Semestrální práce 4ST101

Provedl jsem analýzu datasetu Stroke Prediction Dataset

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

Jako přílohu ho uvádím jako soubor stroke\_data.csv

Data obsahují 12 unikátních atributů:

1. id: unikátní identifikátor
2. gender: “Male”, “Female”, “Other”
3. age: dokončený věk pacienta
4. hypertension: “0”, “1” (má nebo nemá vysoký krevní tlak)
5. heart\_disease: “0”, “1” (má nebo nemá nemoci srdce)
6. ever\_married: “0”, “1” (v manželství nebo nikoliv)
7. work\_type: druh pracovního úvazku
8. Residence\_type: městský nebo vesnický typ bydlení
9. avg\_glucose\_level: průměrná úroveň glukózy v krvi
10. bmi: index tělesné hmotnosti
11. smoking\_status: stávající/bývalý kuřák, nekuřák
12. stroke: měl pacient mozkovou mrtvici nebo nikoliv

*Cílem* této práce je zjistit nejvýznamnější proměnnou pro předpověď mozkové mrtvice na základě stažených dat. Pro tento účel budeme používat různé klasifikační algoritmy, které postupně budeme měřit maticí záměn a výpočtem průměrné přesnosti predikce.

Postup je následující:

* stažení potřebných balíčků ke práci s daty
* stažení dat z csv-souboru
* zakódování kategorických proměnných pro urychlení práce (int to boolean)
* transformace bmi ze stringu v double (číselnou hodnotu)
* vymazaní odlehlých pozorování pro zjednodušení EDA a normalizaci vstupů pro predikci
* rozdělení základních dat na trénovací a testovací část pro simulaci běhu modelu na nových datech
* standardizace dat, abych vyloučil negativní faktory z výpočtu Euklidovské vzdálenosti mezi různorodými daty.
* záměna NaN hodnot na nejbližší pomocí algoritmu KNN
* pracujeme s nerovnoměrnými výsledky pro dostání mrtvice, musíme je vyrovnat pro korektní běh algoritmů
* zkusíme použít binární logistickou regresi pro vyhodnocení, do vzorce jako závislé proměnné označíme age, bmi a avg\_glucose\_level. Nastavíme pravděpodobnost dostání mrtvice jako >0.5. Náš model dobře klasifikuje případy “0” (hodnotí každý případ jako “0”) a nedokáže predikovat “1”, přesto míra predikce je velmi vysoká a nereprezentativní, kvůli nerovnoměrným datům. ROC křivka (operační charakteristika přijímače) odpovídá f(x) = x, AUC teda je 0,5. To znamená, že predikční síla modelu se blíží nule.
* Použijeme 3 techniky pro rebilancování dat: undersampling, oversampling a SMOTE (míchání dvou minulých metod s váhami pro data). S logického hlediska by moc nebylo vhodné použít oversampling v praxi, jelikož nechceme uměle zvýšit počet pozitivních případů mrtvice ve zkoumané populaci. Nejlepší výsledek za tří dává SMOTE (Accuracy = 0.784), a proto ho použijeme v následných modelech.
* Zkusíme použít SVM model, který by klasifikoval data podle optimální polohy křivky, která by je rozdělila do dvou stran podle hodnoty “0” nebo “1”. Model se moc nevylepšil, ale by se dalo ještě použít křížovou validaci, která je už nad rámec této semestrálky.
* Balanced Bagging Classifier je už modelem založeným na vyhodnocení rozhodovacích stromů a kombinovaní predikcí z podmnožin datasetů. Výkonnost algoritmu se ale moc zlepšila.
* Jako poslední použijeme Random Forest Classifier s vážením tříděním. Používá podobný přístup s rozhodovacími stromy, ještě vybírá náhodná data ze vzorků. Vypočteme koeficienty pro vážení dat, což jsou v podstatě poměrem výskytů “0” nebo “1”. Počet stromů byl vybrán podle obecných doporučení (n = 1000). Po vyhodnocení dostaneme dobrý výsledek cca 91,5 % korektních predikcí. Přičemž ukazatel class.error je rovnoměrně rozdělen pro pozitivní a negativní případy mrtvice u pacientů, co znamená , že model je méně závislý na počtu dat a bilanci a poměrně dobře predikuje každý z případů.
* Toto jsem zajistil pomocí nastavení parametru “mrty”, jako počtu vybraných proměnných po každé iteraci algoritmu. Podle výsledku můžeme konstatovat, že největší předpovídací schopnost má věk pacienta, až potom bmi a úroveň glukózy.

**Závěr:** Z datasetu s nerovnoměrným rozdělením případů mrtvice jsem dokázal vytěžit zbilancována data pro účely trénování modelu pomocí 3 metod, metodologicky nejvhodnější z ní byla SMOTE. Pak pomocí Random Forest Classifier jsem dokázal jednoznačně klasifikovat prediktory mozkové mrtvice, i když predikovat nemoc v první řadě podle věku a až potom podle bmi a úrovně glukózy není vítěznou strategií. Z důvodů nedostatku jiných proměnných, jsem nemohl dospět ke smysluplnějšímu výsledku.