисаних група. Врло је битно разликовати зависне од независних варијабли. Приликом дефинисања и одређивања евентуалних разлика, користе се подаци који су већ добијени одређеном анализом, процедуром. Оно што је изузетно код мултиваријационе дискриминационе анализе јесте да проналази које то независне варијабле најбоље раздвајају опсервације из датог узорка у подгрупе. Тиме се одређује и допринос и статистичка значајност сваке независне варијабле. Како би овај метод могао да се примени, неопходно је да се да су испуњене и задовољене одређене претпоставке како би добијени резултати били веродостојни.

Дакле основне претпоставке мултиваријационе дискриминационе анализе су:

- између независних варијабли не постоји мултиколинеарност
- независне варијабле имају мултиваријациону нормалну расподелу
- унапред су дефинисане групе опсервација
- матрице варијанси и коваријанси независних варијабли између група су једнаке
- веза између варијабли је линеарна
- величине група се не разликују много

У овом случају имамо n -димензиони вектор опсервација које потичу из једне од могућих група. Тих група има g , те на основу тога може се формирати $\min(n, g - 1)$ дискриминационих функција. Оно што треба нагласити јесте да није свака од добијених дискриминационих функција статистички значајна.

Уколико постоје две статистички значајне дискриминационе функције, то се може представити на одређеном дијаграму растурања, чиме се избегава проблем раздвајања g група на n -димензионих опсервација. Сам приказ опсервација у дискриминационом простору је значајан јер се уз помоћ њега може визуализовати веза између група, њихова блискост или пак удаљеност, затим њихово груписање, постојање одређених одступања итд.

У случају више група, поступак за одређивање дискриминационих функција је сукцесиван тачније прво се одређује прва дискриминациона функција, која максимизира релативан однос варијација између и унутар група. „Након тога свака наредна дискриминациона функција одређена је уз услов максимизирања односа преосталих варијација између и унутар група (након издвајања претходних дискриминационих функција) уз додатни услов да су њени дискриминационих скорови некорелисани са дискриминационим скоровима свих претходно одређених дискриминационих функција.“ [2]

Дакле посматрамо сада случај где се популација састоји од g група, $\pi_k^*, k = 1, 2, \dots, g$, и где је постављено за циљ да се уз помоћ n -димензионог вектора X изврши раздвајање ових група. Средињу случајног n -димензионог вектора X , ако је пореклом из k -те популације онда означавамо са $\mu_k = E(X | \pi_k^*), k = 1, 2, \dots, g$. Претпоставља се да су коваријационе матрице свих група међусобом једнаке (Σ).

Општа средина свих група добија се на основу средина μ_k

$$\mu = \frac{1}{g} \sum_{k=1}^g \mu_k. \quad (2.37)$$

„Како би се дефинисао Фишеров дискриминациони критеријум за случај више група уводи се матрица сума узајамних производа одступања средина група од опште средине.“ [2]

Та матрица ће бити означена са

$$B_0 = \sum_{k=1}^g (\mu_k - \mu)(\mu_k - \mu)'. \quad (2.38)$$

Потом се прави линерна комбинација елемената помоћ n -димензионог вектора X , тачније $Z = a'X$, чији су очекивана вредност и варијанса дефинисани на следећи начин

$$\mu_{kZ} = E(Z) = a'E(X | \pi_k^*) = a'\mu_k, k = 1, 2, \dots, g, \quad (2.39)$$

$$\sigma_Z^2 = Var(Z) = Var(a'X) = a'Cov(X)a = a'\Sigma a. \quad (2.40)$$

Општа средина линеарних комбинација добија се на основу средина μ_{kZ}

$$\mu_Z = \frac{1}{g} \sum_{k=1}^g \mu_{kZ} = a'\mu. \quad (2.41)$$

И сада се може дефинисати Фишеров дискриминациони критеријум за више група

$$\begin{aligned} \frac{\sum_{k=1}^g (\mu_{kZ} - \mu_Z)^2}{\sigma_Z^2} &= \frac{\sum_{k=1}^g (a'\mu_k - a'\mu)^2}{a'\Sigma a} = \frac{a'(\sum_{k=1}^g (\mu_k - \mu)(\mu_k - \mu)')a}{a'\Sigma a} \\ &= \frac{a'B_0a}{a'\Sigma a}. \end{aligned} \quad (2.42)$$

Фишеров дискриминациони критеријум представља релативан однос варијација између и унутар група. „Коефицијенте линеарне комбинације бирамо тако да се постигне максимална вредност овог критеријума уз ограничење да је $a' \Sigma a = 1$.“ [2]

Класификација опсервација у случају две групе захтева израчунавање средишње тачке између првог и другог центроида (дакле дискриминационих скорова по групама), па се у зависности од стационарања дискриминационог скорa нове опсервације у односу на ту средишњу тачку и врши алокација опсервација у групе. Уопштавајућо овај случај, користимо се Махаланобисовим растојањем опсервације у односу на центроид сваке групе, те алокацију нове опсервације вршимо спрам најмање раздаљине, тј нова опсервација ће бити алоцирана у групу од чијег центроида има најмању раздаљину.

2.4 Аналогија са регресијом

Мултиваријациона дискриминациона анализа је врло слична вишеструкој линеарној регресији, тачније преузима исти задатак као вишеструка линеарна регресија приликом предвиђања исхода, али основна разлика је у томе што код мултиваријационе дискриминационе анализе зависна варијабли је категоријска, а код вишеструке регресије зависна варијабли је скалирана.

У линеарној регресији⁶ је укупна сума квадрата $\sum (Y_i - \bar{Y})^2$ подељена је на регресијску суму квадрата $\sum (\hat{Y}_i - \bar{Y})^2$ и на суму квадрата резидуала $\sum (\hat{Y}_i - Y_i)^2$. Тачније укупна сума квадрата једнака је :

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (\hat{Y}_i - Y_i)^2. \quad (2.43)$$

Циљ регресије јесте управо процена параметара који минимизују суму квадрата резидуала.

Можемо затим уочити да у дискриминационој анализи је укупна сума квадрата $\sum (Z_i - \bar{Z})^2$ подељена је на суму квадрата између група⁷ $\sum (\bar{Z}_j - \bar{Z})^2 = (\bar{Z}_0 - \bar{Z})^2 + (\bar{Z}_1 - \bar{Z})^2$ и на суму квадрата унутар групе⁸ $\sum (Z_{ij} - \bar{Z}_j)^2 = (Z_{i0} - \bar{Z}_0)^2 + (Z_{i1} - \bar{Z}_1)^2$. Приказ тога је на *Слици 3*.

Једначина којом се то описује слична је једначини (2.43):

$$\sum (Z_i - \bar{Z})^2 = \sum (\bar{Z}_j - \bar{Z})^2 + \sum (Z_{ij} - \bar{Z}_j)^2, \quad (2.44)$$

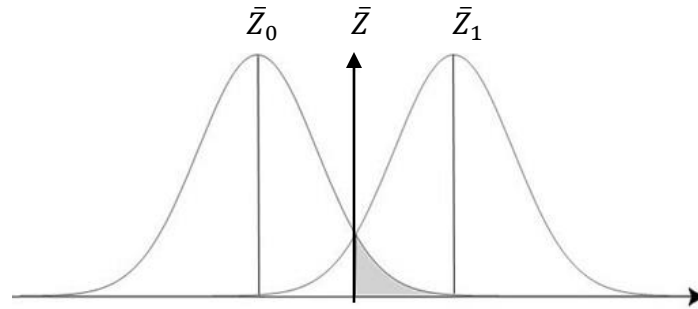
при чему ознаке коришћене у (2.44) представљају:

i	- индивидуални случај
j	- j -та група
Z_{ij}	- индивидуални дискриминациони скор
$\bar{Z}$	- средња вредност дискриминационих скорова
$\bar{Z}_j$	- средња вредност дискриминационих скорова j -те групе (центроид)

⁶ За више информација погледати [4]

⁷ Between Group SS= BSS

⁸ Within Groups SS = WSS



Слика 3. Приказ дискриминационе анализе две групе са центроидима
(извор: Charles M., Friel Ph.D., „Notes on Discriminant Analysis“, [16])

Циљ у овом случају је процена параметара који минимизују суму квадрата унутар групе.

Метод дискриминационе анализе као и метод мултиваријационе регресионе анализе и мултиваријационе анализе варијансе спада у групу метода зависности. Иако се кључно разликују, карактеристично за сва три метода јесте да формирају предвиђање зависне променљиве на основу скупа независних променљивих.

Оно што знамо јесте да је код регресионе анализе линеарна комбинација је формирана од мерљивих променљивих које су квантитативне. Такође, креће се од претпоставке да зависна променљива има нормалну расподелу и да је мерљива, а да независне променљиве представљају фиксне вредности у поновљеним узорцима. Док у дискриминационој анализи, већ нам је познато да се линеарна комбинација формира од мерљивих променљивих, али да је зависна променљива немерљива, тј. она је квалитативне природе. Зависна променљива је овде фиксна, док независне променљиве имају нормалну расподелу. „Циљ регресионе анализе јесте предвиђање средине зависне променљиве на основу скупа независних променљивих, док смо код дискриминационе анализе заинтересовани да одредимо линеарну комбинацију независних променљивих тако да се минимизира вероватноћа погрешне класификације опсервација у једну од група.“ [2]

У суштини методе су наизглед веома сличне али им је другачија основа којом приступају проблему. Слична ситуација се дешава и приликом поређења анализе варијансе и дискриминационе анализе, где знајући сада да су код дискриминационе анализе зависна променљива је квалитативна, а независне променљиве су квантитативне, код анализе варијансе је скроз супротна ситуација.

Логистичка регресија моделира условну вероватноћу $P(Z|X)$, да Z припада одређеној категорији у зависности од променљиве X . „Она је дакле облик регресије која омогућава предикцију зависне варијабле (дискретне) на основу континуираних и дискретних предиктора. Логистичка регресија се најчешће користи када релација између зависне (дискретне) варијабле и предиктора није линеарна.“[43]

Разматрамо алтернативни приступ моделирању овог проблема. У том приступу моделира се расподела променљивих X одвојено, сваке у одговарајућој класи. Затим добијене вероватноће користимо у Бајесовој формули (2.1) којом добијамо процену за дату условну вероватноћу.

Уколико се ради о нормалној расподели онда је овај модел сличан форми логистичке регресије.

Зашто нам је потребан још један метод као што је дискриминациона анализа? Постоји више разлога као што су на пример:

- Када су класе добро раздвојене, параметри процене за модел логистичке регресије су изненађујуће нестабилни, док код линеарне дискриминационе анализе тих проблема нема
- Уколико постоји више од две одговарајуће класе и ако расподела променљиве X за скоро сваку класу нормална, дискриминациони модел је стабилнији од модела логистичке регресије.

2.5 Фазе дискриминационе анализе

Дискриминациона функција служи како би пронашли варијабле због којих постоји велика разлика међу групама. Сама функција како би постигла жељени циљ, изводи дискриминационе коефицијенте којима се ове разлике смањују или пак нестају.

Коефицијенти дискриминационе функције су парцијални коефицијенти који одражавају јединствени допринос сваке варијабле у класификацији група зависних варијабли. Дискриминациони резултат који донекле јесте променљива, може се добити за сваки случај заменом коефицијената са вредностима у одговарајуће независне варијабле. Структурално коефицијенти су корелација између дате независне варијабле и дискриминационих резултата. Уколико је већа вредност коефицијената, већа је и веза између независне варијабле и дискриминационе функције.

Да би читава процедура дискриминационе анализе била испраћена, за формиране групе по појединим варијаблама рачунају се средње вредности, и на основу свих појединачних средњих вредности добија се средња оцена на нивоу читаве групе коју називамо *центроид*. Група центроида представља су средње дискриминационе оцене за сваку групу зависних варијабли, за сваку од дискриминационих функција. За две групе зависних варијабли постоји једна дискриминациона функција. Центроиди су иначе смештени у дводимензионалном простору.

Један од појмова који прати појам дискриминационе функције јесте Z вредност, као резултат те функције. Z вредност је у функцији дискриминације одређена вредношћу дискриминационог скорa и израчунава се за сваку опсервацију.

Израчунавање дискриминационе Z вредности представља заправо низ мера једноструке опсервације која укомпонована чини ту вредност. Формула којом се представља то је:

$$Z = a_1X_1 + a_2X_2 + \dots + a_nX_n, \quad (2.45)$$

где су a_i коефицијенти који се добијају на основу варијабли X_i уз које стоје.

Дискриминационе Z вредности се рачунају тако да максимизирају варијансу међу групама, а минимизирају је у оквиру групе, па уколико је варијанса међу групама релативно велика у односу на варијансу унутар групе онда кажемо да је дискриминациона функција добро раздвојила групе.

Уколико бисмо посматрали упрошћен случај, где се анализа своди на две групе опсервација, како бисмо утврдили која је опсервација вероватнија користи се критична дискриминантна вредност. Добијамо је тако што релацију сваке Z вредности израчунамо у односу на преломну вредност Z -а.

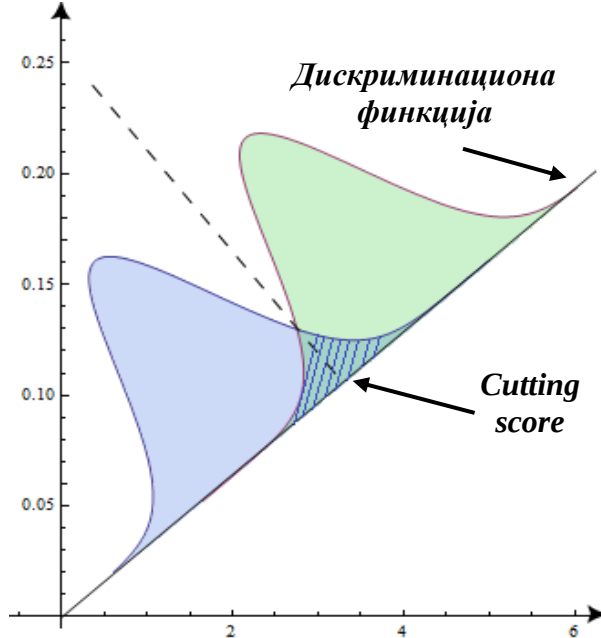
Како бисмо извели формулу за естимацију, уводимо следеће параметре:

S_{ij}	- оцена елемената i -тог реда, j -те колоне коваријационе матрице
N_1	- број опсервација у првој групи
N_2	- број опсервација у другој групи
a	- појединачне опсервације из прве групе ($a = 1, \dots, N_1$)
b	- појединачне опсервације из друге групе ($b = 1, \dots, N_2$)
$\bar{X}_i$	- средина i -те варијабле за посматрану групу
$\bar{X}_j$	- средина j -те варијабле за посматрану групу

Вишеструка линеарна анализа креће од претпоставке да су узорачке коваријационе матрице⁹ за све групе исте те се за то користи наредна формула:

$$S_{ij} = \frac{[\sum_{a=1}^{N_1} (X_{ai} - \bar{X}_i)(X_{aj} - \bar{X}_j) + \sum_{b=1}^{N_2} (X_{bi} - \bar{X}_i)(X_{bj} - \bar{X}_j)]}{N_1 + N_2 + N - 2}. \quad (2.46)$$

На бази вредности Z -а врши се класификација варијабли у унапред дефинисане групе, или пак поређење варијабли у односу на просечну вредност групе или независне варијабле. Можемо да дефинишемо да је центроид и вредност око које се концентришу Z вредности у једној групи. Уколико посматрамо две групе, онда на основу центроида обе групе формирамо Cutting score тј. вредност која раздваја групе. На основу односа између Z вредности и Cutting score одређујемо где нам која варијабла припада.



Слика 4. Cutting score и дискриминациона функција (графички приказ Станковић М., Стаменковић М., „Дискриминациона анализа“, [7])

⁹ Узорачку коваријациону матрицу дефинишемо преко матрице центрираних података и матрице узајамних производа центрираних података.

Шрафирани део на графику, на *Слици 4.* показује преклапање међу групама, и може се уочити да је то мала површина спрема целе две групе. Нормала која је нацртана испрекиданом линијом повучена је на дискриминациону функцију баш у тачки Cutting вредности, те она најбоље раздваја те групе.

Како је већ објашњен појам дискриминационе анализе, неопходно је проћи и кроз сваку фазу процеса како би се приближили више практичним могућностима овог поступка. Поступак обухвата шест фаза.

I. Фаза

Ова фаза се заснива на самом налажењу циљева дискриминационе анализе. Углавном се за циљеве узимају следећи статистички аспекти:

- постојање статистички значајних разлика између одређених независних варијабли
- идентификација независних варијабли које доприносе класификацији
- дефинисање процедура којима се врши класификација објеката у конкретне групе узимајући у обзир Z вредност

II. Фаза

У овој фази се спроводе следећа три поступка.

- Обично дефинишемо да су зависне варијабле дескриптивног карактера, а независне варијабле су нумеричког карактера.
- Уколико посматрамо величину групе за класификацију користи се правило да величина групе мора да буде већа од броја независних варијабли, које су укључене у анализу. (предлаже и истиче да величина узорка буде максимално 20, односно да број обсервација за сваку од идентификованих независних варијабли буде двадесет, будући да рад са мањим узорком носи и већу вероватноћу да ће добијени резултати и њихова веродостојност бити нарушени и нерепрезентативни.)
- Сврха саме дискриминационе анализе јесте подела узорка на групе. Постоје две групе које дефинишемо: узорак за анализирање и узорак валидације. Прва група служи за дефинисање дискриминационе функције, док се друга користи за тестирање валидности дефинисане дискриминационе функције. Овај метод је познат под називом Cross-validation приступ. У пракси не постоје стриктна правила око односа величина узорка за анализирање и узорка валидације у поступку поделе узорка.

III. Фаза

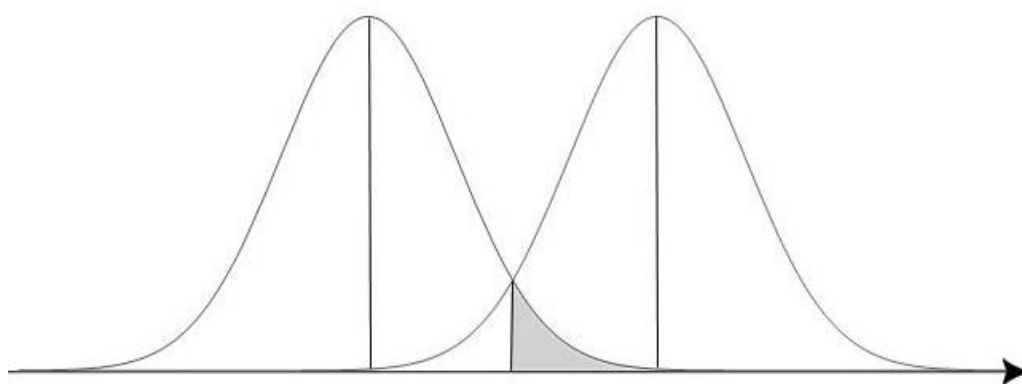
Да би дискриминациона анализа могла да се спроведе, врши се тестирање основних претпоставки дискриминационе анализе као што су нормална расподела независних варијабли, линеарност и одсуство мултиколинераности.

IV. Фаза

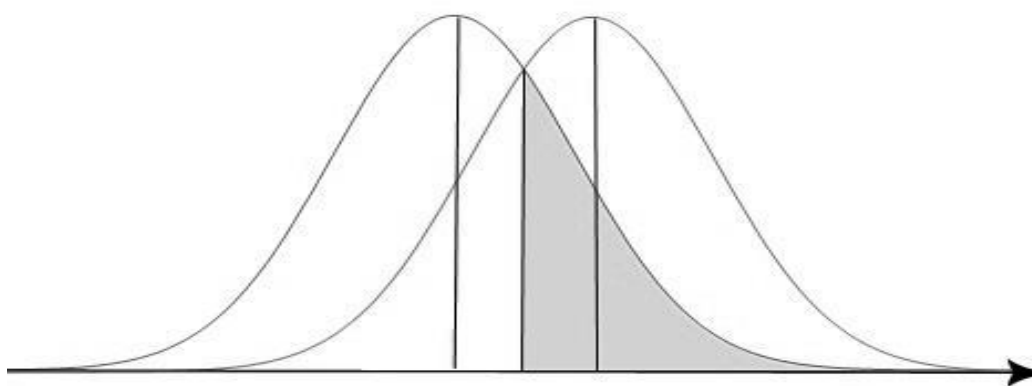
У овој фази утврђујемо дискриминациону функцију на основу два метода као што су: симултана(истовремена) процена и постепена процена.

- Симултана процена узима у обзир све независне варијабле истовремено, и при томе не посматра појединачно дискриминационе моћи тих варијабли.
- Постепена процена узима у обзир сваку посебно независну варијаблу, уводећи их једну по једну. Цео процес се започиње увођењем једне по једне варијабле, увођењем прво оне варијабле чија је дискриминациона моћ највећа и тако редом. Овим начином увођења се постиже елиминација небитних варијабли које не доприносе јачању дискриминационе моћи комплетне функције.

Дискриминациона моћ представља заправо способност дискриминационе функције да на валидан, ефикасан начин изврши класификовање по групама. (Слика 5.) Како би се уочио статистички значај дискриминационе моћи те функције користе се различити тестови у зависности од метода који је коришћен за извођење те функције.



Приказ ефикасне функције



Приказ неефикасне функције

Слика 5. Приказ ефикасне и неефикасне функције
(извор: Станковић М., Стаменковић М., „Дискриминациона анализа“, [7])

Уколико је у питању истовремена процена користе се:

- *Wilk`s lambda*
- *Hotelling`s trace*
- *Pillai`s criterion*
- *Roy`s largest root*

ако је у питању постепена процена користе се:

- *Mahalanobis D^2*
- *Chi-square test*

Критеријуми за статистичко закључивање се огледају у неколико мултиваријационих статистика које су нам доступне у програму¹⁰ како би били извршени тестови значајности који се односе на разне ефекте који утичу на променљиве. Те мултиваријационе статистике уколико је у питању истовремена процена су: Wilk's Lambda, Hotelling's Trace, Pillai's trace и Roy's largest root .

А од ових шест тестова биће посебно истакнут *Wilk's lambda*.

Wilk's lambda је статистички тест коришћен у мултиваријантној анализи да испита да ли постоје разлике између средина идентитета група субјеката у комбинацији зависних варијабли. То је дакле и кључна статистика којом се тестира да ли постоји или не постоји веза између независних и зависних варијабли. Када је вредност Wilks' Lambda за функцију је мала, функција је значајна.

Wilk's lambda статистика може да се трансформише у статистику која има приближно F расподелу. То олакшава да се израчуна *p*-вредност. Често аутори се пре одлуче да презентују F-вредност и степене слободе, пре него што се одлуче за вредност Wilk's lambda. Постоји велики број алтернативних статистика које се могу користити за обављање сличних задатака (Pillai's criterion, Roy's Largest Root).

Односи детерминанти се формирају за тестирање главних ефеката и интеракције уколико се користи Wilk's lambda. Ти односи се представљају формулом

$$\Lambda = \frac{\sum (Z_{ij} - \bar{Z}_j)^2}{\sum (Z_i - \bar{Z})^2}. \quad (2.47)$$

Дакле, Wilk's lambda Λ (2.47) је однос суме квадрата унутар групе (варијација унутар групе) спрам тоталне суме квадрата (где су урачунате и варијације између група и варијације унутар групе).

Hotelling's Trace је однос ефекта варијације између група и варијације унутар групе. Pillai's criterion је комбинација ефекта одступања. Wilk's lambda, Hotelling's Trace, Roy's Largest Root су често моћнији од Pillai's criterion (уколико постоји више од једне димензије).

Wilk's lambda се користи за тестирање нулте хипотезе која тврди да су очекиване средње вредности свих независних варијабли једнаке. Уколико се величина узорка смањује онда се неједнакости појављују, а претпоставка о хомогености варијансе коваријационе матрице¹¹ је нарушена, тако да је тада битан Pillai's criterion у погледу поузданости.

¹⁰ Односи се на било који статистички програм, као што су нпр. СПСС, СТАТИСТИКА.

¹¹ Коваријациона матрица је заправо очекивана вредност случајне симетричне матрице (она представља производ случајних вектора одступања од средине). Користи се за меру варијабилитета.

Дефинисање и формирање класификационе матрице

Класификациона матрица је једноставан табеларни приказ посматраних узорака. Како би постојала добра предикција неопходно је да на дијагонални елементи матрице буду вредности које су високе, а елементи изван дијагонале требају да имају вредности које су близу 0 или пак једнаки 0.

Процес формирања класификационе матрице почиње применом Cross-validation приступа (објашњен у делу 2.5 у Фази II), па се процес наставља поделом укупног узорка на две групе, узорка за анализирање и на holdout узорак.

Узорак за анализирање користимо за дефинисање дискриминационе функције, док је holdout узорак из популације који се не користи у дефинисању функције, већ се користи за тестирање ње.

Пошто постоје одређена одступања у класификацији, тачније уколико су поједини чланови подељених група погрешно класификовани, онда да би се спречило то користе се оцене тачности поступака класификовања.

Карактеристични корен

Корени карактеристичног полинома се називају карактеристични корени матрице. Скуп свих тих корена одређују спектар матрице. Да би модел био добар, мора да има више од једног карактеристичног корена. У дискриминационој анализи постоји један карактеристични корен за сваку дискриминациону функцију посебно. Што је већи карактеристични корен, јача је дискриминациона моћ функције.

Релативни проценат од дискриминационе функције је карактеристични корен подељен са сумом свих карактеристичних вредности свих дискриминационих функција у моделу.

Карактеристични корен је однос између: тоталне суме квадрата и варијација унутар групе. Величина карактеристичног корена нам указује на способности дискриминационе функције у раздвајању између група.

Карактеристични корени ће увек бити наведени у опадајућем редоследу, пошто решење у дискриминационој анализе захтева да на прво место иде најкориснија дискриминациона функција.

$$\lambda = \frac{\sum (\bar{Z}_j - \bar{Z})^2}{\sum (Z_{ij} - \bar{Z}_j)^2} \quad (2.48)$$

Канонска корелација

Уз помоћ карактеристичног корена може се извести један користан статистички индикатор. У питању је канонска корелација η .

Канонска корелација је мера повезаности између две мултидимензионалне варијабле. Висока вредност подразумева висок степен повезаности између те две варијабле и обрнуто.

$$\eta = \sqrt{\frac{\lambda}{1 + \lambda}} \quad (2.49)$$

Одређивање Cutting score-a

Стављањем у однос дискриминационе Z вредности сваког појединачног узорка и Cutting score-a врши се класификовање узорака у унапред дефинисане групе. Постоје различити начини за рачунање Cutting score, спрема тога да ли су групе истих или различитих величина:

групе једнаких величина

$$Z_{CE} = \frac{\bar{Z}_0 + \bar{Z}_1}{2}$$

групе неједнаких величина

$$Z_{CU} = \frac{N_1 \bar{Z}_0 + N_0 \bar{Z}_1}{N_0 + N_1} \quad (2.50)$$

где су:

$\bar{Z}_0$ и $\bar{Z}_1$	- центроиди прве групе и друге групе
N_0 и N_1	- број чланова прве групе и друге групе
Z_{CE} и Z_{CU}	- Cutting score за групе истих и различитих величина

Одређивање критеријума заснованих на случајности

Критеријуми засновани на случајности се формирају у зависности од тога да ли се ради о групама једнаких или неједнаких величина. Уколико су у питању групе неједнаких величина, за критеријум се узима процентуална вредност величине веће групе и тај критеријум се назива максимални "случајни" критеријум. Постоји и прецизнији метод користи се пропорционални "случајни" критеријум, чији је образац за утврђивање дат

$$C_{pro} = p^2 + (1 - p)^2, \quad (2.51)$$

у коме p вредност представља величину веће групе изражене у процентима. Завршни елемент процеса, вршимо поређење добијеног hit ratio са утврђеним критеријумом заснованим на случајности, да бисмо утврдили тачност и ефикасност поступка класификовања испитаника по групама.

Да би ефикасност датог поступка оценили позитивно потребно је да вредност hit ratio буде најмање за $1/4$ већа од вредности датог критеријума заснованог на случајности.

V. и VI. Фаза

Ове фазе односе се на евалуацију добијених резултата.

Сам процес интерпретације добијених резултата подразумева разматрање дискриминационе функције како би се дошло до утврђивања значаја и дискриминационе моћи сваке независне варијабле које смо користили за вршење дискриминације посматраних објеката по унапред идентификованим групама. Разликујемо три метода који се предлажу за утврђивање релативног значаја независних варијабли:

- Стандардизована дискриминантна тежина,
- Структура корелација
- Појединачне F вредности.

Од ова три метода најбољим се показао трећи метод чија се поента интерпретације огледа се у утврђивању апсолутних величина F вредности за сваку независну варијаблу и њиховом рангирању. Уколико се деси да је велика F вредност, онда та независна варијабла поседује значајну дискриминациону моћ и обратно.

2.6 Проблеми у примени дискриминационе анализе

Изузетно је битно нагласити да постоје одређени проблеми уколико дође до напуштања претпоставки. Код сваког статистичког процеса неопходно је утврдити да ли су одређени параметри статистички значајни. Процедурама статистичког тестирања значајности, утврђујемо колико нам је дискриминационих функција потребно за даљу анализу и колико променљивих ћемо укључити. Не узимајући то у обзир, вектор X је садржао све расположиве променљиве, међутим природно се поставља питање да ли су неопходне све променљиве како би се извршило најоптималније раздвајање група или алокација опсервација. Такође је неопходно и разматрати да ли су те променљиве квалитативне или квантитативне природе.

Један од начина да се утврди висина доприноса неке променљиве у раздвајању група јесте да се посматра дискриминациони коефицијент који стоји уз променљиву. Дакле, уколико је коефицијент одређене променљиве близак нули, онда ту променљиву можемо изоставити из дискриминационе функције.

Осим овог начина, може се приступити и са становишта „раздвајања две групе где је на основу Фишеровог дискриминационог приступа добијен резултат да је максимална вредност Фишеровог дискриминационог критеријума једнака квадрату Махаланобисовог одстојања између две групе.“ [2] Уколико изостављањем одређених променљивих нова добијена вредност Махаланобисовог одстојања буде непромењена или се тек мало разликује од претходне, онда се долази до закључка да изостављене променљиве заправо и не доприносе раздвајању група.

Међутим исто као што проверавамо колико постојећа променљива доприноси, тако исто можемо и да уведемо додатну променљиву поред раније укључених променљивих. Овај поступак се реализује коришћењем Вилксове ламбде Λ_j .

Узимамо да имамо j променљивих, а тест статистика за један увећан број променљивих $(j + 1)$ записује се:

$$\frac{n - g - j}{g - 1} \left(\frac{\Lambda_j}{\Lambda_{j+1}} - 1 \right) : F_{(g-1), (n-g-j)}, \quad (2.52)$$

при чему g представља број популација. Само у случају великих узорака и уколико $j + 1$ променљива није статистички значајна, онда ова статистика има F распоред (2.52) при чему $(g - 1), (n - g - j)$ представљају степене слободе. „Недостатак ове F статистике огледа се у томе што показује тенденцију повећања раздвајања добро раздвојених група, уместо да побољшава раздвајање слабије раздвојених група. Уколико p није велико, предлага се поређење резултата раздвајања добијених преко функција које су рачунате на основу свих подскупова променљивих.“ [2]

Такође, пошто је опште познато да је метода дискриминационе анализе заправо метода зависности, и знамо да је зависна променљива квалитативног типа, интересује нас шта уколико је у независне променљиве укључена квалитативна променљива, или шта уколико су све независне променљиве тог типа. У тим случајевима се саветује да се примењује логистичка дискриминациона анализа.

Ако се деси да су све независне променљиве квалитативне, или да постоји мешовит случај да су неке од независних променљивих квалитативне, тада није разумно претпоставити нормалност распореда случајног вектора X . У том случају ФишEROVA дискриминациона функција даје слабе резултате класификације. Дакле, уколико се не може претпоставити нормална расподела онда се препоручује коришћење логистичке дискриминационе анализе.

2.7 SPSS активност - дискриминациона анализа

SPSS¹² је софтверски пакет коришћен за статистичку анализу. Бави се многим врстама статистичких прорачуна, као што је нпр. предвиђање за идентификацију група. У те статистичке прорачуне спадају факторска анализа, кластер анализа и дискриминациона анализа.

Дискриминациона анализа гради интуитивни модел за припадање групи. Модел се састоји од дискриминационе функције (или се састоји са више од две групе, или од скупа дискриминационих функција).

Битно је напоменути да груписана променљива може имати више од две вредности. Вредности за груписану променљиву морају бити цели бројеви, међутим, и потребно је да се одреде минималне и максималне вредности. Случајеви са вредностима које излазе из те граничне линије (минималне / максималне вредности) се искључују из анализе.

У SPSS-у класификација посматраних случајева постиже се израчунавањем условних вероватноћа. Наравно програм избацује у виду табела резултате, а у позадини ових рачуна, између осталог стоји Бајесова теорема (видети део 2.1.1.). За сврхе груписања Бајесова теорема се записује :

$$P(G_i | D) = \frac{P(D | G_i) P(G_i)}{\sum P(D | G_i) P(G_i)}, \quad (2.53)$$

при чему је означено¹³ са :

D	- дискриминациони скор (Z)
$P(G_i D)$	- вероватноћа да уколико случај има дискриминациони скор D , онда је смештен у i -ту групу
$P(D G_i)$	- вероватноћа да уколико је у питању i -та група, онда има дискриминациони скор D
$P(G_i)$	- вероватноћа да је случају групи G_i

Статистички подаци који се могу израчунати уз помоћ SPSS-a:

- за сваку променљиву: средња вредност, стандардне девијације, униваријантна АНОВА.
- за сваку анализу: Вох's M, унутар група корелације матрица, у оквиру групе коваријанса матрице, матрица коваријансе одвојених група, укупна матрица коваријансе.

¹² Име софтвера оригинално на енглеском је Statistical Package for the Social Sciences (SPSS).

¹³ Ови подаци се иначе могу прочитати у табели Casewise Statistic у СПСС-у.

- за сваку каноничку дискриминациону функцију: карактеристични корен, проценат варијансе, каноничке корелације, Wilks lambda, Chi-square test, Махаланобисова удаљеност, најмањи F количник.
- за сваки корак: приоритетна вероватноће, Фишера функција коефицијената, нестандардизована функција коефицијената, Wilks lambda за сваку каноничку функцију.

У пар корака уз помоћ програма долазимо до одговора на питање: које су променљиве статистички значајне, а које треба избацити.

На самом почетку приликом коришћења овог програмског пакета треба одабрати начин за уношење независне варијабле:

- Унети независне променљиве заједно. Истовремено се уносе све независне варијабле које задовољавају критеријуме толеранције.
- Користити поступни метод. На тај начин се контролише унос и избацивање одређене променљиве, која није значајна. Дакле, када постоји много предиктора поступни метод може бити врло користан, зато што аутоматски бира "најбоље" варијабле да се користе у моделу.

Одабиром поступног метода или степене дискриминационе анализе модел се дефинише корак по корак. Оцењена варијабла укључује се у модел и поступак започиње из почетка. О избору варијабли у степеној дискриминационој анализи F вредност има пресудан значај. F вредност поједине варијабле показује њену статистичку значајност у дискриминацији између група. Такође, F вредност је и мера јединственог доприноса поједине варијабле у поступку предикције припадности групи неког објекта.

Једна од ствари које су претходно обрађене теоријски јесу коефицијенти функције, а доступне опције за коефицијенте су Фишера класификација коефицијената и нестандардизовани коефицијенти. SPSS приказује Фишерову класификацију функције коефицијената који се могу користити директно за класификацију и приказује коефицијенте нестандардизоване дискриминационе функције.

Дакле методи који се могу одабрати за обраду података су:

- **Wilks lambda** - тест који је у делу (2.5) објашњен детаљније. Променљива се уноси у дискриминациону функцију на основу тога колико смањује вредност Wilks lambda. На сваком кораку, варијабла која минимизира Wilks lambda биће убачена у модел.
- **Необјашњена варијанса** - на сваком кораку, варијабла која минимизира збир необјашњиве варијације између група је унета.
- **Махаланобисова удаљеност** - је статистички тест који идентификује случајеве као постојеће екстремне вредности једне или више независних варијабли.

- **Најмањи F количник** - променљива се уписује у модел ако је ниво значајности њене F вредности мања од улазне вредности (унесене слободно у програму), а уклања се из модела ако је ниво значајности већи од вредности за уклањање. Унос мора бити мањи од уклањања, али обе вредности морају бити позитивне. На сваком кораку, променљива са највећом F вредношћу, која премашује критеријуме за улазак се додаје моделу.
- **Box's M тест** - тестира претпоставку о хомогености коваријансе матрица. Дакле у питању је нулта хипотеза која гласи: варијационе односно у овом случају коваријационе матрице две или више група су једнаке. Овај тест је веома осетљив на испуњавање претпоставке мултиваријантне нормалности. Дискриминациона анализа функције је јака, чак и када није испуњена претпоставка о хомогености варијанси, под условом да подаци не садрже важне аутлајере. Када је узорак велик, мала одступања од хомогености ће бити значајна, због чега се Box's M мора представити кроз логаритме детерминанти матрице коваријансе. Box's M тест се не може извести, ако постоје мање од две несингуларне групе коваријационих матрица.
- **Структура матрице** - табела у SPSS-у са називом „структура матрице“ показује корелацију сваке варијабле са сваком дискриминационом функцијом. Корелације идентификују највеће апсолутне корелације повезане са сваком дискриминационом функцијом
- **Statistical significance, Sig** - статистичка значајност одређене варијабле се потврђује уколико је мања од 0,05.
- **Карактеристични корен** - табела са карактеристичним коренима пружа информације о ефективности дискриминационих функција.

3 Основе Data mining-а и његових основних техника

„Подели на владај - Divide et impera“ (латинска изрека)

Развој Data mining-а почео је током педесетих година прошлог века, али је своју експанзију и примену досегао тек са развојем нове технологије. У почетку методе које га данас описују нису називане Data mining техникама, иако су се као такве примењивале за велики број научних истраживања и експериментисања. Изузетним напретком на пољу информационе технологије, све веће количине података су могле да се похрањују у базама (нпр. предузећа), тако да се управо због тога повећала потреба за напретком и све чешћим коришћењем Data mining техника у сврхе пословног и комерцијалног одлучивања.

Дакле, „Data mining подразумева коришћење софистицираних алата за анализу, а они могу укључивати статистичке моделе, математичке алгоритме, методу машинског учења, базе података и сл.“[14]

Data mining јесте и процес којим проналазимо одређене узорке у базама података које користимо као знања корисна за наше будуће предикције. Као основне технике Data mining-а можемо издвојити стабла одлучивања, правила асоцијалције, неуронске мреже, метода најближег суседа, алгоритам најближе тачке, генетичко учење итд. Продукциона правила, листе одлучивања и стабла одлучивања су примери разумљивог начина представљања емпиријског знања.

3.1 Стабла одлучивања

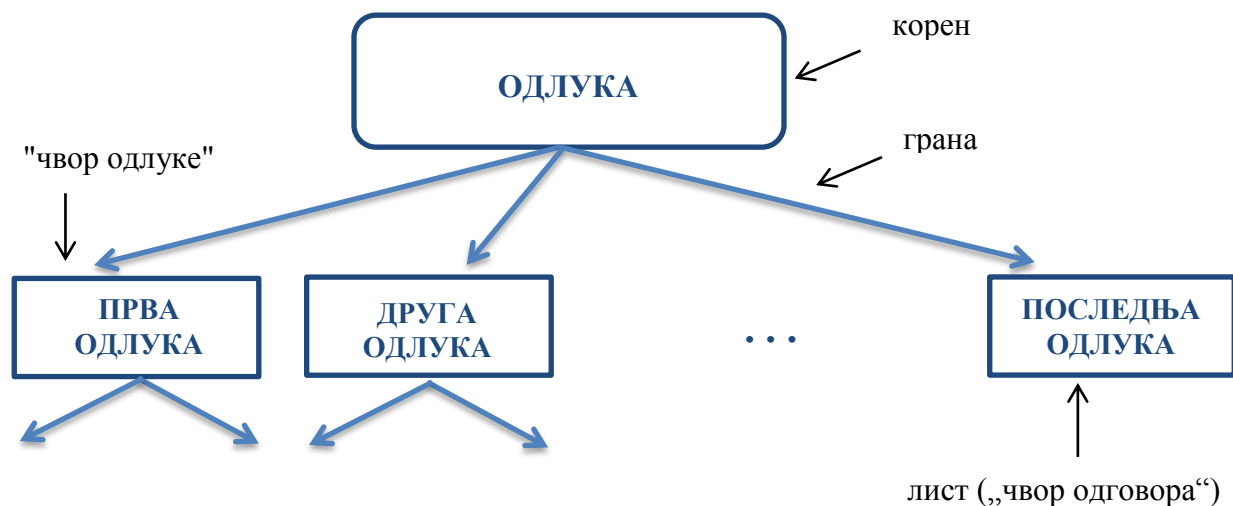
Стабло¹⁴ одлучивања (Decision Tree) је једна од најчешће коришћених и практичних метода за индуктивно закључивање. То је метод који се користи приликом решавања проблема у финансијама, банкарству, маркетингу, осигурању.. Оно што је изузетно код овог метода јесте предвиђање исхода категоризацијом података, сагледавањем стратегија и стања.

Уколико се испуне услови питања за класификацију или правила за категоризацију података онда се гранањем долази до стабла одлучивања. Питањима се категоризују подаци у подскупове, који су хомогенији него виши скуп. Уколико постоје два одговора на постављено питање, тада настају два подскупа као одговори на питање (и тиме се добија бинарно стабло). „Колико питање има одговора толико ће подскупова настати. Самим тим врши се класификација појединих података.“ [14]

¹⁴Стабло(дрво) је хијерархијска структура. Постоји више врста стабала али овде се описује хијерархијско или коренско стабло.

Приликом стварања стабла одлучивања битно је поставити право питање. Од питања зависи колико ће бити хомогени подскупови који настају. Стабла одлучивања се доста разликују у моделима, с обзиром да немају исте базе података на основу којих настају. Дакле, стабла одлучивања се заснивају на методу „подели па владај“, где се скупови деле на подскупове и сам процес је рекурзиван све док не дође се до једног примера којим се врши класификација знања. Елементи знања код стабла одлучивања су чворови и гране.

Један чвор се издваја и чини врх хијерархије („корен“). Свака грана повезује наредни чвор са његовим јединственим претходником. Уколико се од неког чвора даље не грана хијерархија онда се ти чворови називају листови („чворови одговора“). Сви остали чворови су "чворови одлуке". Једно стабло одлучивања представљено на *Слици 6*. Дакле, стабла одлучивања класификују случајеве сортирајући их редом од корена стабла до листа-чвора. Сваки чвор у дрвету представља тест за неки атрибут датог случаја.



**Слика 6. Сlikовити приказ изградње једног стабла одлучивања
(извор: приказ аутора)**

Иако су развијена различита стабла одлучивања, обухватајући различите могућности и услове, ипак овај метод учења је генерално изложен проблемима са следећим карактеристикама:

- Случајеви представљени као пар: атрибут-његова вредност. Најлакша ситуација за стабло одлучивања је када сваки атрибут прима мали број дисјунктних могућих вредности. Међутим, проширења основног алгоритма омогућују руковање реалним нумеричким вредностима атрибута.
- Дисјунктивни описи могу бити неопходни. Стабло одлучивања је подложно грешкама (било да су грешке у класификацији, или грешке код вредности атрибута). Чак могу бити садржани атрибути чије вредности не знамо.

Треба и нагласити да стабло одлучивања прати неизвесност, какве ће последице бити проузроковане овим одлукама.

3.2. Правила асоцијације

Уколико се располаже великом базом података, како би се откриле жељене променљиве и њихов однос са осталим променљивима- користе се правила асоцијације као добро познат метод. Намењена су проналажењу јаких закона¹⁵ откривеним у базама података коришћењем различитих мера интересантности.

На пример: Ракеш Агравал¹⁶ увео је правила асоцијације како би пратио и откривао регуларности унутар скупа продуката у трансакцијама података великих размера. Све ове трансакције су биле бележене уз помоћ система у супермаркетима. Ту се управо огледа изузетна примена правила асоцијације, где се уз помоћ ње управља информацијама које су врло корисне за опстајање на тржишту роба и услуга.

Осим ове наведене примене, асоцијативна правила су изузетно распрострањена у разним сферама технологије у коју спадају претраживање, коришћење светске мреже, биоинформатике, детекције упада итд.

3.3. Метода најближег суседа

Метода најближег суседа једна од најстаријих и најједноставнијих техника која се примењује. Базира се на начину рада који је доста сличан људском начину размишљања. Овај поступак заснива се на налажењу минималне удаљености.

Овде је битно одредити ко је најближи сусед посматраном објекту. Полази се од претпоставке да уколико неки податак има слична својства, значи да ће се слично и понашати, а из тога произилази да је најближи сусед. „Како би метода била што тачнија, потребно је у бази података наћи што сличнији податак (за који је потребно што тачније познавати понашање), што захтева велике количине података.“ [14]

Код методе најближег суседа, модел заправо представља податке за учење, тачније да би се креирао модел не постоји процес учења. Појавом новог податка, уз помоћ алгорита анализирају се сви подаци у бази како би се нашла највише одговарајућа подгрупа случајева за тај нови податак. Након тога врши се предвиђање. Проблеми се наравно могу појавити уколико су подгрупе лоше класификоване.

¹⁵ Концепт јаких закона се најчешће примењује како би се одбацили непотребни предиктори, за више информација о њима погледати

¹⁶ Ракеш Агравал /Rakesh Agrawal/ је индијски научник рачунарских наука и технички сарадник на Microsoft Search Labs. Познат је по развијању основних data mining концепата .

3.4. Неуронске мреже

У самом називу ове технике, уочавамо да постоји одређена корелација са функционисањем људског мозга. Дакле, као што се људски мозак током усвајања нових података, процеса тј. током учења, служи меморијом, на основу које извлачи одређене претпоставке, тако и овај метод неуронске мреже предвиђа одређена дешавања у систему. Заправо, цео процес функционише тако да прво се обрађују подаци са већ познатим исходима, односно излазним вредностима, тачније они се дају неуронској мрежи. Затим неуронска мрежа служећи се вештачком интелигенцијом меморише и уочава правила и обрасце на основу ових података. Са овим се постиже да се гомила података, која постоји у базама (нпр. предузећа), класификује, селекује и истражи на веродостојан начин.

За одређене врсте проблема као што је учење како да се интерпретирају комплексни стварном свету сензорски подаци, вештачке неуронске мреже су једне међу најефикаснијим методама учења.

Студије о вештачкој неуронској мрежи су инспирисане великим делом запажањем да је биолошки систем за учење изграђен од веома сложених мрежа међусобно повезаних неурона. У грубој аналогији, вештачке неуронске мреже су изграђене од густо повезаног скупа једноставних јединица, при чему свака јединица представља број реалних вредности улазних јединица и производи реално вреднован аутпут. Како би се развио осећај за ову аналогију, размотримо неколико чињеница из неуробиологије. Људски мозак, на пример, је процењено да садржи густо повезану мрежу од око 10¹¹ неурона. Неуронска активност се исказује преко везе са другим неуронима. Најбржи неурон има просечну брзину око 10⁻³ секунди – што је прилично споро у односу на компјутерску брзину 10⁻¹⁰ секунди. Међутим ипак људи су у стању да доносе изненађујуће сложене одлуке, изненађујуће брзо. На пример, потребно је око 10⁻¹ секунди до визуелно препознамо своју мајку. Ово запажање је довело многе да спекулишу да информационо-процесно-прерађивачке способности се врше преко мноштво неурона. Једна од мотивација за вештачку неуронску мрежу је да искористи ту врсту паралелног процесуирања на основу дистрибуиране репрезентације.

Међутим, ова метода спада у једне од најкомпликованијих метода (како за употребу, тако и за примену), али уз помоћ ће добијају се најтачнији модели.

4 Примена мултиваријантне дискриминационе анализе у процесу ревизије

4.1 Процес ревизије

Сам процес ревизије представља реализацију одређених активности од стране ревизора, које су предвиђене планом и програмом. Те активности су донекле доста јединствене, и зависе од начина рада ревизора и од облика пословања клијента.

Дакле, ангажованост ревизора зависи од тога колико је поуздан систем интерне контроле које то предузеће има. Како би одредили поузданост интерне контроле ревизори посматрају како сам систем функционише, како су изведене одређене трансакције (продаја, куповина, производња и сл.). Током разматрања трансакција на ревизорима је да описују стање система у радним папирима које касније користе приликом формирања мишљења. Оно што је битно нагласити јесте да радни папири садрже и финансијски извештај који је био предмет ревизије.

Ревизија представља дакле спровођење поступака ради прибављања ревизијских доказа о износима и обелодањивањима датим финансијским извештајима. Одабрани поступци су засновани на ревизорском просуђивању, укључујући процену ризика постојања погрешних исказа у финансијским извештајима, насталих услед криминалне радње или грешке. Приликом процене ризика, ревизор сагледава интерне контроле релевантне за састављање и истинито приказивање финансијских извештаја зарад осмишљавања ревизијских поступака који су одговарајући у датим околностима, али не и циљу изражавања мишљења о делотворности интерних контрола правног лица. Ревизија такође укључује оцену адекватности примењених рачуноводствених политика и оправданости рачуноводствених процена извршених од стране руководства, као и оцену опште презентације финансијских извештаја.

Након спроведених процедура ревизије, ревизор формира мишљење о финансијским извештајима. Мишљење лиценцираног овлашћеног ревизора представља потврду да ли финансијски извештаји пружају истинит и поштен преглед финансијског положаја правног лица у складу са релевантним законским оквиром, као и да ли су годишњи финансијски извештаји у складу са посебним прописима који уређују пословање правног лица, ако је то предвиђено тим прописима.[29]

Мишљење ревизора може бити:

- *позитивно,*
- *мишљење са резервом,*
- *негативно,*
- *уздржано мишљење.*

Позитивно мишљење, дакле без резерви, даје се када је ревизор сигуран да финансијски извештаји дају истиниту и објективну слику пословања. Уколико постоји покренут спор против посматраног предузећа, чији је исход неизвестан, онда ревизор даје такозвано *позитивно мишљење са скретањем пажње*.

Мишљење са резервом или квалификовано мишљење да повољан сигнал о финансијским извештајима, али уз одређене резерве (модификације) које могу бити класификоване као материјалне, али не и фундаменталне. Резерва у овом типу мишљења се изражава са нагласком: „осим за“, чиме се истиче да постоје неке неизвесности или неслагања.[1]

Негативно мишљење се изражава када су ефекти неслагања од таквог материјалног значаја и утицаја на финансијске извештаје, да ревизор жели на неки начин да овим мишљењем упозори да услед непридржавања и непримењивања рачуноводствених стандарда финансијска позиција, финансијски резултат и готовински ток нису коректно приказани у финансијском извештају.

Уздржавање од мишљења се издаје када ревизор не може да изврши ревизију у обиму који му омогућава да уопште формира мишљење о финансијском извештају.[1] Дакле, овлашћени ревизор уколико није у могућности да изрази мишљење, може се уздржати од давања истог.

4.2 Финансијски извештаји предузећа дуванске индустрије у Републици Србији

Дуванска индустрија је једна од старијих грана индустрије, која осим што се бави узгајањем и прерадом дувана бави се и производњом дуванских производа. У нашој земљи цигарете тренутно производе Дуванска индустрија Ниш у власништву компаније Philip Morris, Дуванска индустрија Врање, коју је купио British American Tobacco, Japan Tobacco International Сента, Дуванска индустрија Чока и Monus у Инђији, али Monus нисмо уврстили у рад јер је мало правно лице и не подлаже обавези вршења ревизије финансијских извештаја.

Стога анализираћемо примену мултиваријационе дискриминационе анализе на подацима који су везани за четири претходно наведена дуванска предузећа.

Подаци које користимо су из 2011, 2012. и 2013. године и званично су објављени у складу са законом на сајту Агенције за привредне регистре (www.apr.gov.rs).

Крећући се кроз годишње податке, тачније уз помоћ одређених позиција из биланса стања применићемо мултиваријациону дискриминациону анализу.

Позиције које узимамо као значајне за нашу даљу анализу су: „Нематеријална улагања“, „Залихе“, „Потраживања“, „Капитал“, „Обавезе из пословања“.

АКТИВА	
А. СТАЛНА ИМОВИНА (002+003+004+005+009)	001
III. НЕМАТЕРИЈАЛНА УЛАГАЊА	004
Б. ОБРТНА ИМОВИНА (013+014+015)	012
I. ЗАЛИХЕ	013
III. КРАТКОРОЧНА ПОТРАЖИВАЊА, ПЛАСМАНИ И ГОТОВИНА (016+017+018+019+020)	015
1. Потраживања	016
ПАСИВА	
А. КАПИТАЛ (102+103+104+105+106-107+108-109-110)	101
Б. ДУГОРОЧНА РЕЗЕРВИСАЊА И ОБАВЕЗЕ (112+113+116)	111
III. КРАТКОРОЧНЕ ОБАВЕЗЕ (117+118+119+120+121+122)	116
3. Обавезе из пословања	119

Узимајући у обзир пет салда рачуна из биланса стања наведених дуванских индустрија на основу претходних година (2013., 2012. и 2011.), посматрамо која од ових позиција највише утиче на сам суд ревизора.

Дуванска индустрија Ниш – Philip Morris

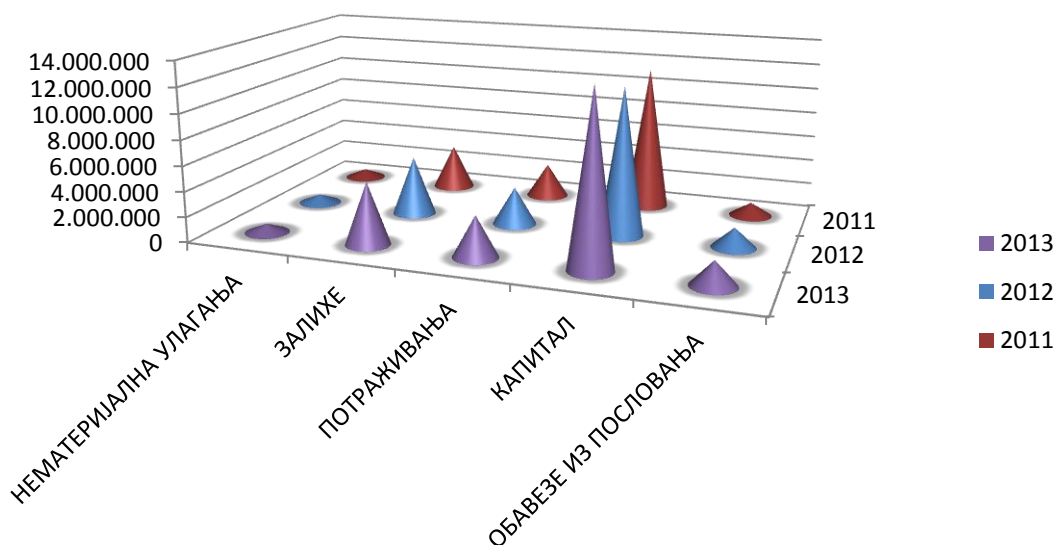
	НЕМАТЕРИЈАЛНА УЛАГАЊА	ЗАЛИХЕ	ПОТРАЖИВАЊА	КАПИТАЛ	ОБАВЕЗЕ ИЗ ПОСЛОВАЊА
2013	689.511	4.942.207	3.251.774	13.572.140	1.900.303
2012	682.472	4.691.889	2.990.171	11.741.260	1.585.287
2011	697.514	3.546.548	2.704.902	11.509.543	1.060.225

Табела 1. Подаци из биланса стања (у 000 динара) Дуванске индустрије Ниш– Philip Morris (извор: www.apr.gov.rs)

Током све три године мишљење ревизора је било позитивно.

Оно што се може уочити посматрајући билансне позиције за три године је да је највећи износ имао капитал, који је током година апроксимативно одржавао исти ниво. Док са друге стране „Нематеријална улагања“ бележе најниже износе.

Дуванска индустрија Ниш-Phillip Morris



Графикон 1. Подаци из биланса стања (у 000 динара) Дуванске индустрије Ниш– Philip Morris (извор: приказ аутора)

Дуванска индустрија Врање – British American Tobacco

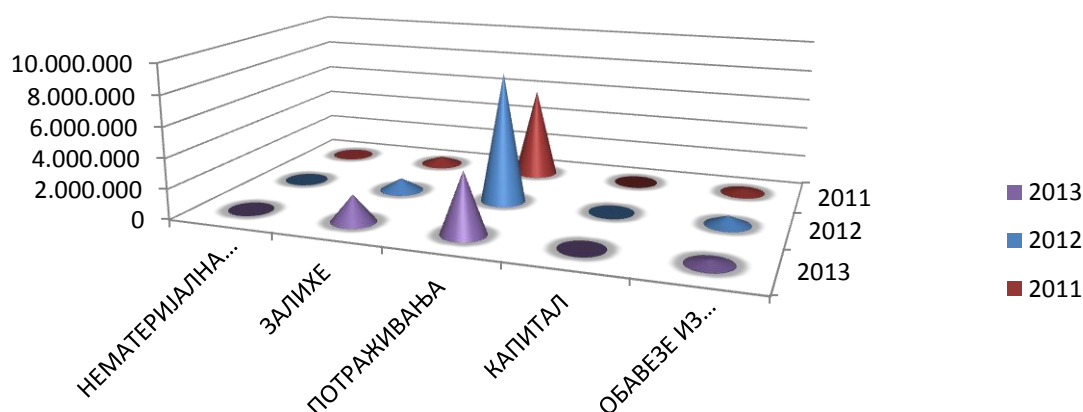
	НЕМАТЕРИЈАЛНА УЛАГАЊА	ЗАЛИХЕ	ПОТРАЖИВАЊА	КАПИТАЛ	ОБАВЕЗЕ ИЗ ПОСЛОВАЊА
2013	0	1.793.333	4.066.425	0	306.898
2012	0	825.242	8.624.124	0	596.446
2011	25.414	512.890	5.999.750	0	279.154

Табела 2. Подаци из биланса стања (у 000 динара) Дуванске индустрије Врање- British American Tobacco (извор: www.apr.gov.rs)

Током све три анализиране године мишљења ревизори су изразили позитивно мишљење са скретањем пажње. Скретање пажње ревизора се односи на остварене Нето губитке друштва, односно изражава се сумња у способност Друштва да послује у складу са начелом сталности пословања.

На *Графикону 2.* може се уочити да капитал не постоји, као и да је најдоминантнија билансна ставка „Потраживања“, за ово Друштво.

Дуванска индустрија Врање - British American Tobacco



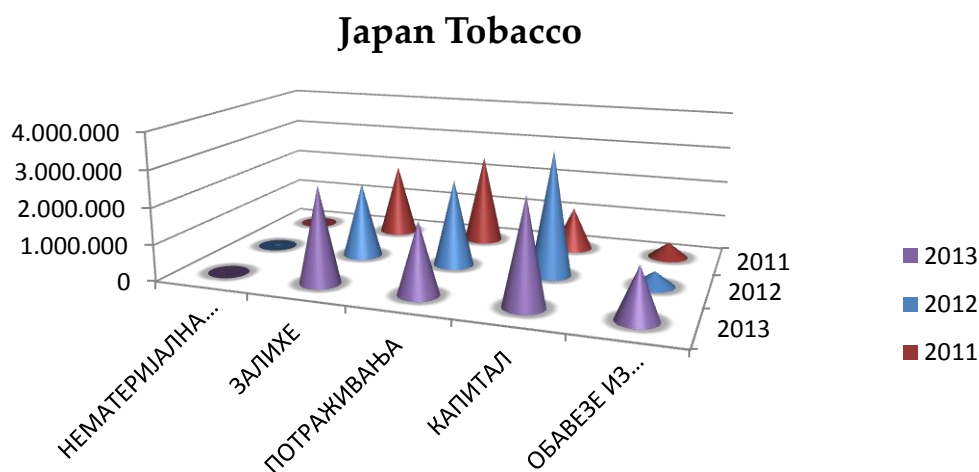
Графикон 2. Подаци из биланса стања (у 000 динара) Дуванске индустрије Врање- British American Tobacco (извор: приказ аутора)

Дуванска индустрија Јапан Тобасо

	НЕМАТЕРИЈАЛНА УЛАГАЊА	ЗАЛИХЕ	ПОТРАЖИВАЊА	КАПИТАЛ	ОБАВЕЗЕ ИЗ ПОСЛОВАЊА
2013	1.174	2.638.659	1.957.554	2.844.475	1.411.948
2012	1.549	2.071.759	2.397.732	3.407.903	407.498
2011	1.806	2.010.706	2.492.189	1.190.490	423.179

Табела 3. Подаци из биланса стања (у 000 динара) Дуванске индустрије Јапан Тобасо
(извор: www.apr.gov.rs)

ЈТ International а.д. Сента је предузеће које сваке године има губитке, који као такви представљају велику одбитну ставку укупног капитала. Због такве ситуације независни ревизор сваке године издаје позитивно мишљење са скретањем пажње. Сама способност даљег пословања овог Привредног друштва зависи од финансијске подршке од стране оснивача ЈТ International Holding B.V., Холандија. Осим тога, ово Привредно друштво обавља и значајан број пословних трансакција са повезаним правним лицима, а притом не исказује у финансијском извештају ефекте трансферних цена на обрачунати порез на добитак. С обзиром да ефекти трансферних цена нису познати до датума састављања финансијског извештаја, ревизори се оградају од процене да ли постоје због тога материјално значајне последице.



Графикон 3. Подаци из биланса стања (у 000 динара) Дуванске индустрије Јапан Тобасо (извор: приказ аутора)

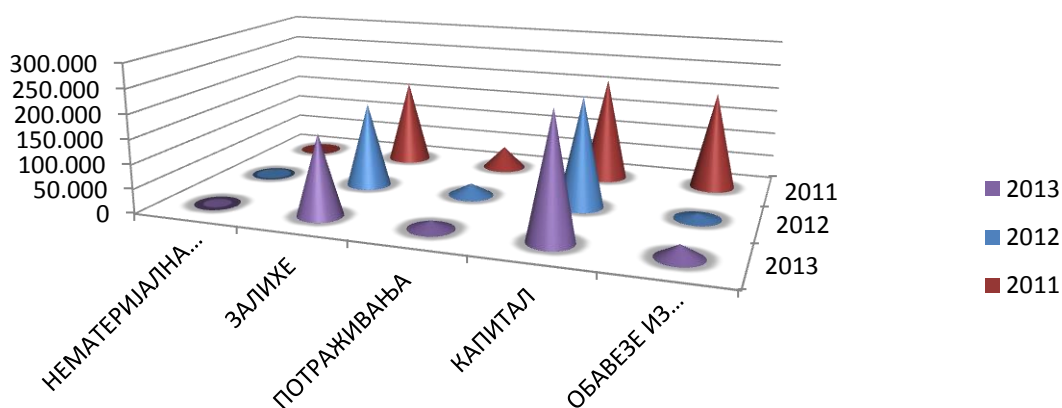
Дуванска индустрија Чока

	НЕМАТЕРИЈАЛНА УЛАГАЊА	ЗАЛИХЕ	ПОТРАЖИВАЊА	КАПИТАЛ	ОБАВЕЗЕ ИЗ ПОСЛОВАЊА
2013	589	165.770	15.946	252.948	25.755
2012	595	175.684	24.100	224.923	11.332
2011	601	175.684	45.766	214.785	201.696

Табела 4. Подаци из биланса стања (у 000 динара) Дуванске индустрије Чока (извор: www.apr.gov.rs)

Оно што се може уочити читајући извештаје ревизора везано за ово привредно друштво, јесте да из године у годину независни ревизор оцењује да финансијски извештаји приказују истинито и објективно финансијски положај а.д. Чока. Против друштва се води више судских спорова чије исходе није могуће предвидети, међутим наведена чињеница не квалификује мишљење ревизора, те је ревизорски извештај за све три посматане године издат као извештај са позитивним мишљењем.

Дуванска индустрија Чока



Графикон 4. Подаци из биланса стања (у 000 динара) Дуванске индустрије Чока (извор: приказ аутора)

4.3 Примена мултиваријантне дискриминационе анализе на примеру у SPSS-у

У овом раду мултиваријантна дискриминациона анализа се примењује у процесу ревизије, тако што се за груписање опсервација користи мишљење ревизора које је на основу историјских података, у случају Дуванске индустрије или „позитивно“ или „позитивно са скретањем пажње“, а варијабле које се користе за раздвајање датих опсервација су позиције преузете на случајан начин из биланса стања тих предузећа. Дискриминациона функција је линеарна комбинација варијабли, у овом случају линеарна комбинација условљена подацима добијеним из биланса стања.

Практична примена ове анализе рађена је на подацима четири наведена предузећа. Анализа је базирана на претходне три године за сва предузећа заједно. Анализирани подаци су наведени у *Табели 5*.

год.	НЕМАТЕРИЈА. УЛАГАЊА	ЗАЛИХЕ	ПОТРАЖИВАЊА	КАПИТАЛ	ОБАВЕЗЕ ИЗ ПОСЛОВА.	мишљење ревизора
2013	689.511	4.942.207	3.251.774	13.572.140	1.900.303	1
	0	1.793.333	4.066.425	0	306.898	5
	1.174	2.638.659	1.957.554	2.844.475	1.411.948	5
	589	165.770	15.946	252.948	25.755	1
2012	682.472	4.691.889	2.990.171	11.741.260	1.585.287	1
	0	825.242	8.624.124	0	596.446	1
	1.549	2.071.759	2.397.732	3.407.903	407.498	5
	595	175.684	24.100	224.923	11.332	1
2011	697.514	3.546.548	2.704.902	11.509.543	1.060.225	1
	25.414	512.890	5.999.750	0	279.154	1
	1.806	2.010.706	2.492.189	1.190.490	423.179	5
	601	175.684	45.766	214.785	201.696	1

Табела 5. Подаци из биланса стања за сва четири предузећа (извор: www.apr.gov.rs)

Мишљење ревизора је замењено са бројевима како би се лакше у програму управљало са подацима. Сваки број означава одређено мишљење:

1	позитивно (без резерве)
2	мишљење са резервом
3	негативно
4	уздржавање од мишљења
5	позитивно мишљење са скретањем пажње

У суштини, на нивоу сва четири предузећа можемо уочити да преовлађује позитивно мишљење. (Слика 7.)



Слика 7. Структура ревизорских мишљења (извор: приказ аутора)

Уношењем Табеле 5. у SPSS-у добијени су следећи параметри.

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
НЕМАТЕРИЈАЛНА УЛАГАЊА	,829	2,066	1	10	,181
ЗАЛИХЕ	,995	,051	1	10	,825
ПОТРАЖИВАЊА	,998	,020	1	10	,890
КАПИТАЛ	,931	,746	1	10	,408
ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	,997	,029	1	10	,868

Табела 6. Тестови једнакости група средњих вредности/ Tests of Equality of Group Means (израчунате у SPSS-у)

Дакле, што је мања вредност теста Wilks' Lambda за одређену променљиву, то значи да је та независна варијабла више значајна за дискриминациону функцију, јер доприноси мањој варијацији унутар групе. Спрам тога можемо да закључимо да су током све три године најзначајнију улогу имала „Нематеријална улагања“, пошто је вредност Wilks' Lambda тада најнижа и износи 0.829. Значајност „Нематеријална улагања“ можемо потврдити и *F* вредношћу, који је мера јединственог доприноса поједине варијабле.

Covariance Matrices					
Correlation	НЕМАТЕРИЈАЛНА УЛАГАЊА	ЗАЛИХЕ	ПОТРАЖИВАЊА	КАПИТАЛ	ОБАВЕЗЕ ИЗ ПОСЛОВА.
НЕМАТЕРИЈАЛНА УЛАГАЊА	1,000	,970	,015	,982	,830
ЗАЛИХЕ	,970	1,000	,098	,977	,926
ПОТРАЖИВАЊА	,015	,098	1,000	-,029	,167
КАПИТАЛ	,982	,977	-,029	1,000	,869
ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	,830	,926	,167	,869	1,000

Табела 7. Приказ корелационе матрице/ Covariance Matrices (израчуната у SPSS-у)

Унутар-група корелациона матрица показује корелације између предиктора, тачније између независних варијабли којима описујемо модел. Понекад у истраживању није довољна информација о корелацији само између две опсервиране варијабле, већ нас интересује и на који то начин више варијабли међусобно утичу једна на другу. После посматрања односа свих парова варијабли утврди њихова међусобна корелација, израђује се матрица корелације. Редовима и колонама матрице представљене су посматране варијабле, а податак који се добија на пресеку (укрштању) одређеног реда и колоне представља коефицијент корелације између варијабли у одговарајућем реду и колони. Дијагонални елементи корелационе матрице су једнаки 1 (пошто је свака варијабла сама са собом у потпуној корелацији). У Табели 7. може се уочити да је матрица симетрична – дакле подаци изнад и испод дијагонале за исти пар варијабли су идентични. [36]

Из корелационе матрице уочава се у којој мери су у корелацији две појединачне варијабле, затим које варијабле у међусобном односу имају највећи или најмањи коефицијент корелације.

Посматрајући *Табелу 7.* видимо да највећи коефицијент корелације између варијабли „Капитал“ и „Нематеријална улагања“ и износи 0,982. Сазнање о корелацији између ове две варијабле помаже нам да с већом сигурношћу предвидимо на који начин ће се мењати вредност друге варијабле уколико се промени прва. Дакле постоји узрочно-последична веза између њих, док на пример за „Капитал“ и „Потраживања“ се то не може рећи, с обзиром да њихова корелација износи –0,029.

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	13,319 ^a	100,0	100,0	,964
a. First 1 canonical discriminant functions were used in the analysis.				

Табела 8. Карактеристични корен/ Eigenvalues (израчунат у SPSS-у)

Што је већи карактеристични корен функције, више је објашњено о варијанси зависне варијабле кроз ту функцију и већа је дискриминациона моћ функције. Оно што се може уочити у *Табели 8.* јесте да карактеристични корен (λ) дискриминационе функције износи 13,319. Проценат објашњене варијансе¹⁷ који је представљен дискриминационом функцијом износи 100 %.

Такође *Табелом 8.* добијамо и податак о каноничкој корелацији (Canonical Correlation). Каноничка корелација је мера повезаности дискриминационе функције и зависне варијабле (у овом случају то је мишљење ревизора), а износи 0,964. У теорији је наведено да се означава са η .

Уврштавајући у формулу под бројем (2.49) можемо потврдити износ каноничке корелације

$$\eta = \sqrt{\frac{\lambda}{1 + \lambda}} = \sqrt{\frac{13,319}{1 + 13,319}} = \sqrt{0,9301} = 0,964. \quad (4.1)$$

¹⁷ Проценат објашњене варијансе је мера која описује колико добро регресиона линија „фитује“ стварне податке. Циљ регресије је да предвиди вредности функције за поједине вредности независних променљивих од којих зависи та функција. Пошто је реч о стохастичким везама не може се тачно предвидети вредност функције, па се зато као могуће решење тражи регресиона права (крива) која ће најмање одступати од емпиријских података.

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	0,07	19,962	5	0,001

Табела 9. Wilks' Lambda (израчунао у SPSS-у)

Пошто видимо у Табели 9. да је $p < 0.05$, значи да коришћен модел одговара унетим подацима, тј. описује их добро. Са df су означени степени слободе, и износе $df = (n - 1)$, при чему је n укупан број варијабли. Уочавамо да Wilks' Lambda тест износи $\Lambda = 0,07$. Дакле, подсетимо се Wilks' Lambda је тест статистика којом се тестира да ли постоји или не постоји веза између независних и зависних варијабли. Када је вредност Wilks' Lambda теста за функцију је мала као што је у овом примеру, онда је функција значајна. Уколико бисмо заменили то у израз који знамо:

$$\Lambda = (1 - \eta^2) = (1 - 0,964^2) = (1 - 0,92) = 0,07, \quad (4.2)$$

можемо уочити да управо тај број се може извести и преко каноничке корелације.

Standardized Canonical Discriminant Function Coefficients	
	Function 1
НЕМАТЕРИЈАЛНА УЛАГАЊА	5,596
ЗАЛИХЕ	-9,168
ПОТРАЖИВАЊА	0,367
КАПИТАЛ	1,006
ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	2,921

Табела 10. Стандардизовани коефицијенти дискриминационе функције/ Standardized Canonical Discriminant Function Coefficients (израчунао у SPSS-у)

Стандардизовани коефицијенти дискриминационе функције имају исту улогу као бета коефицијенти у вишеструкој регресији: они указују на релативну важност независних варијабли у предвиђању зависних. Омогућују да се упореде променљиве, које су мерене различитим скалама (нивоима). Коефицијентима са великим апсолутним вредностима одговарају променљиве са већим дискриминационим способностима. На основу те табеле потврђујемо закључак да у овом моделу значајнију улогу преузимају нематеријална улагања с обзиром да је вредност коефицијента који стоји уз њих 5,596. Међутим, поред независне променљиве „залихе“ стоји коефицијент -9.168 , што значи да ова променљива такође доста утиче на саму дискриминациону функцију али у смислу да повећањем ње смањује се вредност функције.

Користећи одређене вредности за независне променљиве можемо израчунати вредност дискриминационе функције, а да тумачење тог резултата се добија из *Табеле 13.*, везане за групу центроида.

Structure Matrix	
	Function
	1
НЕМАТЕРИЈАЛНА УЛАГАЊА	,125
ЗАЛИХЕ	-,020
ПОТРАЖИВАЊА	,012
КАПИТАЛ	,075
ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	,015

Табела 11. Структура матрице/ Structure Matrix (израчуната у SPSS-у)

Ова *Табела 11.* структуралне матрице, показује корелацију између варијабли и дискриминационих резултата добијених дискриминационом функцијом. Дискриминациона функција је дакле настала као линеарна комбинација независних варијабли. Што је већа апсолутна вредност коефицијената ове матрице, већи је дискриминаторни утицај варијабли на зависну променљиву.

Canonical Discriminant Function Coefficients	
	Function
	1
НЕМАТЕРИЈАЛНА УЛАГАЊА	0,00001887
ЗАЛИХЕ	-0,00000511
ПОТРАЖИВАЊА	0,00000014
КАПИТАЛ	0,00000019
ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	0,00000434
(Constant)	2,64940495

Табела 12. Нестандардизовани коефицијенти дискриминационе функције/ Canonical Discriminant Function Coefficients (израчунати у SPSS-у)

Нестандардизовани коефицијенти дискриминационе функције се користе у једначини за израчунавање предикције зависне варијабле. Уколико се јави негативна вредност, она указује на то да између варијабли постоји негативна повезаност тачније повећањем вредности независне променљиве смањује се вредност зависне променљиве. У *Табели 12.* можемо уочити вредности нестандардизованих коефицијената, па на основу тога, можемо написати једначину за дискриминациону функцију:

$$f = 0,00001887 * x_1 - 0,00000511 * x_2 + 0,00000014 * x_3 + 0,00000019 * x_4 + 0,00000434 * x_5 + 2,64940495. \quad (4.3)$$

У формули (4.3) променљиве x_1 , x_2 , x_3 , x_4 и x_5 представљају независне променљиве редом из Табеле 10.

Functions at Group Centroids	
Мишљење	Function
	1
1	2,356
5	-4,711

Табела 13. Центроиди, група центроида (израчунати у SPSS-у)

Центроиди су средње дискриминационе оцене за сваку групу. Ова табела се користи за успостављање „тачке сечења“ како би се класификовали случајеви. Ако су две групе једнаких величина, најбоља „тачка сечења“ је на пола пута између вредности функција центроида група. Ако су групе неједнаке, оптимална „тачка сечења“ (или „Cutting score“) представља пондерисани просек две вредности. Тачније на овом примеру:

$$Cutting\ score = \frac{2,356 - 4,711}{2} = -1,178. \quad (4.4)$$

Дакле, пошто имамо две ситуације када је мишљење ревизора позитивно или је дато са скретањем пажње, уколико је добијена вредност дискриминационе функције, када се замене одређене вредности независних променљивих, испод Cutting score тј. испод $-1,178$, онда ће бити мишљење са скретањем пажње, а уколико је изнад онда ће бити сигурно позитивно.

Prior Probabilities for Groups			
МИШЉЕЊЕ	Примарне	Случајеви коришћени у анализама	
		Непондерисан	Пондерисан
1	,500	8	8,000
5	,500	4	4,000
Total	1,000	12	12,000

Табела 14. Примарне вероватноће за групе/ Prior Probabilities for Groups (израчунате у SPSS-у)

Примарне вероватноће се користе у класификацији. Одређивање вероватноће припадности групама, дефинисаним на основу зависних варијабли, неопходно је ако постоје доста различите величине група. У овом случају видимо да због постојања две групе, примарна вероватноћа износи 50%.

Classification Function Coefficients		
	МИШЉЕЊЕ	
	1	5
НЕМАТЕРИЈАЛНА УЛАГАЊА	-0,000007033	0,0000
ЗАЛИХЕ	0,000001855	0,00003797
ПОТРАЖИВАЊА	0,0000003693	-0,0000006129
КАПИТАЛ	0,00000007917	-0,000001251
ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	-0,000001232	-0,00003191
константе	-1,811	-28,858
Фишерава линеарна дискриминациона функција		

Табела 15. Класификација коефицијената функције/ Classification Function Coefficients (израчуната у SPSS-у)

Два скупа (један за сваку зависну групу) нестандардизованих линеарних дискриминационих коефицијената су израчунати, и могу се користити за класификовање случајева.

Group Statistics					
		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
1	НЕМАТЕРИЈАЛНА УЛАГАЊА	262087,00	354330,233	8	8,000
	ЗАЛИХЕ	1879489,25	2131085,785	8	8,000
	ПОТРАЖИВАЊА	2957066,63	3099611,769	8	8,000
	КАПИТАЛ	4689449,88	6310568,632	8	8,000
	ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	707524,75	728872,028	8	8,000
5	НЕМАТЕРИЈАЛНА УЛАГАЊА	1132,25	798,196	4	4,000
	ЗАЛИХЕ	2128614,25	360413,669	4	4,000
	ПОТРАЖИВАЊА	2728475,00	921891,425	4	4,000
	КАПИТАЛ	1860717,00	1557030,968	4	4,000
	ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	637380,75	518941,789	4	4,000
Total	НЕМАТЕРИЈАЛНА УЛАГАЊА	175102,08	310489,983	12	12,000
	ЗАЛИХЕ	1962530,92	1714798,754	12	12,000
	ПОТРАЖИВАЊА	2880869,42	2521582,252	12	12,000
	КАПИТАЛ	3746538,92	5286121,773	12	12,000
	ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	684143,42	642424,616	12	12,000

Табела 16. Group Statistics (табела израчуната у SPSS-у)

У Табели 16. просечна вредност нематеријалних улагања везаних за групу са позитивним мишљењем ревизора је 262087,00 што је веће од просечног износа „Нематеријалних улагања“ за групу са мишљењем са скретањем пажње за чак 260954,75. Оно што можемо закључити спрам овога јесте да (са одређеном вероватноћом) уколико предузеће има тако ниску вредност „Нематеријалног улагања“, да ће ревизор дати мишљење са скретањем пажње. Но наравно то не мора да важи стриктно и у сваком случају (нити у сваком моделу). Битно је нагласити да одређени услови и одређене корелације утичу такође на формирање група. Посматрајући даље Табелу 16. уочавамо да су све ставке у склопу групе 1 (групе са позитивним мишљењем ревизора) већих вредности од оних у оквиру групе 5 (групе са мишљењем са скретањем пажње).

Casewise Statistics								
	Case Number	Actual Group	Highest Group				Discriminant Scores	
			Predicted Group	P(D>d G=g)		P(G=g D=d)	Squared Mahalanobis Distance to Centroid	Function 1
				p	df			
Original	1	1	1	,487	1	1,000	,482	1,661
	2	5	5	,926	1	1,000	,009	-4,618
	3	5	5	,404	1	1,000	,696	-3,877
	4	1	1	,703	1	1,000	,145	1,975
	5	1	1	,195	1	1,000	1,681	1,059
	6	1	1	,892	1	1,000	,019	2,220
	7	5	5	,650	1	1,000	,206	-5,166
	8	1	1	,618	1	1,000	,248	1,858
	9	1	1	,013	1	1,000	6,142	4,834
	10	1	1	,843	1	1,000	,039	2,554
	11	5	5	,636	1	1,000	,224	-5,185
	12	1	1	,742	1	1,000	,109	2,685

Табела 17. Casewise Statistics (израчунато у SPSS-у)

Удаљеност од центроида групе мери се Махаланобисовом раздаљином (Mahalanobis Distance to Centroid). Ова Табела 17. корисна је за идентификацију аутлајера. Сам појам аутлајера често се везује за одређену грешку, девијацију али аутлајер може да носи и изузетно значајну информацију за посматрани модел, а у том случају назива се „значајна тачка“. Зато је откривање аутлајера један од битних корака у нпр. Data mining анализи. Само проналажење аутлајера се управо заснива на мерама удаљености, у овом конкретном случају на Махаланобисовој удаљености.

Да би се идентификовали аутлајер посматрамо колону са Махаланобисовом удаљеношћу и тражимо случај у ком је она већа од критичне вредности.

Критична вредност за Махаланобисову удаљеност d^2 је она вредност која може да постигне одређени ниво статистичког значаја, с обзиром на број променљивих које су укључене у обрачун. Конкретно, овде се користи уграђена функција $IDF.CHISQ(p, df)$, како би се израчунала критична вредност са степенима слободе који су једнаки броју независних варијабли. У нашем примеру она износи 15,09. Анализирајући Табелу 17. уочавамо да ниједна Махаланобисова удаљеност не прелази критичну вредност, дакле аутлајер не постоји.

Classification Results ^a					
		МИШЉЕЊЕ	Predicted Group Membership		Total
			1	5	
Оригинал	Count	1	8	0	8
		5	0	4	4
	%	1	100,0	,0	100,0
		5	,0	100,0	100,0
a. 100,0% оригиналних груписаних случајева исправно су класификовани.					

Табела 18. Резултати класификације(израчунати у SPSS-у)

Табела 18. се користи за процену колико добро дискриминациона функција класификује. Дакле даје информације о стварном припадању одређеној групи наспрам предвиђеног чланства. Овде се правилно разврстано 100% случајева (hit ratio¹⁸). Табела показује колико прецизно су објекти класификовани у тим групама. постоје два атрибута који је одликују: осетљивост и специфичност. Осетљивост: веома осетљив тест значи да постоји неколико лажних негативних резултата. Специфичност: врло специфичан тест значи да постоји неколико лажно позитивних резултата.

¹⁸ Hit ratio је проценат успешности класификовања.

4.4 Примена мултиваријантне дискриминационе анализе у програмском алату STATISTICA

STATISTICA је статистички и аналитички софтверски пакет, развијен од стране StatSoft. Овај програмски пакет обухвата анализу података, управљање подацима, статистичке тестове, проналажење података и процедуре за визуелизацију података.

Слично као и у SPSS-у, овим програмом добијамо резултате приказане у виду табела, слика, графика.

Конкретно, користе се подаци наведени у Табели 5. Њиховим уношењем и одабиром метода дискриминационе анализе, добијамо прво Табелу 19. у којој су приказани резултати по позицијама у билансу стања. Црвеним програм означи статистички значајне варијабле, тачније оне које највише доприносе дискриминацији група. Може се уочити да за „Нематеријална улагања“, „Залихе“ и „Обавезе из пословања“ Partial Lambda¹⁹ су најнижих вредности спрам осталих, а да *p*-вредности су мање од 0,05 што значи да стварно јесу статистички значајне (одбацује се нулта хипотезу што значи да постоји статистичка значајност раздвајања група и да очекиване средње вредности свих независних варијабли нису једнаке). *F* вредност одређене варијабле показује њену статистичку значајност у дискриминацији између група, и то је мера јединственог доприноса варијабле у поступку предикције припадности групи неког објекта, где се може уочити да је највећа управо за „Залихе“, „Нематеријална улагања“ па потом за „Обавезе из пословања“.

		Discriminant Function Analysis Summary				
		No. of vars in model: 5; Grouping: Misljenje (2 grps)				
		Wilks' Lambda: ,06984 approx. F (5,6)=15,982 p< ,0021				
N=12		Wilks' Lambda	Partial Lambda	F-remove (1,6)	p-value	Toler. 1-Toler. (R-Sqr.)
	Nematerijalna ulaganja	0,153882	0,453849	7,22027	0,036194	0,018753 0,981247
	Zalihe	0,479963	0,145509	35,23446	0,001020	0,010930 0,989070
	Potrazivanja	0,075966	0,919343	0,52640	0,495439	0,643150 0,356850
	Kapital	0,071074	0,982631	0,10606	0,755736	0,018440 0,981560
	Obaveze iz poslovanja	0,146855	0,475566	6,61654	0,042206	0,066100 0,933900

Табела 19. Резиме дискриминационе анализе (израчунат у програмском алату STATISTICA)

Tolerance представља коефицијент детерминације између зависна варијабла са осталим варијаблама које су укљученим у модел. То је дакле мера редунданције, тј. "сувишности" одређене варијабле. Ако је Tolerance неке од променљивих које су ушле у једначину, којом је описан модел, мања од вредности подразумеване толеранције²⁰ која износи 0,01 ,

¹⁹ Partial Lambda је мера појединачног доприноса одређене варијабле дискриминационој моћи модела. Овај параметра функционише на следећи начин : „Што је мања вредност коефицијента то је већи појединачни допринос варијабле у дискриминационој функцији.“

²⁰ Минимална вредност Tolerance која може бити наведена у програму је 1.00E – 25 (тј број са 24 нула иза децималног зареза). Међутим то је изузетно ниска вредност и не препоручује се, већ се користи број 0,01.

то значи да та варијабла је 99% сувишна у једначини. У овом примеру то није случај, с обзиром да све имају Tolerance већи од 0,01.

Group	Classification Matrix Rows: Observed classifications Columns: Predicted classifications		
	Percent Correct	G_1:1 p=,66667	G_2:5 p=,33333
G_1:1	100,0000	8	0
G_2:5	100,0000	0	4
Total	100,0000	8	4

Табела 20. Класификациона матрица (израчуната у програмском алату STATISTICA)

Класификациона матрица садржи информације о броју и постотку правилно класификованих случајева у свакој групи. Ова матрица сортира све случајеве из модела у категорије, спрам тога да ли је предвиђена вредност одговара стварној вредност. Сви случајеви у свакој категорији се онда преброје, а износи су приказани у матрици.

Редови у матрици (Табела 20.) представљају предвиђене вредности за модел, док колоне представљају стварну вредност.

Класификација матрица је важан алат за процену резултата предвиђања, јер олакшава да се схвате и преброје ефекти погрешних предвиђања. Прегледом износа и процената у свакој ћелији ове матрице, може се аутоматски видети колико често је модел предвиђао тачно. Ту видимо да је 100% случајева добро разврстано, и да се поклапају информације о стварном припадању одређеној групи и информације о њиховом предвиђеном чланству.

Case	Squared Mahalanobis Distances from Group Centroids Incorrect classifications are marked with *		
	Observed Classif.	G_1:1 p=,66667	G_2:5 p=,33333
1	G_1:1	4,29595	44,42499
2	G_2:5	52,09452	3,46909
3	G_2:5	44,87213	6,72167
4	G_1:1	2,12224	46,68526
5	G_1:1	5,24856	36,87013
6	G_1:1	5,70828	53,72985
7	G_2:5	62,69846	6,33225
8	G_1:1	2,22645	45,13151
9	G_1:1	8,36075	93,33558
10	G_1:1	2,13320	54,87442
11	G_2:5	57,65649	1,01939
12	G_1:1	2,36217	56,96424

Табела 21. Квадрирано Махалобисово растојање (спрам центроида групе) (израчунато у програмском алату STATISTICA)

Предвиђање припадности групи се врши и уз помоћ Махаланобисове удаљености. Ова удаљеност је доста слична квадрату Еуклидске удаљености посматраног случаја (објекта) од центроида за сваку групу. Класификација се врши тако да се објекат додељује групи спрам Махаланобисове удаљености (објекта од центроида). У овом примеру постоје две групе G_1:1 и G_2:5 где прва представља случај када ће мишљење ревизора бити позитивно, а друга случаја када ће бити мишљење са скретањем пажње.

Variable	Pooled Within-Groups Correlations				
	Nematerijalna ulaganja	Zalihe	Potrazivanja	Kapital	Obaveze iz poslovanja
Nematerijalna ulaganja	1,00	0,97	0,02	0,98	0,83
Zalihe	0,97	1,00	0,10	0,98	0,93
Potrazivanja	0,02	0,10	1,00	-0,03	0,17
Kapital	0,98	0,98	-0,03	1,00	0,87
Obaveze iz poslovanja	0,83	0,93	0,17	0,87	1,00

Табела 22. The Pooled Within-group Correlations-Обједињене унутар групе корелације (израчунате у програмском алату STATISTICA)

The Pooled Within-group Correlation matrix је дакле матрица којом се представља корелација између свих варијабли.(видети у Табели 22.) Оно што се може уочити јесте да подаци добијени у програмском алату STATISTICA (Табела 22.) и у SPSS-у (Табела 7.) су идентични.

Misljenje	Means					
	Nematerijalna ulaganja	Zalihe	Potrazivanja	Kapital	Obaveze iz poslovanja	Valid N
G 1:1	262087,0	1879489	2957067	4689450	707524,8	8
G 2:5	1132,3	2128614	2728475	1860717	637380,8	4
All Grps	175102,1	1962531	2880870	3746539	684143,4	12

Табела 23. Средње вредности података из биланса стања (груписани спрам мишљења ревизора) (израчунате у програмском алату STATISTICA)

У Табели 23. дата је средња вредност варијабли у свакој групи посебно. То се заправо добија пресеком реда и колоне. Подаци добијени овом табелом се такође покапају са већ добијеним подацима у SPSS-у.

У последњем, маргиналном реду табеле ставка под називом All Groups (All Grps) маргиналном реду, представља просечне вредности варијабли (укупно, без припадности групи). Док су у Табели 24. наведене вредности одсупања од очекиваних вредности варијабли. Ту се могу уочити да постоје знатне разлике у вредностима између прве и друге групе.

	Standard Deviations (provucikrozrad)					
Misljenje	Nematerijalna ulaganja	Zalihe	Potrazivanja	Kapital	Obaveze iz poslovanja	Valid N
G_1:1	354330,2	2131086	3099612	6310569	728872,0	8
G_2:5	798,2	360414	921891	1557031	518941,8	4
All Grps	310490,0	1714799	2521582	5286122	642424,6	12

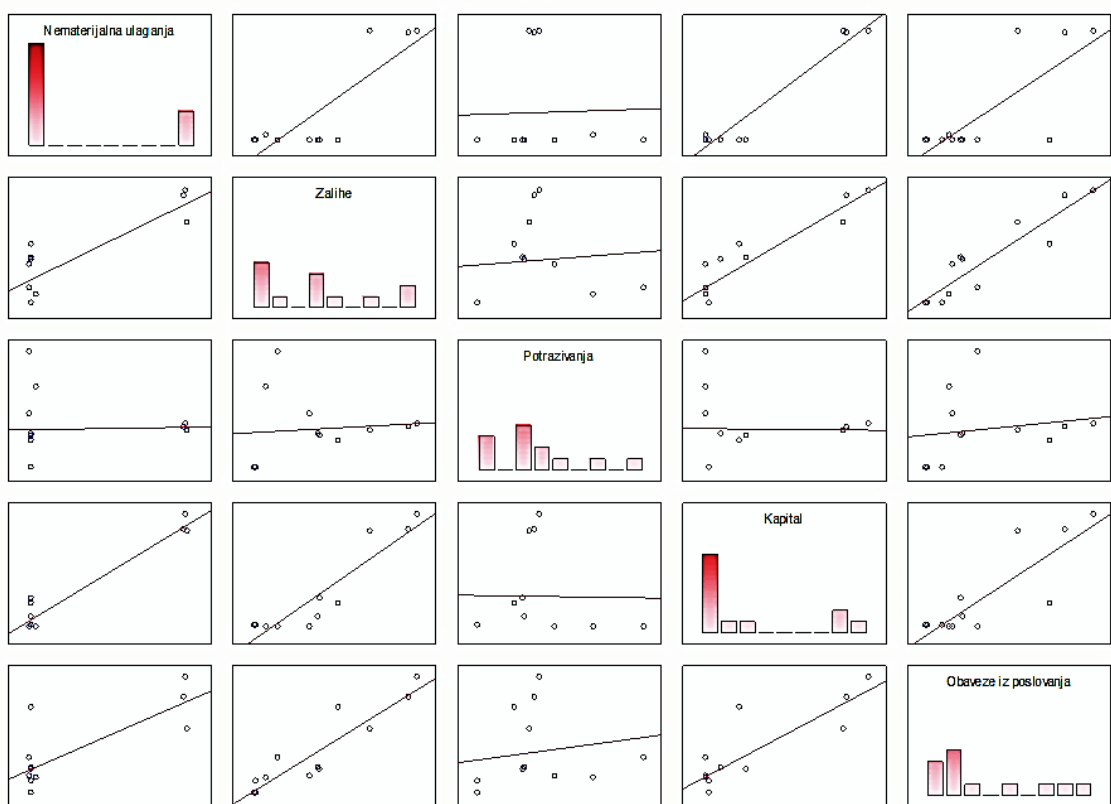
Табела 24. Стандардна девијација података из биланса стања (груписани спрам мишљења ревизора) (израчуната у програмском алату STATISTICA)

Variable	Correlations (Total)				
	Nematerijalna ulaganja	Zalihe	Potrazivanja	Kapital	Obaveze iz poslovanja
Nematerijalna ulaganja	1,00	0,85	0,03	0,97	0,78
Zalihe	0,85	1,00	0,09	0,92	0,92
Potrazivanja	0,03	0,09	1,00	-0,02	0,17
Kapital	0,97	0,92	-0,02	1,00	0,85
Obaveze iz poslovanja	0,78	0,92	0,17	0,85	1,00

Табела 25. Корелације (израчунато у програмском алату STATISTICA)

Табелом 25. дата је корелациона матрица, којом се приказује у којој мери су у корелацији две појединачне варијабле, дакле које варијабле у међусобном односу имају висок, а које пак низак коефицијент корелације. Наравно већ је напоменуто раније да је у питању симетрична матрица, са јединицама на дијагонали, чиме се управо потврђује да је свака варијабла у релацији са самом собом.

На Слици 8. приказане су управо и графички корелације између независних промениљивих (са аутлајерима који су се у тим корелацијама јавили).

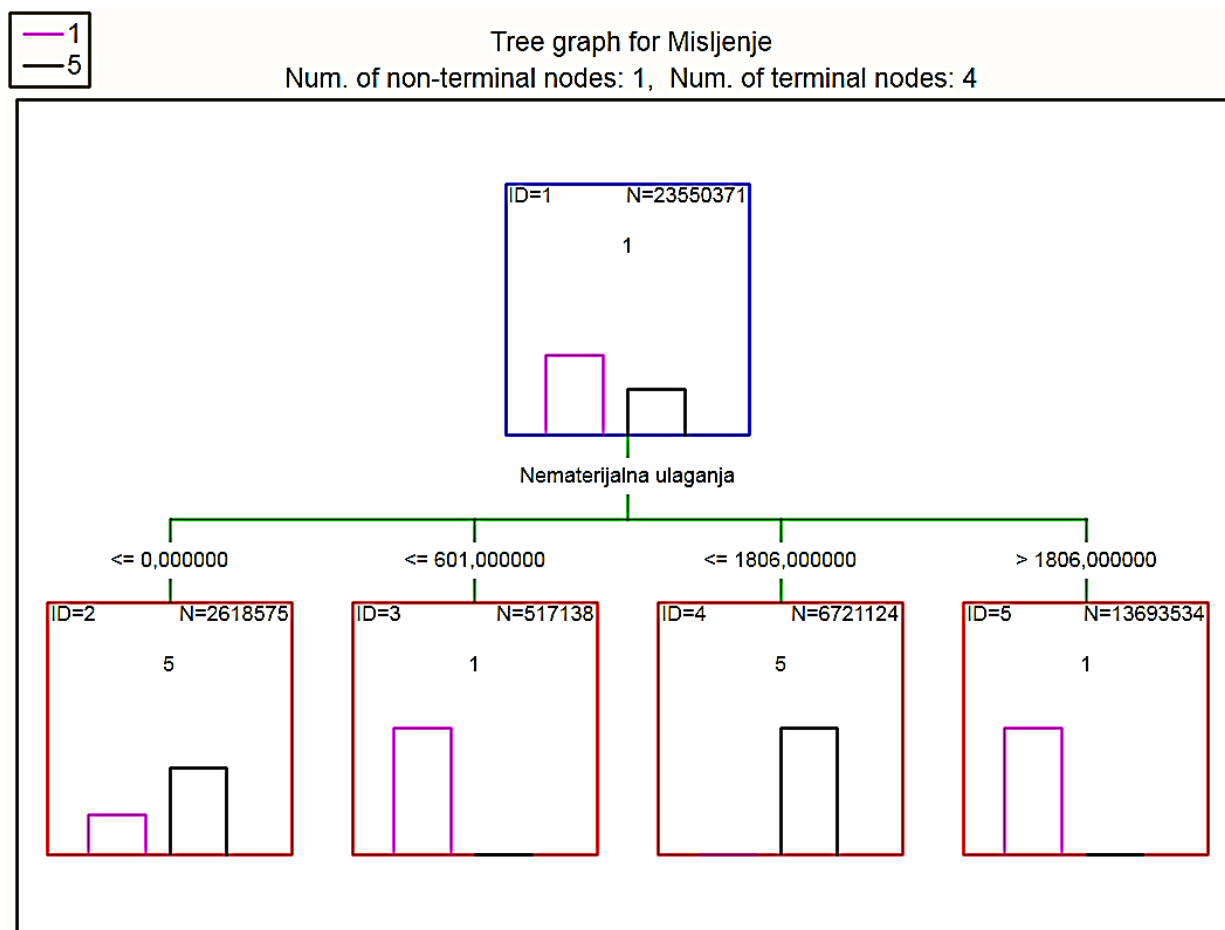


Слика 8. Корелација свих варијабли (Нацрт добијен у програмском алату STATISTICA)

4.5 Примена стабла одлучивања

Дељењем скупова на подскупове добијамо рекурзиван процес, којим се долази до једног примера уз помоћ којег се врши класификација ентитета. Чиме је описан и процес настанка стабла одлучивања. Теоријски део којим је детаљније описан овај процес погледати у делу [2.10].

У случају вредносне анализе пет салда биланса стања, добијамо статистички значајне варијабле спрам којих поставља питање за даљу категоризацију података. Управо користећи се том информацијом и статистичким програмом, добијено је стабло одлучивања представљено на *Слици 9*. Може се уочити да постоји дакле један главни чвор и четири подчвора. У зависности од износа „Нематеријалних улагања“, програм је у правоугаоницима писао „1“ ако је мишљење позитивно и „5“ уколико је са скретањем пажње. Па видимо са *Слике 9*., уколико је „Нематеријално улагање“ (изражено у 000 динара) једнако нули или улази у интервал од 601 до 1806 онда ће мишљење бити са скретањем пажње, у осталим случајевима ће бити позитивно.

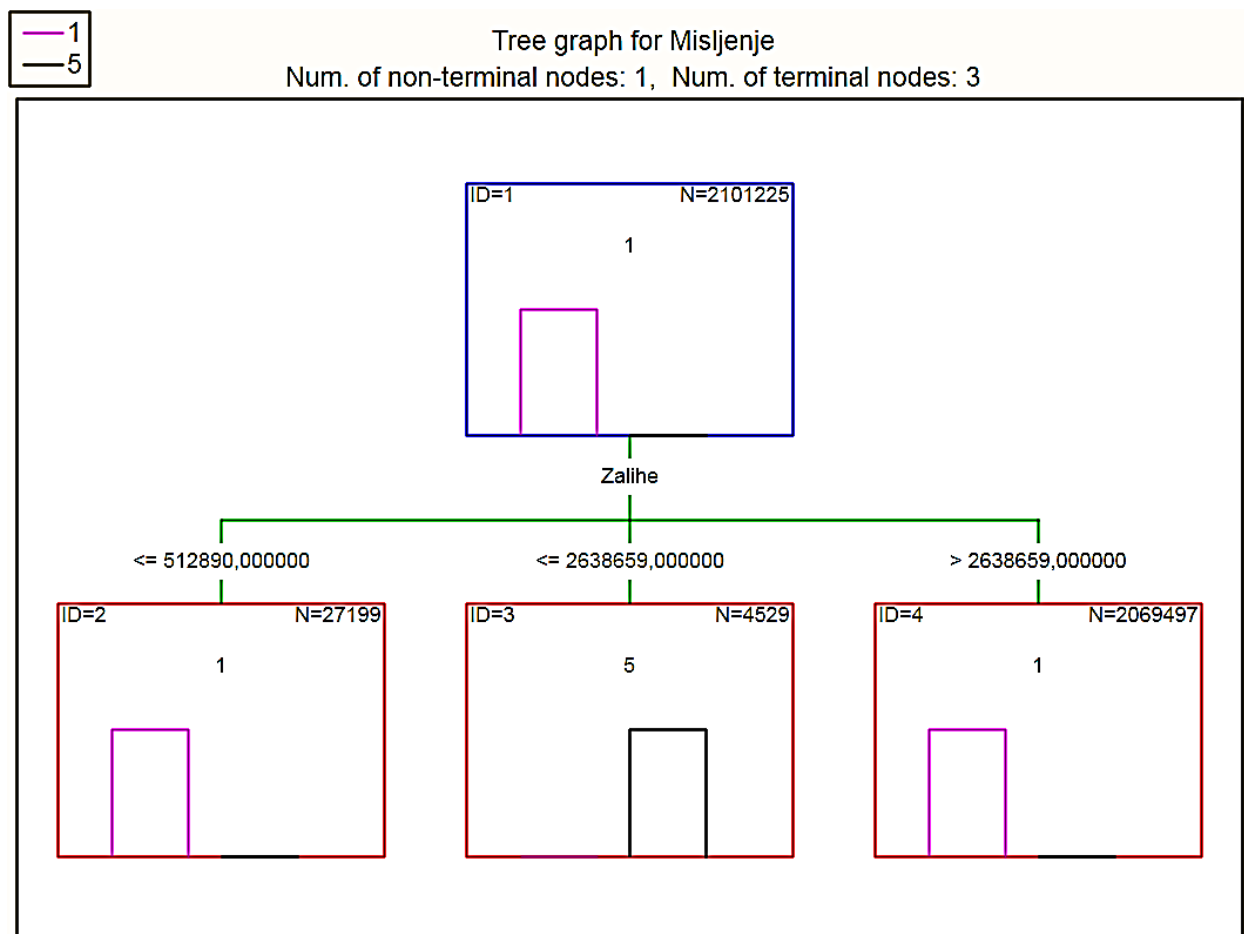


Слика 9. Стабло одлучивања на основу Нематеријалних улагања
(Нацрт добијен у програмском алату STATISTICA)

Оно што се такође може уочити јесте да у првом чвору, унутар датог чвора јесу нацртаним правоугаоницима представљене процентно групе зависних променљивих. Са бојама су одвојене група везане за различита мишљења, а висином правоугаоника уочавамо колико пута се јавља одређена група. Дакле, и у сваком наредном чвору („чвору одговора“) је управо са тим правоугаоницима презентовано којој групи ће припасти посматран објектат спрам посматране независне променљиве.

Стабло одлучивања обезбеђује геометријски оквир за организовање одлука. Свако стабло одлуке нам пружа одређену алтернативу да се одлучимо за наредни корак. Осим „Нематеријалних улагања“, стаблом одлучивања презентујемо и ставку „Залихе“ и „Обавезе из пословања“.

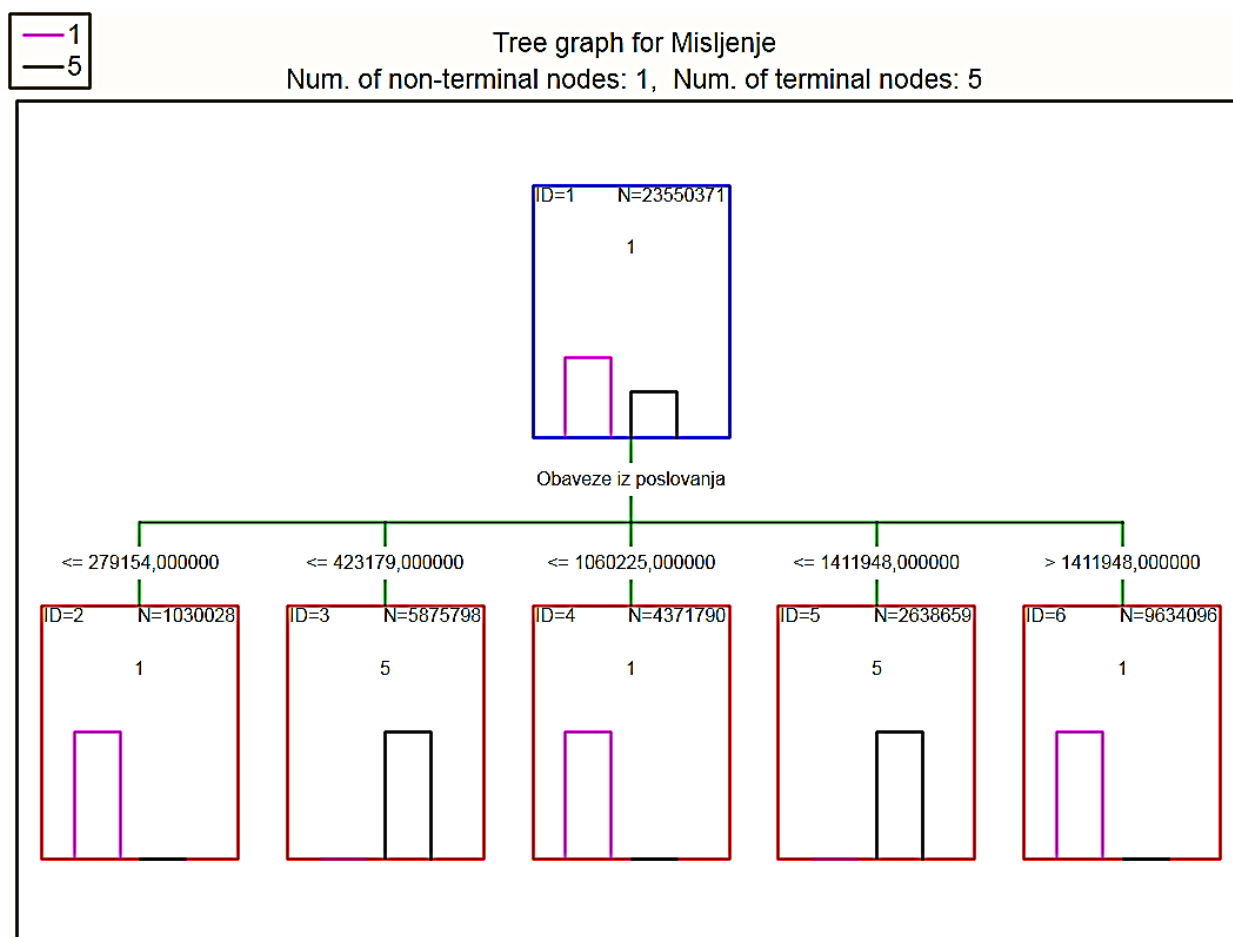
На *Слици 10.* све се креће наспрам „Залиха“ где овог пута вредности које су мање од 2638659 динара а веће од 512890 (изражено у 000 динара) смештају се у групу мишљења са скретањем пажње.



Слика 10. Стабло одлучивања на основу Залиха (Нацрт добијен у програмском алату STATISTICA)

Интерпретација стабла одлучивања (Слика 11.) нам указује да постоји укупно шест чворова, тачније један коренски чвор и осталих пет чворова одговора, на које се он дели.

У овом случају, посматрајући гранање спрема „Обавеза из пословања“ увиђамо да уколико вредности су веће од 279154 а мање од 423179 (изражено у 000 динара) и уколико су веће од 1060225 а мање од 1411948 (изражено у 000 динара) онда ће бити исказано позитивно мишљење са скретањем пажње (у осталим случајевима, тј. уколико се нађе у неком од других интервала биће дато позитивно мишљење).



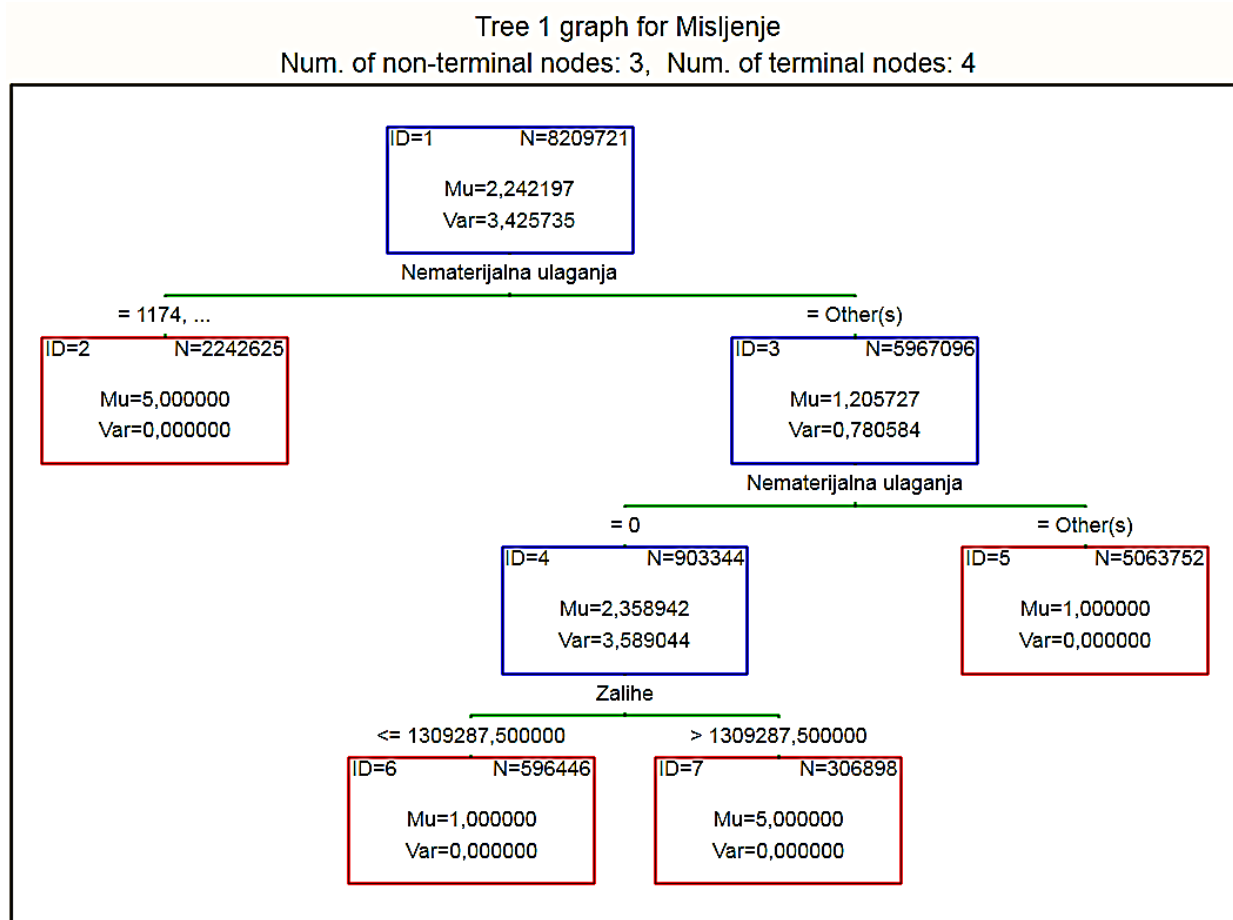
Слика 11. Стабло одлучивања на основу Обавеза из пословања
(Нацрт добијен у програмском алату STATISTICA)

За претходна три стабла одлучивања чији су нацрти добијени у програмском алату STATISTICA, коришћен је метод CHAID (Chi square Automatic Interaction Detector). Компоненте које се користе у овом методу могу бити континуалне, редне и номиналне. CHAID стабло је стабло одлучивања које је изграђено дељењем подскупа простора на два или више чвора одговора у више наврата, почевши са целим скупом података. Но наравно то није једини метод којим се могу презентовати резултати. Поред CHAID метода постоји и C&RT (Classification and Regression Tree) метод.

Сврха ове анализе путем стабла-изградње алгоритма јесте да утврди скуп *ако-онда* логичких услова који дозвољавају прецизне процене за класификацију посматраног објекта (исказа). Ствара интерактивна регресиона стабла за континуалне и категоријалне предикторе; гради оптималну структуру стабла како би се предвиделе континуалне зависне варијабле.

У већини случајева, интерпретација резултата сажета у стаблу је врло једноставна. Ова једноставност је корисна не само за потребе брзог класификовања нових запажања (није тешко проценити један или два логичка услова), већ и за одређивање много једноставнијег модела којим се објашњава зашто су запажања су класификована или предвиђена на одређен начин.

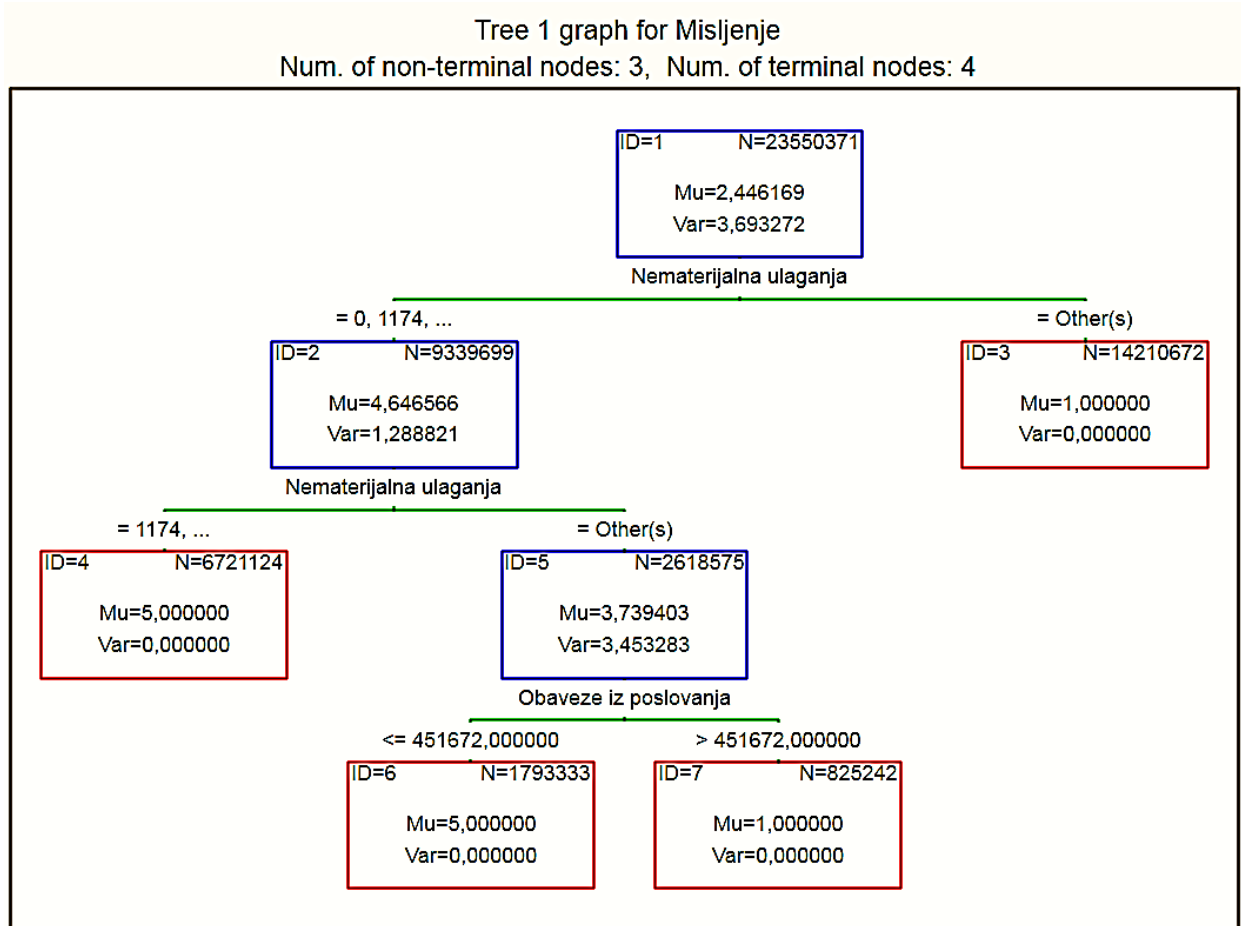
Примена C&RT (Classification and Regression Tree) метода види се на *Слици 12*. Сада се добија стабло са гранама, где свака грана јесте управо одређена ако-онда логичким условима. У саму анализу узети су за опсервацију статистички значајне независне варијабле. Прва два гранаћа условљена су била износима „Нематеријалног улагања“, а последња грана је зависила од износа „Залиха“.



**Слика 12. Стабло одлучивања - C&RT метод
(Нацрт добијен у програмском алату STATISTICA)**

Са *Слике 12.* видимо да уколико „Нематеријална улагања“ имају износ 1174,... (изражено у 000 динара) онда ће бити изражено мишљење са скретањем пажње а другом случају, тачније спрам друге гране може се доћи или до позитивног или до мишљења са скретањем пажње али у зависности од износа „Залиха“. Уколико је вредност „Залиха“ мања или једнака са 1309287,5 (изражено у 000 динара) онда ће бити дато позитивно мишљење, а уколико је веће од 1309287,5 (изражено у 000 динара) онда ће бити дато мишљења са скретањем пажње.

А уколико уместо износа „Залиха“ узмемо за условни параметар „Обавезе из пословања“, као једну од статистички значајних променљивих, добијамо стабло приказано *Сликом 13.* Ту можемо уочити да уколико су износи „Обавеза из пословања“ мањи или једнаки са 451672 (изражено у 000 динара) онда ће бити изражено мишљење са скретањем пажње у осталим случајевима биће дато позитивно мишљење.

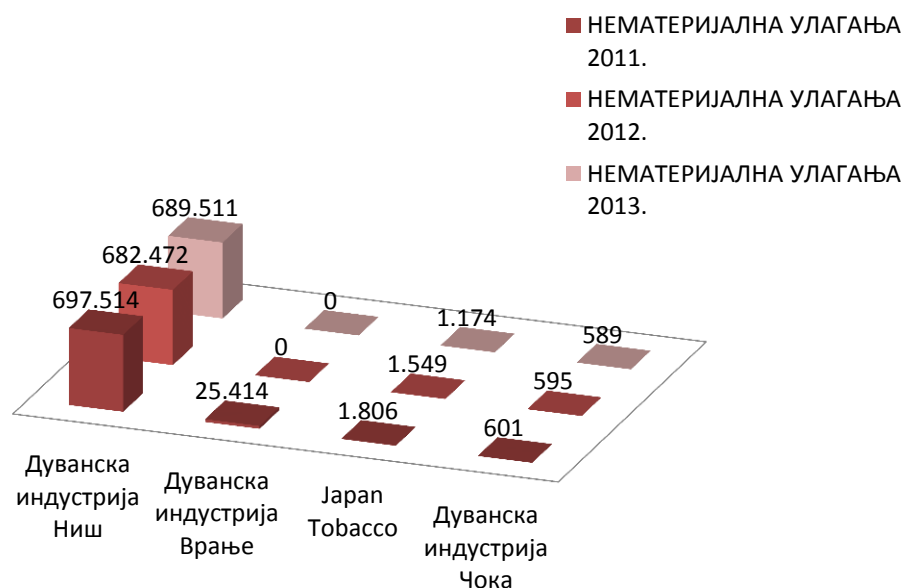


Слика 13. Стабло одлучивања - C&RT метод, други услов (Нацрт добијен у програмском алату STATISTICA)

5 Резултати примене

Дакле, употребом мултиваријационе дискриминационе анализе, тачније коришењем ње уз помоћ неког статистичког програма (нпр. SPSS, STATISTICA,...) може се предвидети какво ће бити мишљење ревизора везано за пословање одређеног предузећа.

Компјутерском обрадом података закључили смо да су статистички најзначајније позиције у билансу стања били „Нематеријална улагања“, „Залихе“ и „Обавезе из пословања“.



**Графикон 5. „Нематеријална улагања“ током година
(за сва четири дуванска предузећа), (извор: приказ аутора)**

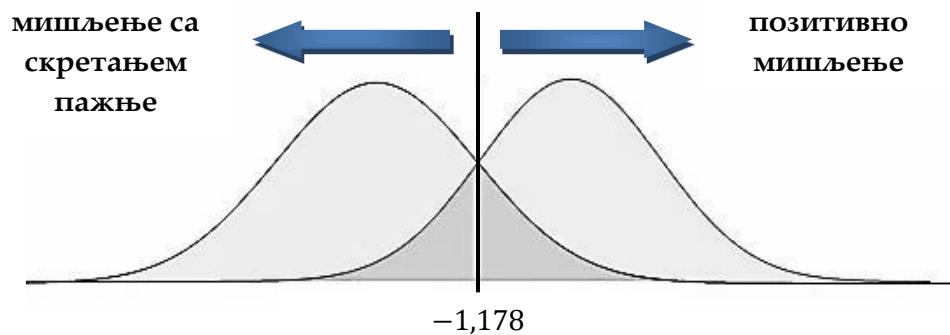
Посматрајући за три године (Графикон 5.) може се уочити да већих разлика у износу „Нематеријалних улагања“ унутар самих предузећа нема, осим за „Дуванску индустрију Врање“ где је 2011. године „Нематеријално улагање“ износило 25.414.000 динара, а у наредним годинама је било 0 динара.

Међутим, уколико посматрамо пласман дуванске индустрије на територији Србије, можемо уочити да „Дуванска индустрија Ниш“ има знатно већа улагања од осталих предузећа.

Даљом обрадом, добили смо и коефицијенте дискриминационе функције којима смо дошли до додатне информације о значајности независних варијабли. С обзиром да дискриминациона анализа се труди да груписањем података предвиди будућа дешавања, томе у прилог иду променљиве са што већим дискриминационим способностима, уз које стоје коефицијенти са великим вредностима.

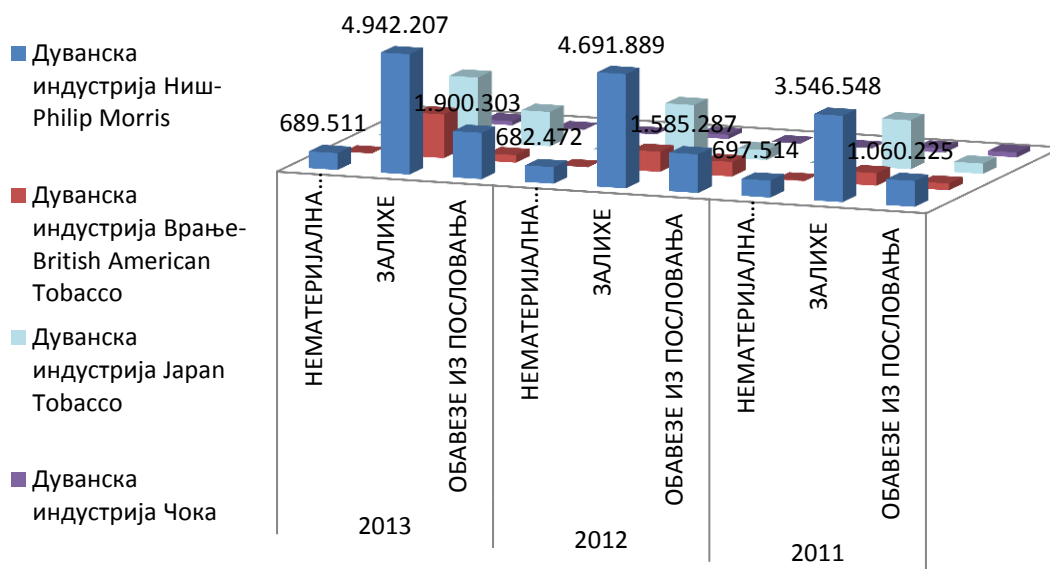
Центроиди су средње дискриминационе оценое за сваку групу, којима смо дефинисали „тачке сечења“ за класификацију случајева. Тиме заправо долазимо до сржи дискриминационе анализе. Дакле, пошто у овом обрађеном примеру имамо две ситуације када је мишљење ревизора позитивно или је дато са скретањем пажње, ми можемо предвидети на основу вредности „тачке сечења“ каквог ће карактера бити мишљење ревизора.

Уколико је добијена вредност дискриминационе функције, заменом одређених варијабли са вредностима из биланса стања, испод „Cutting score“ тј. испод $-1,178$, онда ће бити мишљење са скретањем пажње а уколико је изнад те вредности онда ће бити сигурно позитивно.



Слика 14. Подела на групе добијеном „тачком сечења“
(извор: приказ аутора)

С обзиром да коришењем програмског алата STATISTICA добили смо резултате којима је назначено да су осим позиције „Нематеријална улагања“ статистички значајне и позиције „Залихе“ и „Обавезе из пословања“ приказаћемо то на Графикону 6.



Графикон 6. Нематеријална улагања, залихе и обавезе из пословања током година
(за сва четири дуванска предузећа), (извор: приказ аутора)

Са *Графикана 6.* уочава се да ставка „Залихе“ код Дуванске индустрије „Ниш“, има знатно веће вредности спрам осталих предузећа, и да су ти износи расли са годинама од 2011. године где је износ био 3.546.548 па све до 2013. године где је износ био 4.942.207.

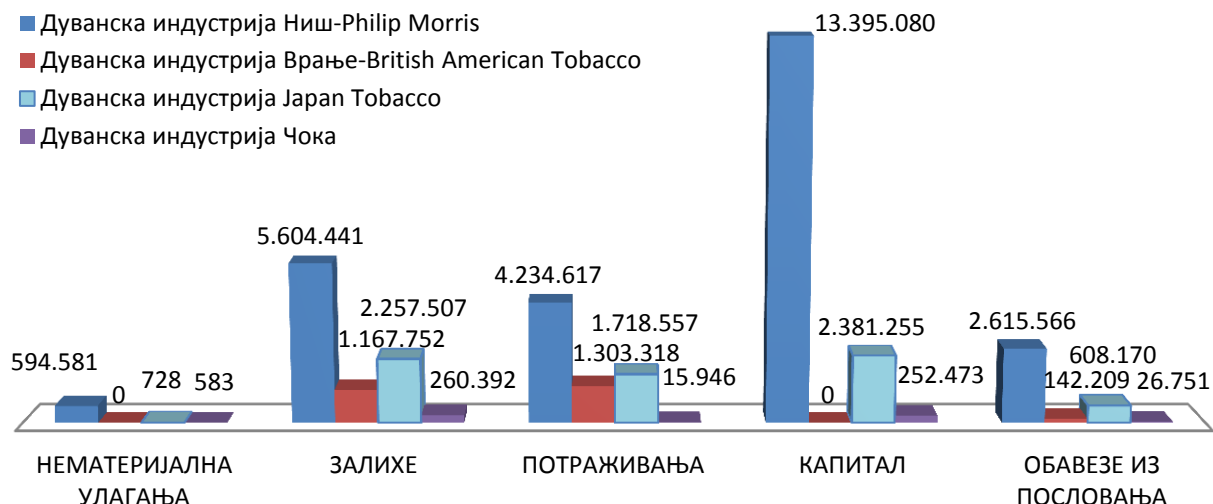
На основу добијених вредности неоспорно је за предикцију будућих исхода укључити, осим вредности значајних билансних позиција, и мишљења ревизора из претходних година (*Графикон 7.*).



**Графикон 7. Мишљење ревизора за свако дуванско предузеће током година,
(извор: www.apr.gov.rs)**

На основу *Графикана 7.* са одређеном нагађањем можемо извршити предикцију наредне године. Али да ли бисмо погрешили уколико бисмо очекивали да ће се сва мишљења ревизора за 2013. годину поновити и у 2014. години.

Резултати примене мултиваријационе дискриминационе анализе најбоље се могу огледати применом добијених резултата на наредну 2014. годину и на биланс стања претходно наведених предузећа.



Графикон 8. Подаци из биланса стања, за сва четири предузећа, за 2014. годину
(извор: www.apr.gov.rs)

Уколико посматрамо из угла предузећа, уз помоћ дискриминационе функције која је већ израчуната на основу историјских података, оно је у могућности да предвиди мишљење ревизора.

У делу о примени мултиваријационе дискриминационе анализе у SPSS-у, добијена је управо та дискриминациона функција изражена формулом (4.3):

$$f = 0,00001887 * x_1 - 0,00000511 * x_2 + 0,00000014 * x_3 + 0,00000019 * x_4 + 0,00000434 * x_5 + 2,6494.$$

Променљиве x_1 , x_2 , x_3 , x_4 и x_5 представљају независне променљиве, тачније представљају одговарајуће износе из биланса стања за 2014. годину, који су наведени *Графикону 8*. Вредности дискриминационе функције могу се израчунати за свако предузеће. Одговарајуће променљиве се замењују у формулу (4.3), а износи променљивих дати су у *Табели 26*.

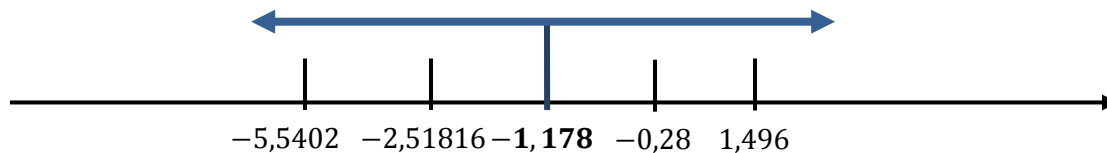
	x_1	x_2	x_3	x_4	x_5
Дуванска индустрија Ниш-Philip Morris	594.581	5.604.441	4.234.617	13.395.080	2.615.566
Дуванска индустрија Врање-British American Tobacco	0	1.167.752	1.303.318	0	142.209
Дуванска индустрија Јапан Tobacco	728	2.257.507	1.718.557	2.381.255	608.170
Дуванска индустрија Чока	583	260.392	15.946	252.473	26.751

Табела 26. Вредности независних променљивих (за 2014. годину)
(извор: www.apr.gov.rs)

Користећи неки од програма израчунава се да је вредност дискриминационе функције за Дуванска индустрија Ниш-Philip Morris једнака $-0,28008202$, за Дуванску индустрију Врање-British American Tobacco је $-2,51816114$, за Дуванску индустрију Јапан Tobacco је $-5,54022918$, а за Дуванску индустрију Чока је $1,49609974$.

Затим поредимо добијену вредност дискриминационе функције са претходно израчунатим „Cutting score“. Уколико је вредност функције за одређено предузеће испод $-1,178$, онда ће бити мишљење са скретањем пажње а уколико је изнад те вредности онда ће бити сигурно позитивно.

У случају ових предузећа стање је следеће:



Слика 15. Вредности дискриминационих функција представљене на бројеној правој (извор: приказ аутора)

Дакле, Дуванска индустрија Јапан Tobacco и Дуванска индустрија Врање-British American Tobacco могу да очекују мишљење са скретањем пажње, а Дуванска индустрија Ниш-Philip Morris и Дуванска индустрија Чока позитивно мишљење. Након што је извршена ревизија за 2014. годину, резултати који су јавно објављени током 2015. године потврђују резултате који су предвиђени дискриминационом анализом. За Дуванску индустрију Чока дато је првобитно мишљење са скретањем пажње због одређених судских спорова чији исход није био извршен. Међутим у априлу 2015. године у извештају другог ревизора исказано је позитивно мишљење као што је и било предвиђено анализом.

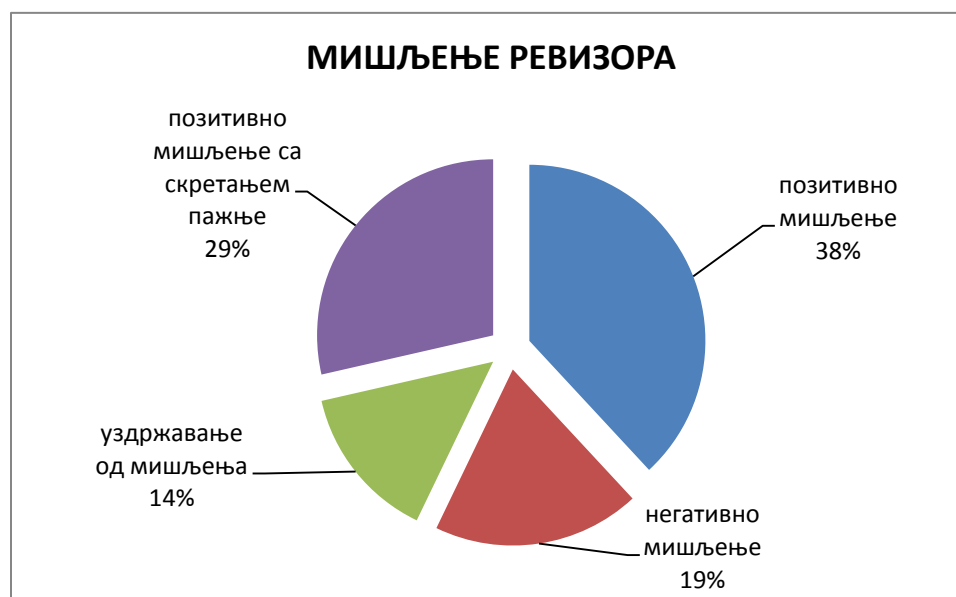
5.1 Примена анализе са додатим варијаблама

Како би сама дискриминациона анализа била што веродостојнија неопходно је да број случајева у најмањој групи буде већи од броја независних варијабли. Због тога у рад се уводе још три вештачке променљиве (насумично одређене) којима ће бити употпуњен тај услов. Дакле, посматрамо већ постојећу *Табелу 5.* која је допуњена са следећим износима:

год.	НЕМАТЕРИЈА. УЛАГАЊА	ЗАЛИХЕ	ПОТРАЖИВАЊА	КАПИТАЛ	ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	мишљење ревизора
2013	0	2.395	4.011	50.527	3.500	4
	106	2.767	65	47.903	12.111	5
	120	456.000	574	58.225	58.742	3
2012	0	2.353	5.500	60.980	4.105	4
	0	2.767	401	47.903	7.855	4
	2	123.001	12.356	69.874	48.478	3
2011	0	2.302	3.399	10.145	1.023	3
	0	3.549	401	47.903	4.852	5
	420	1.250	41.526	477	7.485	3

Табела 27. Накнадно додати износи варијабли (извор: приказ аутора)

Можемо уочити да додајући и податке из *Табеле 27.* добијамо дискриминациону анализу за више класификационих група. Осим постојећих анализираних мишљења ревизора, спрам реалних података, сада се уводе као могућности негативно мишљење и уздржавање од мишљења.



Слика 16. Процентуално изражавање одређеног мишљења (извор: приказ аутора)

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
HEMATEPИЈAЛHА УЛAГAЊA	,722	2,187	3	17	,127
ЗAЛИХE	,751	1,874	3	17	,172
ПOTPAЖИBАЊA	,709	2,322	3	17	,112
КАПИТАЛ	,769	1,707	3	17	,203
ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	,740	1,994	3	17	,153

Табела 28. Tests of Equality of Group Means (израчунаме у SPSS-у)

Анализирање Табеле 28. своди се на Табелу 6. , дакле што је мања вредност теста Wilks' Lambda за одређену променљиву, то значи да је та независна варијабла више значајна за дискриминациону функцију, спрам тога највећи допринос имају „Потраживања“ пошто је вредност Wilks' Lambda тада најнижа.

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1,388 ^a	70,5	70,5	,762
2	,572 ^a	29,1	99,6	,603
3	,009 ^a	,4	100,0	,093
a. First 3 canonical discriminant functions were used in the analysis.				

Табела 29. Eigenvalues (израчунамо у SPSS-у)

У Табели 29. може се уочити да постоје три дискриминационе функције и да свака има свој карактеристични корен (λ) и они износе редом 1,388, 0,572 и 0,009. И знамо да уколико је већи карактеристични корен функције, утолико је више објашњено о варијанси зависне варијабле кроз ту функцију, спрам тога може се закључити да прва дискриминациона функција има већу дискриминациону моћ од друге две функције. Каноничка корелација је мера повезаности дискриминационе функције и зависне варијабле и она такође потврђује претходан закључак о првој дискриминационој функцији је њена каноничка корелација износи 0,762 ,а код остале две 0,603 и 0,093.

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	,264	20,643	15	,149
2 through 3	,630	7,150	8	,521
3	,991	,133	3	,988

Табела 30. Wilks' Lambda за три дискриминационе функције (израчунамо у SPSS-у)

Пошто видимо у Табели 30. да није $p < 0.05$ код све три дискриминационе функције, значи да се прихвата нулта хипотеза да је сума разлике квадрата између група једнака нули, и да не постоји значајна веза између зависних и независних варијабли. До овог закључка могуће да је дошло због погрешно насумично извршеног одабира износа независних варијабли. Но, треба и нагласити да је у теоријском делу претходно написано да може да се деси да није свака од добијених дискриминационих функција статистички значајна. Даљом анализом података долази се до потврде претходно утврђеног.

Structure Matrix			
	Function		
	1	2	3
НЕМАТЕРИЈАЛНА УЛАГАЊА	,445*	,440	-,273
ЗАЛИХЕ	,213	,778*	,016
ПОТРАЖИВАЊА	,127	,733*	-,229
КАПИТАЛ	,207	,715*	-,128
ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	,325	,519*	-,126
*. Largest absolute correlation between each variable and any discriminant function			

Табела 31. Structure Matrix за три дискриминационе функције (израчунато у SPSS-у)

Табелом 31. приказује се корелација између варијабли и дискриминационих резултата добијених дискриминационим функцијама. Апсолутна вредност коефицијената матрице указује на њихов дискриминаторни утицај, стога предиктор „Нематеријална улагања“ највишег је утицаја на првој дискриминационој функцији, а „Залихе“, „Потраживања“, „Капитал“ и „Обавезе из пословања“ су највишег утицаја на другој дискриминационој функцији.

Canonical Discriminant Function Coefficients			
	Function		
	1	2	3
НЕМАТЕРИЈАЛНА УЛАГАЊА	,0000091073439	-,0000047034957	-,0000196475176
ЗАЛИХЕ	-,000002187688	0,0000006706619	-,0000017799006
ПОТРАЖИВАЊА	0,000000284045	0,0000002863201	0,000000282395
КАПИТАЛ	0,000000098335	0,0000001982433	,0000016596762
ОБАВЕЗЕ ИЗ ПОСЛОВАЊА	,0000020740318	-0,000000452312	-0,000000293107
(Constant)	,0976164209791	-1,020778382396	,0796700771392

Табела 32. Canonical Discriminant Function Coefficients (израчунати у SPSS-у)

Из Табеле 32. можемо уочити вредности нестандардизованих коефицијената за сваку једначину дискриминационе функције:

$$\begin{aligned}
 f_1 &= 0,0000091073439 * x_1 - 0,000002187688 * x_2 + 0,000000284045 * x_3 \\
 &\quad + 0,000000098335 * x_4 + 0,0000020740318 * x_5 + 0,0976164209791, \\
 f_2 &= -0,0000047034957 * x_1 + 0,0000006706619 * x_2 + 0,00000028632 * x_3 + \\
 &\quad 0,0000001982433 * x_4 - 0,000000452312 * x_5 - 1,020778382396, \\
 f_3 &= -0,000019647517 * x_1 - 0,0000017799006 * x_2 + 0,000000282395 * x_3 \\
 &\quad + 0,0000016596762 * x_4 - 0,000000293107 * x_5 + 0,0796700771392.
 \end{aligned}
 \tag{5.1}$$

Вредности променљивих x_1 , x_2 , x_3 , x_4 и x_5 узимамо из Табеле 26. и Табеле 27.

Functions at Group Centroids			
Мишљење	Function		
	1	2	3
1	1,141	,463	-,004
3	-,152	-,926	-,129
4	,109	-1,010	,163
5	-1,475	,504	,010

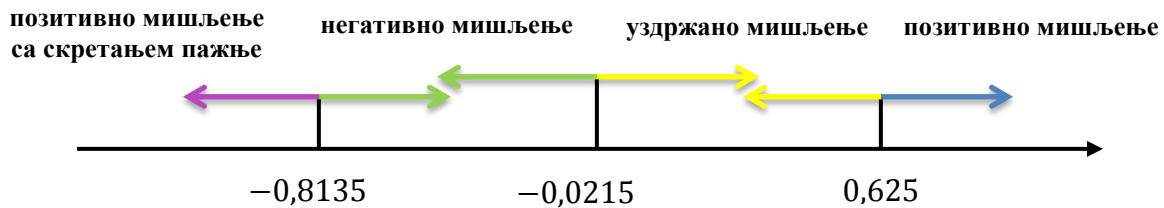
Табела 33. Functions at Group Centroids (израчунамо у SPSS-у)

Познато је уколико постоје две групе онда се користи Фишерава дискриминациона функција за алокацију опсервација у једну од група. „Класификација опсервација у случају две групе захтевала је израчунавање средишње тачке између првог и другог центроида (средине дискриминационих скорова по групама).“ [2] У том случају, посматрајући положај дискриминационог скор опсервиране варијабле у односу на центроид, алокација се врши у прву или другу групу. Нова опсервација биће алоцирана спрам краћег растојања од центроида одређене групе. „У општем случају користи се Махаланобисово растојање опсервације од центроида сваке групе, па нову опсервацију алоцирамо у групу од чијег је центроида ова најмање удаљена.“ [2]

Тако се за овај пример одређује се средишња или пресечна тачка („Cutting score“) спрам које се врши алокација. С обзиром да постоје четири групе, постојаће три пресечне тачке дате у наставку. (5.2) Посматраћемо шта би било да је значајна прва дискриминациона функција. У то случају добијамо

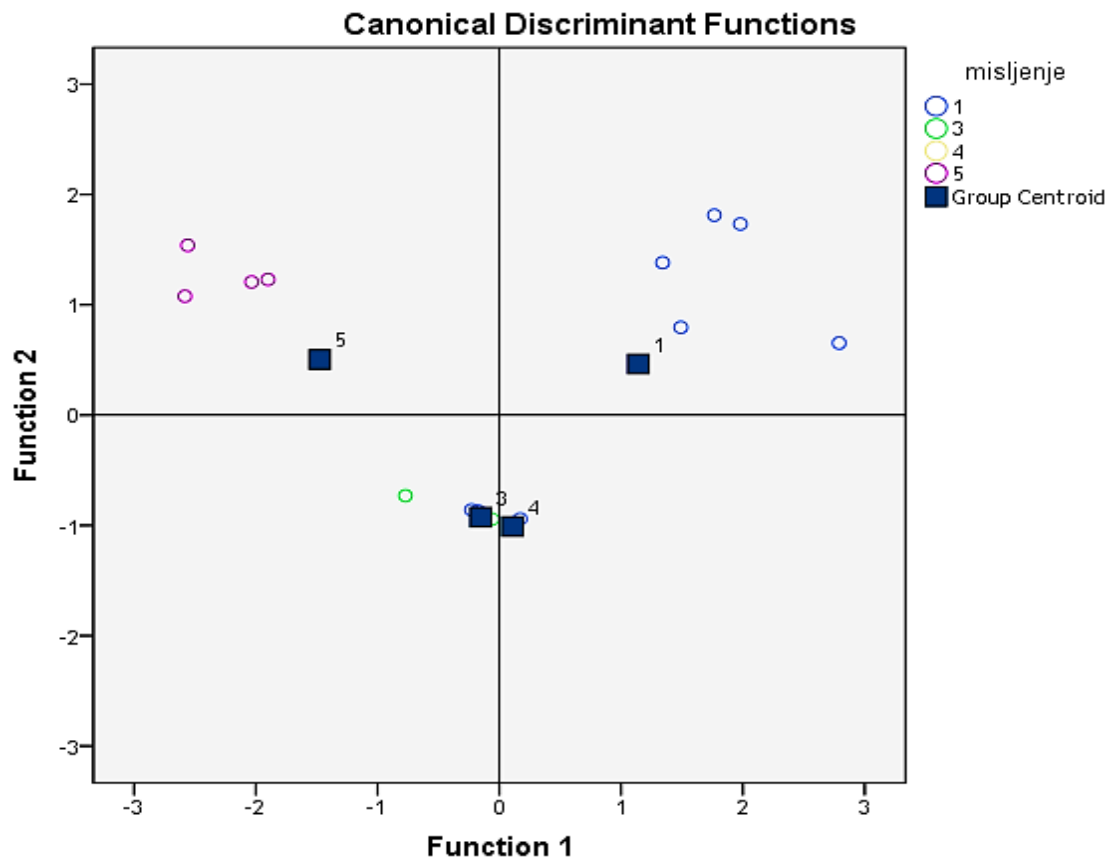
$$\begin{aligned}
 Cutting\ score_1 &= \frac{1,141 + 0,109}{2} = 0,625, \\
 Cutting\ score_2 &= \frac{-0,152 + 0,109}{2} = -0,0215, \\
 Cutting\ score_3 &= \frac{-0,152 - 1,475}{2} = -0,8135.
 \end{aligned}
 \tag{5.2}$$

С обзиром да имамо четири могуће групе тачније ситуације када је мишљење ревизора позитивно, са скретањем пажње, негативно или је уздржавајуће, па уколико је добијена вредност дискриминационе функције, када се замене одређене вредности независних променљивих (датих Табелом 26. и Табелом 27.), испод $Cutting\ score_1 = 0,625$, онда ће бити уздржано мишљење, а уколико је изнад те вредности онда ће бити сигурно позитивно. Затим уколико је вредност дискриминационе функције испод $Cutting\ score_2 = -0,0215$ ревизор ће негативно мишљење, а уколико је изнад уздржаће се од мишљења. И уколико је вредност функције испод $Cutting\ score_3 = -0,8135$ биће дато мишљење са скретањем пажње у супротном ће дати негативно мишљење. (Слика 16. и Слика 17.)



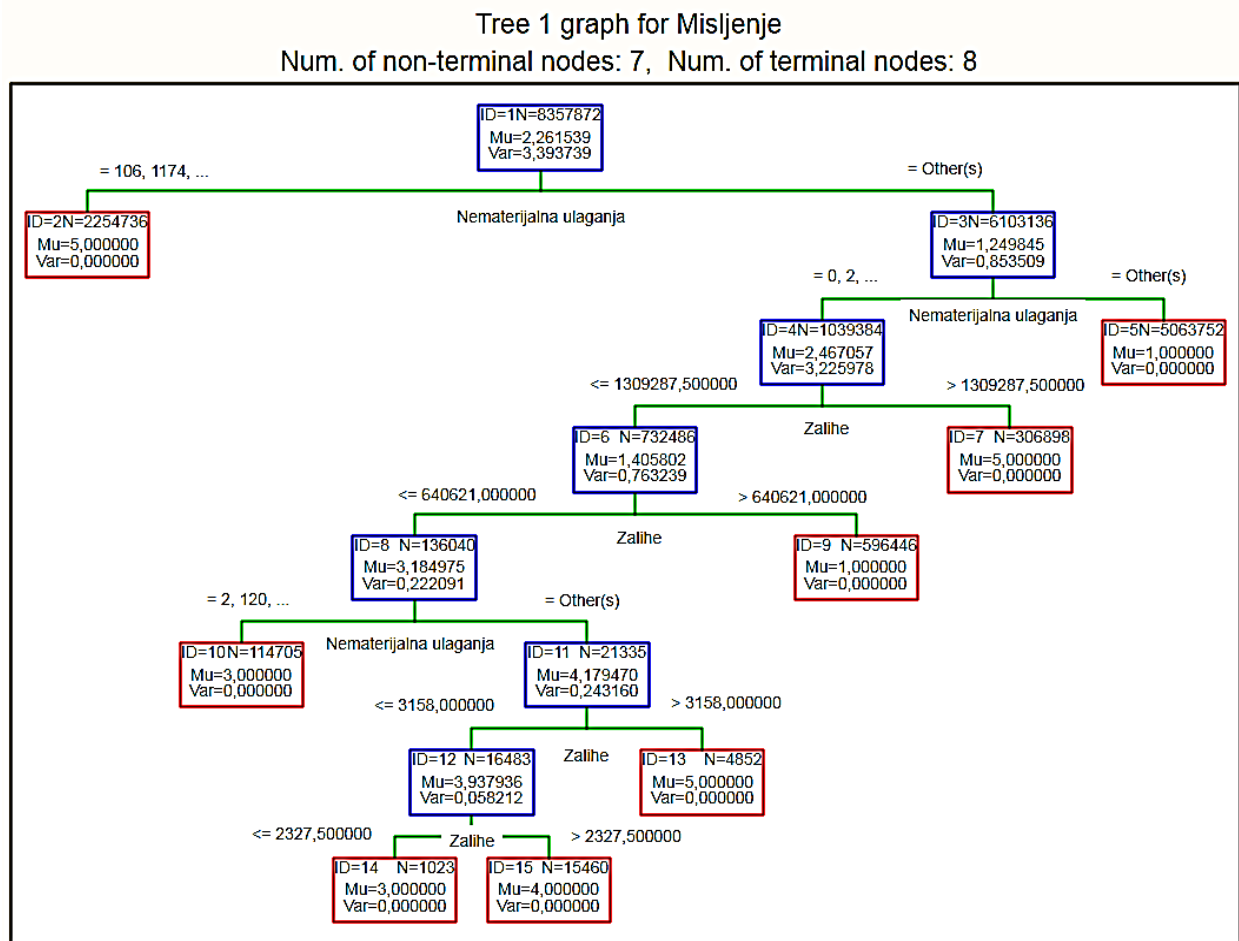
Слика 16. Износ „Cutting score-a“ (извор: приказ аутора)

Напомена: због случајно одабраних додатих варијабли и мишљења која су им насумично додељена, постоје одређена одступања.



Слика 17. Приказ центроида по групама (Нацрт добијен у SPSS-у)

Уколико постоји више класификационих група и стабло одлучивања има више грана. У том случају добија се једно од могућих стабала приказаних на *Слици 18*.



Слика 18. Стабло одлучивања - C&RT метод, са 15 чворова (Нацрт добијен у програмском алату STATISTICA)

Дакле, убацивањем још могућих класификационих група у пример (подељене спрам ревизоровог мишљења) добијамо стабло одлучивања које се у првом моменту грана спрам износа „Нематеријалних улагања“ где уколико је износ = 106,1174, ... (изражено у 000 динара) онда ће бити мишљење са скретањем пажње, а друга грана даје остале могућности где осим „Нематеријалних улагања“ битну улогу имају и „Залихе“. Уколико су „Нематеријална улагања“ = 0,2, ... (изражено у 000 динара) онда се посматра гранање спрам износа „Залиха“ који ако је > 1309287,500 (изражено у 000 динара) онда ће бити мишљење са скретањем пажње. За износ ≤ 1309287,500 (изражено у 000 динара) долази се до даљег гранања са више могућности где се даје позитивно мишљење ако је вредност „Залиха“ већа од 640621,000 (изражено у 000 динара) или се за мање од тог износа гранање наставља наспрам „Нематеријалног улагања“. И на самом крају долази се опет до условљавајућих „Залиха“ где се уколико је износ већи од 2327,5000 (изражено у 000 динара) уздржава од мишљења или у супротном даје негативно.

6 Закључак

У данашњем свету који непрекидно и незауостављиво напредује у технолошком смислу, и похрањује свакодневно светску базу података, принуђени смо на налажење метода како бисмо што боље искористили тренутак и место. Изузетно је битно да усавршавањем техника и метода, уз одређену прецизност, у стању смо да предвидимо какав ће исход бити у будућности. Предвидивост. Како предвидети сутра, како пар година. Како предвидети исходе како поступити тако да последице и грешке у одлукама буду минималне. У том случају ми, на основу историјских података тражимо неку доследност у понашању, на основу које увиђамо шта можемо да очекујемо.

У овом раду пажња је била посвећена примени мултиваријационе дискриминационе анализе, као једном таквом методу уз помоћ којег манипулишемо подацима. Сама срж јесте управо у правој класификацији опсервација. Ослањајући се на дискриминациону анализу у процесу ревизије добијени су одређени значајни резултати. Но, наравно не можемо занемарити чињеницу да је тешко одредити прави узорак за одређену анализу како би избегли девијантност те анализе. Сваки резултат, исход, треба узети са резервом тражећи нове, боље применљивије технике.

Овај рад, састоји се из више делова који чине једну компактну целину. Дакле, прво је у рад уведена теоријско-математичка подршка, на којој се заснива наредни део у ком је примењена дискриминациона анализа у пословању дуванске индустрије на простору Србије.

Посматрајући финансијске извештаје из претходних година циљ нам је био да предвидимо на основу различитих аспеката и позиција какво ће пословање бити и какво ће мишљење дати ревизор. Наведене информације су изузетно битне за свако привредно друштво, јер оно као субјекат који се котира на пословном тржишту има за циљ да буде што боље тј. да изнађе и искористи могућност избегавања негативног резултата. Дискриминациона анализа управо пружа ту могућност предвидивости исхода, уз коришћење неког од програмских алата. Могућност примене те анализе огледа се у добијању дискриминационе функције, која се формира на основу постојећих података и која би требала да се обнавља и уопуњује за сваку наредну годину. Како би се употпунила и усавршила примена, треба мењати и бирати различите предикторе у различитим комбинацијама, наспрам претходних.

Конкретно, овде су за предикторе коришћене следеће позиције из биланса стања: „Нематеријална улагања“, „Залихе“, „Потраживања“, „Капитал“ и „Обавезе из пословања“. Применом дискриминационе анализе показали су се, као статистички значајне варијабле, „Нематеријална улагања“, „Залихе“ и „Обавезе из пословања“. Пративши те варијабле, добијени су одређени резултати на којима су вршене предикције. Такође, дате варијабле су уношене у дискриминациону функцију, како бисмо рачунањем вредности дате функције били у могућности да прецизно алоцирамо припадност одређеној групи.

Групе зависних варијабли овде су чинила мишљења ревизора. Спрам података који су званично објављени у складу са законом на сајту Агенције за привредне регистре, имали смо приступ финансијским извештајима дуванских предузећа. Ту је уочено да су мишљења ревизора варијирала између „позитивног“ и „позитивног мишљења са скретањем пажње“, па је класификација вршена на основу тог постојећег стања.

У зависности од износа предиктора, односно независних варијабли, дат је и пример такозваног надзираног учења путем стабла одлучивања. Уз помоћ стабла одлучивања, добија се логичко-последична веза којом се долази до груписања. Стабло одлучивања обезбеђује геометријски оквир за организовање одлука и пружа одређену алтернативу да се одлучимо за наредни корак.

Оно што повезује мултиваријантну дискриминациону анализу и стабло одлучивања јесте способност за проналажење скривених знања. Управо је то оно као чему сваки учесник на тржишту тежи - да читањем јавно доступних података увиди нешто пре других, или да предвиди нешто наизглед неочекивано.

Међутим, треба напоменути да уколико је нарушен неки од услова дискриминационе анализе (као што су: непостојање мултиколинearности између независних варијабли, заступљеност независне мултиваријационе нормалне расподеле, да се матрице варијанси и коваријанси независних варијабли између група једнаке...) онда она не може да се примени и на њено место долазе неке друге технике.

Литература

- [1] Андрић М., Крсмановић Б., Јакшић Д., *“Ревизија – теорија и пракса”*, Економски факултет Суботица, 2009.
- [2] Ковачић Ј. З., *“Мултиваријациона анализа”*, Економски факултет, Београд, 1994.
- [3] Лозанов-Црвенковић З., *“Статистика”*, Универзитет у Новом Саду, Природно математички факултет, Нови Сад, 2012.
- [4] Maddala G. S., *“Introduction to Econometrics”*, 2nd ed., Macmilan, 1992.
- [5] Tabachnick B., Fidell L., *“Using Multivariate Statistics”*, fifth ed., Pearson Education, 2007.
- [6] Edward I. A., *“Financial Ratios Discriminant Analysis and Prediction of Corporate Bankruptcy”*, The Journal of Finance, 1968.
- [7] Станковић М., Стаменковић М., *„Дискриминациона анализа“*, Студијско истраживачки рад, Универзитет у Крагујевцу, Економски факултет у Крагујевцу, 2009.
- [8] Huberty C., Olejnik S., *„Applied MANOVA and Discriminant Analysis“*, 2nd ed., John Wiley & Sons, Inc., Hoboken, New Jersey, 2006 .
- [9] Kirkos E., Spathis C., Manolopoulos Y., *“Data Mining for the detection of fraudulent financial statements”*, Expert Systems with Applications, 2007.
- [10] Koskivaara E., *“Artificial neural in Auditing: State of the Art”*, Turku School of Economics and Business Administration, Turku, Finland, 2003.
- [11] James G., Witten D., Hastie T., Tibshirani R., *“An Introduction to Statistical Learning with Applications in R”*, Springer, New York, 2013.
- [12] Holdsworth K. D., Tan S. S., Chung K., *„Insolvency Prediction Model Using Multivariate Discriminant Analysis and Artificial Neural Network for the Finance Industry in New Zealand“*, Vol. 3 (1), International Journal of Business and Management, The Publishing Center of IJBM, Perth Australia, 2008.
- [13] Живковић М., *„Алгоритми“*, Математички факултет, Београд, 2008.

- [14] Кијевчанин В., Грачанин Ш., „*Data Mining*“, Самостални истраживачки рад, Универзитет у Крагујевцу, Економски факултет у Крагујевцу, 2009.
- [15] Drury C., „*Management Accounting for Business*“, third ed., Thomson Learning, London, 2005.
- [16] Charles M., Friel Ph.D., „*Notes on Discriminant Analysis*“, Criminal Justice Center, Sam Houston State University, 2008.
- [17] Clemen R.T., Reilly T., „*Making Hard Decisions with Decision Tools*“, 2nd revised edition, Chapter 3, Duxbury Press, 2001.
- [18] Sharma, S., „*Applied Multivariate Techniques*“, John Wiley & Sons, Inc., 1996.
- [19] Ћирић-Рајтер Д., „*Вероватноћа*“, Универзитет у Новом Саду, Природно-математички факултет, Нови Сад, 2009.
- [20] Благојевић-Папић Н., *Презентација предавања са курса Актуарство*, доступно на <http://www.vps.ns.ac.rs/Materijal/mat13331.pdf>
- [21] Јевремовић В., *Презентација другог поглавља са курса Вероватноћа и статистика*, Математички факултет, Београд, 2013.
- [22] Ћирић-Рајтер Д., *Белешке са курса Стохастичка анализа*, Природно-математички факултет, Нови Сад, 2013.
- [23] Јовановић С., „*Теорија стохастичке интеграције*“, Математички факултет, Београд, 2011.
- [24] Илић М., Решовић С., „*Кластер анализа*“, Приступни рад, Универзитет у Крагујевцу, Економски факултет у Крагујевцу, 2009.
- [25] <http://www-01.ibm.com/software/analytics/spss/>
- [26] [www.researchgate.net/publications.PublicPostFileLoader.html?id=54eb12afef97130f298b4576&key=4e2dde10-a8b1-4b14-af47-1ba56fe681f4](http://www.researchgate.net/publications/PublicPostFileLoader.html?id=54eb12afef97130f298b4576&key=4e2dde10-a8b1-4b14-af47-1ba56fe681f4)
- [27] www.apr.gov.rs
- [28] <https://www.ifac.org/>
- [29] http://www.nbs.rs/export/sites/default/internet/latinica/20/zakoni/rac_revizija
- [30] https://en.wikipedia.org/wiki/Test_set
- [31] <http://web.efzg.hr/dok/OIM/mdarabos/4-Stablo%20odlu%C4%8Divanja>
- [32] https://sr.wikipedia.org/sr/алоритам_две_најближе_тачке
- [33] <https://www.msu.edu/user/sw/statrev/strv211.htm>

- [34] <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4262615/>
- [35] <http://www.kor.rs/>
- [36] <https://sh.wikipedia.org/wiki/Korelacija>
- [37] <http://www3.nd.edu/~busiforc/handouts/DataMining/dataminingdefinitions.html>
- [38] [http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/TwoGroupHatcoDiscriminant Analysis doc html/035 Overall significance of the discriminant function s .html](http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/TwoGroupHatcoDiscriminantAnalysis%20doc%20html/035%20Overall%20significance%20of%20the%20discriminant%20function%20s%20.html)
- [39] <http://documentation.statsoft.com/STATISTICAHelp.aspx?path=Gxx/InteractiveTrees/InteractiveTreesCRTCHAIDExample>
- [40] <http://documents.software.dell.com/Statistics/Textbook/Discriminant-Function-Analysis>
- [41] <http://www.singipedia.singidunum.ac.rs/attachment.php?s=9e9299b3a0fdc2a9be29dcd649891799&attachmentid=651&d=1275660397>
- [42] https://en.wikipedia.org/wiki/Ronald_Fisher
- [43] <http://supa.pharmacy.bg.ac.rs/assets/21648>.

Списак табела, графикана и слика

Списак табела

Табела 1. Подаци из биланса стања (у 000 динара) Дуванске индустрије Ниш–Philip Morris (извор: www.apr.gov.rs)

Табела 2. Подаци из биланса стања (у 000 динара) Дуванске индустрије Врање-British American Tobacco (извор: www.apr.gov.rs)

Табела 3. Подаци из биланса стања (у 000 динара) Дуванске индустрије Јарап Tobacco (извор: www.apr.gov.rs)

Табела 4. Подаци из биланса стања (у 000 динара) Дуванске индустрије Чока (извор: www.apr.gov.rs)

Табела 5. Подаци из биланса стања за сва четири предузећа (извор: www.apr.gov.rs)

Табела 6. Тестови једнакости група средњих вредности/ Tests of Equality of Group Means (израчунаме у SPSS-у)

Табела 7. Приказ корелационе матрице/Covariance Matrices(израчуната у SPSS-у)

Табела 8. Карактеристични корен/ Eigenvalues (израчунае у SPSS-у)

Табела 9. Wilks' Lambda (израчунае у SPSS-у)

Табела 10. Стандардизовани коефицијенти дискриминационе функције/ Standardized Canonical Discriminant Function Coefficients (израчунае у SPSS-у)

Табела 11. Структура матрице/ Structure Matrix (израчуната у SPSS-у)

Табела 12. Нестандардизовани коефицијенти дискриминационе функције/ Canonical Discriminant Function Coefficients (израчунае у SPSS-у)

Табела 13. Центроиди, група центроида (израчунае у SPSS-у)

Табела 14. Примарне вероватноће за групе/ Prior Probabilities for Groups (израчунае у SPSS-у)

Табела 15. Класификација коефицијената функције/ Classification Function Coefficients (израчуната у SPSS-у)

Табела 16. Group Statistics (табела израчуната у SPSS-у)

Табела 17. Casewise Statistics (израчунато у SPSS-у)

Табела 18. Резултати класификације(израчунати у SPSS-у)

Табела 19. Резиме дискриминационе анализе (израчунат у програмском алату STATISTICA)

Табела 20. Класификациона матрица (израчуната у програмском алату STATISTICA)

Табела 21. Квадрирано Махалонобисово растојање (спрам центроида групе) (израчунато у програмском алату STATISTICA)

Табела 22. The Pooled Within-group Correlations-Обједињене унутар групе корелације (израчунате у програмском алату STATISTICA)

Табела 23. Средње вредности података из биланса стања (груписани спрам мишљења ревизора) (израчунате у програмском алату STATISTICA)

Табела 24. Стандардна девијација података из биланса стања (груписани спрам мишљења ревизора) (израчуната у програмском алату STATISTICA)

Табела 24. Стандардна девијација података из биланса стања (груписани спрам мишљења ревизора) (израчуната у програмском алату STATISTICA)

Табела 25. Корелације (израчунато у програмском алату STATISTICA)

Табела 26. Вредности независних променљивих (за 2014. годину) (извор:www.apr.gov.rs)

Табела 27. Накнадно додати износи варијабли (извор: приказ аутора)

Табела 28. Tests of Equality of Group Means (израчунате у SPSS-у)

Табела 29. Eigenvalues (израчунато у SPSS-у)

Табела 30. Wilks' Lambda за три дискриминационе функције (израчунато у SPSS-у)

Табела 31. Structure Matrix за три дискриминационе функције (израчунато у SPSS-у)

Табела 32. Canonical Discriminant Function Coefficients (израчунати у SPSS-у)

Табела 33. Functions at Group Centroids (израчунато у SPSS-у)

Списак графикана

Графикон 1. Подаци из биланса стања (у 000 динара) Дуванске индустрије Ниш–Philip Morris (извор: приказ аутора)

Графикон 2. Подаци из биланса стања (у 000 динара) Дуванске индустрије Врање-British American Tobacco (извор: приказ аутора)

Графикон 3. Подаци из биланса стања (у 000 динара) Дуванске индустрије Јарап Tobacco (извор: приказ аутора)

Графикон 4. Подаци из биланса стања (у 000 динара) Дуванске индустрије Чока (извор: приказ аутора)

Графикон 5. „Нематеријална улагања“ током година (за сва четири дуванска предузећа), (извор: приказ аутора)

Графикон 6. Нематеријална улагања, залихе и обавезе из пословања током година (за сва четири дуванска предузећа), (извор: приказ аутора)

Графикон 7. Мишљење ревизора за свако дуванско предузеће током година, (извор: www.apr.gov.rs)

Графикон 8. Подаци из биланса стања, за сва четири предузећа, за 2014. годину (извор: www.apr.gov.rs)

Списак слика

Слика 1. Шема анализе мултиваријационих података(извор: „Applied MANOVA and Discriminant Analysis“,[8])

Слика 2.Приказ дискриминационе анализе две групе (извор: Ковачић Ј. З. „Мултиваријациона анализа“, [2])

Слика 3. Приказ дискриминационе анализе две групе са центроидима (извор: Charles M., Friel Ph.D., „Notes on Discriminant Analysis“, [16])

Слика 4.Cutting score и дискриминациона функција (графички приказ Станковић М., Стаменковић М., „Дискриминациона анализа“,[7])

Слика 5. Приказ ефикасне и неефикасне функције (извор: Станковић М., Стаменковић М., „Дискриминациона анализа“,[7])

Слика 6.Сликовити приказ изградње једног стабла одлучивања (извор: приказ аутора)

Слика 7. Структура ревизорских мишљења (извор: приказ аутора)

Слика 8.Корелација свих варијабли (нацрт добијен у програмском алату STATISTICA)

Слика 9. Стабло одлучивања на основу Нематеријалних улагања (Нацрт добијен у програмском алату STATISTICA)

Слика 10. Стабло одлучивања на основу Залиха (Нацрт добијен у програмском алату STATISTICA)

Слика 11. Стабло одлучивања на основу Обавеза из пословања (Нацрт добијен у програмском алату STATISTICA)

Слика 12. Стабло одлучивања - C&RT метод (Нацрт добијен у програмском алату STATISTICA)

Слика 13. Стабло одлучивања - C&RT метод, други услов (Нацрт добијен у програмском алату STATISTICA)

Слика 14. Подела на групе добијеном „тачком сечења“ (извор: приказ аутора)

Слика 15. Вредности дискриминационих функција представљене на бројеној правој (извор: приказ аутора)

Слика 16. Износ „Cutting score-a“ (извор: приказ аутора)

Слика 17. Приказ центроида по групама (Нацрт добијен у SPSS-у)

Слика 18. Стабло одлучивања - C&RT метод, са 15 чворова (Нацрт добијен у програмском алату STATISTICA)

Биографија



Миљана Вукоје рођена је 3.10.1991. године у Новом Саду. Завршила је Основну школу „Доситеј Обрадовић“ у Новом Саду 2006. године, као ђак генерације. Затим је уписала Гимназију „Светозар Марковић“, коју завршава 2010. године као носилац Вукове дипломе.

Те исте године уписује основне студије на Природно-математичком факултету смер Примењена математика, модул Математика финансија. У року завршава трогодишње студије 2013. године и потом уписује мастер студије Примењене математике на

Природно-математичком факултету у Новом Саду.

Завршно са јунским роком 2015. године положила је све испите предвиђене планом и програмом мастер студија и тако стекла право на одбрану мастер рада.

Током студирања бавила се креативним промовисањем науке, волонтерским радом, медијским залагањем у емисији едукативног типа, и певањем у хору.

Последње године мастер студија, освојила је прво место на такмичењу студената природно-математичких факултета из региона, са радом „Примена линеарног програмирања у војне сврхе“.

Нови Сад, 2015. године

Миљана Вукоје

УНИВЕРЗИТЕТ У НОВОМ САДУ

ПРИРОДНО – МАТЕМАТИЧКИ ФАКУЛТЕТ

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број:

РБР

Идентификациони број:

ИБР

Тип документације: Монографска документација

ТД

Тип записа: Текстуални штампани материјал

ТЗ

Врста рада: Мастер рад

ВР

Аутор: Миљана Вукоје

АУ

Ментор: др Наташа Спахић

МН

Наслов рада:

Примена мултиваријантне дискриминационе анализе у процесу ревизије

НР

Језик публикације: српски (ћирилица)

ЈП

Језик извода: српски / енглески

ЈИ

Земља публикавања: Република Србија

ЗП

Уже географско подручје: Војводина

УГП

Година: 2015.

ГО

Издавач: Ауторски репринт

ИЗ

Место и адреса: Трг Доситеја Обрадовића 4, Нови Сад

МА

Физички опис рада: (6/101/43/33/18/8/0)

(број поглавља/страна/лит. цитата/табела/слика/графика/прилога)

ФО

Научна област: Математика- финансије

НО

Научна дисциплина: Ревизија

НД

Предметна одредница/Кључне речи: финансијски извештај, ревизија, мишљење ревизора, Бајесова теорема, ФишEROVA дискриминациона функција, мултиваријантна дискриминациона анализа, data mining, стабло одлучивања

ПО

УДК

Чува се: Библиотека департмана за математику и информатику, Природно-математички факултет, Трг Доситеја Обрадовића 4, Нови Сад

ЧУ

Важна напомена: Нема

ВН

Извод:

Главни циљ овог мастер рада јесте приказивање могућности примене мултиваријантне дискриминационе анализе у процесу ревизије. Радом се повезује мишљење ревизора о финансијским извештајима предузећа за производњу дувана (зависна варијабла) са неколико позиција финансијског извештаја наведених предузећа (независне варијабле). Примена мултиваријантне дискриминационе анализе у процесу ревизије извршена је на основу јавно обелодањених података који се налазе у званичним финансијским извештајима и извештајима ревизора предузећа за производњу дувана у Републици Србији. На крају, након прецизног вредновања добијених резултата, извршено је разматрање да ли мултиваријантна дискриминациона анализа може бити адекватан методолошки апарат у процесу ревизије.

ИЗ

Датум прихватања теме од стране ННвећа:

ДП

Датум одбране: 2015.

ДО

Чланови комисије:

Председник: др Наташа Крејић, редовни професор, Природно-математички факултет, Универзитет у Новом Саду

Ментор: др Наташа Спахић, ванредни професор, Природно-математички факултет, Универзитет у Новом Саду

Члан: др Сања Рапајић, ванредни професор, Природно-математички факултет, Универзитет у Новом Саду

UNIVERSITY OF NOVI SAD
FACULTY OF NATURAL SCIENCES & MATHEMATICS
KEY WORDS DOCUMENTATION

Accession number:

ANO

Identification number:

INO

Document type: Monograph documentation

DT

Type of record: Textual printed material

TR

Contents code: Master thesis

CC

Author: Miljana Vukoje

AU

Mentor: Nataša Spahić, Ph.D.

MN

Title: The application of multivariate discriminant analysis in the audit process

TI

Language of text: Serbian (Cyrillic)

LT

Language of abstract: en/s

LA

Country of publication: Serbia

CP

Locality of publication: Vojvodina

LP

Publication year: 2015.

PY

Publisher: Author's reprint

PU

Publ. place: Novi Sad, Department of Mathematics and Informatics, Faculty of Science and Mathematics, University of Novi Sad, Trg Dositeja Obradovica 4

PP

Physical description: (6/101/43/33/18/8/0)

(number of sections/pages/references/tables/pictures/graphs/appendices)

PD

Scientific field: Mathematics- Finance

SF

Scientific discipline: Audit

SD

Subject Key words: financial statement, audit, auditor's opinion, Bayes theorem, Fisher discriminant function, multivariate discriminant analysis, data mining, decision tree

SKW

UC

Holding data: The Library of the Department of Mathematics and Informatics, Faculty of Science and Mathematics, University of Novi Sad, Trg Dositeja Obradovića 4, Novi Sad

HD

Note:

N

Abstract: The main goal of this master thesis is to show the possibilities of applying multivariate discriminant analysis in the audit process. The document associate the auditor's opinion on the financial statements of the companies for the production of tobacco (the dependent variable) with several positions of the financial statements of

these companies (independent variables). The application of multivariate discriminant analysis in the audit process was conducted based on publicly resolved data, which are contained in the official financial statements and reports of auditors of companies for tobacco production in the Republic of Serbia. Finally, after a great evaluation of the results, it was performed consideration, whether multivariate discriminant analysis can be appropriate methodological equipment in the audit process.

AB

Accepted on Scientific board on:

AS

Defended:

DE

Thesis Defend board:

DB

President: Nataša Krejić Ph. D., Full professor, Faculty of Science and Mathematics, University of Novi Sad

Mentor: Nataša Spahić Ph. D., Associate Professor, Faculty of Science and Mathematics, University of Novi Sad

Member: Sanja Rapajić Ph. D., Associate Professor, Faculty of Sciences and Mathematics, Novi Sad