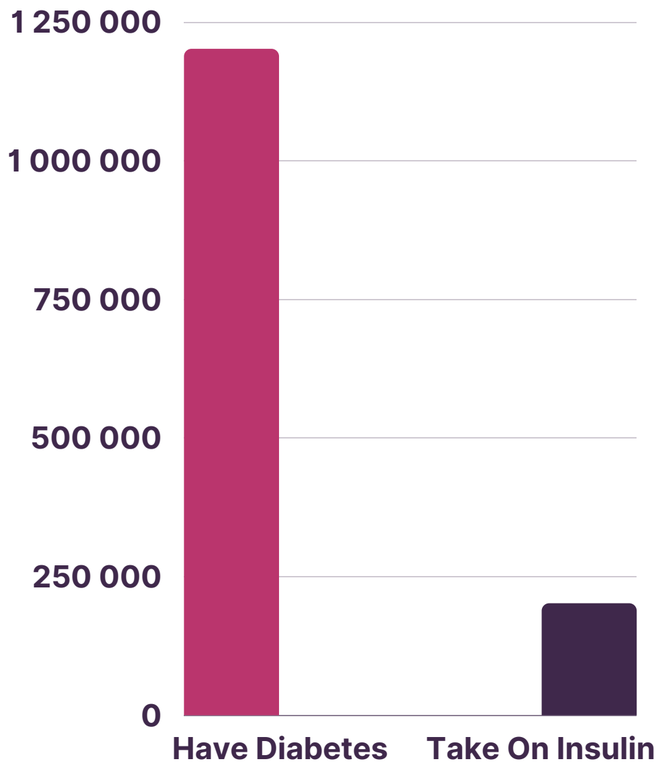
AI COURSE PROJECT

DIABETES PREDICTION

AUTHORS:
DENYS HERASYMUK & YAROSLAV MOROZEVYCH



Проблема

- Великий ризик захворіти.
- Тяжко ідентифікувати допоки не захворієш.
- Небезпечний у поєднанні з іншими захворюваннями.

- Існує спосіб ідентифікації діабету шляхом аналізу багатьох показників.
- Цього можна досягти завдяки AI.
- Спосіб застосування моделей є безмежним.

Що було зроблено

1. Дослідження проблемної області.
2. EDA аналіз & Feature Engineering. Визначення найважливіших фіч за допомогою RandomForest.
3. Балансування датасетів. Використання SMOTE та ROS методів оверсемплінгу.
4. Використання ML моделей: SVM, RandomForest, DecisionTree, XGBoost.
5. Використання DL моделей: власна NN модель, MLPClassifier.

DiabetesRisk

Фіча, створена в процесі Feature Engineering'a.

Кореляції з іншими фічами:

- Pregnancies (0.39)
- Glucose (0.69)
- BloodPressure (0.31)
- SkinThickness (0.39)
- BMI (0.4)
- Age (0.5)
- Outcome (0.53)

Що було зроблено

1. Дослідження проблемної області.
2. EDA аналіз & Feature Engineering. Визначення найважливіших фіч за допомогою RandomForest.
3. Балансування датасетів. Використання SMOTE та ROS методів оверсемплінгу.
4. Використання ML моделей: SVM, RandomForest, DecisionTree, XGBoost.
5. Використання DL моделей: власна NN модель, MLPClassifier.

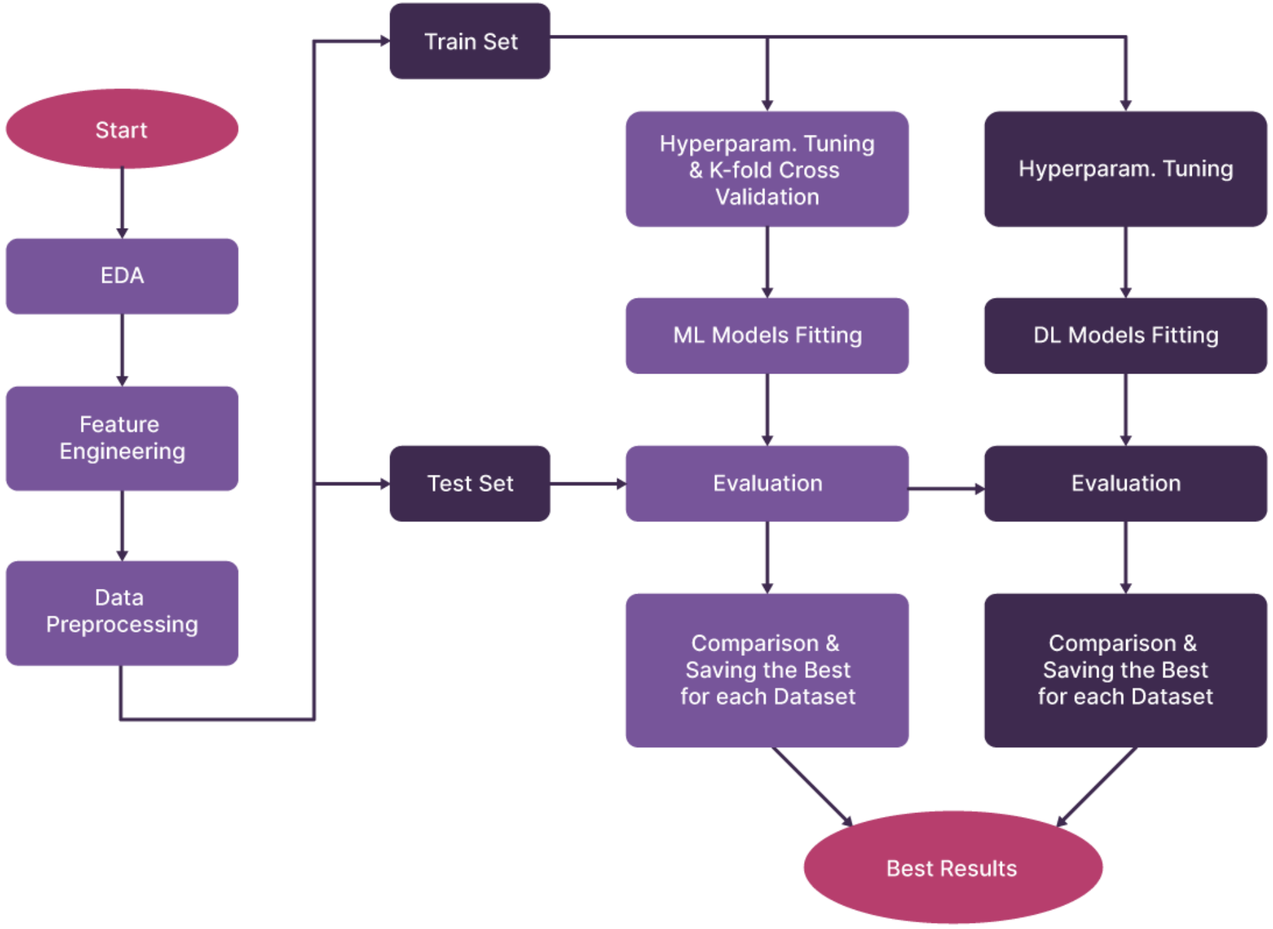
Pima Indians Diabetes Dataset

- Сконцентрований на медичних аналізах.
- 10 фіч - 768 рядків.
- Особливості: числові фічі, підходить під 4 типи діабету.
- Найкращі фічі / Кореляція: Age (0.31), BMI (0.3), Glucose (0.46), SkinThickness (0.26).

Ranchi Diabetes 2019 Dataset

- Сконцентрований на веденні здорового чи нездорового способу життя.
- 17 фіч - 951 рядок.
- Особливості: здебільшого, категоріальні фічі, підходить під 3 типи діабету.
- Найкращі фічі / Кореляція: Age (0.57), highBP (0.37), RegularMedicine (0.6), Stress (0.21), Pregnancies (0.22).

Підхід до вирішення задачі



Досягнуті метрики

	Dataset_Name	Model_Name	F1_Score	Accuracy_Score
0	Pima Indians, Original	SVC	0.824121	0.842593
6	Pima Indians, Original	MLPClassifier	0.837594	0.851852
1	Pima Indians, ROS	XGBClassifier	0.851852	0.861111
7	Pima Indians, ROS	MLPClassifier	0.781618	0.796296
2	Pima Indians, SMOTE	XGBClassifier	0.839465	0.851852
8	Pima Indians, SMOTE	NN with 2 hidden layers	0.814114	0.824074
3	Ranchi, Original	DecisionTreeClassifier	0.933299	0.944751
9	Ranchi, Original	NN with 2 hidden layers	0.934019	0.944751
4	Ranchi, ROS	RandomForestClassifier	0.948286	0.955801
10	Ranchi, ROS	MLPClassifier	0.909500	0.922652
5	Ranchi, SMOTE	XGBClassifier	0.923175	0.933702
11	Ranchi, SMOTE	NN with 2 hidden layers	0.908586	0.922652

Висновки

1. **Дослідження проблемної області** є досить корисним, особливо в задачах пов'язаних з медициною

2. **Балансування датасету** добре покращує результати моделей, проте слід проводити ретельні тестування

3. **Feature engineering** може підвищити результати краще, навіть ніж більш глибокий hyper-parameter fine-tuning

4. Найсильніші **ознаки, що добре корелюють з діабетом:** "Pregnancies", "Glucose", "SkinThickness", "Insulin", "BMI", "DiabetesPedigreeFunction", "Age", "RegularMedicine", "Stress", "BPLevel"