Logistická a ML-based regresní modely

Sergii Babichev

Univerzita Jana Evangelisty Purkyně v Ústí nad Labem

sergii.babichev@ujep.cz

Zadání úlohy 1

- Úkolem je vytvořit modely binární klasifikace na základě datasetů **Diabetes** a **Framingham**.
- Použité metody:
  - Logistická regrese
  - Rozhodovací strom (Decision Tree)
  - Random Forest
- Vyhodnocení modelů pomocí ROC analýzy a Confusion Matrix.

Postup řešení

1. Načtěte dataset a proveďte jeho prozkoumání (např. pomocí `df.info()` a `df.describe()`).
2. Předzpracujte data:
   - Normalizujte numerické atributy pomocí **StandardScaler**.
   - Zpracujte chybějící hodnoty (např. odstranění nebo imputace).
   - Rozdělte data na trénovací a testovací sadu (70:30).
3. Vytvořte a natrénujte tři modely:
   - Logistickou regresi.
   - Rozhodovací strom.
   - Random Forest.
4. Proveďte predikce na testovacích datech.
5. Vyhodnoťte modely pomocí následujících metrik:
   - **Confusion Matrix** – porovnání skutečných a predikovaných hodnot.
   - **ROC křivka a AUC** – srovnání výkonu modelů.
   - **Přesnost, citlivost, F1-skóre** – výpočet klíčových metrik.
6. Diskutujte výsledky a srovnejte výkonnost modelů.

Zadání úlohy 2

- Úkolem je vytvořit regresní modely na základě datasetu **Wine Quality (Red)**.
- Použité metody:
  - Rozhodovací strom (Decision Tree)
  - Random Forest
- Vyhodnocení modelů pomocí metrik MSE, RMSE a R^2.

Postup řešení

1. Načtěte dataset a proveďte jeho prozkoumání (např. pomocí `df.info()` a `df.describe()`).
2. Předzpracujte data:
   - Normalizujte numerické atributy pomocí **StandardScaler**.
   - Zpracujte chybějící hodnoty (např. odstranění nebo imputace).
   - Rozdělte data na trénovací a testovací sadu (70:30).
3. Vytvořte a natrénujte dva modely:
   - Rozhodovací strom.
   - Random Forest.
4. Proveďte predikce na testovacích datech.
5. Vyhodnoťte modely pomocí následujících metrik:
   - **MSE (Mean Squared Error)** – průměrná kvadratická chyba.
   - **RMSE (Root Mean Squared Error)** – odmocněná průměrná kvadratická chyba.
   - **R^2 skóre** – míra vysvětlené variance modelem.
6. Diskutujte výsledky a srovnejte výkonnost modelů.