Eksploratorna analiza za projekt iz Strojnog učenja:

Predviđanje uspjeha učenika na temelju demografskih i socio-ekonomskih faktora

Grupa Epsilon

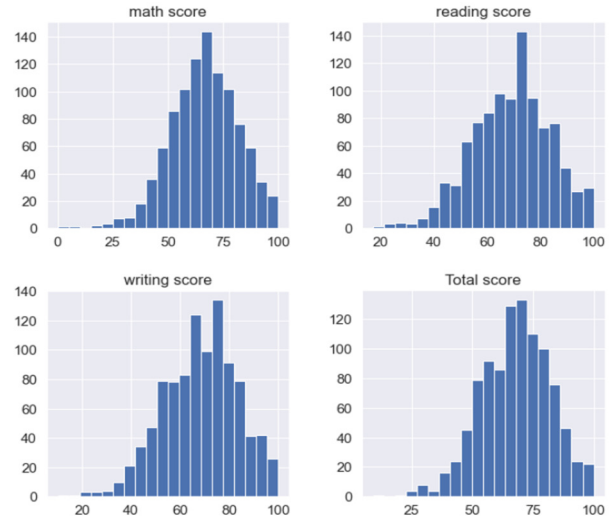
Uvod

Promatramo utjecaj odabranih atributa studenata na njihov rezultat na ispitima. Želimo naučiti prognozirati uspjeh studenata na testovima iz pojedinih disciplina, na temelju njihovog spola, rase, dosad stečenog stupnja obrazovanja i slično. Data set koji koristimo je tablica anonimnih studenata u kojem za svakog piše redom: spol, rasa, stupanj obrazovanja, kvaliteta ručka, pripreme prije polaganja, te rezultati iz triju ispita, matematike, čitanja i pisanja. Objavljeno na Kaggleu „Students Performance in Exams“, https://www.kaggle.com/spscientist/students-performance-in-exams.

Cilj i hipoteze

Cilj nam je pronaći (naučiti) funkciju koja za dani niz ulaznih podataka za nekog studenta dati dobru predikciju koliki broj bodova bi taj ili takav student mogao ostvariti na ispitu.

Naš skup podataka sadrži informacije o 1000 studenata i studentica. Budući da želimo učiti funkciju koja iz prvih 5 stupaca prognozira preostala tri (ili samo zbroj preostala tri), a ta tri stupca poprimaju samo diskretne vrijednosti, i to prvi 2 (spol), drugi 5 (rasna grupa), treći 5 (stupanj obrazovanja) te četvrti i peti također samo 2 (ručak i priprema),



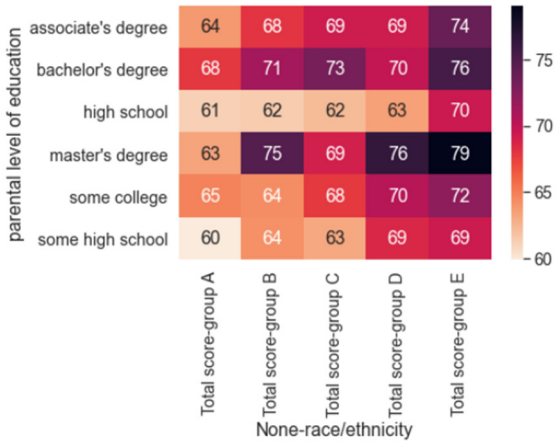
Slika 1 Histogram bodova učenika

što znači da svih mogućih kombinacija ovih atributa ima konačno mnogo (200) i to manje nego što imamo redaka u tablici (odnosno studenata), znamo da sigurno imamo par studenata (točnije, mora postojati bar jedna petorka studenata) koji imaju sve navedene atribute iste. Međutim, oni vjerojatno neće imati iste bodove na ispitima.

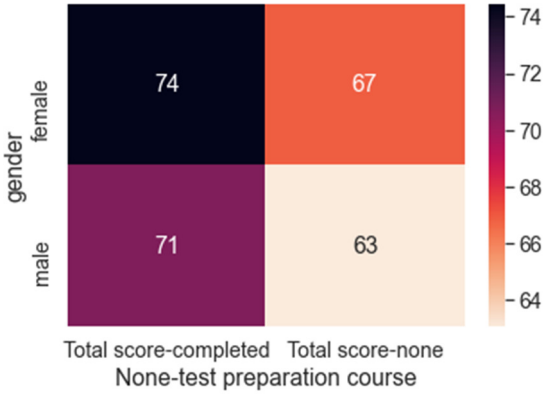
Imajući to u vidu, htjeli bismo da nam funkcija osim nekakvog očekivanog broja bodova (vjerojatno prosjeka za neki još veći skup studenata) vraća i neku ocjenu greške, to jest očekivano odstupanje od očekivanog broja bodova (opet, vjerojatno varijanca).

Na primjer, za dvije hipotetske grupe studenata koje bi imale sve atribute iste osim pripreme, očekivali bi bolji (prosječni) rezultat grupe koja je imala pripreme, ali i da većina broja bodova bude bliska tom prosjeku,

odnosno da svi imaju otprilike podjednaki broj bodova, a za grupu koja nije imala pripreme, osim što bi imala manji prosjek, očekujemo i veća odstupanja od tog prosjeka, da bude studenata s jako dobrim rezultatom i jako lošim.



Slika 2 Kumulativni prikaz rezultata ovisnosti vrste obrazovanja roditelja s rezultatima ispita - grupirano po rasi/etnicitetu



Slika 3 Kumulativni prikaz rezultata ovisnosti spola s rezultatima ispita - grupirano po odabiru priprema za ispit

Pregled dosadašnjih istraživanja

Osim očitog pristupa u kojem se iz poznatih atributa zaključiti nešto o postignutom rezultatu, neka istraživanja istraživala su i donekle obrnuti smjer. Znajući rezultate ispita i možda neke atribute dati neku predikciju za neki nepoznati atribut. Konkretno, radio se prediktor koji za dane rezultate sva tri ispita daje spol tog studenta. Jedno takvo istraživanje navodi i vrlo visoku preciznost za svoj naučeni prediktor. Čak 92,5% [2]. Jedna stvar koju smo primijetili u datasetu, za koju smatramo da je svakako pridonijela tako visokoj preciznosti u tom istraživanju je to da su studentice u prosjeku bolje napisale testove iz čitanja i pisanja, dok su studenti bolje napisali ispit iz matematike.

Za predikciju nekih drugih atributa nismo mogli naći radove. To je i razumljivo jer drugi atributi nemaju takvo svojstvo (ili bar nije toliko očito) da na neka dva rezultata djeluju suprotno (na jedno pozitivno, na drugo negativno) kao što smo to uočili za spol.

Preostala istraživanja uglavnom rade ono što ćemo i mi raditi, odnosno uče davati prognozu za broj bodova za poznate atribute.

Materijali, metodologija i plan istraživanja

Data set ćemo podijeliti na skupove za učenje i treniranje.

Za odabir skupa za učenje i skupa za testiranje željeli bismo osim nasumičnog izbacivanja napraviti i nekakvo „pametno“ izbacivanje. U skup za testiranje stavili bi točno jedan redak iz tablice za svaku od 200 kombinacija atributa. Tako bi između ostalog osigurali i testiranje svih kombinacija koje možda ne bi imali da smo samo nasumično izbacili 200 redaka.

Za funkciju koja daje procjenu rezultata na ispitima mjerit ćemo prosječno odstupanje od stvarnog rezultata na testnom skupu. Po tom prosjeku ćemo mjeriti uspješnost naučenog prediktora, želimo da on bude što manji.

Prvo ćemo probati vidjeti kolko podaci linearno koreliraju s rezultatima ispita. Razvit ćemo modele ovisno o tome kakva je korelacija podataka, i isprobati različite modele poput logističke regresije ,slučanih šuma i SVMa, te proučiti koji daju najbolje rezultate na testnom skupu.

Mada postoji bolji modeli od slučajnih šuma za te skupove, slučajne šume bi u našem kontekstu dobro došle jer imamo puno kategoričkih podataka pa bi prirodno podijelile naš set na različite podskupove.

Očekivani rezultati predloženog projekta

Prediktor koji želimo naučiti je funkcija iz diskretnog skupa u numerički, pa samim time očekujemo da slika prediktora vjerojatno neće pokrivati cijelu kodomenu, te će vjerojatno biti neprecizna za neke izolirane slučajeve. Kao što je već bilo rečeno, za svaku kombinaciju atributa studenata vjerojatno će postojati dva studenta s takvom kombinacijom, ali s vjerojatno različitim uspjehom na ispitima. To znači da ćemo već na skupu za treniranje morati imati odstupanja.

Očekujemo da ćemo imati otprilike iste ili slične rezultate kao i na dosadašnjim istraživanjima. Mogli bi očekivati neku malenu prednost prediktora koji je naučen na testnom skupu s „pametnim“ izbacivanjem, u odnosu na onaj s nasumičnim.

Popis literature

[1] https://www.kaggle.com/spscientist/students-performance-in-exams

[2] https://www.kaggle.com/erkamk/gender-prediction-visualization-92-5-accuracy