**LUNG CANCER PREDICTION**

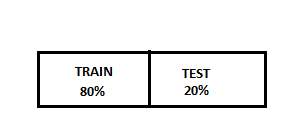
U daljem tekstu nalazi se šema plana izrade projektnog zadatka.

Obrada nepotrebnih vrednosti:

U našoj bazi imamo 1000 uzoraka i 26 obeležja .Obzirom da indeks i ID pacijenta obeležja ne igraju značajnu ulogu u našem slučaju,smatramo da možemo samo odbaciti kolone gde se nalaze ta obeležja.

Nakon što smo završili sa odbacivanjem obeležja koja nećemo koristiti nadalje, možemo se pozabaviti i sa direktnim odbacivanjem uzoraka, ali zbog jako dobrog prvobitnog stanja baze, nije bilo potrebno odbaciti nijedan uzorak jer nemamo nedostajuće vrednosti.

Nakon ovoga pristupamo podeli podataka na trening i test skup. Odvojićemo 80% podataka za trening, a 20% na test skup – nasumičnim odabirom.



Nakon podele pristupamo odabiru algoritama za rešavanje problema.

Za smanjenje dimenzionalnosti koristićemo PCA algoritam.

Zadržaćemo maksimalnu varijansu u podacima i očuvati što više informacija u novom-smanjenom prostoru obeležja.

Skup za trening ćemo deliti na 10 delova za unakrsnu validaciju (GridSearchCV)

Evaluacija najboljih modela na test skupu.

Biraćemo najbolje modele iz svakog algoritma i dobijene rezultate ćemo prikazati grafički.

LOGISTIČKA REGRESIJA

Prvi pristup koji koristimo je Logička regresija i ovo su koraci po kojima ćemo raditi. Nakon prikupljanja, pretprocesiranja i podele podataka na trening i test skup, pristupamo fitovanju i evaluaciji modela.

Parametri koje koristimo u ovom algoritmu su: fit\_intercep , class\_weigh, solver i multi\_class.

KNN ALGORITAM

Naredni algoritam koji koristimo je KNN algoritam:

1. Obučavamo skup prema najboljem izboru

2. Vršimo predikciju

3. Računamo preciznost

4. Računamo matricu konfuzije

Parametri:

gridSearch – idealan broj suseda

n\_neighbors – broj suseda koje razmatramo(parameter k)

weights – značaj svakog od k najbližih suseda

metric – funkcija rastojanja(minkowski)

NAIVNI BAJESOV KLASIFIKATOR

Poslednji algoritam je Naivan Bajesov algoritam, a njegovi koraci su sledeći:

1. Računamo verovatnoće

*P*(*C*∣*X*)=*P*(*X*)*P*(*X*∣*C*)⋅*P*(*C*)​

2. Nezavisni atributi

3. Vršimo predikciju

4. Odabir klase sa najvećom verovatnoćom