# Taiwanese Bankruptcy Prediction

## 1. Uvod i opis zadatka

Cilj ovog rada je analiza skupa podataka "Taiwanese Bankruptcy Prediction" i kreiranje prediktivnih modela koji mogu klasifikovati da li će kompanija bankrotirati ili ne na osnovu finansijskih pokazatelja. Problem je definisan kao binarna klasifikacija.

## 2. Opis podataka

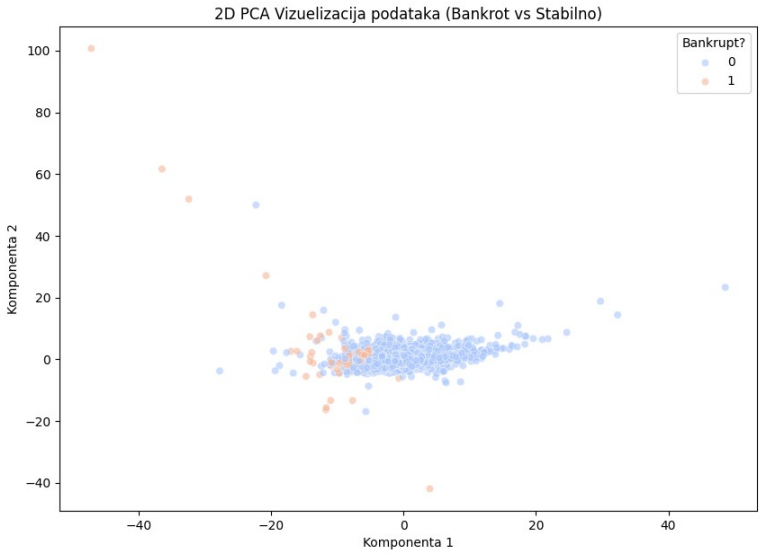
Podaci su preuzeti iz UCI Machine Learning repozitorijuma.

- **Izvor:** Ekonomske regulative Tajvanske berze.
- **Broj instanci:** 6819.
- **Broj atributa:** 95 finansijskih atributa + 1 ciljna promenljiva.
- **Ciljna promenljiva:** Bankrupt? (1 = Bankrot, 0 = Nije bankrot).
- **Struktura:** Podaci su izrazito neuravnoteženi (nebalansirani), gde je broj bankrotiranih firmi znatno manji od stabilnih, što predstavlja glavni izazov u modelovanju.

## 3. Metodologija i obrada podataka (Preprocesiranje)

Za obradu podataka korišćen je programski jezik Python sa bibliotekama pandas, sklearn i imblearn. Sprovedeni su sledeći koraci:

1. **Čišćenje podataka:** Provereno je postojanje nedostajućih vrednosti.
2. **Vizuelizacija:** Izvršena je redukcija dimenzionalnosti na 2D prostor koristeći PCA (Principal Component Analysis) radi vizuelnog uvida u separabilnost klasa.



3. **Skaliranje:** Svi numerički atributi su normalizovani koristeći StandardScaler (srednja vrednost 0, devijacija 1), što je neophodno za algoritme poput SVM i KNN.
4. **Balansiranje klasa:** S obzirom na mali broj primera bankrota, na trening skupu je primenjena **SMOTE** (Synthetic Minority Over-sampling Technique) metoda kako bi se veštački generisali primeri manjinske klase i izbalansirao odnos.
5. **Redukcija atributa:** Kreiran je poseban podskup podataka gde je metodom SelectKBest (ANOVA F-value) izdvojeno 10 najuticajnijih finansijskih pokazatelja.

**4. Korišćeni algoritmi**

U skladu sa zahtevima, primenjeno je 5 različitih algoritama mašinskog učenja:

1. **Logistic Regression:** Kao osnovni linearni model.
2. **Decision Tree:** Zbog interpretabilnosti pravila odlučivanja.
3. **Random Forest:** Kao ansambl metoda za veću stabilnost.
4. **K-Nearest Neighbors (KNN):** Algoritam baziran na instancama.
5. **Support Vector Machine (SVM):** Za pronalaženje optimalne hiperravni.

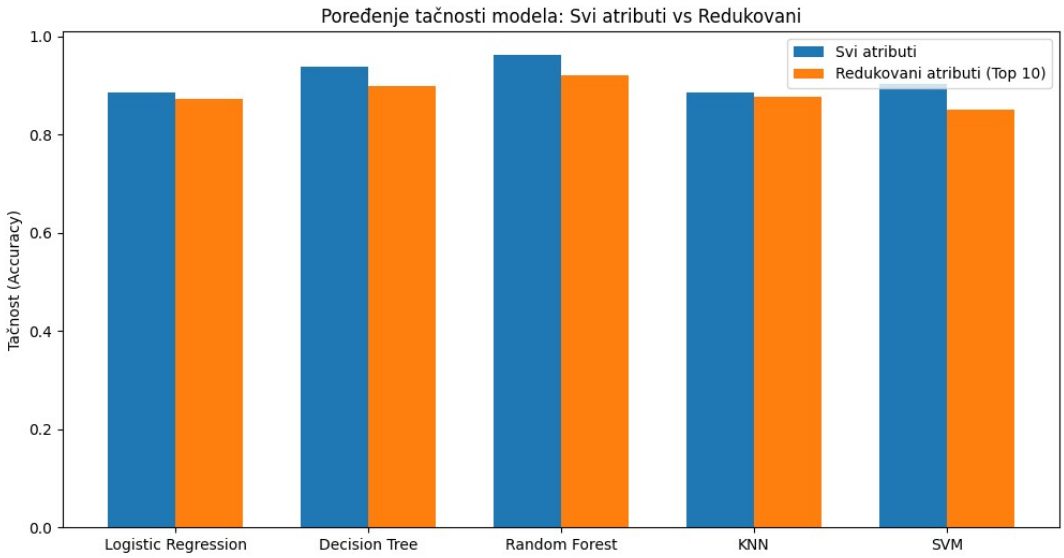
**5. Analiza rezultata**

Modeli su trenirani u dva scenarija:

1. Sa svim atributima (95 atributa).
2. Sa redukovanim skupom (10 atributa).

**Tabela rezultata (Tačnost / Accuracy):**

| **Algoritam** | **Svi atributi** | **Redukovani atributi** |
|---|---|---|
| Logistic Regression | 0.8866 | 0.8719 |
| Decision Tree | 0.9394 | 0.8998 |
| Random Forest | 0.9624 | 0.9218 |
| KNN | 0.8866 | 0.8763 |
| SVM | 0.9037 | 0.8500 |




**Diskusija:**

**Problem nebalansiranih podataka:** Iako **Random Forest** na punom skupu podataka ima najveću ukupnu tačnost (96.24%), njegov odziv (Recall) za detekciju bankrota je nizak (0.55). To znači da model propušta skoro polovinu firmi koje će zapravo bankrotirati, klasifikujući ih kao stabilne.

- **Efekat redukcije atributa:** Zanimljivo je da su modeli na redukovanom skupu (samo 10 atributa) pokazali **bolju sposobnost detekcije bankrota** (veći Recall), iako je ukupna tačnost blago opala.
- Na primer, **Logistic Regression** na redukovanom skupu ima Recall od **0.85**. To znači da ovaj model uspešno detektuje 85% svih bankrota, što je u praksi mnogo korisnije od visoke tačnosti koja dolazi od pogađanja većinske klase.
- **Trade-off (Preciznost vs. Odziv):** Primećujemo da modeli sa visokim Recall-om (LogReg, SVM na redukovanom skupu) imaju nisku preciznost (oko 0.15 - 0.18). To znači da modeli često dižu "lažnu uzbunu" (predviđaju bankrot tamo gde ga nema), ali je to prihvatljiva cena kako bi se detektovala većina stvarnih rizičnih slučajeva.

## 6. Zaključak

Cilj ovog rada bio je kreiranje modela za predikciju bankrota tajvanskih kompanija. Kroz proces obrade podataka, balansiranja klasa (SMOTE) i selekcije atributa, došli smo do sledećih zaključaka:

1. **Finansijski indikatori:** Bankrot se može uspešno predvideti praćenjem malog broja ključnih parametara, prvenstveno **ROA (povrat na imovinu)** i **Debt ratio (zaduženost)**. Korišćenje svih 96 atributa nije neophodno i može dovesti do preprilagođavanja (overfitting).
2. **Izbor modela:** Za ovaj specifičan problem, gde je "skuplje" propustiti bankrot nego lažno optužiti stabilnu firmu, **Logistic Regression sa redukovanim skupom atributa** se pokazala kao najkorisniji model. Iako nema najveću ukupnu tačnost, ima najveći **Recall (85%)**, što je ključno za upravljanje rizikom.
3. **Upravljanje podacima:** Balansiranje podataka je bio kritičan korak. Bez primene tehnika poput SMOTE-a, modeli bi težili da sve klasifikuju kao "stabilno" zbog malog broja primera bankrota u originalnom skupu.

Možemo zaključiti da je mašinsko učenje moćan alat za finansijsku analizu, ali da se uspeh modela ne sme meriti samo prostom tačnošću (Accuracy), već se mora uzeti u obzir specifičnost problema (Recall vs. Precision).