

Laboratorijska vježba br. 11 Korištenje Automatskih Modela za Dubinsku Analizu Podataka

Uputstvo za izradu laboratorijske vježbe

Izrada laboratorijske vježbe vrši se u formi izvještaja koja je data u nastavku. Potrebno je popuniti sva polja data u izvještaju, odgovoriti na pitanja i dodati tražene slike. Nije dozvoljeno brisati postojeća, niti dodavati nova polja.

Izvještaj sa izradom laboratorijske vježbe pretvorene u PDF dokument šalje se na e-mail adresu odgovornog asistenta grupe za laboratorijske vježbe.

NAPOMENA: Izvještaj se radi samostalno. Rad u paru ili grupi nije dozvoljen.

Informacije o studentu

Ime i prezime: Haris Masovic
Broj indexa: 1689/17993
Grupa za laboratorijske vježbe: Utorak 17:00

Zadatak 1. (Priprema podataka)

Preuzeti dataset koji se nalazi na sljedećem linku: [link](#)

Pokrenuti RapidMiner Studio okruženje, a zatim na početnom ekranu odabrati opciju **Turbo Prep**, kao što je prikazano na slici ispod.

Start with



Blank Process

Start a new process from scratch in the design view.



Turbo Prep

Prepare your data interactively: transform, clean and combine data sets.



Auto Model

Build and optimize models using automated machine learning.

Odabrati opciju **Load Data**, kao što je prikazano na slici ispod.

Data Sets

+ LOAD DATA

No Data Selected

Add new data sets on the left. Details for the selected data are shown below. You can change the data with the following actions. ⓘ



TRANSFORM



CLEANSE



GENERATE



PIVOT



MERGE

Uploadovati prethodno preuzeti dataset **2d_dataset_small.csv**.

Nakon uploada, odabrati opciju **Charts**, kao što je prikazano na slici ispod.

Dubinska Analiza Podataka

2d_dataset_small

Add new data sets on the left. Details for the selected data are shown below. You can change the data with the following actions. ⓘ

✕ TRANSFORM ✎ CLEANSE 📊 GENERATE ∑ PIVOT ➡ MERGE

MODEL CHARTS CREATE PROCESS HISTORY ...

Odabrati sljedeće opcije:

Plot type: Scatter / Bubble

X-Axis column: post_weight

Value column: user_hatred_speech_rating

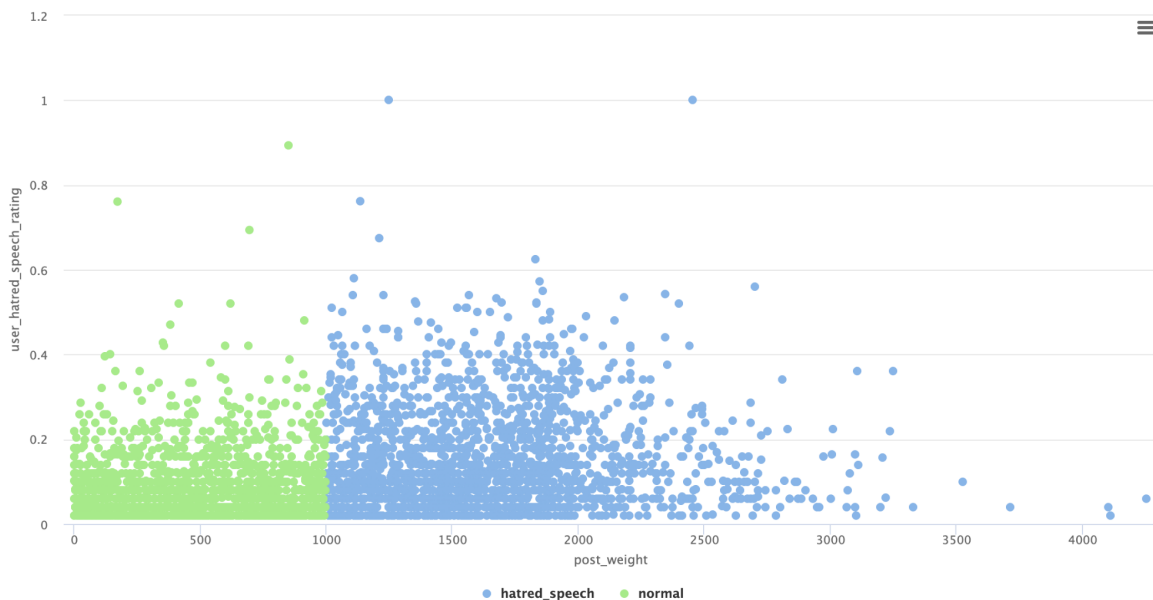
Color: category

Prikazati izgled vizualizacije podataka:

2d_dataset_small

Configure the desired chart with the settings on the left. ⓘ

CANCEL



Izvršiti kratku analizu prikaza. Da li su opsezi na x i y-osi proporcionalni? Da li ima vidljivih outliera i na kojem dijelu grafika se nalaze? Da li postoje određene kombinacije x i y vrijednosti za koje ne postoje tačke na grafiku?

- Opsezi na x i y osi nisu proporcionalni
- Ima vidljivih outliers i nalaze se na desnom dijelu grafa
- Postoje određene kombinacije x i y vrijednosti za koje ne postoje tačke na grafiku (3900, 0.8 npr)

Vratiti se na prošli ekran i sada odabrati opciju **Cleanse**, kao što je prikazano na slici ispod.

Dubinska Analiza Podataka

2d_dataset_small

Add new data sets on the left. Details for the selected data are shown below. You can change the data with the following actions. ⓘ

✕ TRANSFORM **CLEANSE** 📊 GENERATE Σ PIVOT ➡ MERGE

MODEL CHARTS CREATE PROCESS HISTORY ...

Odabrati opciju **Auto Cleansing**. Nakon toga izvršiti sljedeće korake:

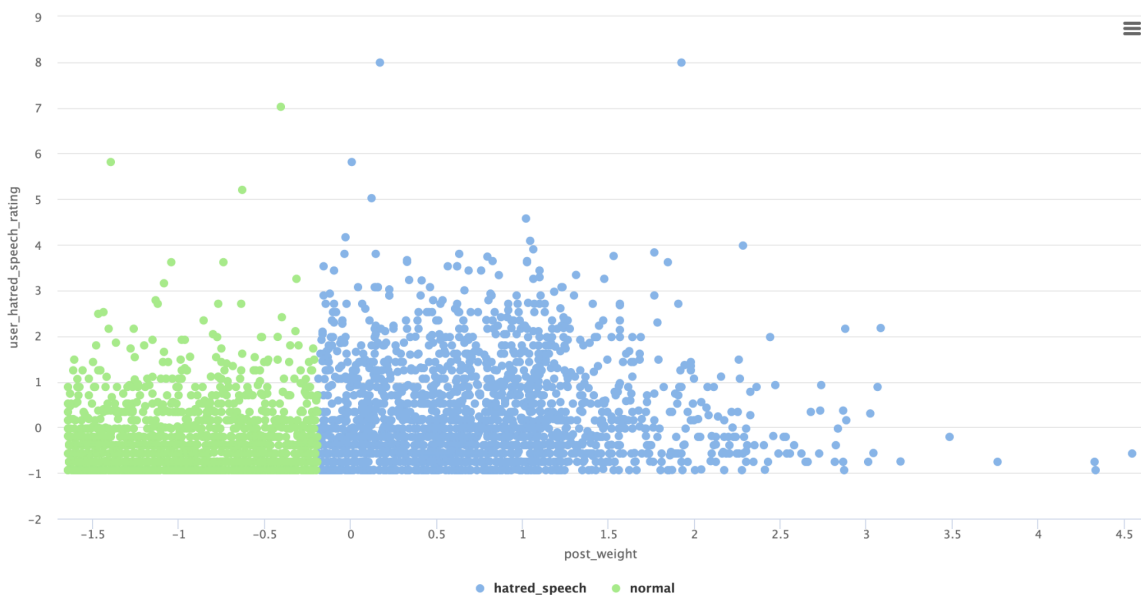
- Za target kolonu odabrati kolonu **category**.
- Odabrati opciju **Keep original** pri upitu da li se želi mijenjati struktura kolona.
- Odabrati opciju **Perform normalization**. Ne odabirati opciju za vršenje PCA.
- Pokrenuti automatsko prečišćavanje podataka.
- Nakon prečišćavanja, odabrati opciju **Commit cleanse**.

Prikazati izgled *dataseta* nakon završetka prečišćavanja podataka:

2d_dataset_small

Configure the desired chart with the settings on the left. ⓘ

CANCEL



Izvršiti kratku analizu prikaza. Šta se promijenilo u strukturi *dataseta*? Da li *dataset* još uvijek sadrži tri kolone i kakve su sada njihove vrijednosti? Koje su moguće prednosti, a koji nedostaci ovog prečišćavanja podataka?

- Struktura *dataseta* se promijenila tako da je izvršena normalizacija kolona, osim target kolone
- Dataset i dalje sadrži tri kolone, a njihove vrijednosti (osim target kolone) su normalizovane
- Prednosti ovog preciscavanja podataka su jednostavnost i normalizovanost vrijednosti u kolonama, nedostatak može predstavljati nepoznavanje sadržaja kolona odnosno možda postoji drugi bolji metod za neku specifičnu kolonu

Sada se vratiti na prethodni ekran i odabrati više opcija ..., kao što je prikazano na slici ispod.

2d_dataset_small

Add new data sets on the left. Details for the selected data are shown below. You can change the data with the following actions. ⓘ

✂ TRANSFORM ✂ CLEANSE 📊 GENERATE ∑ PIVOT ➡ MERGE

MODEL CHARTS CREATE PROCESS HISTORY



Odabrati opciju **Export**, a zatim opciju **Repository**. Spasiti prečišćeni dataset u **Local repository**, pritom mu dajući ime po izboru.

Zadatak 2. (Pronalazak outlieria u podacima)

Pokrenuti RapidMiner Studio okruženje, a zatim na početnom ekranu odabrati opciju **Auto Model**, kao što je prikazano na slici ispod. Opciji za kreiranje automatskog modela moguće je pristupiti i bez restartovanja okruženja, na gornjem srednjem dijelu ekrana.

Start with



Blank Process

Start a new process from scratch in the design view.



Turbo Prep

Prepare your data interactively: transform, clean and combine data sets.

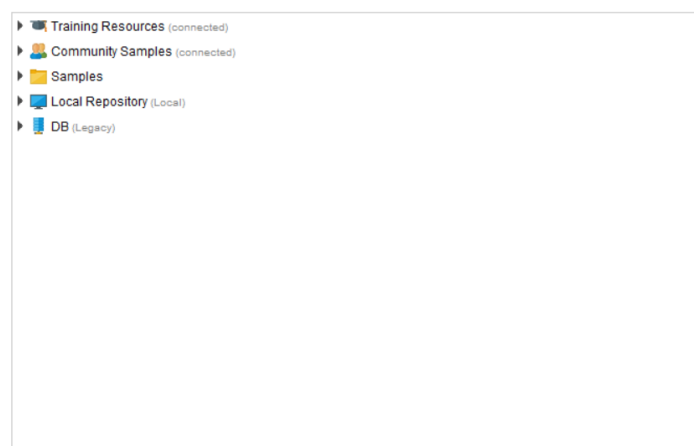


Auto Model

Build and optimize models using automated machine learning.

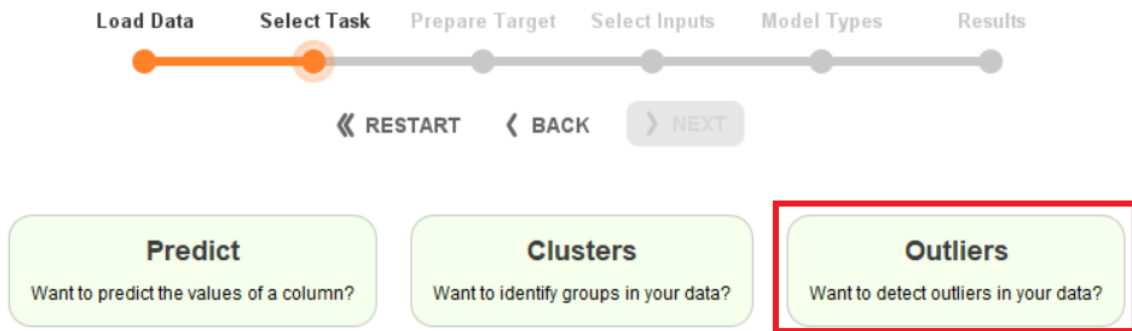
Provjeriti da li u folderu **Local Repository** već postoje dva dataseta (izvorni **2d_dataset_small** i prethodno sačuvani normalizovani dataset). Ukoliko izvorni dataset nije dostupan, odabrati opciju za importovanje postojećih podataka koji će se koristiti kao ulaz za željeni klasifikator, kao što je prikazano na slici ispod.

Select Data for a New Model

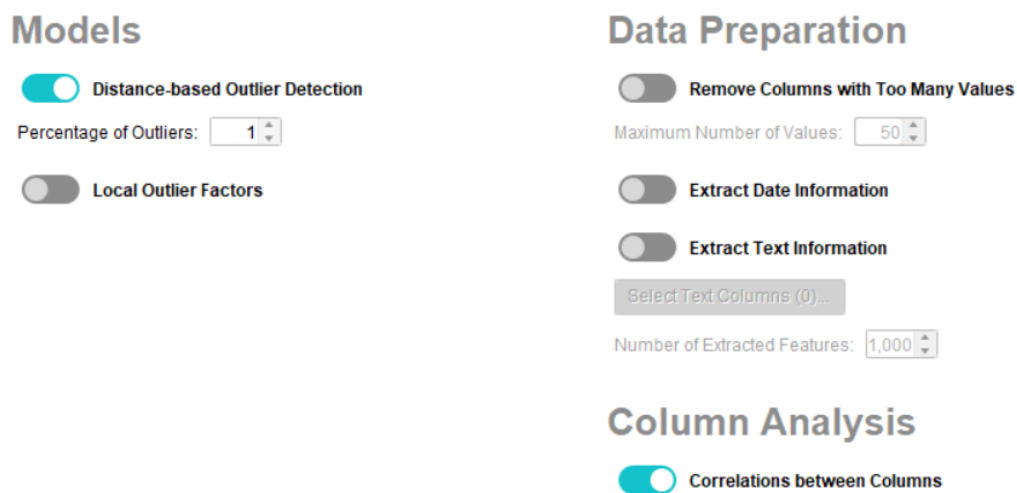


IMPORT NEW DATA

Odabrati izvorni dataset, a zatim kliknuti na opciju **Next**. Zatim odabrati opciju **Outliers**, kao što je prikazano na slici ispod.



U sljedećem koraku odabrati **Select All**, kako bi se razmatrale sve kolone pri detekciji outliera. Nakon toga podesiti postavke tako da izgledaju kao na slici ispod.



The screenshot shows the configuration options for the Outliers detection model. It is divided into three sections: Models, Data Preparation, and Column Analysis.

- Models:**
 - ☒ Distance-based Outlier Detection
 - Percentage of Outliers: 1
 - ☐ Local Outlier Factors
- Data Preparation:**
 - ☐ Remove Columns with Too Many Values
 - Maximum Number of Values: 50
 - ☐ Extract Date Information
 - ☐ Extract Text Information
 - Select Text Columns (0)...
 - Number of Extracted Features: 1,000
- Column Analysis:**
 - ☒ Correlations between Columns

Kliknuti na opciju **Run** kako bi se pokrenulo traženje outliera u datasetu. Izvršavanje bi trebalo trajati oko 3 minute.

Prikazati tabelu podataka sa kolonom **outlier** koja pokazuje da li je instanca označena kao outlier ili ne:

<div> <div>Load Data</div> <div>Select Task</div> <div>Prepare Target</div> <div>Select Inputs</div> <div>Model Types</div> <div>Results</div> </div> <div> <div>RESTART</div> <div>BACK</div> <div>OPEN PROCESS</div> <div>EXPORT</div> </div>				
Distance-based Outliers – Outlier Data				
Row No.	outlier	category	post_weight	user_hatred_speech_rating
1	true	hatred_speech	1213.200	0.674
2	true	hatred_speech	3107.300	0.360
3	true	hatred_speech	3250.556	0.362
4	true	normal	121.714	0.396
5	true	hatred_speech	3526	0.100
6	true	hatred_speech	1831.200	0.625
7	true	hatred_speech	2440.250	0.420
8	true	hatred_speech	3237.600	0.220
9	true	hatred_speech	1136.688	0.761
10	true	hatred_speech	1112.500	0.580
11	true	normal	600.889	0.420
12	true	hatred_speech	3207.500	0.156
13	true	hatred_speech	2400.333	0.520
14	true	hatred_speech	2699.714	0.560
15	true	hatred_speech	2344.062	0.543
16	true	hatred_speech	1107.765	0.540

Koliki broj instanci je identificiran kao *outlier*? Da li vrijednosti u kolonama za indeks govora mržnje i težinu posta kod ovih instanci imaju neke specifične vrijednosti u odnosu na druge instance? Objasniti odgovor.

- 36 instanci je identificirano kao outlier
- Vrijednosti u kolonama za težinu posta nemaju neke specifične vrijednosti u odnosu na druge instance, dok za indeks govora mržnje ovaj koeficijent je povećan u odnosu na ostale non-outlier vrijednosti

Odabrati opciju **Restart**. Nakon toga kao izvorne podatke selektovati prethodno spašeni normalizovani dataset i ponoviti isti postupak traženja outliera u podacima.

Prikazati tabelu podataka sa kolonom **outlier** koja pokazuje da li je instanca označena kao *outlier* ili ne:

<div> <div>Load Data</div> <div>Select Task</div> <div>Prepare Target</div> <div>Select Inputs</div> <div>Model Types</div> <div>Results</div> </div>				
<div> <div>RESTART</div> <div>BACK</div> <div>OPEN PROCESS</div> <div>EXPORT</div> </div>				
Distance-based Outliers – Outlier Data				
Row No.	outlier ↓	category	post_weight	user_hatred_speech_rating
1	true	hatred_speech	0.121	5.020
2	true	hatred_speech	2.878	2.159
3	true	hatred_speech	3.087	2.175
4	true	normal	-1.468	2.486
5	true	hatred_speech	3.488	-0.210
6	true	hatred_speech	1.020	4.574
7	true	hatred_speech	1.907	2.706
8	true	hatred_speech	3.068	0.884
9	true	hatred_speech	0.009	5.812
10	true	hatred_speech	-0.026	4.164
11	true	normal	-0.770	2.706
12	true	hatred_speech	3.024	0.303
13	true	hatred_speech	1.849	3.617
14	true	hatred_speech	2.285	3.982
15	true	hatred_speech	1.767	3.831
16	true	hatred_speech	-0.033	3.799

Koliki broj instanci je identificiran kao *outlier*? Da li vrijednosti u kolonama za indeks govora mržnje i težinu posta kod ovih instanci imaju neke specifične vrijednosti u odnosu na druge instance? Objasniti odgovor.

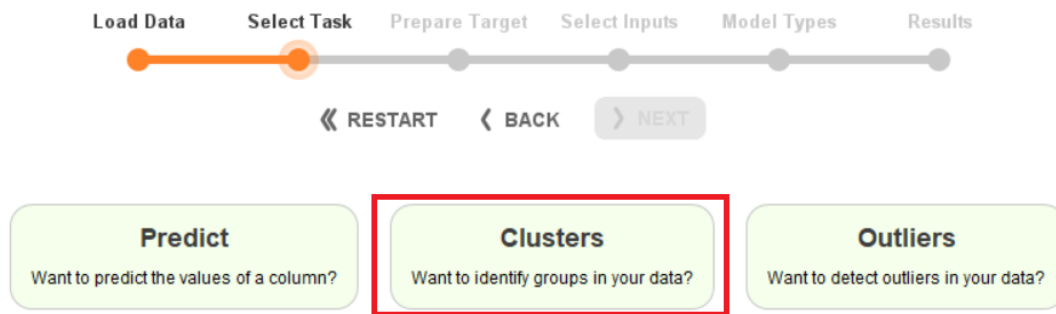
- Identificirano je 36 outliers
- Vrijednosti u kolonama zatezinu posta nemaju neke specifične vrijednosti u odnosu na druge instance, dok za za indeks govora mržnje ovaj koeficijent je povećan u odnosu na ostale non-outlier vrijednosti

Da li je broj instanci koje su identificirane kao *outlieri* isti za normalizovani i izvorni *dataset*? Šta to govori o postupku detekcije *outliera*? Šta sam broj detektovanih instanci govori o strukturi *dataseta*?

- Broj outliera je isti za normalizovani i izvorni dataset
- Govori da detekcija outliera ne ovisi o reprezentaciji podataka i normalizaciji
- Sam broj detektovanih instanci govori da je na 3582 instanci samo 36 outliers, sto je i realno za očekivati, samim tim može se zaključiti da je fino balansiran dataset po pitanju outliera.

Zadatak 3. (Clustering podataka)

Ponovo odabrati opciju **Restart** i selektovati izvorni **2d_dataset_small** dataset. Sada umjesto opcije za detekciju outliera selektovati opciju **Clusters**, kao što je prikazano na slici ispod.



Ponovo selektovati sve kolone dataseta koje će se koristiti pri vršenju clusteringa podataka, a zatim podesiti postavke tako da izgledaju kao na slici ispod.

Models

☒ **k-Means Clustering**
Number of Clusters:

☐ **x-Means Clustering**
Minimal Number of Clusters: Maximal Number of Clusters:

Data Preparation

☐ **Remove Columns with Too Many Values**
Maximum Number of Values:

☐ **Extract Date Information**

☐ **Extract Text Information**
Selected Text Columns (0)...

Number of Extracted Features:

☐ **Automatic Feature Selection**
Additional Time (in Minutes):

Final Feature Set should be

Column Analysis

☒ **Correlations between Columns**

Kliknuti na opciju **Run** kako bi se pokrenuo clustering podataka. Izvršavanje bi trebalo trajati par sekundi.

Prikaz informacija o broju *clusters* i broju instanci koje se u njima nalaze (*Summary* stranica).

k-Means – Summary

Number of Clusters: 2

Cluster 0

1,632

post_weight is on average **54.95%** smaller, **user_hatred_speech_rating** is on average **42.31%** smaller

Cluster 1

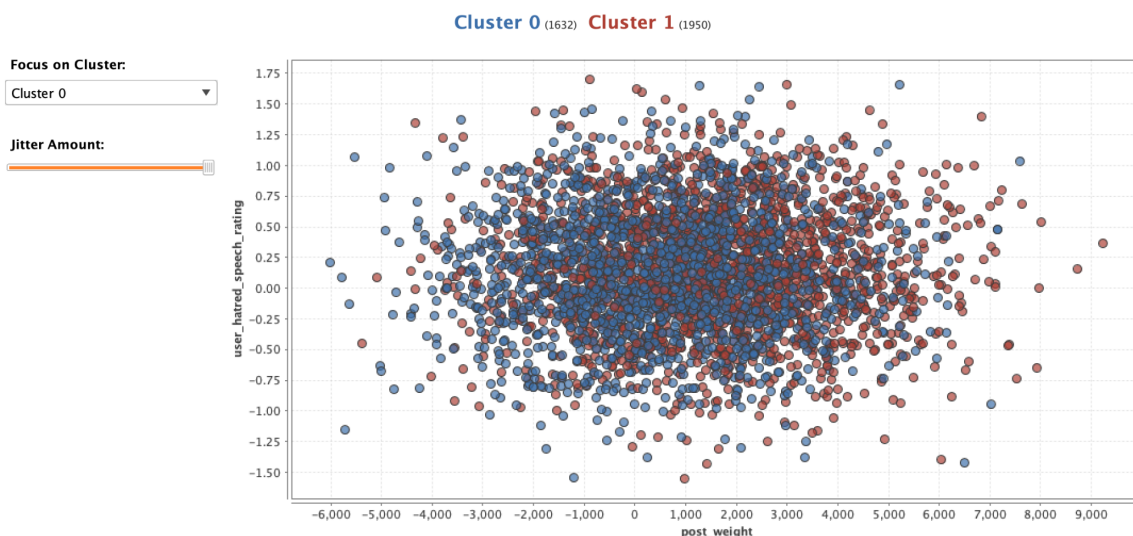
1,950

post_weight is on average **45.99%** larger, **user_hatred_speech_rating** is on average **35.41%** larger

Odabrali opciju **Scatter plot** s lijeve strane ekrana, a zatim povećati parametar **Jitter amount** na maksimum.

Prikaz grafika sa clusteriziranim podacima:

k-Means – Scatter Plot

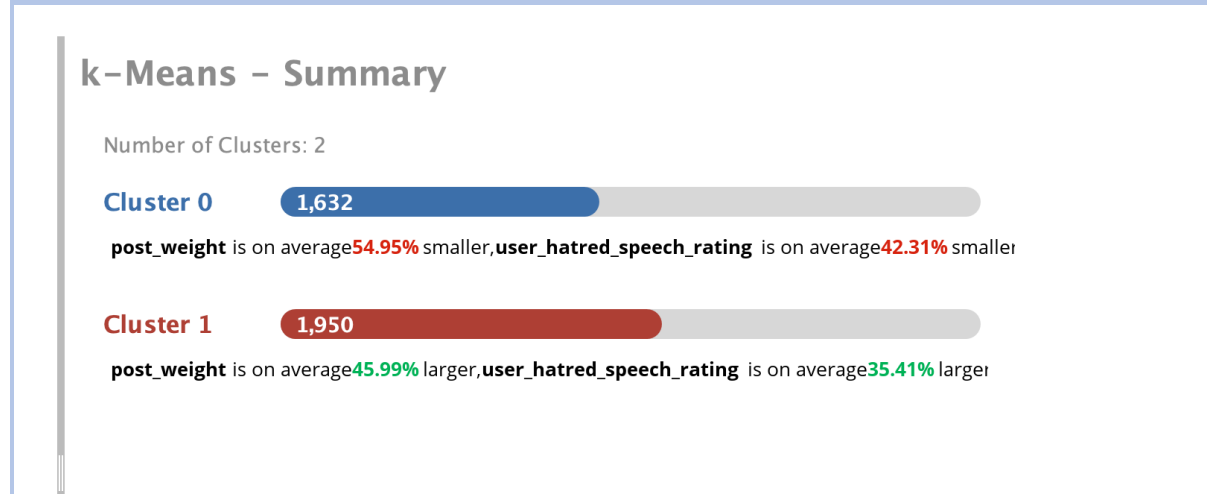


Da li se prethodno detektovani *outlieri* jasno ističu na grafiku i gdje se nalaze? Da li je moguće jasno razgraničiti granice *cluster*a, ili su tačke izmiješane u njima? Šta to govori o formi podataka? Objasniti odgovor.

- Prethodno detektovane *outliere* možemo jasno istaknuti na grafiku, nalaze se oko glavnih dva *clusters*
- Nije moguće jasno razgraničiti granice *cluster*e, tačke su izmiješane
- Govori da struktura podataka nije fino podesena da se clustering odradi tj. da bi se mogao imati bolji prikaz *cluster*a

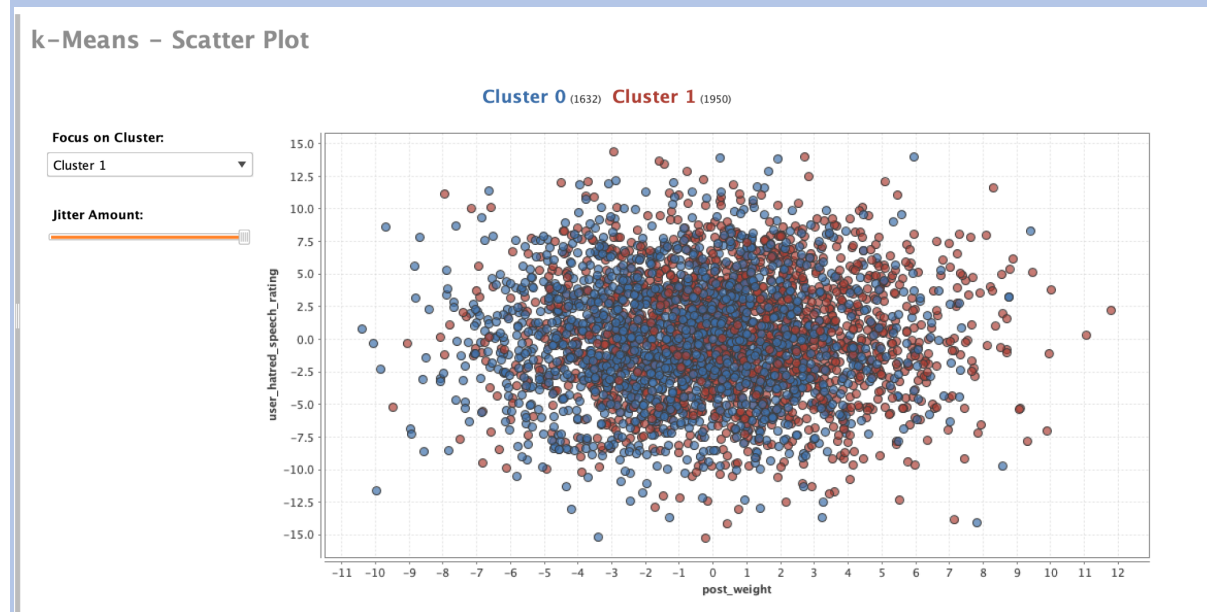
Odabрати opciju **Restart**. Nakon toga kao izvorne podatke selektovati prethodno spašeni normalizovani dataset i ponoviti isti postupak formiranja clustera.

Prikaz informacija o broju *cluster*a i broju instanci koje se u njima nalaze (*Summary* stranica).



Odabрати opciju **Scatter plot** s lijeve strane ekrana, a zatim povećati parametar **Jitter amount** na maksimum.

Prikaz grafika sa *clusteriziranim* podacima:



Da li se prethodno detektovani *outlieri* jasno ističu na grafiku i gdje se nalaze? Da li je moguće jasno razgraničiti granice *cluster*a, ili su tačke izmiješane u njima? Šta to govori o formi podataka? Objasniti odgovor.

Dubinska Analiza Podataka

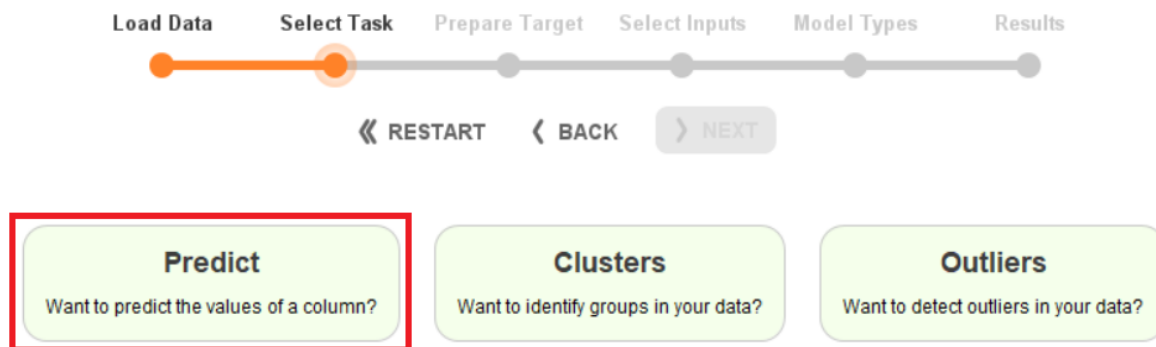
- Prethodno detektovane outliere mozemo jasno istaknuti na grafiku, nalaze se oko glavna dva clusters
- Nije moguće ni sad jasno razgraniciti granice clustera, tačke su izmjesane
- Govori da struktura podataka nije fino podesena da se clustering odradi tj. da bi se mogao imati bolji prikaz clustera

Da li postoje razlike između *cluster*a u izvornom *datasetu* i u normalizovanom *datasetu*? Kakve su razlike, ukoliko ih ima? Nad kojim podacima je *clustering* dao bolje rezultate i zašto? Objasniti odgovor.

- Ne postoje razlike između clustera u različitim datasetovima
- Rezultati su isti, neovisno o normalizovanosti podataka tj. neovisno o datasetu

Zadatak 4. (Klasifikacija podataka)

Ponovo odabrati opciju **Restart** i selektovati izvorni **2d_dataset_small** dataset. Sada umjesto opcije za detekciju outliera selektovati opciju **Predict**, kao što je prikazano na slici ispod. Kliknuti na kolonu **category** kako bi se specificiralo da je to kategorička kolona.



U sljedećem koraku (*Prepare target*) ostaviti izvorne postavke. Nakon toga ponovo selektovati sve kolone (dimenzije na osnovu kojih će se vršiti klasifikacija).

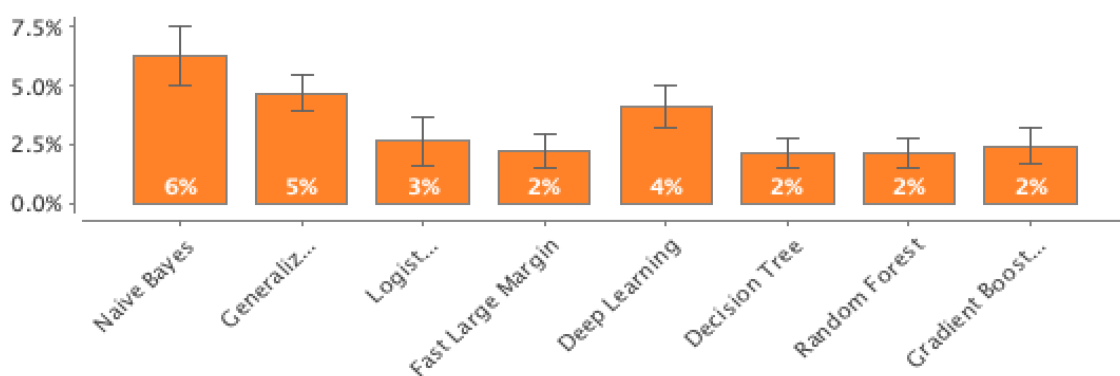
Da li kolona **post_weight** ima status žute boje i zašto? Šta znači visok koeficijent korelacije? Objasniti odgovor.

- To znači da vrijednosti te varijable utiče dosta na rezultate predikcije, odnosno da je u visokoj korelaciji sa target varijablom tj. rezultat dosta ovisi od te varijable.

U narednom koraku odabrati sve klasifikatore **osim SVM** (8 modela) i odabrati opciju **Run**. Vrijeme izvršavanja za svaki pojedinačni klasifikator trebalo bi biti između 10 s i 60 s.

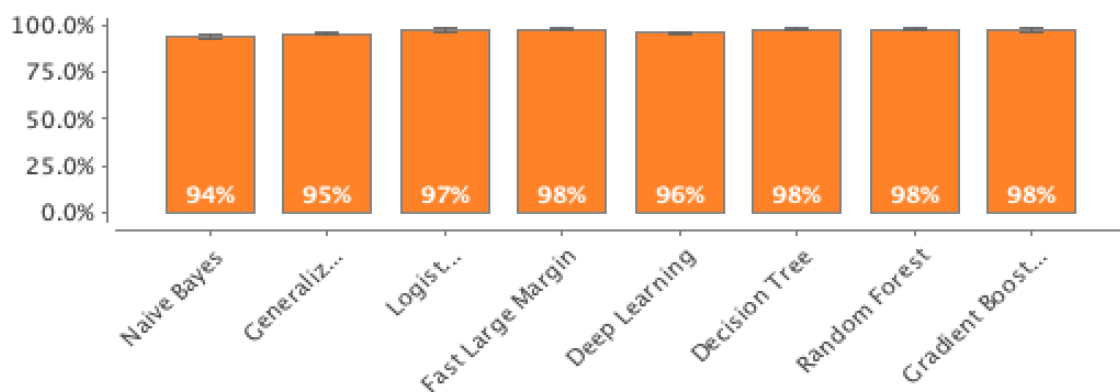
Prikaz **classification error** grafika na *Overview* stranici:

Classification Error



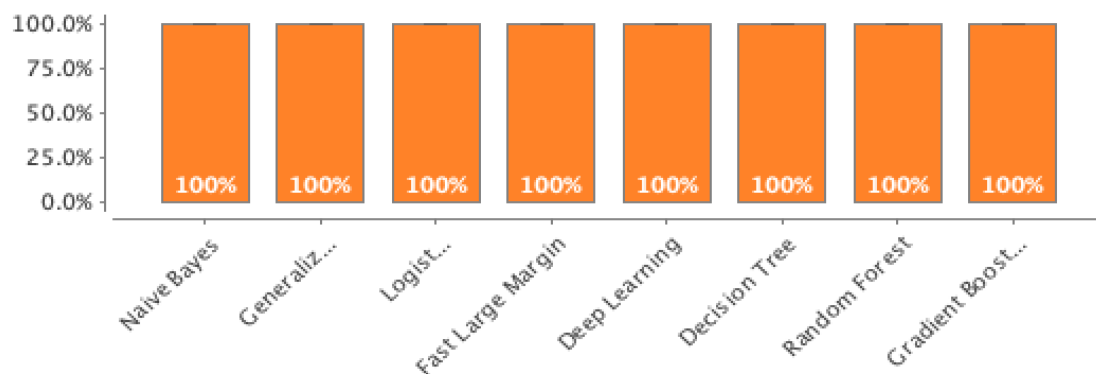
Prikaz **accuracy** grafika na *Overview* stranici:

Accuracy



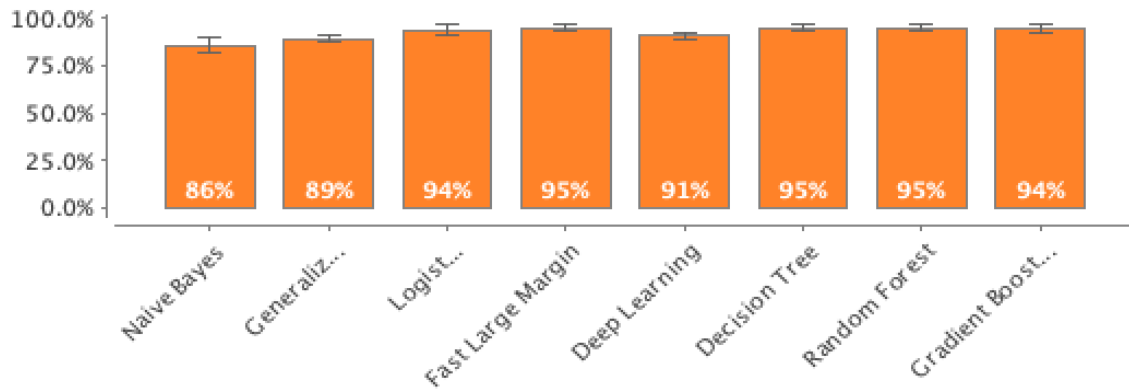
Prikaz **sensitivity** grafika na *Overview* stranici:

Sensitivity



Prikaz **specificity** grafika na *Overview* stranici:

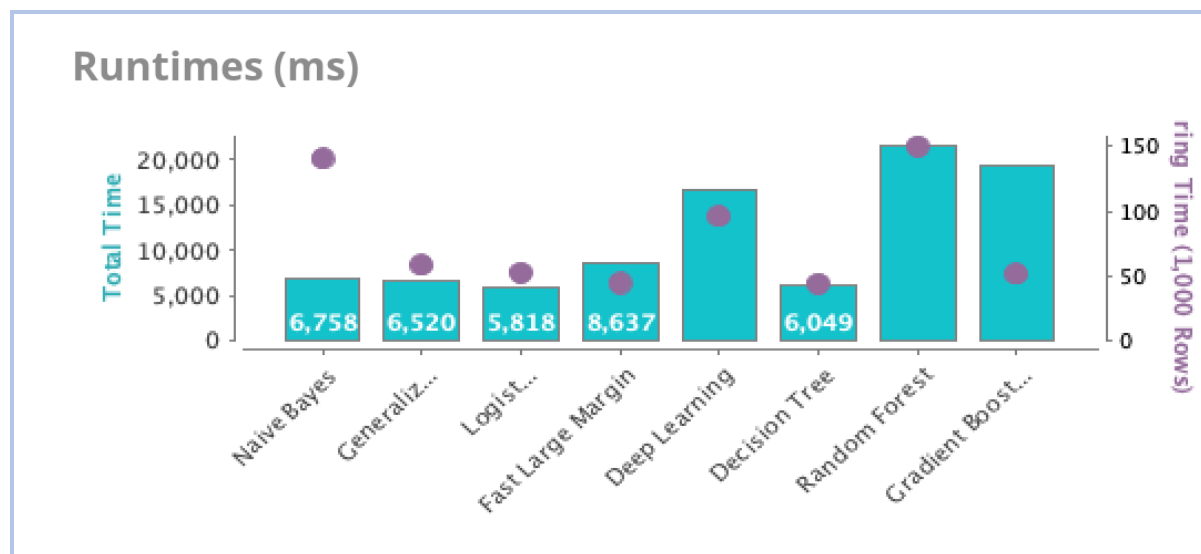
Specificity



Šta svi prethodni grafici nedvosmisleno prikazuju? Kakva je razlika između senzitivnosti i specifičnosti modela? Zbog čega jedno ima vrijednost 100%, a drugo promjenjivu vrijednost? Obrazložiti odgovor.

- Classification error predstavlja grešku u predikciji za svaki navedeni model
- Accuracy predstavlja preciznost tj. tačnost klasifikacije za svaki navedeni model
- Sensitivity predstavlja mjeru procenta koji kaže za sve vrijednosti koje su pozitivne i pozitivno su klasificirane, odnosno taj odnos
- Specificity predstavlja mjeru procenta koji kaže za sve vrijednosti koje su negativne i negativno su klasificirane, odnosno taj odnos
- Razlika između sensitivity i specificity predstavlja target vrijednost, odnosno jedan gleda positive values, drugi negative values
- Sensitivity ima 100% vrijednost zato što je model sve pozitivne vrijednosti pogodio kao pozitivne, a specificity ima promjenjivu zato što od modela do modela zavisi, da li je model pogodio negativne vrijednosti kao negativne zapravo

Prikaz **runtimes (ms)** grafika na *Overview* stranici:



Povezati vrijeme izvršavanja sa tačnošću klasifikatora. Da li je najbrži klasifikator i najmanje tačan? Da li je najsporiji klasifikator najtačniji? Da li uopće postoji korelacija između tačnosti klasifikatora i vremena izvršavanja? Objasniti odgovor.

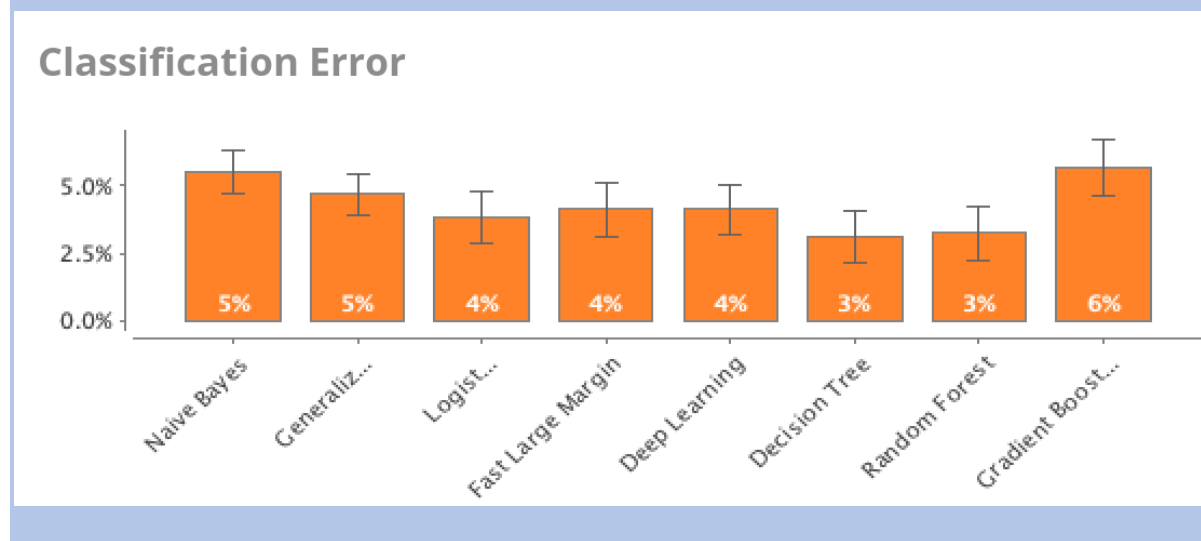
- Najbrži klasifikator nije najmanje tačan
- Najsporiji klasifikator je jedan od najtačnijih (jer ih ima više sa 98%)
- Ne postoji korelacija između tačnosti klasifikatora i vremena izvršavanja

Koji klasifikator(i) ima(ju) najveću tačnost, a koji najmanje vrijeme izvršavanja?

- Fast Large Margin, Decision Tree, Random Forest, Gradient Boost imaju najveće tačnosti
- Najmanje vrijeme izvršavanja ima Logistic Regression

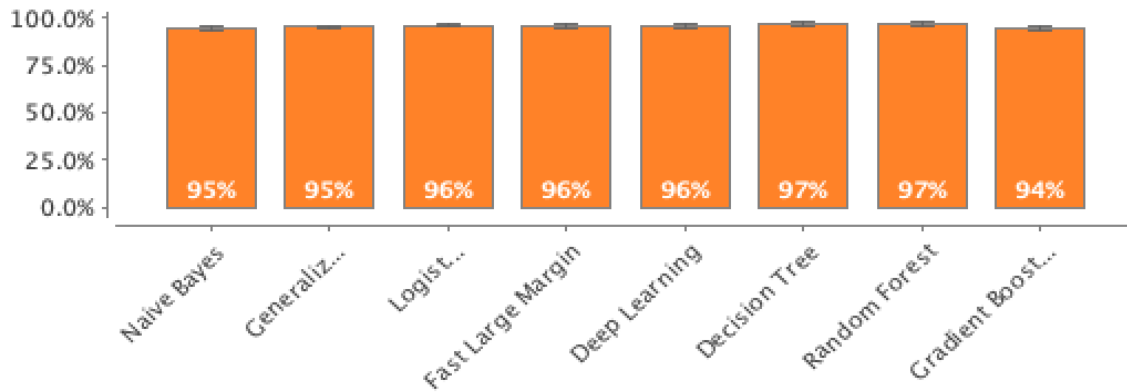
Odabrali opciju **Restart**. Nakon toga kao izvorne podatke selektovati prethodno spašeni normalizovani dataset i ponoviti isti postupak klasifikacije podataka.

Prikaz **classification error** grafika na *Overview* stranici:



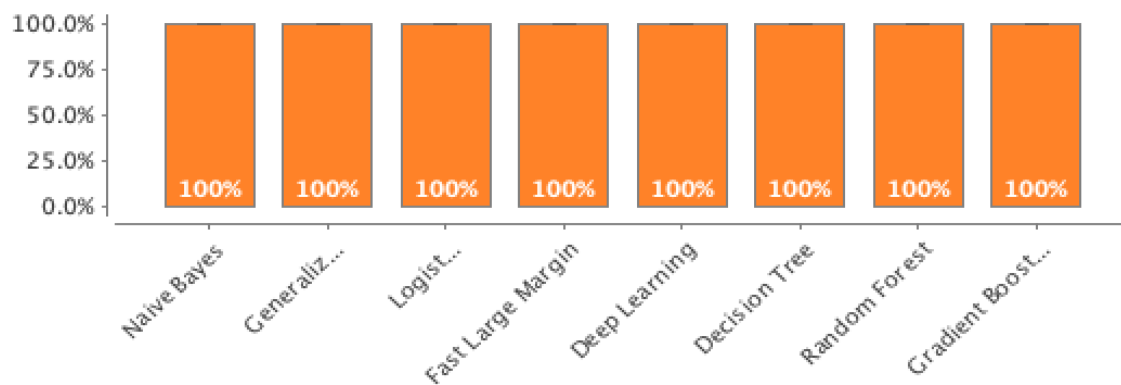
Prikaz **accuracy** grafika na *Overview* stranici:

Accuracy



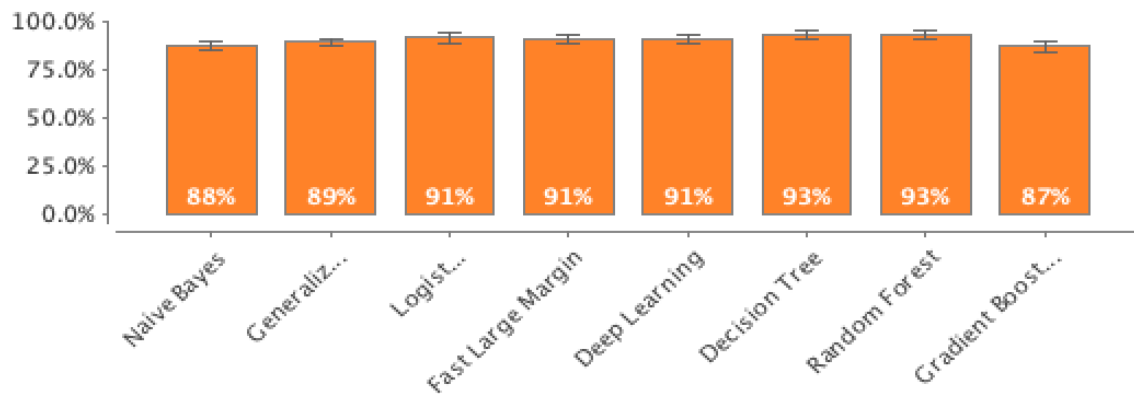
Prikaz **sensitivity** grafika na *Overview* stranici:

Sensitivity



Prikaz **specificity** grafika na *Overview* stranici:

Specificity



Šta svi prethodni grafici nedvosmisleno prikazuju? Kakva je razlika između senzitivnosti i specifičnosti modela? Zbog čega jedno ima vrijednost 100%, a drugo promjenjivu vrijednost? Obrazložiti odgovor.

- Classification error predstavlja grešku u predikciji za svaki navedeni model
- Accuracy predstavlja preciznost tj. tačnost klasifikacije za svaki navedeni model
- Sensitivity predstavlja mjeru procenta koji kaže za sve vrijednosti koje su pozitivne i pozitivno su klasificirane, odnosno taj odnos
- Specificity predstavlja mjeru procenta koji kaže za sve vrijednosti koje su negativne i negativno su klasificirane, odnosno taj odnos
- Razlika između sensitivity i specificity predstavlja target vrijednost, odnosno jedan gleda positive values, drugi negative values
- Sensitivity ima 100% vrijednost zato što je model sve pozitivne vrijednosti pogodio kao pozitivne, a specificity ima promjenjivu zato što od modela do modela zavisi, da li je model pogodio negativne vrijednosti kao negativne zapravo

Prikaz **runtimes (ms)** grafika na *Overview* stranici:

Runtimes (ms)



Povezati vrijeme izvršavanja sa tačnošću klasifikatora. Da li je najbrži klasifikator i najmanje tačan? Da li je najsporiji klasifikator najtačniji? Da li uopće postoji korelacija između tačnosti klasifikatora i vremena izvršavanja? Objasniti odgovor.

- Najbrzi klasifikator nije najmanje tačan
- Najsporiji klasifikator je jedan od najtačnijih (jer ih ima više sa 97%)
- Ne postoji korelacija između tačnosti klasifikatora i vremena izvršavanja

Koji klasifikator(i) ima(ju) najveću tačnost, a koji najmanje vrijeme izvršavanja?

- Decision Tree, Random Forest imaju najveće tačnosti
- Najmanje vrijeme izvršavanja ima Gradient Boost

Usporediti rezultate dobivene nad izvornim *datasetom* i nad normalizovanim podacima. Postoji li razlika između tačnosti i brzine klasifikacije? Da li se išta promijenilo u pojedinačnim ili ukupnim rezultatima? Obrazložiti odgovor.

- Postoji razlika, za prvih par algoritama je brzina povećanja, a ujedno i tačnost u odnosu na prvi dataset. Pojedinačni rezultati za određene modele su povećani, a sa druge strane par zadnjih modela u drugom slučaju je smanjena tačnost.

Da li se može izvesti zaključak da je normalizacija podataka dovela do poboljšanja rezultata? Obrazložiti odgovor.

- Normalizacija podataka nije dovela toliko do poboljšanja rezultata, ali jeste do smanjenja vremena izvršavanja za prvih par algoritama.