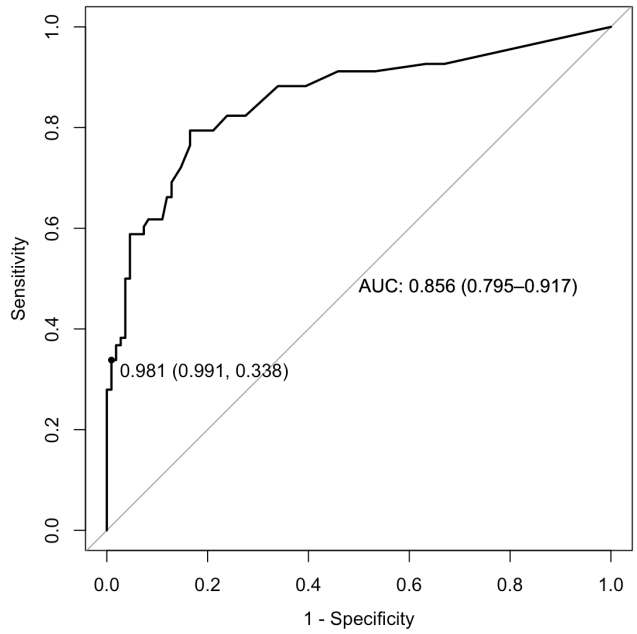
Naive Bayes klasifikator, Titanic dataset

```
prevalence = count(Survived=='Yes')/N = 0.3841
```

FN.cost	threshold	accuracy	sensitivity	specificity
0.005	0.9812137	0.7401130	0.3382353	0.9908257
0.500	0.4425405	0.8192090	0.7941176	0.8348624
1.000	0.4425405	0.8192090	0.7941176	0.8348624
1.500	0.4425405	0.8192090	0.7941176	0.8348624
2.000	0.4425405	0.8192090	0.7941176	0.8348624
2.500	0.4425405	0.8192090	0.7941176	0.8348624
3.000	0.4425405	0.8192090	0.7941176	0.8348624
3.500	0.4425405	0.8192090	0.7941176	0.8348624
4.000	0.4425405	0.8192090	0.7941176	0.8348624
4.500	0.2080834	0.7853107	0.8235294	0.7614679
5.000	0.2080834	0.7853107	0.8235294	0.7614679

FN cost = 0.005



Comment [JJ1]: manje 'vrednujemo' FN greške, na račun većeg 'vrednovanja' FP grešaka (kao u slucaju klasifikacije email poruka na spam/not-spam).
Razlog: hoćemo da klasifikator pravi što manje FP grešaka (čak i ako to znači da ćemo imati više FN grešaka)

Konkretno, postavljanjem FN cost na 0.005, kažemo klasifikatoru da FN greške maltene ignoriše, i time ga maltene potpuno usmeravamo da smanjuje FP greške.

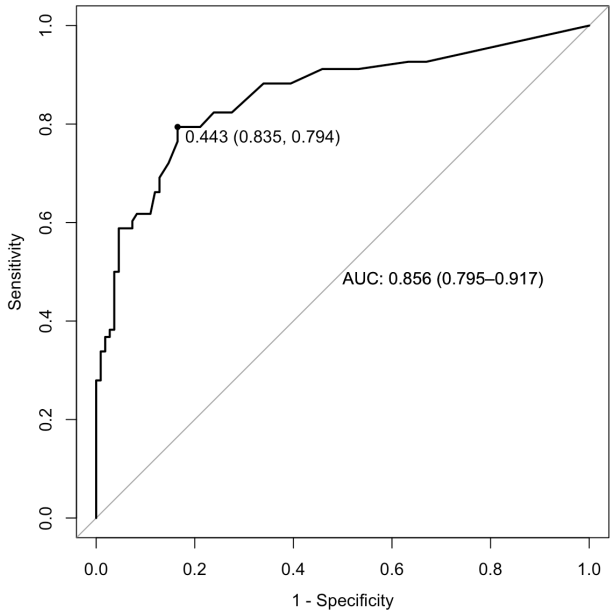
S obzirom da je Specificity = TN / (TN+FP), smanjenjem FP grešaka, povećavamo Specificity, što se ovde i vidi

Comment [JJ2]: FP i FN se jednako 'vrednuju', odnosno pridružujemo im isti trošak i time usmeravamo klasifikator da podjednako smanjuje i jedne i druge

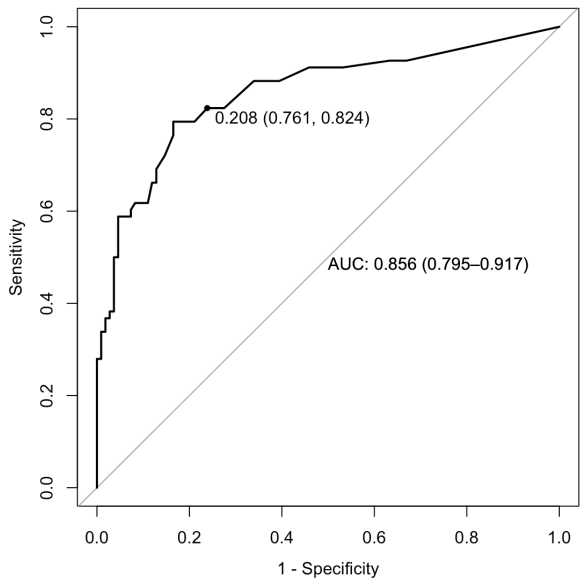
Comment [JJ3]: Konkretno, postavljanjem FN cost na 4.5, kažemo klasifikatoru da nam je jako bitno da eliminiše FN greške - jedna FN greška nosi isti trošak kao 4.5 FP greške; time usmeravamo klasifikator da pravi što manje FN grešaka

S obzirom da je Sensitivity = TP / (TP+FN), smanjenjem FN grešaka, povećavamo Sensitivity, što se ovde i vidi

FN cost > 0.1 && FN cost < 4



FN cost = 4.5



Decision Tree klasifikator, Titanic dataset

prevalence = 0.3841

FN.cost	threshold	accuracy	sensitivity	specificity
0.005	0.8288256	0.8248588	0.5588235	0.9908257
0.500	0.3166667	0.8248588	0.7205882	0.8899083
1.000	0.3166667	0.8248588	0.7205882	0.8899083
1.500	0.3166667	0.8248588	0.7205882	0.8899083
2.000	0.3166667	0.8248588	0.7205882	0.8899083
2.500	0.3166667	0.8248588	0.7205882	0.8899083
3.000	0.3166667	0.8248588	0.7205882	0.8899083
3.500	0.1806075	0.7966102	0.7352941	0.8348624
4.000	0.1806075	0.7966102	0.7352941	0.8348624
4.500	0.1806075	0.7966102	0.7352941	0.8348624
5.000	0.1806075	0.7966102	0.7352941	0.8348624

Formula za određivanje optimalne vrednosti za threashold:

$$\min(\ (1 - sensitivity)^2\ +\ r \times (1 - specificity)^2)$$

$$r = \frac{1 - prevalence}{FN.cost \times prevalence}$$