Klasifikacija spam i ham emailova

Agenda

- Uvod
- Skup podataka
- Pretprocesiranje podataka
- Metodologija
- Naive Bayes
- Support Vector Classifier (SVC)
- Feedforward Neural Network
- Zaključak

Uvod..

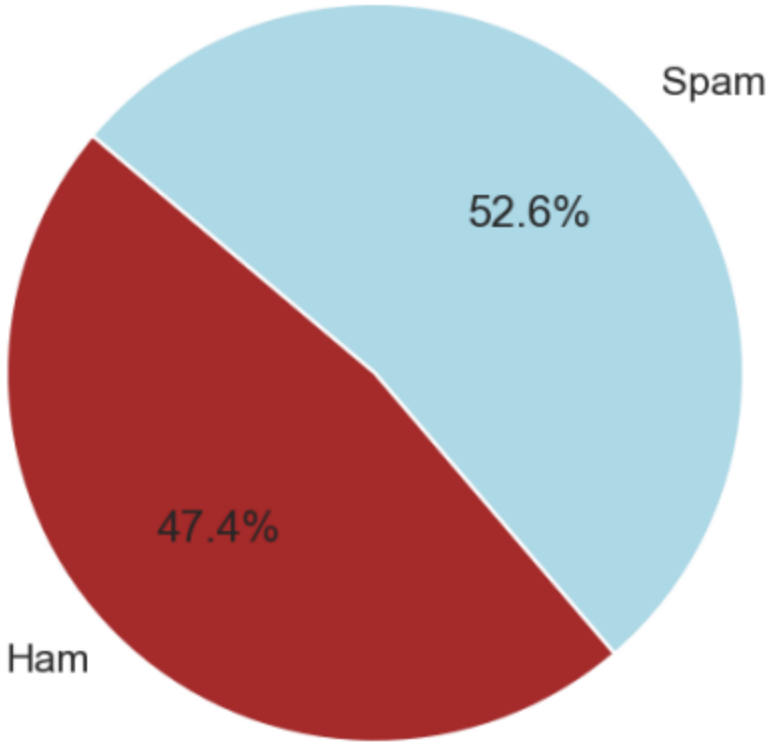
Uvod

- Email je jedan od glavnih načina poslovne komunikacije.
- Spam poruke predstavljaju veliku pretnju bezbednosti.
- Problem predstavlja identifikacija karakteristika koje razlikuju neželjene email poruke od legitimnih korišćenjem mašinskog učenja.

Skup podataka...

Skup podataka...

- Koristi skup podataka koji sadrži 83,446 zapisa emailova koji su označeni kao spam ili ne spam. Ovaj skup je formiran kombinovanjem 2007 TREC Public Spam Corpus i Enron-Spam Dataset.
- Svaki email je označen sa '1' ako je klasifikovan kao spam, dok je označen sa '0' ako je legitimna poruka.



Predprocesiranje podataka...

- Proces predprocesiranja podataka emailova sastojao se iz izvlačenja stop reči, pretvaranja teksta u lowercase, uklanjanja drugih znakova I vektorizacije teksta.
- Proces predprocesiranja klasifikacije tih emailova na odredjene kategorije sastojao se samo u ucitavanju numericke vrednosti koja reprezentuje kategoriju.

Metodologija

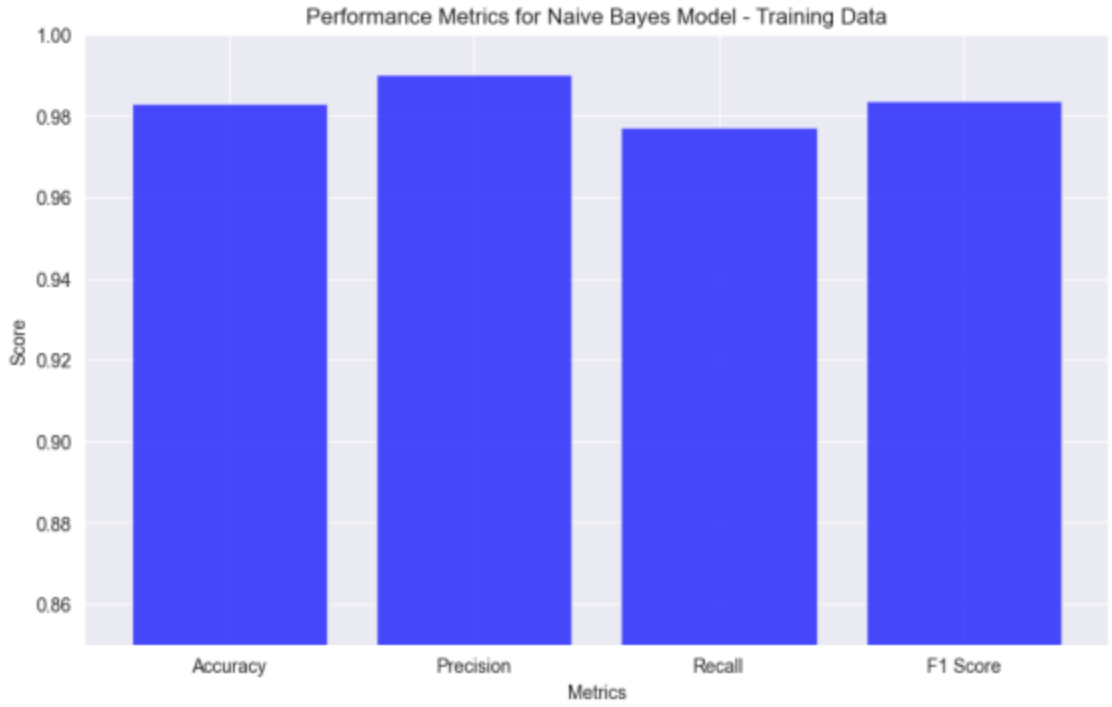
- Tokom rešavanja ovog problema testirali smo performanse klasifikacije različitih modela:
  - Naive Bias
  - Support Vector Classifier (SVC)
  - Feedforward Neural Network
- Svaki od modela sproveli smo kroz faze obrade, treniranja, testiranja I konfigurisanja radi otkrivanja adekvatnog modela za naše potrebe I problem

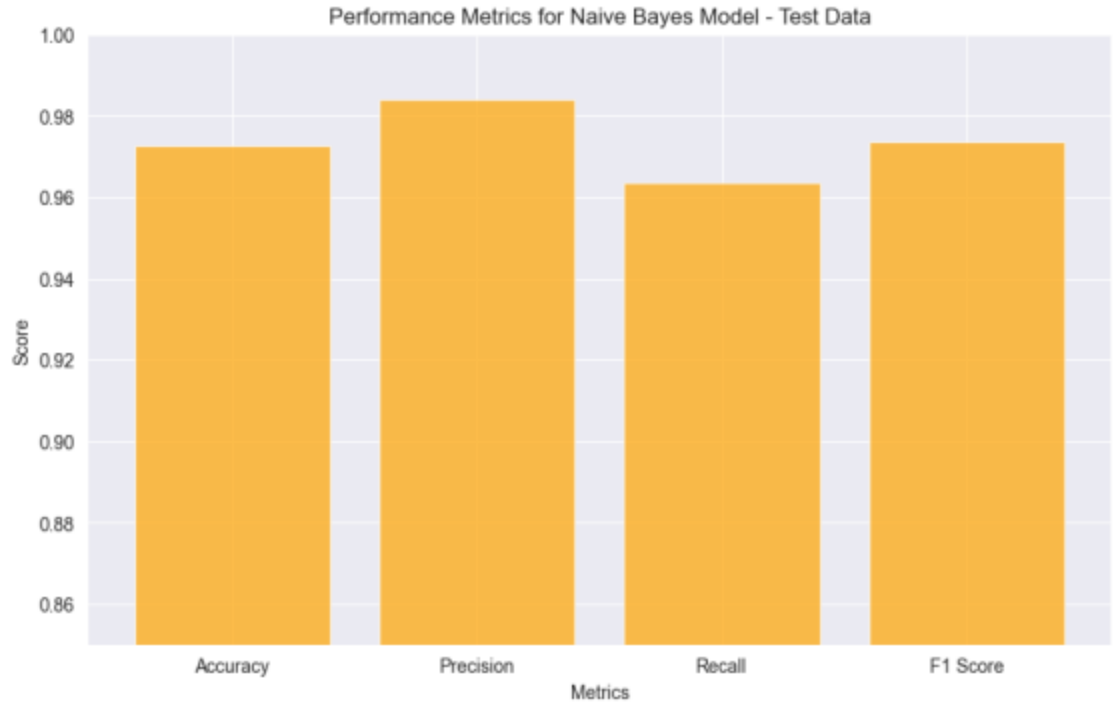
Naive Bayes

- Osnovna ideja iza naivnog Bayesa je primena Bayesovog teorema sa "naivnom" pretpostavkom da su svi atributi nezavisni jedni od drugih, što često nije realnost, ali olakšava računanje.
- Naivni Bayes je efikasan za rad sa velikim skupovima podataka i često se koristi za klasifikaciju teksta (kao što su spam filteri), medicinske dijagnoze, detekciju prevara, i druge zadatke.

Naive Bayes

- Sa defaultnom konfiguracijom nakon treniranja modela nad velikim skupom podataka dobijamo sledeće rezultate:

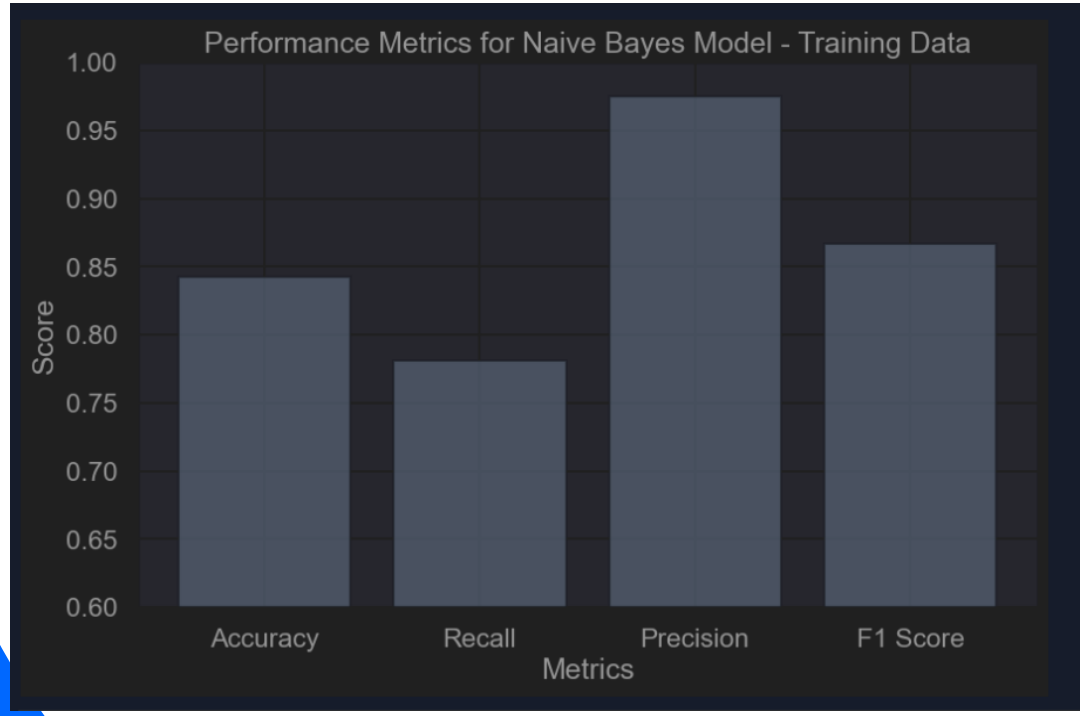


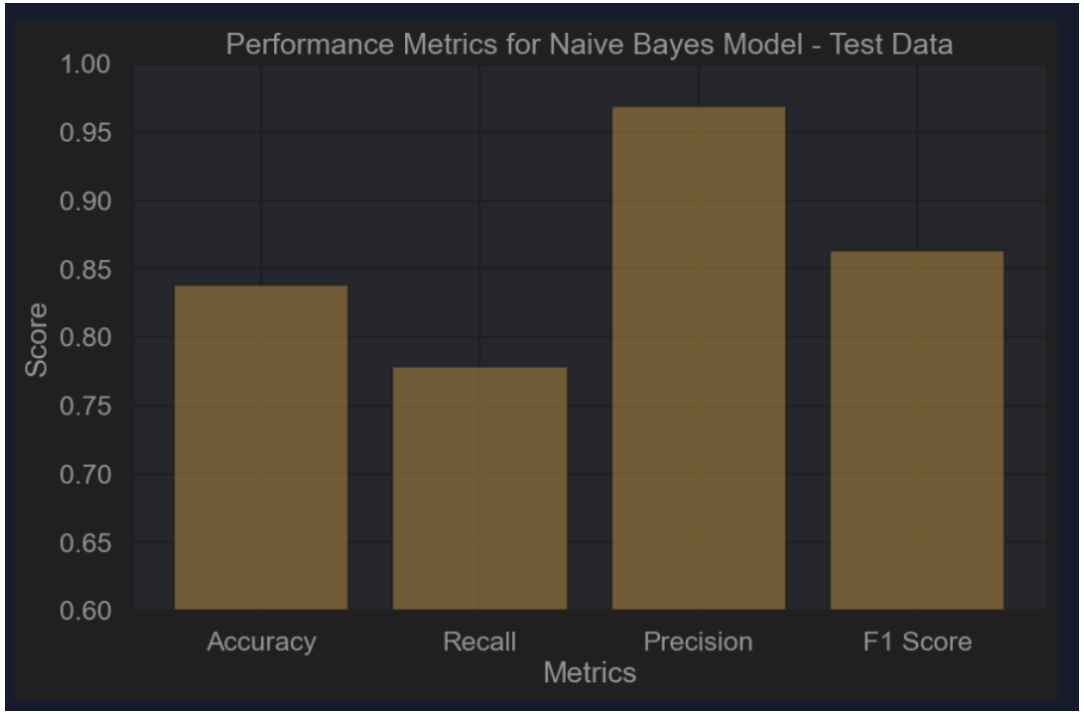


- Sve metrike su u proseku 97% nakon treniranja sa velikim skupom podataka.

Naive Bayes

- Nakon smanjenja skupa podataka na 100 primeraka rezultati metrika su:

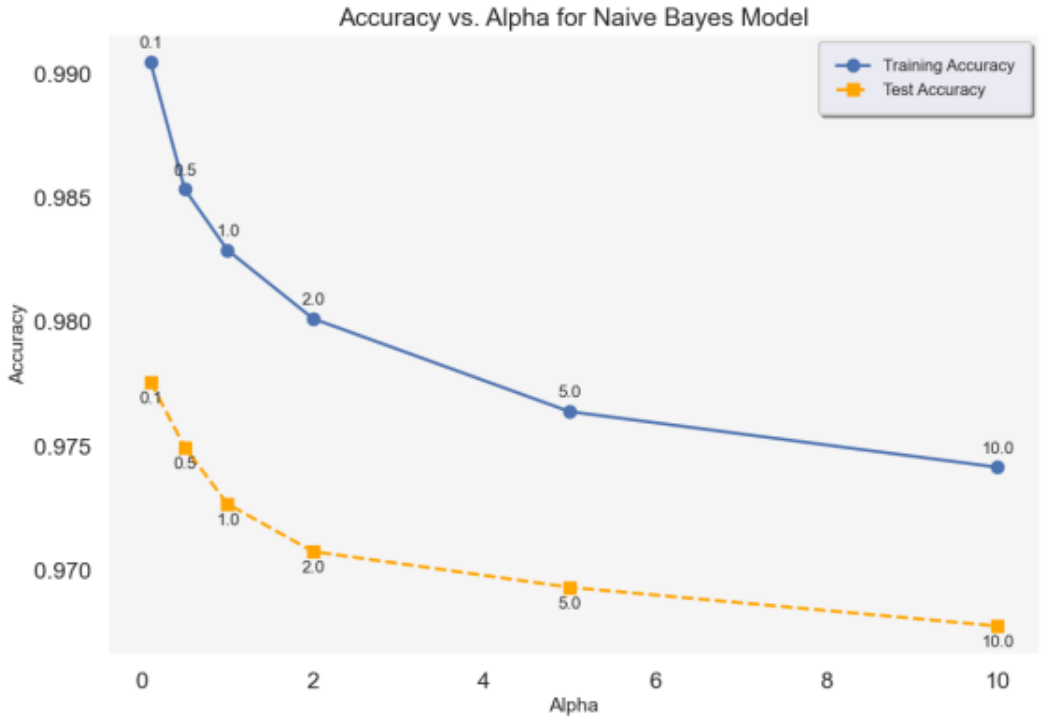




- Uočavamo da su metrike u proseku 85%, nakon daljeg istraživanja tek ispod 50 primeraka ulaznih primeraka performanse padaju ispod 50%.

Naive Bayes

- Konfigurisanjem alfa hiperparametara odredjujemo
- Visoka alfa (high alpha) -> underfitting (slaba prilagođenost). Dodajemo velike vrednosti svemu i time razređujemo signal u podacima.
- Niska alfa (low alpha) -> overfitting (pretjerana prilagođenost).



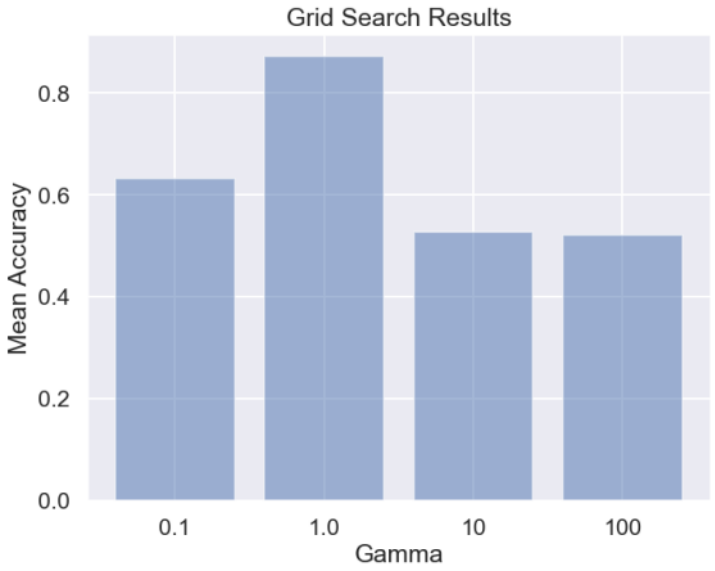
- Na osnovu grafikona, vrednost alfa oko 1.0 čini se kao dobar balans između overfittovanja i underfittovanja. U ovom trenutku, razlika između tačnosti na trening i test skupu je minimalna, a obe tačnosti su relativno visoke.

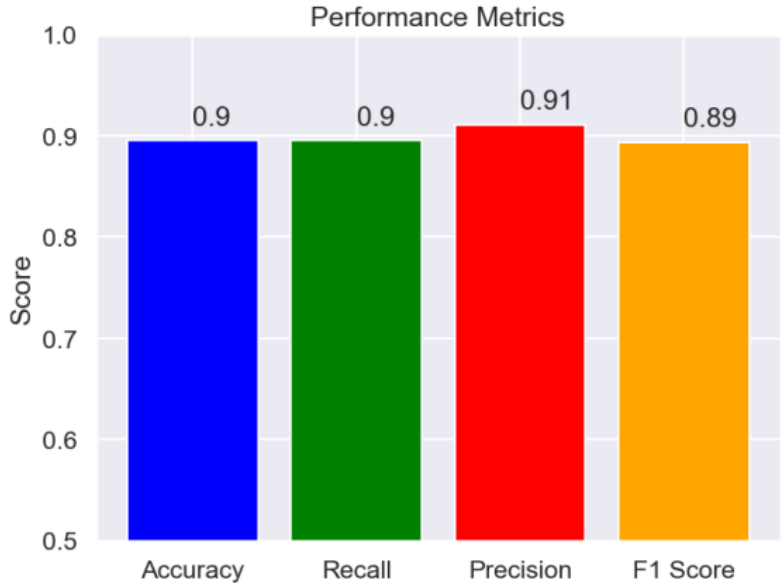
Support Vector Classifier (SVC)

- Nadgledani algoritam mašinskog učenja koji se koristi za klasifikaciju i regresiju. Najčešće se primenjuje za binarnu klasifikaciju, ali može se proširiti i na višeklasnu klasifikaciju.
- Osnovni principi:
  - **Hiper-ravan za separaciju**
  - **Support Vectors (potporni vektori)**
  - **Margin**
  - **Kernel trik**

Support Vector Classifier (SVC)

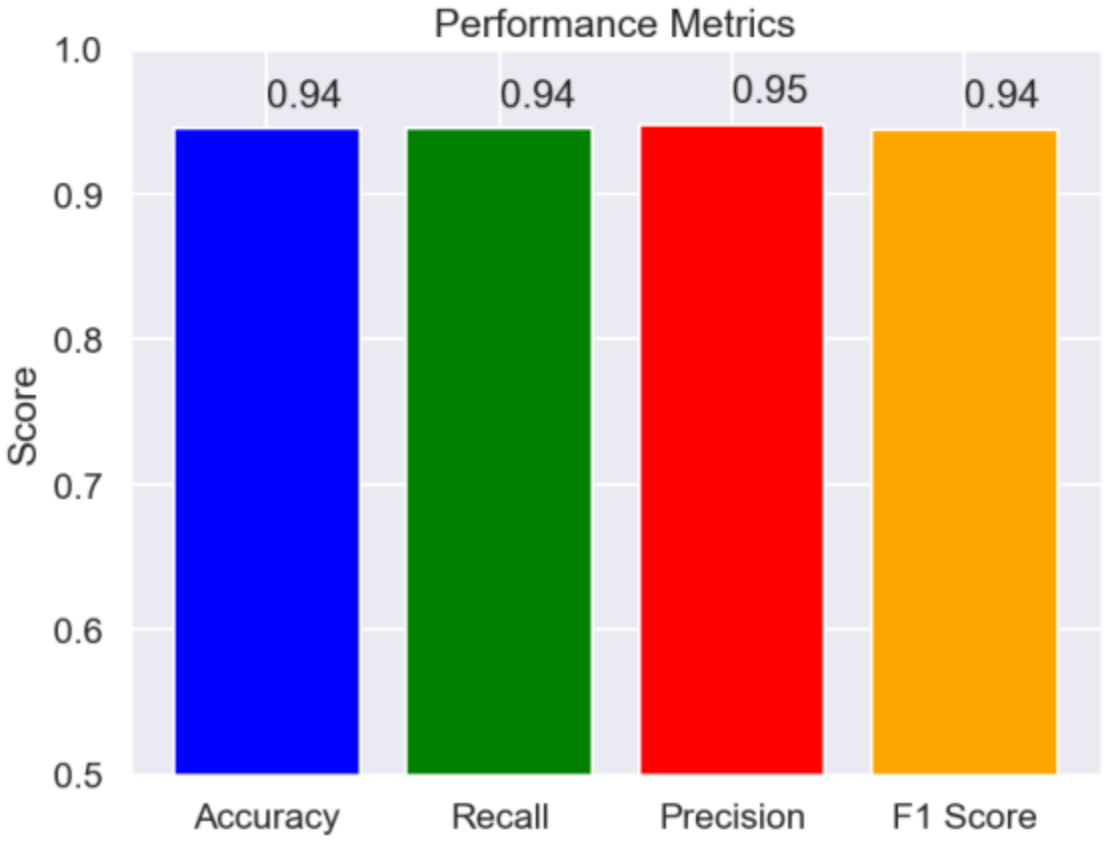
- Jedna od mana SVC-a je zahtevnost za resursima pa smo za testiranje ovog modela suzili skup podataka s kojim radimo.
- Pomoću GridSearcha tražili smo najadekvatniji svc model, modifikovanjem **samo** Gamma vrednosti.





Support Vector Classifier (SVC)

- Zatim pomoću GridSearcha tražili smo najadekvatniji svc model, modifikovanjem Gamma vrednosti I C parametra.
- Zapažamo da dodavanjem konfiguraciji istraživa nje C parametra poboljšali smo performanse najboljeg modela za 5%.

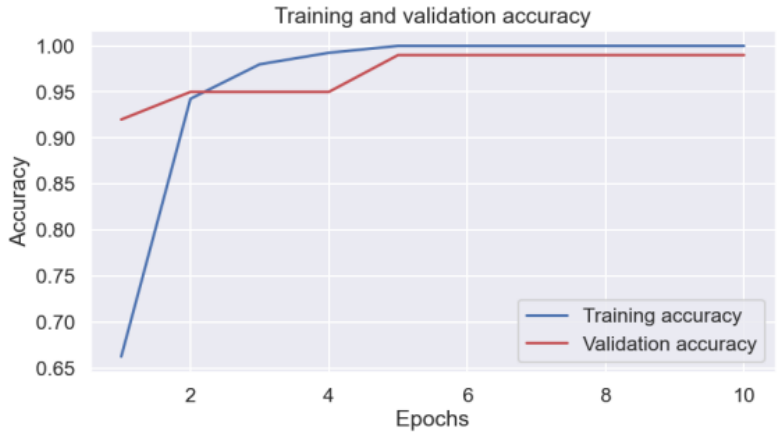


Feedforward Neural Network (FNN)

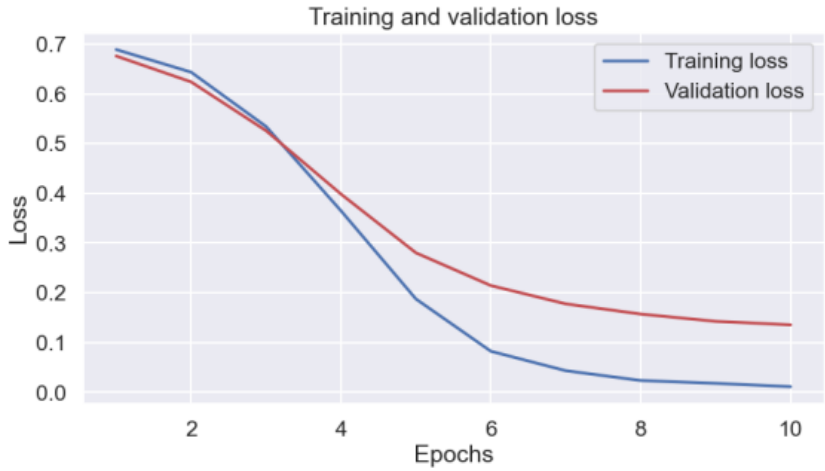
- jedan od najosnovnijih tipova veštačkih neuronskih mreža, koji se koristi za različite zadatke u mašinskom učenju, uključujući klasifikaciju i regresiju. Kao što ime sugeriše, u ovoj mreži informacije se kreću samo u jednom smeru – napred, od ulaznog sloja, preko skrivenih slojeva, do izlaznog sloja.
- Prednosti FNN-a:
  - **Jednostavnost**
  - **Univerzalni aproksimator**
- Mane FNN-a:
  - **Prilagođavanje hiperparametara**
  - **Tendencija ka overfittingu**

Feedforward Neural Network (FNN)

- Tačnost na oba skupa brzo raste i stabilizuje se na visokim vrednostima.
- Male razlike između trening i validacione tačnosti pokazuju da model nije previše prilagođen.



- Gubitak na trening i validacionom skupu brzo opada, ukazujući na efikasno učenje modela.
- Mala razlika između trening i validacionog gubitka na kraju treniranja sugeriše dobru generalizaciju.



Zaključak

- **Naive Bayes (~95%):**
  - **Prednosti:** Brz i efikasan za trening, idealan za velike skupove podataka.
  - **Nedostaci:** Niža tačnost zbog pretpostavke nezavisnosti karakteristika.
- **Support Vector Classifier (SVC)(~96%):**
  - **Prednosti:** Visoka tačnost sa optimalno podešenim hiperparametrima.
  - **Nedostaci:** Spor za vrlo velike skupove podataka, zahteva pažljivo podešavanje.
- **Feedforward Neural Network (FNN)(~100%):**
  - **Prednosti:** Najviša tačnost, uči složene obrasce.
  - **Nedostaci:** Dugo treniranje, zahteva fino podešavanje hiperparametara.

Hvala na pažnji

Ognjen Gligorić SV79-2021