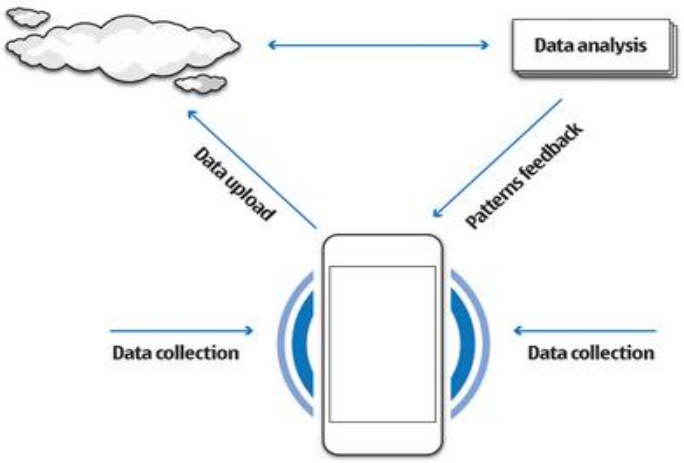
Predikcija demografskih atributa korisnika mobilnih telefona

Cilj projekta:

Cilj projekta je konstrukcija modela za predikciju pola korisnika na osnovu praćenja njegove aktivnosti na mobilnom telefonu. Korisnici i njihovi kontakti su anonimni. U logovima je zabeleženo vreme kada na telefonu pokreće aplikacije, obavlja i prima pozive/poruke, menja GSM ćeliju, približna rastojanja koja pređe, merenja sa senzora ubrzanja...



Slika 1: Prikupljanje podataka sa mobilnog telefona

Zadatak:

Obučiti modele za automatsko prepoznavanje pola korisnika (labele: 1 - žena, 2 - muškarac) na osnovu izdvojenih obeležja iz logova mobilnog telefona. Izabrati bar tri različita algoritma za predikciju (neuralne mreže, SVM, random forest, KNN ...). Za evaluaciju modela i odabir parametara algoritma koristiti kros-validaciju (10-fold i leave-one-out). Zadatak uraditi i na redukovanom skupu obeležja.

Podaci:

Podaci su deo Nokia MDC - Mobile Data Challenge-a (https://www.idiap.ch/dataset/mdc). Skup podataka predviđen za zadatak predikcije demografskih atributa sadrži 80 korisnika i obuhvata informacije o pozivima, porukama, lokacijama, aplikacijama, itd. Njihov ukupan broj sumiran je na slici 2.

Data type	Quantity
Calls (in/out/missed)	240,227
SMS (in/out/failed/pending)	175,832
Photos	37,151
Videos	2,940
Application events	8,096,870
Calendar entries	13,792
Phone book entries	45,928
Location points	26,152,673
Unique cell towers	99,166
Accelerometer samples	1,273,333
Bluetooth observations	38,259,550
Unique Bluetooth devices	498,593
WLAN observations	31,013,270
Unique WLAN access points	560,441
Audio samples	595,895

Slika 2: Podaci iz Nokia MDC skupa

Za 78 korisnika su dostupne informacije o polu i oni su analizirani za potrebe projekta. Podaci u izvornom obliku za svakog korisnika se nalaze u logovima. Za potrebe projekta iz logova je izdvojeno 40 obeležja. Izdvojena obeležja su opisana u tabeli 1.

Drugi deo zadatka se odnosi na redukovani skup obeležja. Selekcija obeležja izvršena je ReliefF algoritmom i na taj način je 40 obeležja redukovano na 16. Podaci i odgovarajuća obeležja se nalaze u fajlovima mdc-gender-all-features.mat i mdc-gender-selected-features.mat. Vrednosti svagog od obeležja su dodatno normalizovane deljenjem sa srednjom vrednošću po svim uzorcima.

Tabela 1: Izdvojena obeležja iz logova mobilnih telefona

Naziv obeležja	Opis
application-t1d	Prosečan broj pokretanja aplikacija u vremenskom intervalu 00-08h radnim danima
application-t2d	Prosečan broj pokretanja aplikacija u vremenskom intervalu 08-16h radnim danima
application-t3d	Prosečan broj pokretanja aplikacija u vremenskom intervalu 16-24h radnim danima

application-t1w	Prosečan broj pokretanja aplikacija u vremenskom intervalu 00-08h vikendom
application-t2w	Prosečan broj pokretanja aplikacija u vremenskom intervalu 08-16h vikendom
application-t3w	Prosečan broj pokretanja aplikacija u vremenskom intervalu 16-24h vikendom
bluetooth-t1d	Prosečan broj bluetooth konekcija u vremenskom intervalu 00-08h radnim danima
bluetooth-t2d	Prosečan broj bluetooth konekcija u vremenskom intervalu 08-16h radnim danima
bluetooth-t3d	Prosečan broj bluetooth konekcija u vremenskom intervalu 16-24h radnim danima
bluetooth-t1w	Prosečan broj bluetooth konekcija u vremenskom intervalu 00-08h vikendom
bluetooth-t2w	Prosečan broj bluetooth konekcija u vremenskom intervalu 08-16h vikendom
bluetooth-t3w	Prosečan broj bluetooth konekcija u vremenskom intervalu 16-24h vikendom
calendar-a-d	Prosečan broj upisanih sastanka u kalendaru radnim danima
calendar-e-d	Prosečan broj upisanih događaja u kalendaru radnim danima
calendar-a-w	Prosečan broj upisanih sastanka u kalendaru vikendom
calendar-e-w	Prosečan broj upisanih događaja u kalendaru vikendom
duration_d	Prosečno trajanje poziva radnim danima
duration_w	Prosečno trajanje poziva vikendom
calls_incoming_rate	Odnos dolazećih poziva prema ukupnom broju svih konekcija u calllog-u
calls_outgoing_rate	Odnos odlazećih poziva prema ukupnom broju svih konekcija u calllog-u
messages_incoming_rate	Odnos dolazećih poruka prema ukupnom broju svih konekcija u calllog-u
messages_outgoing_rate	Odnos odlazećih poruka prema ukupnom broju svih konekcija u calllog-u
prefix	Jedinstven broj prefiksa u calllog-u
country_prefix	Jedinstven broj prefiksa u calllog-u
contacts_avg	Prosečan broj dodatih kontakata u imenik po danu
unique_prefix_avg	Prosečan broj dodatih kontakata sa jedinstvenim regionalnim pefiksom u imenik po danu
unique_country_prefix_avg	Prosečan broj dodatih kontakata sa jedinstvenim državnim pefiksom u imenik po danu
distance_avg	Prosečno rastojanje koje korisnik prelazi u toku dana
gsm_country	Broj gsm ćelija u kojima je boravio korisnik normalizovan brojem dana u kojima je posmatran
gsm_area	Broj država u kojima je boravio korisnik normalizovan brojem dana u kojima je posmatran
gsm_area_change	Prosečan broj promena gsm ćelija u toku dana
media_count_avg	Broj multimedijalnog sadržaja koje je korisnik dodao na telefon normalizovan brojem dana u kojima je posmatran
media_size_avg	Veličina multimedijalnog sadržaja koje je korisnik dodao na telefon normalizovana brojem dana u kojima je posmatran

mediaplay_t1d	Prosečan broj pokretanja multimedijalnog sadržaja u vremenskom intervalu 00-08h radnim danima
mediaplay_t2d	Prosečan broj pokretanja multimedijalnog sadržaja u vremenskom intervalu 08-16h radnim danima
mediaplay_t3d	Prosečan broj pokretanja multimedijalnog sadržaja u vremenskom intervalu 16-24h radnim danima
mediaplay_t1w	Prosečan broj pokretanja multimedijalnog sadržaja u vremenskom intervalu 00-08h vikendom
mediaplay_t2w	Prosečan broj pokretanja multimedijalnog sadržaja u vremenskom intervalu 08-16h vikendom
mediaplay_t3w	Prosečan broj pokretanja multimedijalnog sadržaja u vremenskom intervalu 16-24h vikendom
accel_avg	Prosečna vrednost merenja sa senzora ubrzanja

Predstavljanje rezultata:

1) Opisati potrebu za analizom korisnika mobilnih telefona i moguće probleme sa zaštitom privatnosti.
2) Odabrati tri obeležja i prikazati raspodelu vrednosti za obe klase. Raspodele prikazati histogramima ili ih estimirati nekom od kernel metoda.
3) Analizirati tačnost klasifikacije pomoću matrica konfuzije u zavisnosti od odabranih algoritama i parametara.
4) Uporediti klasifikatore na potpunom i redukovanom skupu obeležja. Izračunati tačnost klasifikacije i za obe klase: preciznost, odziv i F-meru. Nacrtati ROC krive.
5) Navedite svoja zapažanja i moguće razloge za razliku u performansama u različitim eksperimentima.

Odbrana projekta:

Projekat se predaje u vidu pisanog izveštaja i pratećeg programa koji se usmeno brane u dogovorenom terminu. Projekat nosi ukupno 30 poena na ispitu, od čega pisani izveštaj maksimalno 20, a usmena odbrana 10 poena. Usmena odbrana je neophodna. Usmena odbrana se sastoji iz prezentacije izveštaja i prezentacije programa.

Sastavila: Sanja Brdar