UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Miha Matjašič

Podatkovno rudarjenje v športu

Diplomsko delo

Ljubljana, 2012

UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Miha Matjašič

Mentor: doc. dr. Damjan Škulj

Podatkovno rudarjenje v športu

Diplomsko delo

Ljubljana, 2012

Zahvala

Zahvaljujem se mentorju doc. dr. Damjanu Škulju za usmerjanje in vse strokovne nasvete pri izdelavi diplomskega dela. Še posebej bi se rad zahvalil svoji družini za vso podporo in spodbudo pri študiju.

Podatkovno rudarjenje v športu

Diplomsko delo obravnava področje podatkovnega rudarjenja, s pomočjo katerega lahko pridobimo koristen in natančen vpogled v veliko količino podatkov. Hitrejši in učinkovitejši način zbiranja velike količine podatkov, ki je posledica sodobne družbe in sodobne tehnologije, je prispeval k dodatni prepoznavnosti podatkovnega rudarjenja v svetu. Sprva so podatkovno rudarjenje uporabljali le za poslovne namene, kmalu pa se je razširilo tudi na področje športa, kjer ga športne organizacije s pridom uporabljajo z namenom doseganja zmag. Tako dobiva vedno večjo veljavo in uspešnost v športu, kar je prikazano v petem in šestem delu diplomskega dela. V petem delu diplomskega dela sem predstavil praktični primer uporabe podatkovnega rudarjenja v košarkarski ligi NBA, kjer različne ekipe uporabljajo podatkovno rudarjenje za iskanje prednosti v igri pri svojih igralcih in iskanje slabosti v igri pri nasprotnih igralcih. Šesto poglavje pa zajema zbiranje in analizo podatkov v ligi NBA ter odkrivanje odnosov med njimi s pomočjo linearne regresije in metode grozdenja (ang. clustering).

Ključne besede: podatkovno rudarjenje, statistika, algoritmi, tehnike, podatkovno rudarjenje v športu.

Data mining in sport

This thesis deals with the area of data mining, by which we can obtain useful and accurate view of the large amount of data. Faster and more efficient way of collecting large amounts of data as a result of modern society and modern technology, has contributed to additional visibility of data mining in the world. Initially, data mining was used only for business purposes, but it quickly spread to the field of sport, where it benefits has been used from sporting organizations to achieve victories. It is gaining more and more value in sport, which is shown in the second and third part of the thesis. In the fifth part of the thesis I presented a practical example of using data mining in the NBA league, where different teams use data mining to search for the benefits in game of their players, and for finding weaknesses in game of the opposing players. The sixth chapter covers data collection and their analysis in the NBA league and with the help of linear regression and clustering methods, discovery of relationships between them.

Key words: data mining, statistics, algorithms, techniques, data mining in sports.

KAZALO

KAZALO SLIK

KAZALO TABEL

1 UVOD

Analiziranje in zbiranje podatkov ni nekaj novega, kar bi poznali šele kratek čas, pač pa gre za že dolgo uporabljeno tehniko v statistiki. Z uporabo različnih matematičnih formul so statistiki zbrane podatke opisovali, s pomočjo različnih statistik (mediana, aritmetična sredina, variance, standardni odkloni) pa ugotavljali njihove značilnosti. Prav zaradi tega je statistika dolga leta veljala za edino vejo, ki lahko uspešno analizira in ocenjuje vrednosti podatkov. Šele s prihodom digitalnih računalnikov in trdih diskov, kamor so se lahko shranjevali podatki, je prišlo do drastičnih sprememb tako na področju analize podatkov, kot tudi v njihovi količini.

V zadnjem času se je hitrost ustvarjanja in uporabljanja informacij ter podatkov s strani uporabnikov znatno povečala. Razširjena uporaba črtnih kod za večino tržnih produktov, informatizacija večine poslovnih in vladnih transakcij ter napredek v razvoju orodij za zbiranje podatkov so vzrok vedno večje količine podatkov, kar prispeva k ogromnim količinam informacij na spletu. To posledično vpliva na nastanek iz dneva v dan naraščajočega števila podatkovnih baz. Vse to pa je privedlo do tega, da dandanes v množici podatkov težko ločimo za nas pomembne in nepomembne informacije. Lahko rečemo, da je prišlo do eksplozije rasti za uporabnika nepomembnih podatkovnih baz in podatkov, ki je ustvarila nujno potrebo po novih tehnologijah in orodjih, ki bodo lahko z uporabo umetne inteligence samostojno preuredili ogromne količine podatkov v uporabne informacije in znanje. Posledično je podatkovno rudarjenje postalo raziskovalno področje z vedno večjo veljavo.

Če so bile statistike, kot so: *V letu 1900 je znašalo število prebivalcev 1.6 milijarde. Sto let pozneje je število prebivalcev znašalo že 6 milijard. Leta 1906 sta brata Stanley (Francis in Freelan) dosegla svetovni rekord, ko sta s svojim dirkalnim avtomobilom zabeležila hitrost 195km/h. 63 let pozneje je Apollo prvič pristal na Luni, hitrost, s katero je raketa letela v vesolje, je znašala 40.000 km/h*, včasih nekaj zelo impresivnega, lahko trdimo, da to ni skoraj nič v primerjavi z današnjo količino podatkov. Rastoča množica informacij in s tem vedno večja zasičenost uporabnikov je prispevala k razvoju samodejnih tehnik rudarjenja podatkov, ki ga označujemo z izrazom *podatkovno rudarjenje* (ang. Data mining).

Ker je bil koncept podatkovnega rudarjenja zaradi svoje uspešnosti na področju poslovanja vse bolj priljubljen, je hitro pritegnil pozornost različnih organizacij in podjetij. Med njimi so bile tudi športne organizacije, ki so vedno znova iskale novosti pri načinu analiziranja podatkov, saj so jim te omogočile prednosti pred športnimi konkurenti, kar je prineslo tudi velike vsote denarja. Prav zaradi uspešnosti pri pretvorbi ogromnih količin podatkov (met iz igre, odigrano število tekem, poškodbe igralca, prestopi igralca v drugo ekipo, prednosti igralca) v uporabne informacije, je koncept podatkovnega rudarjenja postajal vse bolj priljubljen tudi v športu.

Razvoj podatkovnega rudarjenja je omogočil, da so analize podatkov v sodobnem športu doživele svojo revolucijo. Dolga leta so na podatke v športu gledali le kot zapis dogodkov v igri, ki so ga vodile organizacije ali trenerji ekip. Šele razvoj računalnikov in eksplozija rasti podatkov sta povzročila, da sta objavljanje in analiza podatkov postala dovolj poceni, kar je posledično privabilo pozornost različnih športnih organizacij. Začela se je doba zbiranja in analiziranja podatkov o tekmah oziroma uporaba podatkovnega rudarjenja v športu.

1.1 Zasnova dela

Diplomsko delo je razdeljeno na šest poglavij. V prvih štirih poglavjih so podrobneje predstavljene metode podatkovnega rudarjenja. Peto poglavje predstavlja podatkovno rudarjenje na konkretnem primeru, in sicer podatkovno rudarjenje v košarkarski ligi NBA (ang. National basketball association). Ker lahko uporaba podatkovnega rudarjenja v košarki bistveno pripomore k napovedi določenih lastnosti posameznega igralca (izostanek s tekme, prestop k drugi ekipi, uspešnost igralca v ekipi itd.) in k iskanju prednosti ekipe, sem v šestem poglavju s pomočjo programa SPSS na podatkovni bazi 551 igralcev lige NBA izvedel linearno regresijo in grozdenje (ang. clustering) ter v podatkih poskušal analizirati odnose.

Kot metodo zbiranja podatkov sem uporabil študijo literature primarnih in sekundarnih virov ter analizo na področju podatkovnega rudarjenja v športu.

1.2 Cilji in namen

Moj namen je predstaviti podatkovno rudarjenje, ki ga lahko podjetja, organizacije ali pa uporabniki z ustreznim znanjem v prid izkoriščajo v športu, zato sem si zastavil sledeče raziskovalno vprašanje, ki sem ga preverjal skozi diplomsko delo: *Ali je podatkovno rudarjenje v športu uspešno?*

2 TEORETIČNI DEL

2.1 Podatkovno rudarjenje

Preden začnem z opredelitvijo koncepta podatkovnega rudarjenja, bom najprej podal definicijo konceptov, ki so pomembni za razumevanje obravnavane tematike.

Podatkovna baza pomeni računalniški sistem, v katerem se shranjujejo podatki. Gre za urejeno zbirko podatkov, ki so shranjeni na strežniku. Podatkovna baza uporabniku omogoča hiter dostop do informacij (Beynon-Davies 2004).

Podatkovno skladišče je predmetno usmerjena, integrirana, časovno neomejena zbirka podatkov, namenjena podpori odločanja v procesih poslovanja. Omogoča shranjevanje, organizacijo in upravljanje podatkov (Pujari 2004).

Algoritem pomeni eno od različnih orodij namenjenih obdelavi podatkov, kot npr. nevronske mreže, odločitveno drevo, metoda najbližjih sosedov (Pujari 2004).

V literaturi zasledimo več definicij podatkovnega rudarjenja, ki pa so si med seboj zelo podobne, zato bom v nadaljevanju predstavil le nekatere.

Witten in drugi (2011, 36−37) definirajo podatkovno rudarjenje kot »izločanje pomembnih in uporabnih informacij iz velike količine podatkov na svetovnem spletu«. Etzioni (1996, 1) na drugi strani podatkovno rudarjenje opredeli kot »uporabo različnih tehnik za samodejno

odkrivanje in izločanje uporabnih informacij iz dokumentov in servisov na svetovnem spletu«. Nadalje Chen in drugi (1996, 866) podatkovno rudarjenje imenujejo »odkritje znanja v podatkovnih bazah« (ang. Knowledge discovery in databases) oz. definirajo podatkovno rudarjenje kot »proces izločanja pomembnih in nevsakdanjih informacij iz podatkovnih baz«. Avtorja Han in Kamber (2001, 7) pa podatkovno rudarjenje opredelita kot »raziskave in analize, ki vsebujejo velike količine podatkov shranjenih v podatkovnih bazah, podatkovnih skladiščih ali ostalih informacijskih odlagališčih, z namenom odkritja novih znanj vključno s pomembnimi vzorci in pravili«.

Opredelitve podatkovnega rudarjenja zaključujem z avtorjem Pujari (2001, 44−45): »Podatkovno rudarjenje je iskanje odnosov in globalnih vzorcev, ki obstajajo znotraj podatkovnih baz, skriti med ogromno količino podatkov. Pomembno je namreč, da znamo s pomočjo različnih tehnik podatkovnega rudarjenja (nevronske mreže, regresija, odločitveno drevo) iz ogromne količine podatkov izločiti le informacije in odnose, ki so za nas pomembni.*«*

Kot je razvidno iz zgornjih definicij, podatkovno rudarjenje uporabnikom omogoča veliko koristnih funkcij, med drugim: »Rudarjenje prinaša uporabnikom iskanje skritih vzorcev, povezav, profila obnašanja kupcev ipd., torej ključne konkurenčne prednosti poslovanja: večji obseg prodaje, znižanje stroškov, večje zadovoljstvo strank, pro-aktivno reagiranje na določene situacije itd.« (Finance.si 2007). Poleg slednjega lahko z uporabo podatkovnega rudarjenja prepoznamo odnose med podatki v podjetju. Kot primer vzemimo podatkovno bazo nakupov v trgovini. Če kupci kupujejo izdelke A in izdelke B, kateri izdelek C bodo kupci najverjetneje tudi kupili? Na takšna in drugačna vprašanja nam podatkovno rudarjenje poda odgovore, ki so nam lahko v veliko pomoč pri marketinških strategijah.

Že iz zgornjega primera je razvidno, da podatkovno rudarjenje v veliki meri sloni na statističnih konceptih in metodah. Tudi če se uporablja v drugih metodah, so statistične metode vedno prisotne pri analizi podatkov.

2.2 Razvoj podatkovnega rudarjenja

Podatkovno rudarjenje je rezultat dolgoletnih raziskav na področju informacijskih sistemov. Zametki podatkovnega rudarjenja segajo v čas prvih shranjevanj podatkov na računalnikih. V Tabeli 2.1 prikazujem razvojne faze podatkovnega rudarjenja (Thearling 2010).

Tabela 2.1: Razvojne faze podatkovnega rudarjenja

Faza	Tehnologije (orodja), ki omogočajo posamezne faze	Lastnosti tehnologij
Zbiranje podatkov 1960 (ang. Data Collection)	računalniki, diski	statični podatki
Dostop do podatkov 1980 (ang. Data Access)	relacijske podatkovne baze, strukturirani povpraševalni jezik (SQL)	dinamični podatki, rekordna dostava podatkov
Podatkovna navigacija 1990 (ang. Data Navigation)	OLAP (spletna analitična obdelava podatkov), podatkovna skladišča, multidimenzijske podatkovne baze	dinamični podatki, več nivojska dostava podatkov
Podatkovno rudarjenje 2000 (ang. Data Mining)	napredni algoritmi, masivne podatkovne baze	proaktivna dostava informacij

Leta 1960 se je pričela faza »zbiranja podatkov«, v kateri so zbrane podatke shranjevali na trde diske računalnikov. Podatke so podjetja zbirala statično, s pomočjo anketnih vprašalnikov, kar je bilo dolgotrajno in drago.

Danes je zbiranje podatkov nekoliko drugačno in se pri večjem številu enot izvaja preko spleta, kjer posamezne spletne strani zbrane podatke uporabijo za enostaven izračun povprečij ali vsot iskanih zadetkov, ki jih uporabniki vpišejo v spletni iskalnik. Z izračunom povprečij

ali vsot iskanih zadetkov lahko na primer dobimo odgovore na vprašanja o skupnem ali povprečnem dohodku (podjetja ali posameznika) v nekem obdobju.

Fazi zbiranja podatkov je sledila faza »dostopa do podatkov«, kjer so relacijske podatkovne baze shranjevale podatke v strukturiran format, ki jih je kasneje podjetje uporabljalo za pregled uspešnosti prodaje podjetja. Poznali so tudi strukturiran povpraševalni jezik (SQL), ki je omogočal izdelavo različnih poizvedb med podatki.

V začetku 90-ih let je prišlo do razvoja »podatkovne navigacije«, ki je z različnimi analitičnimi orodji (SPSS) in podatkovnimi bazami omogočala zbiranje in natančnejšo analizo večje količine podatkov (primerjava med podjetji glede na uspešnost prodaje itd.).

Pri vseh treh fazah lahko analize podatkov opravimo le za pretekle dogodke, kar pomeni, da podjetje ne more dobiti povratne informacije s strani podjetja ali posameznika v sedanjosti.

V fazi »podatkovnega rudarjenja« pa je izmenjava informacij v sedanjosti, kar pomeni, da lahko z naprednimi algoritmi analiziramo ogromno količino podatkov v realnem času in na podlagi tega napovemo verjetnost dogodkov v sedanjosti ali prihodnosti.

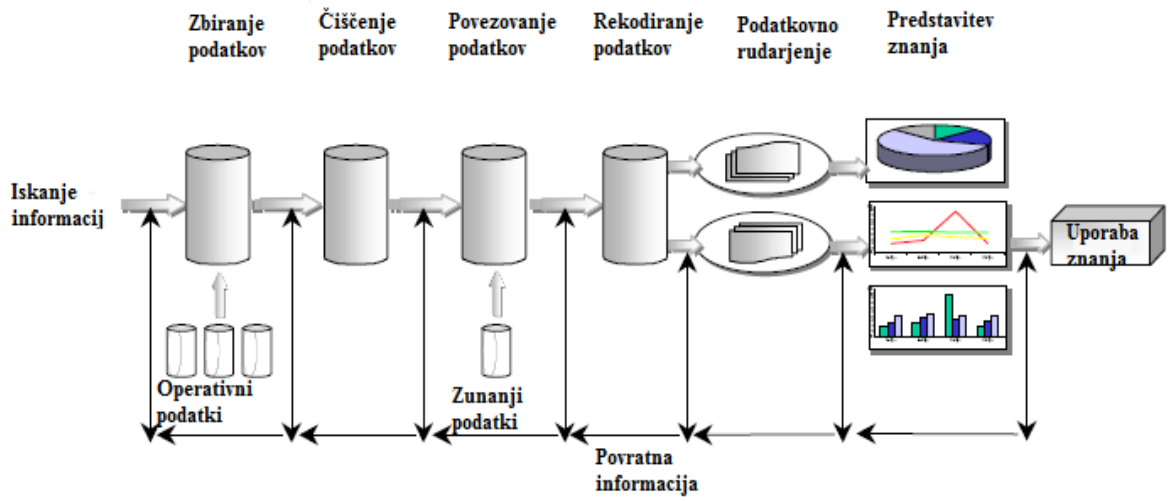
2.3 Potek podatkovnega rudarjenja

Podatkovno rudarjenje uporabniku omogoča odkrivanje novih, zanimivih spoznanj, kot so vzorci, pravila, spremembe in napake v ogromni količini podatkov, shranjenih v podatkovnih bazah. Celoten proces odkrivanja in iskanja za uporabnika pomembnih informacij pa poteka na naslednji način (Velickov in Solomantine 2000):

1. ***Zbiranje podatkov**, kjer so podatki, pomembni za analizo, zbrani iz podatkovne baze.*
2. ***Čiščenje podatkov**, kjer se nepravilni ali manjkajoči podatki odstranijo.*
3. ***Povezovanje podatkov**, kjer se več različnih podatkov poveže v skupen podatek.*
4. ***Rekodiranje podatkov**, kjer se podatke rekodira ali prečisti v oblike, primerne za različne podatkovne algoritme (tehnike).*

5. ***Podatkovno rudarjenje**, ki je ključni proces, v katerem so uporabljene inteligentne tehnike za zbiranje skritih in pomembnih znanj iz podatkov.*
6. ***Predstavitev znanja** (ang. knowledge representation), kjer so uporabljene različne tehnike podatkovnega rudarjenja z namenom predstavitve znanja uporabniku.*

Slika 2.1: Proces pridobivanja znanja



Vir: Velickov in Solomantine (2000).

2.4 Kategoriji podatkovnega rudarjenja

Podatkovno rudarjenje s pomočjo svojih orodij iz podatkovne baze vzame podatek, ga oblikuje v algoritem (odločitveno drevo, nevronske mreže, regresija itd.) in ga predstavi. Natančneje, iz podatkov izloči znanje v obliki vzorcev, ki razlagajo vzroke in posledice, kar pa izkoristimo za pridobivanje znanja. Tako lahko glede na cilje analize podatkovnega rudarjenja oblikujemo dve glavni kategoriji rudarjenja:

- opisno podatkovno rudarjenje in
- napovedno podatkovno rudarjenje.

Opisno podatkovno rudarjenje pomeni iskanje povezav in korelacij (grozdenje), ki opisujejo podatke. S tem odkrijemo podatke, ki so nenavadni, in jih izločimo (Rygielski in drugi 2002). Glavni namen opisnega podatkovnega rudarjenja je na podlagi podatkov poiskati uporabniško razumljive vzorce, ki opisujejo podatke (poiskati podatke, ki so razumljivi uporabniku) (Giudici in Figini 2009)

Opisno podatkovno rudarjenje se pogosto uporablja, ko smo soočeni z naslednjimi vprašanji (Berry in Linoff 2000):

- Kaj je podatek?
- Kakšen je podatek po videzu oz. obliki?
- Ali podatki vsebujejo nenavadne vzorce?
- Kaj nam podatek pove o informaciji, ki nas zanima (npr. o strankah, košarkarski tekmi)?

Napovedno podatkovno rudarjenje pa pomeni gradnjo modelov (odločitveno drevo, nevronske mreže, linearna regresija, logistična regresija itd.) na podlagi podatkov ali spremenljivk. Na podlagi dobljenih rezultatov pa sklepamo, s kolikšno verjetnostjo se bo nekaj zgodilo v prihodnosti, oz. predvidevamo, kakšne bodo neznane vrednosti (Rygielski in drugi 2002). Napovedno podatkovno rudarjenje analizira eno ali več spremenljivk v odnosu z drugimi spremenljivkami, dobljene podatke pa uporabi za napoved dogodkov v prihodnosti (Giudici in Figini 2009).

Preden se odločimo za uporabo napovednega podatkovnega rudarjenja, Edelstein (1999) predlaga, da sledimo naslednji hierarhični lestvici odločitev, ki nam podrobneje predstavi analiziran model in s tem omogoča boljše rezultate rudarjenja:

- Poslovni cilj.
- Vrsta napovedi.
- Vrsta modela.
- Algoritem.
- Produkt.

pomembnost odločitev

Najvišje na lestvici je »**poslovni cilj**«. Vedno se vprašamo *Kaj je namen rudarjenja tega podatka?* Za primer vzemimo iskanje vzorcev v podatkih naših strank. Podatki nam lahko razkrijejo stranke, ki v naši trgovini veliko zapravijo. Na podlagi teh vzorcev lahko naredimo dva modela. Enega, ki bo napovedal verjetnost zapravljanja strank, in drugi model, ki bo odkril stranke, ki najverjetneje ne bodo zapravile nič. Poslovni cilj organizacije določa model in njegov cilj, ki ga bomo uporabili.

Naslednji korak je odločanje na podlagi »**vrste napovedi**«, ki je najbolj primerna: (1) klasifikacija: napoved v katero kategorijo ali skupino spada podatek ali (2) regresija: napoved vrednosti spremenljivke. Če uporabim prejšnji primer, lahko uporabimo klasifikacijo za napoved, katera stranka v trgovini najverjetneje ne bo nič zapravila in regresijo za napoved, koliko bo znašal znesek, ki ga bo stranka v naši trgovini najverjetneje zapravila.

Ko smo določili poslovni cilj in tip napovedi, lahko izberemo »**vrsto modela**«. Uporabimo nevronsko mrežo za izvedbo regresije in odločitveno drevo za klasifikacijo. Lahko pa izbiramo tudi med različnimi statističnimi metodami, kot so: logistična regresija, diskriminantna analiza ali splošni linearni modeli.

Za gradnjo modelov je na voljo velika izbira »**algoritmov«**. Nevronske mreže lahko naredimo z uporabo vzvratnega učenja (ang. backpropagation). Model odločitvenega drevesa pa lahko naredimo na podlagi CART ali CHAID modelov.

In še najmanj pomembna odločitev na lestvici, »**produkt**« podatkovnega rudarjenja. Razlikujemo med tremi tipi produktov podatkovnega rudarjenja. Prvi tip produktov imenujemo »**orodja**« (ang. tools), ki so analitični pripomočki za OLAP (ang. On Line analytical processing). Kratica OLAP označuje programsko orodje, ki omogoča hitro analiziranje velike količine podatkov. OLAP orodja uporabnikom omogočajo interaktivno analizo večdimenzionalnih podatkov iz več vidikov. S serijo hipotez in odnosov, ki jih orodje postavi med podatki v podatkovni bazi, preverja pravilnost ali nepravilnost le-teh (Edelstein 1999).

Drugi tip produktov imenujemo »**čisti**« produkti podatkovnega rudarjenja. Gre za horizontalna orodja, ki se ukvarjajo z različnimi problemi, na primer: reševanje težav v odnosu do strank (ang. customer relationship management problems). Eno izmed takšnih orodji je paket SPSS (ang. Statistical Package for Social Science).

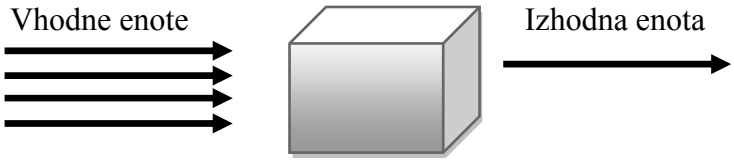
V zadnji podatkovni tip pa spadajo »**analitične aplikacije**«. Gre za aplikacije, ki omogočajo izvajanje specifičnih poslovnih procesov, katerih sestavni del je podatkovno rudarjenje.

2.5 Usmerjeno in neusmerjeno podatkovno rudarjenje

V literaturi zasledimo dva pristopa, ki se uporabljata pri podatkovnem rudarjenju. Prvi pristop se imenuje *usmerjeno podatkovno rudarjenje* (ang. directed data mninig) in drugi pristop *neusmerjeno podatkovno rudarjenje* (ang. undirected data mninig).

Pri **usmerjenem podatkovnem rudarjenju** govorimo o pristopu »od zgoraj navzdol«. Uporabimo ga takrat, kadar točno vemo, kaj iščemo. Primer usmerjenega podatkovnega rudarjenja prikazujemo na Sliki 2.2.

Slika 2.2: Model črne skrinjice (ang. Black box)



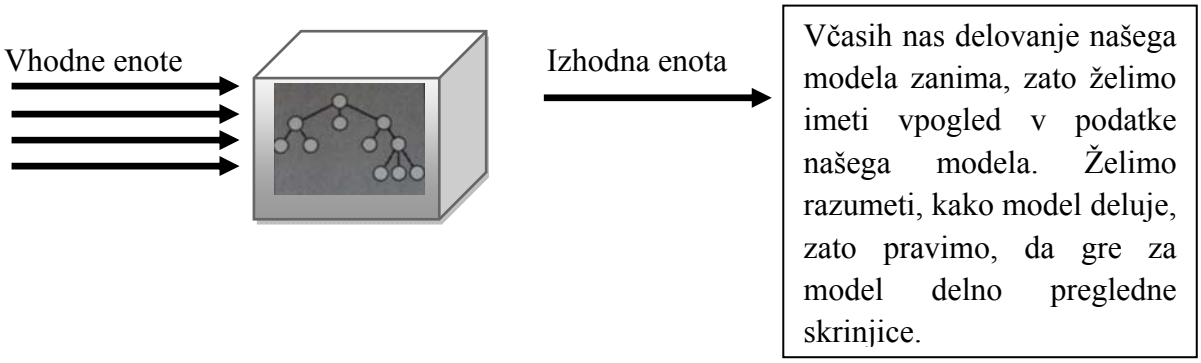
Včasih nas delovanje našega modela ne zanima, kar pojmujemo črna skrinjica, saj želimo dobiti le najboljše rezultate.

Vir: Berry in Linoff (2000).

Model prejme več vhodnih enot in naredi eno izhodno enoto, kar z drugimi besedami pomeni, da uporabnika ne zanima, kaj model počne (ne zanimajo ga kakšni so podatki v modelu), zanima ga le točnost/natančnost dobljenih podatkov (natančnost izhodne enote).

Za pristop **neusmerjenega podatkovnega rudarjenja** pa se odločimo v primeru, ko nas poleg rezultatov zanimajo tudi podatki, ki se nahajajo v našem modelu. Zanima nas delovanje modela (glej Sliko 2.3).

Slika 2.3: Model delno pregledne skrinjice (ang. semitransparent box)



Vir: Berry in Linoff (2000).

3 ALGORITMI IN TEHNIKE PODATKOVNEGA RUDARJENJA

Kaj podatkovno rudarjenje prinaša uporabnikom/podjetju? Uporaba algoritmov in tehnik nam omogoča, da lahko z opisnim ali napovednim podatkovnim rudarjenjem v kratkem času analiziramo ogromno količino podatkov in podamo ugotovitve.

V nadaljevanju bom predstavil nekaj najpogosteje uporabljenih tehnik za odkrivanje znanja v podatkovnih bazah: klasifikacija, grozdenje, nevronske mreže, odločitvena drevesa in metoda voditeljev.

3.1 Klasifikacija

Klasifikacija je najpogosteje uporabljena tehnika podatkovnega rudarjenja, ki za svoje delovanje uporablja množico matematičnih metod (odločitveno drevo, linearno programiranje, nevronske mreže in statistiko). Klasifikacija se uporablja za razvrščanje elementov (v množici podatkov) v vnaprej poznane razrede ali skupine (Ramageri 2010). Natančneje klasifikacija razvrsti vsak element velike količine podatkov v vnaprej določene skupine.

Klasifikacijo lahko uporabimo v primeru, ko nas zanima seznam vseh zaposlenih, ki so bili v preteklosti največkrat odsotni od dela zaradi bolezni. Klasifikacija na podlagi podatkov od odsotnosti predvideva, kateri zaposleni bodo tudi v prihodnosti najverjetneje izostali od dela zaradi bolezni. V tem primeru naredimo klasifikacijo v dve skupini, in sicer lahko naredimo skupini »odsotnost« in »prisotnost«.

Algoritmi klasifikacijskih modelov (Ramageri 2010):

- Odločitveno drevo.
- Nevronske mreže.
- Metoda podpornih vektorjev.

3.2 Grozdenje

Grozdenje (ang. clustering) lahko opredelimo kot tehniko, ki išče pomembne in uporabne grozde primerov (skupine) s podobnimi lastnostmi (Ramageri 2010). Cilj grozdenja je poiskati med seboj različne grozde (skupine), ki vsebujejo zelo podobne podatke.

Z grozdenjem lahko iščemo značilne skupine kupcev (ali so redni ali naključni kupci), lahko iščemo značilnosti pacientov in jih na podlagi grozdov razdelimo v dve skupini (tisti, ki potrebujejo operacijo, in tisti, ki operacije ne potrebujejo).

Za razliko od klasifikacije, pri grozdenju na začetku postopka ne vemo, kakšne skupine bomo imeli in na podlagi katerih lastnosti jih bomo grozdili oz. združevali v skupine. Omenjeno razliko bom pojasnil z naslednjim primerom: v knjižnici imamo kot člani na voljo knjige različnih avtorjev s širokim naborom tem. Radi bi zagotovili, da bodo člani knjižnice brez težav lahko poiskali več knjig različnih avtorjev s podobno tematiko. Z uporabo grozdenja moramo knjige najprej razvrstiti v grozd oz. skupino, jih na podlagi podobnosti tematike označiti z uporabnim imenom ter vse knjige s podobno tematiko postaviti na eno polico. S tem smo članom knjižnice omogočili namesto iskanja knjig po celotni knjižnici hiter dostop do knjig.

3.3 Nevronske mreže

Nevronske mreže lahko kot tehniko uporabimo v več primerih, zlasti pa jih uporabljamo v opisnem ali napovednem podatkovnem rudarjenju. Uporabniku omogočajo učinkovito modeliranje velikih in kompleksnih problemov, ki lahko vsebujejo na stotine med seboj povezanih odvisnih spremenljivk (Giudici in Figini 2009).

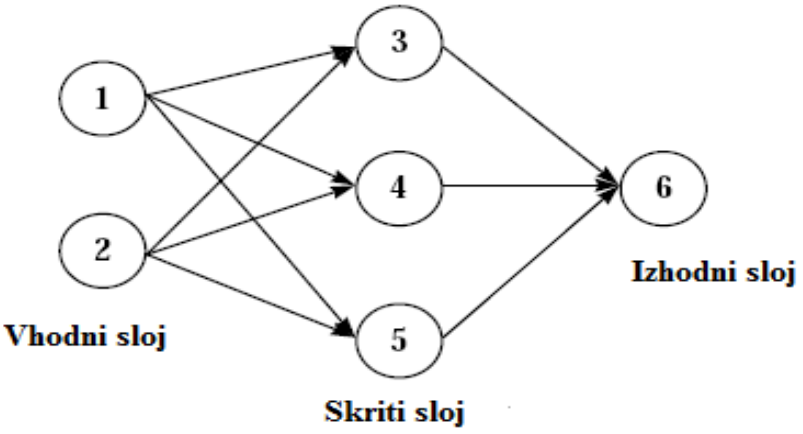
Nevronske mreže lahko opredelimo kot niz osnovnih, med seboj povezanih računskih enot, ki jih imenujemo nevroni (Giudici in Figini 2009).

Uporabimo jih lahko pri tehniki klasifikacije (rezultat tehnike je nominalna spremenljivka) ali pri regresijski analizi (zvezna spremenljivka).

Nevronska mreža je sestavljena iz treh ali več slojev (Edelstein 1999) (glej Sliko 3.1):

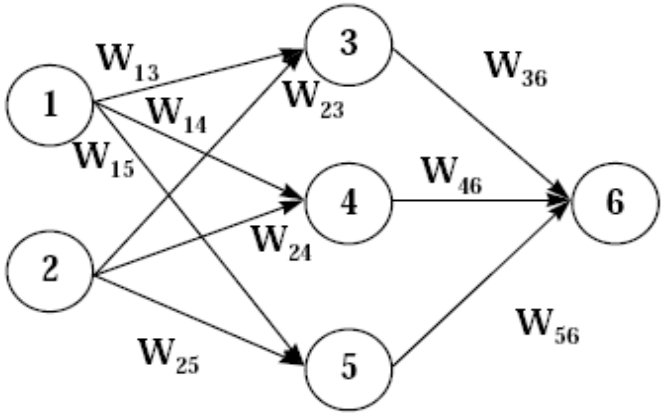
- Začne se z **vhodnim slojem (**ang. input layer), kjer vsako vozlišče predstavlja neodvisno spremenljivko. Vozlišča vhodnega sloja so povezana z vozlišči skritega sloja.
- Preide v **skriti sloj** (ang. hidden layer), kjer so vozlišča skritega sloja povezana tudi z vozlišči izhodnega sloja.
- Konča se z **izhodnim slojem** (ang. output layer), ki je sestavljen iz ene ali več odvisnih spremenljivk.

Slika 3.1: Nevronska mreža z enim skritim slojem



Vir: Edelstein (1999).

Slika 3.2: Delovanje nevronske mreže



Vir: Edelstein (1999).

Ponazoritev delovanja nevronske mreže prikazujem s Sliko 3.2. Številke od 1 do 6 pomenijo vozlišča mreže, *W* pa pomeni povezovalno utež posameznega vozlišča (gre za neznane parametre, katerih vrednosti pridobimo z metodo vzvratnega učenja). Mreža deluje tako, da vsako vozlišče v vhodnem sloju pomnoži s povezovalno utežjo W_{xy} (utež od vozlišča x do vozlišča y), vozlišča združi skupaj, na njih uporabi aktivacijsko funkcijo in vrednost vozlišča prenese na vozlišče v naslednjem sloju. Če zadevo poenostavimo, lahko delovanje mreže predstavimo z naslednjim primerom:

Zanima nas, kakšna je vrednost vozlišča, ki se prenese iz vozlišča 4 na vozlišče 6. To lahko izračunamo po naslednjem postopku:

Aktivacijska funkcija pomnožena z
([W_{14}* vrednost vozlišča 1]+[W_{24}*vrednost vozlišča 2]).

Arhitektura nevronske mreže je sestavljena iz števila vozlišč, skritih slojev in medsebojnih povezav. Pri sestavi mreže je potrebno izbrati število skritih vozlišč, aktivacijsko funkcijo in omejitve na utežeh.

3.4 Odločitvena drevesa

Pri odločitvenih drevesih prikažemo s serijo pravil opazovano vrednost ali razred. Gre za hierarhično prikazovanje vrednosti ali razredov, ki jih na podlagi njihovih vrednosti ustrezno klasificiramo. V odločitvenem drevesu list vozlišča označuje odločitev (ali klasifikacijo), medtem ko vozlišče brez listov označuje lastnost, na podlagi katere se odločamo (barva, velikost itd.). Cilj klasifikacije je po najkrajši poti priti do lista vozlišča, saj tako delo opravimo v najkrajšem možnem času (Berry in Linoff 2000).

Vendar v literaturi zasledimo dve kategoriji odločitvenih dreves, in sicer (Berry in Linoff 2000):

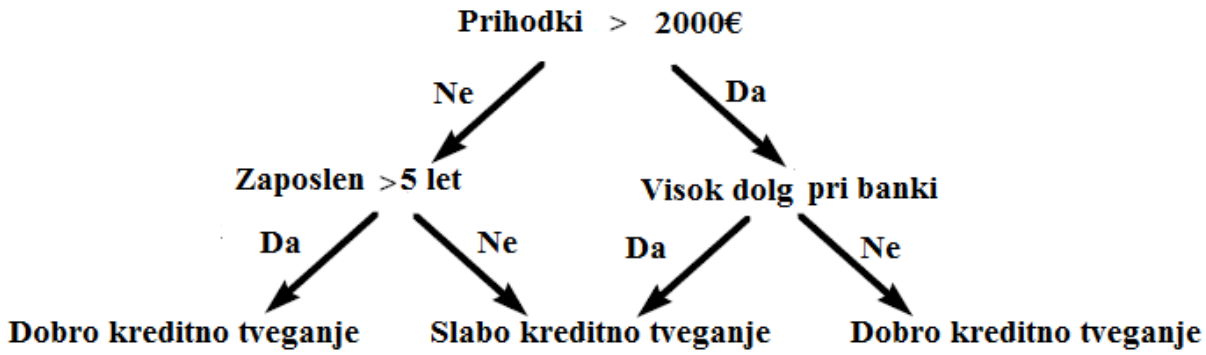
- **Klasifikacijska drevesa**, ki označujejo podatke in jih dodeljujejo ustreznim razredom.

- **Regresijska drevesa**, ki ocenjujejo vrednosti ciljnih spremenljivk s številskimi vrednostmi. Tako lahko regresijska drevesa izračunajo približen znesek, ki ga bo sponzor podelil nekemu društvu.

Vzemimo naslednji primer za ponazoritev klasifikacijskega odločitvenega drevesa:

Banka želi klasificirati odobritev kreditov strankam glede na dobra ali slaba kreditna tveganja (glej Sliko 3.3). Prihodke večje od 2000 € imenujemo vrh vozlišča in določajo poskus, ki se bo izvedel. Izvedemo poskus in kot rezultat dobimo delitev drevesa na dve veji, ki predstavljata enega od možnih odgovorov poskusa. Če ima stranka, ki prosi za kredit, večje prihodke od 2000 € in je dolg, ki ga dolguje banki, visok, banka ne bo odobrila kredita svoji stranki, saj odobritev posojila za banko pomeni slabo tveganje in s tem še višji dolg stranke. Banka bo tako stranko klasificirala v skupino »slaba kreditna tveganja«.

Slika 3.3: Primer klasifikacijskega drevesa

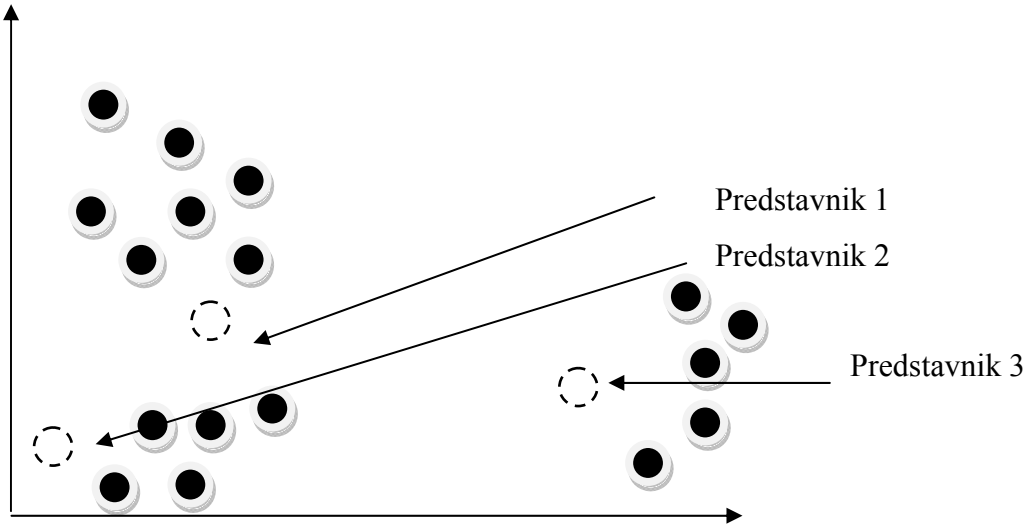


3.5 Metoda voditeljev

Metoda voditeljev je

> *iteracijska metoda, kjer se je potrebno odločiti, v koliko skupin razvrščamo enote. Postopek se začne z vnaprej podano množico predstavnikov posameznih skupin - voditeljev. Metoda priredi enote najbližjim voditeljem, poišče centroide (težišča) tako dobljenih skupin - nove voditelje, zopet priredi enote najbližjim voditeljem itd. Postopek se konča, ko se nova množica voditeljev ne razlikuje od množice voditeljev, dobljene korak pred njo* (Ferligoj 1989, 93).

Slika 3.4: Začetek postopka metode voditeljev (Berry in Linoff 2000)



V zgornjem primeru je razvidno, da smo postopek razvrščanja začeli z izborom treh predstavnikov posameznih skupin.

4 PREDSTAVITEV PODATKOV

Tukaj se lahko vprašamo *Kako zapišemo/predstavimo podatke, pridobljene s pomočjo podatkovnega rudarjenja?* Poglejmo si naslednjo tabelo (glej Tabelo 4.1) (Berry in Linoff 2000):

Tabela 4.1: Tipični zapis podatkov podatkovnega rudarjenja

2610000101	010377	14		A	19,1		14 Spring	Pravilno
2610000102	154566	7		A	19,1		NULL	Pravilno
2610000103	018548	1		B	21,2		71 W. 19.St	Napačno
2610000104	159494	1		S	38,3		3562 Osk	Napačno
2610000105	031511	22		C	56,1		9672 W 142	Napačno
2610000106	131212	45		C	56,1		NULL	Pravilno
2610000180	080897	6		A	19,1		PO, BOX	Napačno
26100001183	123059	3		D	10,0		560 Robson	Pravilno
2610000000	020948	2		S	38,3		222 E. 11th	Pravilno

Vir: Berry in Linoff (2000).

Vsi algoritmi podatkovnega rudarjenja podatke prikažejo z vrsticami in stolpci. Podatki, ki so zapisani v stolpcih zgornje tabele, so naslednji:

- Prvi stolpec vsebuje identifikacijsko številko (ID) stranke.
- Drugi stolpec predstavlja podatke o stranki.
- V tretjem stolpcu je seštevek transakcij, ki so jih stranke opravile.
- V petem in šestem stolpcu so podane vrednosti referenčnih tabel.

- V osmem stolpcu so podane unikatne vrednosti stranke (naslovi prebivališča itd.).
- Deveti stolpec predstavlja cilj podatkovnega rudarjenja oz. predstavlja naše predvidevanje na podlagi dobljenih podatkov.
- Vrstice, ki so prazne, vsebujejo napačne identifikacijske številke strank, zato so prazne (izpuščene).

5 PODATKOVNO RUDARJENJE IN ŠPORT

Tehnike podatkovnega rudarjenja se uspešno uporabljajo v številnih znanstvenih, industrijskih in poslovnih področjih. Področje vrhunskega športa je znano po velikih količinah podatkov, zbranih za vsakega igralca, skupino, igro in sezono. Vedno več športnih organizacij se je pričelo zavedati, da je bogastvo neizkoriščenega znanja v podatkih, ki jih imajo, zato se vedno bolj povečuje zanimanje za tehnike, ki te podatke lahko koristno predstavijo.

Zaradi visoko tekmovalnega okolja in ogromnih količin denarja v športu, so športne organizacije primorane iskati rešitve, na podlagi katerih lahko pridobijo prednost pred drugimi moštvi. Dolga desetletja je prevladovalo prepričanje, da sta športno znanje in napredek športnikov odvisena le od skavtov (ki iščejo nove in talentirane športnike), trenerjev in lastnikov kluba. Šele z napredkom znanja in znanosti pa so se športne organizacije začele vedno bolj zavedati, da so ključni podatki in znanje njihovi igralci. Začelo se je t.i. iskanje praktičnih metod za razvoj znanja. V začetku so bili to statistiki, ki so jih organizacije najele za analizo podatkov in s tem pridobitev znanja. Vendar pa sta denar in tekmovalnost športne organizacije kmalu vodila k iskanju še bolj praktičnih metod zbiranja in analiz podatkov in s tem pripeljala do vedno bolj uveljavljene tehnike podatkovnega rudarjenja v športu.

Svet športa je poznan po veliki količini statističnih metod, ki se uporabljajo za različne analize posameznih igralcev, ekip, tekem itd. Pri vsakem športu obstajajo različni tipi statistik, ki se uporabljajo za analizo podatkov. Tako lahko s podatkovnim rudarjenjem iščemo povezave med igralci ali pa napovemo verjetnost dogodkov, ki se lahko zgodijo (poškodbe, prestopi igralcev, nezadovoljstvo igralcev, doseženo število točk itd.). Če vzamemo primer podatkovnega rudarjenja pri košarki, lahko pri igralcu košarke analiziramo podatke o doseženih točkah, o številu asistenc, o številu ukradenih žog, poškodbah, prestopih, za vsako tekmo posebej, na podlagi tega pa s podatkovnim rudarjenjem napovemo verjetnost njegove poškodbe ali število doseženih točk, asistenc, prekrškov itd. na naslednji tekmi ali v sezoni.

Žal pri nas nisem zasledil literature na temo podatkovnega rudarjenja v športu, se pa zato tovrstna metoda v veliki meri uporablja v tujini, predvsem v ameriških športih.
Če se osredotočim na namen, ki ga podatkovno rudarjenje ima, lahko rečem, da cilj podatkovnega rudarjenja v športu ni prevladovanje nad odločitvami lastnikov, trenerjev in skavtov, temveč je uporaba podatkovnega rudarjenja namenjena kot pripomoček v procesu

odločitve lastnikov kluba, trenerjev in skavtov. Podatkovno rudarjenje je v športu že tako visoko vrednoteno, da lastniki klubov brez konkretnih analiz podatkovnega rudarjenja ne sprejemajo odločitev.

5.1 Podatkovno rudarjenje v športu

Šport je zakladnica različnih podatkov. Ti podatki lahko pokažejo/napovejo individualne kvalitete določenih igralcev, verjetnost dogodkov, ki se bodo na tekmi zgodili, ali pa, kako deluje ekipa kot celota. Pomembno je poznavanje pomembnosti podatkov oz. katere podatke bomo analizirali (uporabili), da bomo prišli do praktičnega znanja, ki nam bo pomagalo pri napredku ekipe. Vsi ti podatki so pridno izkoriščeni s strani organizacij, ki dobljeno znanje uporabljajo kot prednost v primerjavi z drugimi ekipami.

Ker pa različne športne organizacije podatke različno obravnavajo, lahko odnos organizacij do podatkov razdelimo na pet različnih pristopov (Schumaker in drugi 2010):

- Med športnimi podatki in njihovo uporabo ni povezanosti.
- Strokovnjaki (trenerji, skavti, lastniki kluba) z določenega področja skušajo podati napoved dogodkov v športu na podlagi svojih izkušenj.
- Strokovnjaki z določenega področja za napoved dogodkov v športu uporabljajo že zbrane podatke.
- Uporaba statistik v procesu odločanja.
- Uporaba podatkovnega rudarjenja v procesu odločanja.

Prvi tip odnosov organizacij do podatkov je pristop »med športnimi podatki in njihovo uporabo ni povezanosti«. Gre za pristop, ki ga uporabljajo različne amaterske športne organizacije, katerih namen je zabava igralcev in predstavitev športa (npr. rekreacija odraslih, treniranje otrok itd.). Te športne organizacije pogosto zbirajo podatke o svojih igralcih, vendar jih ne analizirajo, saj zbirajo podatke le zaradi tradicije kluba ali pa zaradi beleženja določenih tekem.

Bolj napreden pristop organizacij do podatkov je uporaba strokovnjakov (trenerji, skavti, lastniki kluba) z določenega področja, ki skušajo podati napoved dogodkov v športu na podlagi svojih izkušenj. Odločitve, sprejete na podlagi izkušenj strokovnjakov, ne podpirajo

nobeni konkretni podatki, gre le za domneve strokovnjakov. Takšne odločitve imajo lahko velike posledice na igralca ali ekipo (npr. trener se odloči, da bo zamenjal igralca, ker se mu tako zdi prav; za zamenjavo igralca se je odločil le na podlagi svojega občutka in ne na podlagi konkretnih statistik). Takšne vrste pristop se je uporabljal v preteklosti, ko podatkovno rudarjenje še ni bilo uveljavljeno.

Pristop, ki se uporablja predvsem v organizacijah, ki analizam svojih podatkov ne namenijo veliko denarja, je pristop uporabe strokovnjakov z določenega področja, ki za napoved dogodkov v športu uporabljajo že zbrane podatke preteklih tekem. Podatke, ki so na voljo že nekaj let, trenerji ali pa skavti uporabijo za določene odločitve. Tako imajo igralci, ki so v preteklosti imeli dobre rezultate (visoko povprečje točk na tekmi) prednost pri igranju v ekipi pred igralci, ki so imeli prejšnja leta slabše rezultate ali pa jih v ekipi še ni bilo.

Vedno bolj uveljavljen pristop, ki ga uporabljajo v nogometu, hokeju in košarki, je statistični pristop, ki analizira podatke in ugotavlja pogostosti določenih dogodkov v športu (strategija moštva, igra določenega igralca, število zmag) ali uspešnost celotne ekipe in igralca na posameznik tekmi. Statistika se uporablja kot orodje, ki strokovnjakom olajšuje odločitve.

Zadnji pristop pa v procesu odločanja uporablja tehnike podatkovnega rudarjenja. Ta tehnika se od predhodnih tehnik zelo razlikuje, saj lahko tehnike podatkovnega rudarjenja posplošimo in uvedemo v nove situacije ter iz njih napovemo verjetnost določenih dogodkov. Gre za to, da ne beremo le statistik posameznih igralcev in na podlagi le teh sklepamo o igralčevi uspešnosti na naslednji tekmi, vendar gre za to, da se podatki o ekipi in igralcih vnesejo v različna orodja, ki na podlagi algoritmov odkrivajo skrite vzorce med podatki. Na podlagi tega pa napovedo verjetnosti določenih dogodkov, kot so: uspešnost igralca na naslednji tekmi, strategija moštva na naslednji tekmi, verjetnost, da bo igralec na naslednji tekmi igral itd.

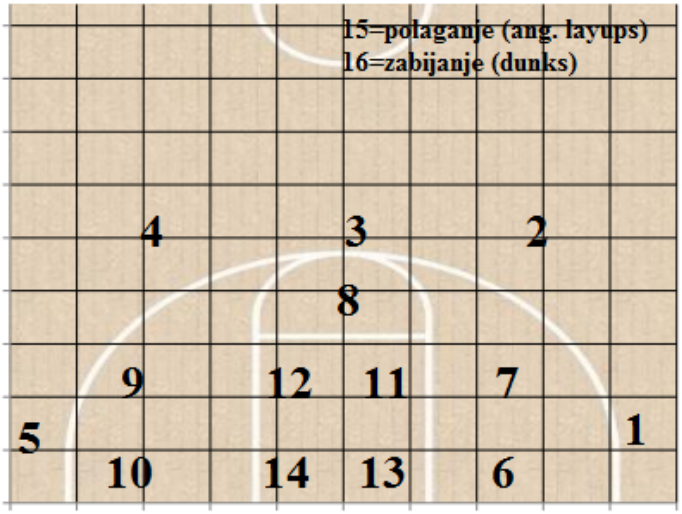
5.2 Primer analize podatkov

Košarkarski podatki (met iz igre, prosti meti, poškodbe, minutaža igralca, asistence, osebne napake itd.), zbrani za več kot 3000 tekem, so organizirani in shranjeni v podatkovni bazi. Ta vsebuje informacije o vsakem dogodku, ki je nastal v kateri koli od več kot 3000 tekem. Tekme so v podatkovni bazi razvrščene po letu in košarkarski sezoni. Na podlagi teh podatkov se podatki igralcev analizirajo na naslednje načine (Schumaker in drugi 2010):

1. Cone meta (ang. shot zones)

Eden od edinstvenih načinov, na katerem lahko analiziramo podatke košarkarske tekme, se imenuje cona meta. Spletna stran 82Games.com, ki je namenjena statističnim analizam, rangiranju in analiziranju posameznih igralcev s pomočjo con meta, analizira odstotek uspešnosti metov igralcev. To naredi na način razdelitve košarkarskega igrišča na 16 področij (glej Sliko 5.1).

Slika 5.1: Polovica analiziranega košarkarskega igrišča



Vir: 82Games.com.

Vsaka od 16 področij ima pri analizi podatkov naslednji pomen (82Games.com):

- Cone 1 do 5 označujejo met vreden tri točke.
- Cone 6 do 10 označujejo met za dve točki.
- Cone 11 do 14 označujejo mete pod obročem.
- Cona 15 označuje polaganje igralca.

- Cona 16 označuje zabijanje igralcev.

Vrednost tovrstne analize podatkov nam pove, na katerih področjih določen igralec najbolje zadeva (ima največji odstotek zadetih metov) in na katerih področjih ima največ zgrešenih metov. To je še posebej koristna informacija za trenerje, saj jim podatki pomagajo postaviti igralca na najboljše mesto.

Za ponazoritev vzemimo naslednji primer:

Tabela 5.1: Cona 1: Levi kot - največ poskusov pri metu za tri točke

Ekipa	Igralec	Met iz igre	Poskus meta iz igre	% Uspešnosti
SA	Bowen	49	108	0.454
TOR	Peterson	38	93	0.409
PHO	Johnson	38	91	0.481
PHI	Korver	34	77	0.442
MIA	E.jones	29	69	0.420

Vir: 82Games.com.

Tabela 5.2: Cona 5: Desni kot - največ poskusov pri metu za tri točke

Ekipa	Igralec	Met iz igre	Poskus meta iz igre	% Uspešnosti
SA	Bowen	46	90	0.511
TOR	Peterson	38	90	0.422
PHO	Johnson	34	89	0.382
PHI	Korver	27	60	0.450
MIA	E.jones	27	59	0.458

Vir: 82Games.com.

V Tabeli 5.1 in 5.2 so zbrane statistike »največ poskusov pri metu za tri točke« iz sezone 2005/06. Iz obeh tabel je razvidno, da je igralec moštva San Antonio Spurs, Bowen, največ svojih poskusov za tri točke vrgel iz cone 1, in sicer 108. Njegova učinkovitost je 45,4 odstotna oz. je od 108 poskusov zadel 49 metov za tri točke. Iz dobljenih rezultatov je razvidno, da Bowen ni med najbolj natančnimi strelci za tri točke, vendar ima v coni 1 dobro uspešnost pri metu za tri točke oz. če bo le imel priložnost, bo največ metov za tri točke izvedel v coni 1. Iz česar sledi, da se bo igralec Bowen (njegova strategija) na tekmah proti nasprotnikom v napadu večino časa zadrževal v (blizu) coni 1, kjer bo poskušal z metom za tri točke. Vendar pa je pri tem potrebno poudariti, da bo zaradi teh podatkov tudi nasprotna ekipa na igralca v coni 1 bolj pozorna in bo zato tudi nasprotnik igral v obrambi bolj agresivno ter Bowenu poskušal preprečiti zadetek za tri točke.

Iz zgornjega primera je razvidno, da statistika ne koristi samo igralčevi ekipi, temveč bo zbrane podatke uporabila tudi nasprotnikova ekipa in na podlagi njih iskala nasprotnikove slabosti.

2. Učinkovitost igralca (ang. Player Efficiency Rating)

PER (ang. Player Efficiency Rating) meri igralčevo učinkovitost na minuto igre. Metoda pri merjenju učinkovitosti igralca upošteva tako pozitivne kot negativne prispevke igralca k ekipni uspešnosti/neuspešnosti skozi celotno sezono. Metoda za merjenje učinkovitosti analizira igralčeve pozitivne prispevke k ekipi, kot so: odstotek zadetih metov iz igre, odstotek zadetih prostih metov, odstotek zadetih metov za tri točke, število asistenc, število blokad, število ukradenih žog in njegove negativne prispevke; odstotek zgrešenih metov, število osebnih napak, število izgubljenih žog. Na podlagi statističnih analiz pa poda igralčevo uspešnost skozi celotno sezono.

3. Plus/Minus rangiranje (ang. Plus/Minus Rating)

Druga metoda oz. način za izračun uspešnosti igralca je sistem plus/minus ocenjevanja, kjer je vsak igralec analiziran na podlagi izračuna doseženega števila točk ekipe z igralcem, minus število doseženih točk nasprotne ekipe. Pozitivna vrednost plus/minus koeficienta pomeni, da z igralcem ekipa igra bolje. Nasprotno pa negativna ocena koeficienta pomeni, da ekipa deluje bolje, če igralec sedi na klopi.

Ponazorimo to z naslednjim primerom: Igralec v ligi NBA vstopi v igro pri rezultatu 80:80. Ko se tekma konča, zmaga igralčeva ekipa z rezultatom 102:80, zato znaša vrednost koeficienta igralca +22. Če pa igralec vstopi v igro pri rezultatu 80:72, kjer vodi igralčeva ekipa za dve točki in je končni rezultat 80:94 za nasprotnikovo ekipo, znaša vrednost koeficienta igralca -16. Vendar pa kritiki očitajo tovrstnemu načinu analize neustreznega vrednotenja igralcev, saj igralcem z visokim številom metov (ki niso nujno vsi zadeti) iz igre daje večje pozitivne vrednosti.
Ko so vse te statistike zbrane, se nekatere ekipe odločijo še za bolj natančnejše analize podatkov, s katerimi bo napoved določenih dogodkov (zmaga ekipe na naslednji tekmi, uspešnost igralca na tekmi, strategija nasprotne ekipe) še bolj verjetna. V ta namen uporabijo orodja, namenjena podatkovnemu rudarjenju.

Za ponazoritev si poglejmo naslednji primer razvrščanja podatkov v orodja podatkovnega rudarjenja in analizo le-teh v ligi NFL (ang. **National Football League)** (Khan 2003):

V študiji z naslovom *Napoved izidov NFL tekem na podlagi nevronskih mrež* je avtor Joshua Khan analiziral sposobnosti nevronskih mrež točne napovedi zmage ali poraza moštva v nogometni ligi. Natančneje je Khan uporabil nevronsko mrežo vzvratnega učenja (ang. backpropagation), da bi dobil vzorce in odnose med podatki, ki bodo napovedali izide prihodnjih tekem na podlagi preteklih predstav ekipe. V ta namen je zbral podatke (statistike) za 208 tekem, ki so jih odigrala naslednja moštva (glej Tabelo 5.3).

Tabela 5.3: Analizirana moštva

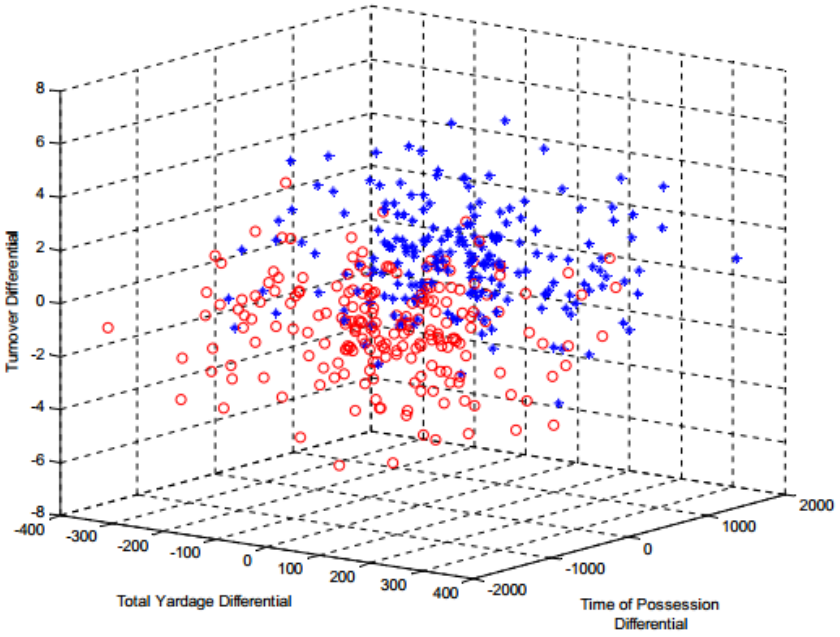
Philadelphia def. Dallas
San Diego def. Detroit
Atlanta def. Carolina
Minnesota def. Seattle
New England def. Jacksonville
New York Jets def. Pittsburgh
Cincinnati def. San Francisco
Oakland def. Baltimore

V algoritem je vključil naslednje spremenljivke:

- Dolžina zavzetega igrišča: spremenljivka zajema uspešnost napadov napadalcev in uspešnost obrambe branilcev obeh ekip na tekmi.
- Razlika zavzetega igrišča: spremenljivka zajema podatke o uspešnosti obeh ekip pri zavzemanju dolžine igrišča in njihove neuspešnosti pri branjenju le-tega.
- Posest žoge: tukaj so bil zbrani podatki o posesti žoge obeh ekip.
- Napadi: sem spadajo podatki o številu ukradenih žog in številu napadov na nasprotnikov gol tako ene kot druge ekipe.
- Teren: podatki o tem, ali je ekipa igrala doma ali v gosteh.

Sestavljene spremenljivke so bile uporabljene z namenom napovedi določenih dogodkov (izidi tekem). Za vsako igro sta bila narejena dva možna izida (ang. output), in sicer izid za ekipo na domačem terenu in izid tuje ekipe.

Slika 5.2: Nevronska mreža skupnih izidov 208 tekem, ki temelji na spremenljivkah: dolžina zavzetega igrišča, razlika zavzetega igrišča in posest žoge.



Vir: Khan (2003).

Zmage moštva so označene z zvezdico in porazi s krogcem.

Rezultati nevronskih mrež so pokazali, da so nevronske mreže predvidele 75 % zmag pravilno. Dobljeni rezultati nevronskih mrež so bili nato primerjani z napovedmi strokovnjakov, ki na spletnem naslovu ESPN.com napovedujejo izide NFL tekem. Prišli so do ugotovitev, da so nevronske mreže bolj pravilno predvidele in napovedale izide tekem kot strokovnjaki iz tega področja, kar nakazuje na to, da so orodja za rudarjenje podatkov, kot so nevronske mreže, ki se lahko uporabljajo ne le za iskanje vzorcev v podatkih, temveč tudi za napovedovanje in predvidevanje v športu, uspešne.

Poglejmo pa si še en uspešen primer podatkovnega rudarjenja. V letu 1997 je trener ekipe Orlando Magic s pomočjo podatkovnega rudarjenja odkril skrite vzorce v zbranih podatkih (uspešnost igralca v igri) igralca Darrella Armstronga. Rezultati podatkovnega rudarjenja so pokazali, da je igralec Darrell zelo koristen za ekipo, saj se verjetnost, da bo ekipa zmagala zelo poveča, ko je Darrell v ekipi. Tako je trener ekipe Darrnellu namenjal vedno večjo minutažo na tekmah, kar je pripeljalo do tega, da je ekipa Orlando v končnici 1997 zmagala dve zaporedni tekmi.

5.3 Orodja, namenjena analizi podatkov

Orodja, namenjena podatkovnemu rudarjenju v športu, kot so: Napredno iskanje (ang. Advaenced Scout), Medsebojna povezava (ang. Synergy Online), B-žoga (BBall), WEKA, Orange, SPSS itd., dobivajo v športu vedno večjo veljavo, saj se vse več športnih organizacij odloči za njihovo uporabo pri analizi ogromne količine podatkov. Uporaba orodij ne koristi samo športnim organizacijam, v veliko pomoč so tako trenerjem kot tudi igralcem samim, ki se s pomočjo aplikacij v igri dodatno izpopolnjujejo.

5.3.1 Napredno iskanje (ang. Advaenced Scout)

Program »Napredno iskanje« je razvilo podjetje IBM (ang. International Business Machines) v sredini 90-ih let kot orodje podatkovnega rudarjenja. Gre za aplikacijo, ki se uporablja v ligi NBA za odkritje zanimivih vzorcev v podatkih košarkarskih tekem. Aplikacija deluje tako, da v podatkih (tekme NBA) odkriva skrite vzorce in trenerjem ekip omogoča podrobnejši vpogled v ekipo (prednosti in slabosti ekipe). Poleg statistik (met iz igre, prosti meti, met za tri točke, izgubljene žoge itd.), ki jih program zbira med tekmo, slednjo tudi posname, kar omogoča podrobnejši pogled same tekme in na ta način tudi boljšo preučitev (pomanjkljivost, kot je npr. zgrešeni meti nasprotnikov) nasprotnikove igre. Predhodno sem omenil, da aplikacija v podatkih odkriva skrite vzorce, kaj so »skriti vzorci v podatkih« pa bom ilustriral z naslednjim primerom (VirtualGold 2012): aplikacija »Napredno iskanje« je na podlagi podatkov trenerju ekipe Orlando Magic pokazala skrite vzorce oz. nekaj kar pred tem niso opazili. Ko sta bila oba igralca Brian Shaw in Darrell Armstrong na parketu, je njun soigralec Penny Hardway igral veliko bolje. Njegova igra je bila bolj učinkovita, meti bolj natančni, imel je več asistenc. Vse to je vplivalo tudi na uspešnost ekipe, saj je začela zmagovati. Ob zamenjavi tako Shawa kot Armstronga z drugima dvema soigralcema pa se je učinkovitost igre Hardwaya poslabšala. Imel je več zgrešenih metov, več izgubljenih žog, tudi ekipa je izgubljala tekme. Če povzamem, je aplikacija pokazala, da so tako igralci kot ekipa v igri bolj uspešni, če sta na parketu prisotna Shaw in Armstrong.

5.3.2 Medsebojna povezava (ang. Synergy Online)

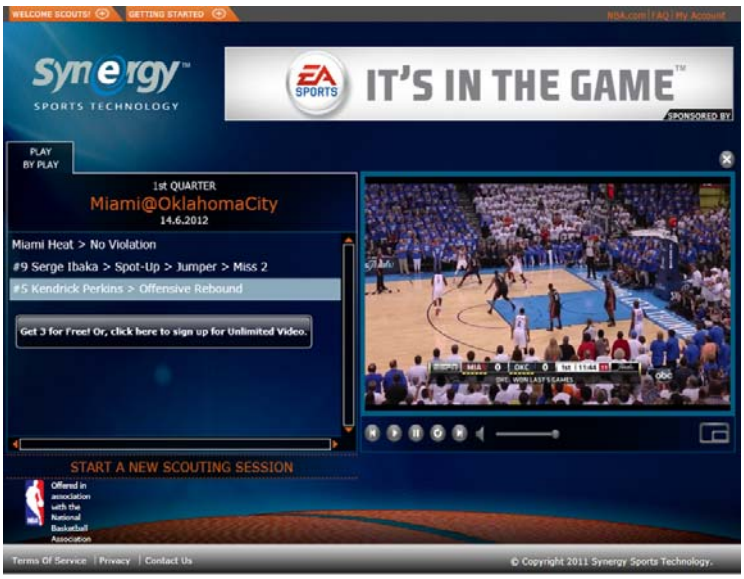
Aplikacija je namenjena za analiziranje košarkarskih tekem na podlagi video predvajanja tekem v živo. Na podlagi tega imajo tako trenerji kot tudi igralci in oboževalci (navijači) vpogled v uspešnost (statistika) posameznega igralca in ekipe že med tekmo (koliko točk je dosegel, koliko žog je izgubil), kar omogoča poizvedbe o uspešnosti igre v realnem času. S tem aplikacija omogoča trenerjem, da med tekmo spremljajo uspešnost svojih igralcev in jim v primeru slabe igre dodatno svetujejo ali pa jih zamenjajo.

Da je aplikacija v veliko pomoč trenerjem in igralcem pa si poglejmo naslednje primere (mysynergysports 2012):

Z aplikacijo »Medsebojna povezava« lahko natančno ugotovimo, kolikokrat je igralec zadel iz nekega mesta na igrišču in njegovo natančnost meta, kar je v veliko korist trenerjem pri postavitvi svojih igralcev na parketu.

Nadalje lahko razčlenimo podatke o igralcu, da ugotovimo, na kateri strani igrišča (levi ali desni) je igralec pri metu bolj natančen. Poleg slednjega lahko ugotovimo, ali bo za določenega igralca v coni bolj verjetno, da bo metal na koš, podal žogo ali pa se premaknil na drugo pozicijo.

Slika 5.3: Aplikacija Synergy Online



Vir: mysynergysports

Iz zgornje slike je razvidno, da aplikacija omogoča zelo podroben vpogled v košarkarsko tekmo. Na levi strani se izpisujejo statistike posameznih igralcev (npr. igralec Serge Ibaka je v skoku vrgel za dve točki in zgrešil), na desni strani pa lahko to opazujemo tudi na posnetku. Iz slednjega lahko razberemo slabosti igralca in njegovo strategijo.

5.3.3 BBall

Gre za orodje, ki temelji na podatkovnem rudarjenju in strojnem učenju. Vključuje obsežen sklop sestavnih delov za pripravo podatkov, filtriranje podatkov, modeliranje podatkov z različnimi algoritmi, vizualizacijo podatkov in še mnogo drugih uporabnih funkcij. Tako lahko s pomočjo programa BBall ugotovimo, kako bi bila v igri uspešna ekipa, če bi v ekipi manjkal najboljši igralec. Program omogoča tudi, da na podlagi podatkov naredimo simulacije tekem in dobimo verjetnost za zmago ekipe v prihodnjih tekmah.

6 EMPIRIČNI DEL

V nadaljevanju predstavljam način zbiranja in analize podatkov na dveh primerih v ligi NBA. Pri prvem primeru sem podatke analiziral z metodo linearne regresije, pri drugem primeru pa so podatki analizirani z metodo grozdenja (ang. clustering).

Najprej sem se lotil analiziranja podatkov s pomočjo linearne regresije, zato sem s pomočjo statističnega paketa SPSS analiziral zbrane podatke na področju košarke, natančneje podatke vseh igralcev v ligi NBA, ki so v sezoni 2011/2012 igrali na tekmah. Podatke za svojo podatkovno bazo sem pridobil na spletni strani databaseBasketball.com, podatki pa so javno dostopni.

Ker se v košarki zbirajo podatki različnih statistik, ki ocenjujejo uspešnost igralca in ekipe, kot na primer:

- število odigranih tekem posameznega igralca, doseženo število točk posameznega igralca (na tekmi in v sezoni), minutaža,
- uspešnost (merjena v številu zadetih metih in odstotkih) pri izvajanju prostih metov, metov za tri točke, metu iz igre,
- število: osebnih napak, asistenc, skokov v napadu in v obrambi, blokad,
- učinkovitost igralca (ang. PER), +/- rangiranje in še več,

je med ogromno količino podatkov potrebno izbrati ustrezne podatke, ki nam bodo kar najbolje (najbolj natančno) odgovorili na zastavljena vprašanja (Berry in Linoff 2000).

6.1 Analiza primera

V košarki trenerji in lastniki klubov beležijo in preučujejo veliko statistik o njihovih igralcih, ki jim nato pomagajo pri njihovih različnih odločitvah, kot so: koliko minut bo posameznik igral na tekmah, na koliko tekmah bo dobil priložnost, ali igralec pripomore k uspešnosti ekipe, kakšna je verjetnost, da se bo igralec na tekmi poškodoval, prednosti igralca itd. Odgovore na ta vprašanja jim ponujajo različni algoritmi podatkovnega rudarjenja, kar v nadaljevanju prikazujem tudi sam.

Želel sem namreč izvedeti, katere spremenljivke so tiste, ki lahko napovedo število minut, ki jih bo posameznik dosegel na tekmah v naslednji sezoni. Če sem bolj natančen me je zanimalo, kako uspešnost igralca, pri metu iz igre na minuto, metu za tri točke na minuto, številu asistenc na minuto, številu skokov v napadu na minuto, številu ukradenih žog na minuto in številu osebnih napak na minuto, vpliva na minutažo igralca. **Da sem dobil uspešnost igralca na minuto sem vse spremenljivke delil s spremenljivko minutaža**. Iz velikega števila spremenljivk sem izbral 7 spremenljivk s katerimi sem preučeval obravnavan primer:

- Minutaza: število minut, ki jih je igralec odigral na tekmah v eni sezoni.
- Met_iz_igre: meti, ki jih je igralec zadel v eni sezoni.
- Met_za_3tocke: zadeti meti za tri točke v eni sezoni.
- Asistence: število asistenc, ki jih je dosegel igralec v eni sezoni.
- Skok_v_napadu: število dobljenih žog igralca pri skoku v napadu v eni sezoni.
- Ukradene_žoge: število žog, ki jih je igralec nasprotnemu igralcu ukradel v eni sezoni.
- Osebne_napake: število osebnih napak, ki jih je igralec naredil v eni sezoni.

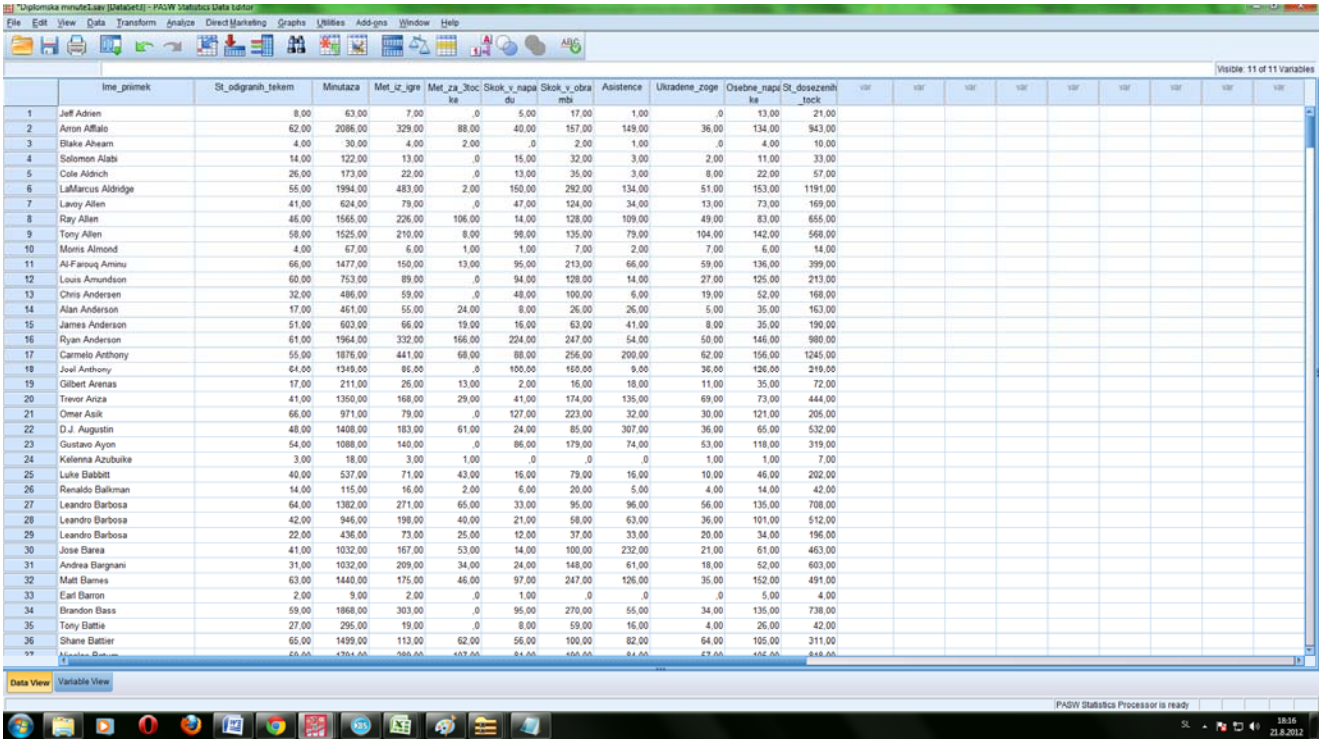
Odvisna spremenljivka minutaža je linearno odvisna od naslednjih spremenljivk:

1. Met_iz_igre/Minutaza.
2. Met_za_3tocke/ Minutaza.
3. Asistence/ Minutaza.
4. Skok_v_napadu/ Minutaza.
5. Ukradene_zoge/ Minutaza.
6. Osebne_napake/Minutaza.

Predpostavljam namreč, da spremenljivke met iz igre na minuto, met za tri točke na minuto, število asistenc na minuto, število skokov v napadu na minuto in število ukradenih žog na minuto, pozitivno vplivajo na spremenljivko minutaža, kar pomeni, da če se vrednost neodvisnih spremenljivk poveča, se poveča tudi vrednost odvisne spremenljivke. Nasprotno pa predpostavljam, da število osebnih napak na minuto negativno vpliva na minutažo igralca. Manjše kot je število osebnih napak na minuto, večje bo število minut, ki jih bo igralec na tekmi igral.

V želji po čim večjem številu podatkov, sem v podatkovno bazo zbral podatke in statistike vseh igralcev lige NBA v sezoni 2011/2012. Podatkovna baza tako zajema podatke o 551 igralcih in njihovih statistikah. Na zbranih podatkih sem nato izvedel linearno regresijo, kjer sem na podlagi neodvisnih spremenljivk skušal napovedati verjetnost oziroma vrednost odvisne spremenljivke, **minutaža**.

Slika 6.1: Podatkovna baza v paketu SPSS



	Ime_priimek	St_odigranih_tekem	Minutaza	Met_iz_igre	Met_za_3tocke	Skok_v_napadu	Skok_v_obrambi	Asistence	Ukradene_zoge	Osebne_napake	St_dosezenih_tock
1	Jeff Adrien	8,00	63,00	7,00	,0	5,00	17,00	1,00	,0	13,00	21,00
2	Arron Afflalo	62,00	2086,00	329,00	88,00	40,00	157,00	149,00	36,00	134,00	943,00
3	Blake Ahearn	4,00	30,00	4,00	2,00	,0	2,00	1,00	,0	4,00	10,00
4	Solomon Alabi	14,00	122,00	13,00	,0	15,00	32,00	3,00	2,00	11,00	33,00
5	Cole Aldrich	26,00	173,00	22,00	,0	13,00	35,00	3,00	8,00	22,00	57,00
6	LaMarcus Aldridge	55,00	1994,00	483,00	2,00	150,00	292,00	134,00	51,00	153,00	1191,00
7	Lavoy Allen	41,00	624,00	79,00	,0	47,00	124,00	34,00	13,00	73,00	169,00
8	Ray Allen	46,00	1565,00	226,00	106,00	14,00	128,00	109,00	49,00	83,00	655,00
9	Tony Allen	58,00	1525,00	210,00	8,00	98,00	135,00	79,00	104,00	142,00	568,00
10	Morris Almond	4,00	67,00	6,00	1,00	1,00	7,00	2,00	7,00	6,00	14,00
11	Al-Farouq Aminu	66,00	1477,00	150,00	13,00	95,00	213,00	66,00	59,00	136,00	399,00
12	Louis Amundson	60,00	753,00	89,00	,0	94,00	128,00	14,00	27,00	125,00	213,00
13	Chris Andersen	32,00	486,00	59,00	,0	48,00	100,00	6,00	19,00	52,00	168,00
14	Alan Anderson	17,00	461,00	55,00	24,00	8,00	26,00	26,00	5,00	35,00	163,00
15	James Anderson	51,00	603,00	66,00	19,00	16,00	63,00	41,00	8,00	35,00	190,00
16	Ryan Anderson	61,00	1964,00	332,00	166,00	224,00	247,00	54,00	50,00	146,00	980,00
17	Carmelo Anthony	55,00	1876,00	441,00	68,00	88,00	256,00	200,00	62,00	156,00	1245,00
18	Joel Anthony	64,00	1349,00	86,00	,0	166,00	166,00	9,00	36,00	126,00	219,00
19	Gilbert Arenas	17,00	211,00	26,00	13,00	2,00	16,00	18,00	11,00	35,00	72,00
20	Trevor Ariza	41,00	1350,00	168,00	29,00	41,00	174,00	135,00	69,00	73,00	444,00
21	Omer Asik	66,00	971,00	79,00	,0	127,00	223,00	32,00	30,00	121,00	205,00
22	D.J. Augustin	48,00	1408,00	183,00	61,00	24,00	85,00	307,00	36,00	65,00	532,00
23	Gustavo Ayon	54,00	1088,00	140,00	,0	86,00	179,00	74,00	53,00	118,00	319,00
24	Kelenna Azubuike	3,00	18,00	3,00	1,00	,0	,0	,0	1,00	1,00	7,00
25	Luke Babbitt	40,00	537,00	71,00	43,00	16,00	79,00	16,00	10,00	46,00	202,00
26	Renaldo Balkman	14,00	115,00	16,00	2,00	6,00	20,00	5,00	4,00	14,00	42,00
27	Leandro Barbosa	64,00	1382,00	271,00	65,00	33,00	95,00	96,00	56,00	135,00	708,00
28	Leandro Barbosa	42,00	946,00	198,00	40,00	21,00	58,00	63,00	36,00	101,00	512,00
29	Leandro Barbosa	22,00	436,00	73,00	25,00	12,00	37,00	33,00	20,00	34,00	196,00
30	Jose Barea	41,00	1032,00	167,00	53,00	14,00	100,00	232,00	21,00	61,00	463,00
31	Andrea Bargnani	31,00	1032,00	209,00	34,00	24,00	148,00	61,00	18,00	52,00	603,00
32	Matt Barnes	63,00	1440,00	175,00	46,00	97,00	247,00	126,00	35,00	152,00	491,00
33	Earl Barron	2,00	9,00	2,00	,0	1,00	,0	,0	,0	5,00	4,00
34	Brandon Bass	59,00	1868,00	303,00	,0	95,00	270,00	55,00	34,00	135,00	738,00
35	Tony Battie	27,00	295,00	19,00	,0	8,00	59,00	16,00	4,00	26,00	42,00
36	Shane Battier	65,00	1499,00	113,00	62,00	56,00	100,00	82,00	64,00	105,00	311,00

6.2 Bivariatna analiza: korelacijski koeficienti med spremenljivkami

Za proučevanje linearne povezanosti med spremenljivkami sem uporabil Pearsonov korelacijski koeficient, kjer sem preverjal, ali med mojimi spremenljivkami obstaja povezanost.

Pearsonov korelacijski koeficient se uporablja za označevanje razmerij dveh naključnih spremenljivk. Z njim merimo moč in smer povezanosti spremenljivk v razponu od -1 do +1. Pozitivne vrednosti kažejo, da sta spremenljivki pozitivno soodvisni, kar pomeni, da se z večanjem ene spremenljivke v povprečju večajo tudi vrednosti druge spremenljivke.

Negativne vrednosti pa pomenijo, da spremenljivki negativno korelirata oziroma nista soodvisni. Z večanjem ene spremenljivke se v povprečju manjšajo vrednosti druge spremenljivke (Schlotzhauer 2007).

Tabela 6.1: Korelacija med spremenljivkami

		St_dosezenih_tock_minuta	Met_iz_igre_minuta	Met_za_3tocke_minuta	Asistence_minuta	Skok_v_napadu_minuta	Osebne_napake_minuta
St_dosezenih_tock_minuta	Pearsonov koeficient	1,000	,951	,327	,218	-,112	-,268
	Značilnost	.	,000	,000	,000	,004	,000
	N	551	551	551	551	551	551
Met_iz_igre_minuta	Pearsonov koeficient	,951	1,000	,191	,168	-,031	-,197
	Značilnost	,000	.	,000	,000	,237	,000
	N	551	551	551	551	551	551
Met_za_3tocke_minuta	Pearsonov koeficient	,327	,191	1,000	,236	-,597	-,309
	Značilnost	,000	,000	.	,000	,000	,000
	N	551	551	551	551	551	551
Asistence_minuta	Pearsonov koeficient	,218	,168	,236	1,000	-,463	-,330
	Značilnost	,000	,000	,000	.	,000	,000
	N	551	551	551	551	551	551
Skok_v_napadu_minuta	Pearsonov koeficient	-,112	-,031	-,597	-,463	1,000	,365
	Značilnost	,004	,237	,000	,000	.	,000
	N	551	551	551	551	551	551
Osebne_napake_minuta	Pearsonov koeficient	-,268	-,197	-,309	-,330	,365	1,000
	Značilnost	,000	,000	,000	,000	,000	.
	N	551	551	551	551	551	551

Pearsonov koeficient pokaže, da med opazovanimi spremenljivkami obstajajo različne korelacije (pozitivne in negativne), ki tudi niso vse statistično značilne.

Negativni Pearsonov korelacijski koeficient in statistično značilni rezultati se kažejo med opazovanima spremenljivkama minutaža in skok v napadu na minuto ter med spremenljivkama minutaža in osebne napake na minuto. Negativna povezanost med spremenljivkami, ki je statistično značilna pomeni, da spremenljivke niso soodvisne in variirajo v nasprotno smer. Negativno korelacijo med spremenljivko minutaža in osebne napake na minuto lahko interpretiramo na naslednji način: igralcu ki bo imel na tekmah višje število osebnih napak na minuto, se bo število minut, ki jih bo na tekmah igral zmanjšalo.

Nasprotno pa se statistično značilna in pozitivna vrednost Pearsonovega korelacijskega koeficienta kaže med odvisno spremenljivko minutaža in neodvisnimi spremenljivkami, met iz igre na minuto, met za tri točke na minuto in asistence na minuto. To pomeni, da

spremenljivke variirajo v isto smer. Igralec, ki bo na tekmah na minuto zadel več metov iz igre, zadel več metov za tri točke in svojim igralcem podal več žog za doseganje zadetkov, bo v igri tudi več igral. Trenerji bodo takšnim igralcem namenili vedno več minut v igri.

6.1 Regresijska analiza

Linearni regresijski model predpostavlja, da obstaja linearni odnos med odvisno spremenljivko in neodvisno spremenljivko (prediktorjem).

Tabela 6.2: Število doseženih točk (ang. summary model)

Model	R	R kvadrat	Prilagojeni R kvadrat	Standardna napaka ocene
1	,538[a]	,289	,281	583,82118

a. Neodvisne spremenljivke: (Konstanta), Osebne_napake_minuta, Met_iz_igre_minuta, Met_za_3tocke_minuta, Asistence_minuta, Skok_v_napadu_minuta, Ukradene_zoge_minuta

Prilagojeni R kvadrat znaša 0.281 in nam pove, koliko variance pojasnijo izbrane neodvisne spremenljivke, ki vplivajo na odvisno spremenljivko (minutaža igralca). V tem razčlenjenem modelu torej neodvisne spremenljivke pojasnijo 28% variance.

Tabela 6.3: Anova

Model		Vsota kvadratov	df	Povprečje kvadratov	F	Signifikanca
1	Regresija	7,538E7	6	1,256E7	36,857	,000[a]
	Rezidual	1,854E8	544	340847,170		
	Skupaj	2,608E8	550			

a. Neodvisne spremenljivke : (Konstanta), Osebne_napake_minuta, Met_iz_igre_minuta, Met_za_3tocke_minuta, Asistence_minuta, Skok_v_napadu_minuta, Ukradene_zoge_minuta

b. Odvisna spremenljivka: Minutaza

O skupni značilnosti regresijskega modela sklepamo na podlagi statistične značilnosti F testa, ki znaša 0,000. Statistična značilnost je manjša od 0,05 – pri takem tveganju lahko trdimo, da med neodvisnimi spremenljivkami in odvisno spremenljivko obstaja linearni odnos.

Tabela 6.4: Regresijski koeficienti

Model		Nestandardizirani koeficienti		Standardizirani koeficienti		
		B	Std. Napaka	Beta	t	Značilnost
1	(Konstanta)	268,825	130,957		2,053	,041
	Met_iz_igre_minuta	5537,359	546,716	,387	10,128	,000
	Met_za_3tocke_minuta	2964,028	1349,314	,102	2,197	,028
	Asistence_minuta	1272,143	503,670	,109	2,526	,012
	Skok_v_napadu_minuta	2148,688	938,351	,117	2,290	,022
	Ukradene_zoge_minuta	-987,320	1483,059	-,025	-,666	,506
	Osebne_napake_minuta	-3560,582	627,329	-,229	-5,676	,000

a. Odvisna spremenljivka: Minutaza

Regresijski koeficient B je smerni koeficient regresijske premice. Pozitivna vrednost regresijskega koeficienta pove, za koliko enot se v povprečju poveča vrednost odvisne spremenljivke, če se vrednost neodvisne spremenljivke poveča za eno enoto. Negativna vrednost regresijskega koeficienta B pa pove, za koliko enot se v povprečju zmanjša vrednost odvisne spremenljivke, če se vrednost neodvisne spremenljivke poveča za eno enoto.

Parametri beta so pri neodvisnih spremenljivkah, met iz igre na minuto, met za tri točke na minuto in asistence na minuto, pozitivni in statistično značilni, kar pomeni, da neodvisne spremenljivke pozitivno vplivajo na odvisno spremenljivko **minutaža**. Če se vrednost spremenljivke met iz igre na minuto poveča za eno enoto, se vrednost odvisne spremenljivke minutaža poveča za 5537,359 enot. Z drugimi besedami, če bo igralec na tekmah v minuti bolj uspešen oz. se bo njegovo število zadetih metov iz igre v minuti povečalo, se bo povečalo tudi število minut, ki jih bo igralec igral na tekmah. Iz tabele 6.5 je razvidno tudi, da bo

igralec, ki se mu bo število asistenc na minuto povečalo za eno enoto, na tekmah dobil tudi za 1272,143 enot večjo minutažo.

Tudi pri spremenljivki osebne napake na minuto lahko svojo predpostavko potrdim. Če se bo vrednost spremenljivke osebne napake na minuto povečala za eno enoto, se bo odvisna spremenljivka minutaža igralca na minuto zmanjšala za 3560,582 enot, kar pomeni, da bo igralec v primeru višjega števila osebnih napak na minuto na tekmah igral manj minut.

Moje predpostavke pa ne držijo za spremenljivko ukradene žoge na minuto (statistična značilnost je 0,582). Statistično neznačilne vrednosti regresijskih koeficientov B v Tabeli 6.5 so mi pokazali, da vpliva neodvisne spremenljivke ukradene žoge na minuto na odvisno spremenljivko minutaža, ne moremo potrditi.

Na podlagi dobljenih rezultatov lahko na zastavljeno vprašanje: »*Katere spremenljivke so tiste, ki lahko napovedo število minut, ki jih bo posameznik odigral na tekmah v naslednji sezoni?*«, podam naslednji odgovor. Če se bodo igralčeve statistike, merjene v minuti, na tekmah povišale (boljši met iz igre na minuto, boljši met za tri točke na minuto, večje število asistenc na minuto in večje število skokov v napadu na minuto), se bo povišalo tudi število minut, ki jih bo igralec na tekmah odigral. Nasprotno pa se bo v primeru zmanjšanja vrednosti neodvisnih spremenljivk zmanjšala tudi minutaža igralca na tekmah.

Linearna regresija lahko zelo pripomore pri odločanju trenerjev na podlagi dobljenih statistik. Igralec, ki bo imel v igri visoke vrednosti statistik (met iz igre, asistence, skok v napadu) na minuto, bo na tekmah dobil veliko priložnosti za igro, poleg tega pa se bo zanimanje drugih moštev za nakup tega igralca dodatno povečalo.

6.2 Grozdenje v košarki

Grozdenje klasificira opazovane spremenljivke v dve ali več med seboj različnih skupin, ki vsebujejo kombinacijo intervalnih spremenljivk. Gre za iterativni proces (iteracije), ki se izvaja toliko časa, dokler se povprečja grozdov v zaporednih korakih ne razlikujejo več. Glavni namen je odkritje sistema za razvrščanje spremenljivk v skupine, katerih člani si delijo

podobne lastnosti. Prav skupine s podobnimi lastnostmi pa nam omogočajo, da lažje napovemo dogodke, ki se bodo zgodili.

Zato sem se odločil, da bom igralce na podlagi spodnjih spremenljivk klasificiral v tri skupine (pri večjem številu skupin so razlike v vrednostih spremenljivk med skupinami zelo majhne), kar prikazujem v tabeli 6.6. Ker me je zanimala uspešnost posameznega igralca v eni minuti, sem tudi v tem primeru vse izbrane spremenljivke delil s spremenljivko minutaža.
Analizirane spremenljivke:

- Met_iz_igre: zadeti meti, ki jih je igralec zadel v eni sezoni.
- Met_za_3tocke: zadeti meti za tri točke v eni sezoni.
- Asistence: število asistenc, ki jih je dosegel igralec v eni sezoni.
- Skok_v_napadu: število dobljenih žog igralca pri skoku v napadu v eni sezoni.
- Skok_v_obrambi: število dobljenih žog igralca pri skoku v obrambi v eni sezoni.
- Ukradene_žoge: število žog, ki jih je igralec nasprotnemu igralcu ukradel v eni sezoni.

Tabela 6.5: Začetni centri grozdov

	Grozd		
	1	2	3
Met_iz_igre_minuta	,12	,15	,00
Met_za_3tocke_minuta	,00	,03	,00
Asistence_minuta	,06	,34	,00
Skok_v_napadu_minuta	,12	,01	,00
Skok_v_obrambi_minuta	,41	,08	,00
Ukradene_zoge_minuta	,12	,02	,00

Tabela prikazuje prvi korak v procesu grozdenja, ki je iskanje števila centrov (ang. K-means). V mojem primeru znaša število centrov 3.

Tabela 6.6: Zgodovina iteracijskih postopkov

Iteracija	Sprememba v centru grozda		
	1	2	3
1	,216	,168	,187
2	,021	,013	,008
3	,010	,003	,007
4	,006	,001	,005
5	,003	,002	,003
6	,002	,001	,002
7	,000	,001	,001
8	,001	,001	,001
9	,000	,000	,000

SPSS je izvedel 9 iteracijskih postopkov, kar pomeni, da se pri 9 zaporednih iteracijah povprečja grozdov ne razlikujejo bistveno.

Tabela 6.7: Končni centri grozdov

	Grozd		
	1	2	3
Met_iz_igre_minuta	,14	,15	,12
Met_za_3tocke_minuta	,01	,03	,03
Asistence_minuta	,04	,17	,06
Skok_v_napadu_minuta	,08	,02	,03
Skok_v_obrambi_minuta	,18	,09	,10
Ukradene_zoge_minuta	,03	,04	,03

Končni centri grozdov so izračunani kot povprečje vsake spremenljivke znotraj končnih grozdov. V prvem grozdu so tako razvrščeni igralci, ki so bili pri metu iz igre na minuto manj uspešni kot igralci v grozdu dva. Imajo pa zato vrednosti spremenljivk, skok v obrambi na minuto (0,18) in skok v napadu na minuto (0,08), največje. To so predvsem igralci, ki igrajo pozicijo centra ali krilnega centra. Igralci na omenjenih pozicijah so v obrambi zadolženi za pobiranje odbitih žog pod košem. Poleg tega pa so uspešni tudi pod košem nasprotnikov, kjer dosežejo večino svojih točk. Metov z razdalje pa se redko poslužujejo.

Iz drugega grozda ugotovimo, da so v njem razvrščeni igralci, ki so bili v sezoni 2011/2012 po uspešnosti pri metu iz igre na minuto (0,15 zadetega meta na minuto), najbolj uspešni.

Razvidno je tudi, da imajo na minuto največ asistenc, in sicer 0,17 asistenc na minuto ter največ ukradenih žog na minuto. Gre za igralce, ki igrajo pozicijo organizatorja (ang. point guard). To so igralci, ki so zadolženi za organizacijo napada, imajo dober pregled nad dogajanjem na igrišču, s svojo spretnostjo pa ukradejo tudi veliko nasprotnikovih žog. Od njih se pričakuje natančnost pri metu iz igre.

V tretjem grozdu pa so razvrščeni igralci, ki so pri metu iz igre na minuto najmanj uspešni. Imajo pa večje število asistenc na minuto in večje število zadetih metov za tri točke na minuto kot igralci v grozdu 1. Uspešni so tudi pri skoku v obrambi na minuto. V tem grozdu so igralci, ki igrajo pozicijo branilca (ang. shooting guard) in pozicijo krila. Pri branilcih gre za igralce, ki so v napadu zadolženi za mete od daleč (meti za tri točke). Poleg metov od daleč pomagajo organizatorjem organizirati igro in z različnimi asistencami razigrati ekipo. Pozicijo krila pa igrajo igralci, ki so v obrambi zadolženi za pobiranje odbitih žog (skok v obrambi), v napadu pa so uspešni tako pri metih za tri točke, kot pri metih pod košem.

Tabela 6.8: Razvrstitev enot v grozde

Grozd	1	198,000
	2	127,000
	3	226,000
Veljavne vrednosti		551,000
Manjkajoče vrednosti		,000

Tabela 6.8 prikazuje delitev enot v grozde. Tako je v prvi grozd razvrščenih 198 igralcev, ki igrajo pozicijo centra ali krilnega centra, v drugi grozd je razdeljenih najmanj igralcev (127), ki igrajo pozicijo organizatorja ter v tretji grozd največ igralcev (226), ki igrajo pozicijo branilca ali pozicijo krila.

Enako kot linearna regresija je tudi tehnika grozdenja zelo učinkovita pri analiziranju podatkov v košarki. S pomočjo grozdenja sem odkril uspešnost posameznih igralcev na minuto v sezoni 2011/2012. Poleg slednjega pa sem tudi ugotovil, v katerem grozdu se nahajajo igralci glede na igralna mesta (organizator igre, branilec, center itd.). Takšne analize podatkov pa lastnikom klubov v primeru odločitve za nakup ali izposojo igralca pridejo še kako prav.

7 SKLEP

Rudarjenje podatkov je analitično orodje, ki omogoča odkrivanje informacij in znanja, pomembnega za pripravo različnih odločitev (poslovnih in športnih). Gre za proces pridobivanja skritih vzorcev iz podatkov, ki se uporablja v poslovanju, bioinformatiki, boju proti terorizmu in vedno bolj tudi v profesionalnem športu. Z uporabo in analizo preteklih podatkov lahko podjetja ali moštva predvidijo prihodnje trende in se na to tudi ustrezno pripravijo.

Vendar pa naloga podatkovnega rudarjenja v športu ni le zbiranje in analiza podatkov, temveč odkrivanje njihovega pomena/znanja in uporaba spoznanj na najboljši možni način. Vse to lahko športnim organizacijam zagotovi konkurenčno prednost v primerjavi z njihovimi nasprotniki. Podatkovno rudarjenje ne koristi samo lastnikom in trenerjem moštva, v veliko pomoč je tudi igralcem samim. Tako lahko igralec s pomočjo podatkovnega rudarjenja pridobi uporabne podatke o njegovih prednostih in slabostih v igri, slednje pa na podlagi dodatnih treningov odpravi in se na ta način dodatno izpopolnjuje v uspešnosti. Rudarjenje v športu se je v nekaj letih tako močno razširilo, da je število trenerjev, ki uporabljajo strojno učenje in simulacijske tehnike za iskanje optimalne strategije ekipe, vedno večje.

V svojem diplomskem delu sem s pomočjo teorije in analize primerov pokazal, da je podatkovno rudarjenje v športu uspešno, ker lahko poda odgovore na različna zastavljena vprašanja. Tako lahko trenerjem omogoča odkrivanje skritih vzorcev v podatkih njihovih igralcev, kot so: število doseženih točk na minuto, odstotek uspešnosti pri metu iz igre, na katerem položaju (levo krilo, ali desno krilo) bo njegov odstotek zadetih metov najbolj natančen, število minut, ki jih bo igralec v naslednji sezoni igral, ali pa, koliko odigranih tekem v sezoni bo imel. Tehnik, na podlagi katerih lahko napovemo vrednosti, je veliko, tako lahko na primer s tehniko grozdenja razvrščamo opazovane enote v grozde, ki jim pripišemo določene lastnosti. Na podlagi podatkov v posameznih razredih pa lahko nato sklepamo o njihovih značilnostih. V primeru, da moštvo potrebuje zelo dobrega igralca, ki igra na mestu organizatorja, lahko vzame podatkovno bazo vseh igralcev, ki so igrali v tekoči sezoni, in na teh podatkih opravi proces grozdenja ter se na podlagi dobljenih vrednosti grozdov odloči, v katerem grozdu so uvrščeni najbolj uspešni igralci, ki igrajo na mestu organizatorja.

8 LITERATURA

1 *82games*. Dostopno prek: http://82games.com/ (24. julij 2012).

2 Berry, Michael J. A. in Gordon S. Linoff. 2000. *Mastering data mining: The art and science of customer relationship management*. New York: John Wiley & Sons, inc.

3 Beynon-Davies, Paul. 2004. *Database Systems third Edition*. Basingstoke: Palgrave Macmillan.

4 Chen, Ming-Syan, Jiawei Han in Philip S. YU. 1996. Data mining: An overview from database perspective. *Knowledge and Data Engineering, IEEE Transactions on* (8): 866−883.

5 *Databasebasketball*. Dostopno prek: http://databasebasketball.com (24. julij 2012).

6 Edelstein, Herbert A. 1999. *Intorduction to data mining and knowledge discovery, third edition*. Potomac: Two Crows Corporation.

7 Etzioni, Oren. 1996. The World Wide Web: quagmire or gold mine? *Communications of the ACM* 39 (11): 65−68.

8 Ferligoj, Anuška. 1989. Razvrščanje v skupine: Teorija in uporaba v družboslovju. *Metodološki zvezki* (4). Dostopno prek: http://www.fdvinfo.net/db/34/1321/Publikacije/Razvrscanje_v_skupine__teorija_in_uporaba_v_druzboslovju/?&p1=24&p2=35&p3=96&p4=102&page=58 (25. julij 2012).

9 --- 1995. *Osnove statistike na prosojnicah*. Ljubljana: samozaložba.

10 *Finance*. 2007. Z rudarjenjem podatkov se želimo nekaj novega naučiti, 12. marec. Dostopno prek: http://www.finance.si/177113 (11. julij 2012).

11 Giudici, Paolo in Silvia Figini. 2009. *Applied data mining for business and industry: Second edition.* New York: John Wiley & Sons, inc.

12 Gujarati, Damodar. 1992. *Essentials of econometrics*. New York: McGraw-Hill.

13 Han, Jiawei in Micheline Kamber. 2001. *Data mining: Concepts and techniques*. London: Academic Press.

14 Khan, Joshua. 2003. *Neural Network Prediction of NFL Football Games*. Dostopno prek: http://homepages.cae.wisc.edu/~ece539/project/f03/kahn.pdf (3. avgust 2012)

15 Lloyd, Smith, Bret Lipscomb in Adam Simkins. 2007. Data mining in sports: predicting CY Young award winners. *Journal of Computing Sciences in Colleges 22* (4): 115−121.

16 *Mysynergysports*. Dostopno prek: http://mysynergysports.com/ (26. julij 2012).

17 *NBA*. Dostopno prek: http://www.nba.com/ (23. julij 2012).

18 Pujari, Arun K. 2001. *Data mining techniques*. Hyderabad: Universities Press.

19 Ramageri, Baharati M. 2010. Data mining techniques and applications. *Indian journal of computer science and engineering* 1 (4): 301−305.

20 Rygielski, Chris, Jyun-Cheng Wang in David C. Yen. 2002. Data mining techniques for custimer relationship management. *Technology in society* 24 (4): 483−502.

21 Schlotzhauer, Sandra. 2007. *Elementary statistics using JMP*. North Carolina: SAS Institute Inc.

22 Schumaker, Robert P., Osama K. Solieman in Chen Hsinchun. 2010. *Sports data mining*. New York: Springer.

23 Thearling, Kurt. 2010. *An introduction to data mining*. Dostopno prek: http://www.thearling.com/text/dmwhite/dmwhite.htm (12. julij 2012).

24 Velickov, Slavco in Dimitri Solomantine. 2000. Predictive data mining: Pratcial examples. *International Institute for Infrastructural, Hydraulic, and Environmental Engineering*. Dostopno prek: http://www.ihe.nl/hi/sol/papers/predictive_dm_cottbus-2.pdf (13. julij 2012).

25 Virtualgold. 2012. *Customer Success Stories*. Dostopno prek: http://www.virtualgold.com/customers_sstories.html (24. julij 2012).

26 Witten, Ian H., Eibe Frank in Mark A. Hall. 2011. *Data mining: Practical machine learning tools and techniques third edition.* Burlington: British Library.

27 Zhengxin, Chen. 2001. *Data mining and uncertain reasoning: An integrated approach*. New York: John Wiley & Sons, inc.