PROBLEM, WHEN THE CLASS 0 IS MAJORITY

PROBLEM:
In the Stroke Prediction dataset from Kaggle, the share of class 1 is ~5%.

KNN always looks for the closest points.

Because class 0 is the majority, the neighborhood is often all zeros → the model predicts 0.

OVERSAMPLING SLOT

```
from imblearn.over_sampling import SMOTE
from collections import Counter

sm = SMOTE(random_state=42)
X_res, y_res = sm.fit_resample(X, y)

print(sorted(Counter(y_res).items()))
```

DATA NORMALISATION

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

ADJUST N_NEIGHBORS PARAMETR

knn = KNeighborsClassifier(**n_neighbors**=3)

MACHINE LEARNING

Final Project Instructions

- **Create INTERNATIONAL teams** (students from different countries where possible).
- **Team Size:** 2 to 4 students.
- **Choose a Team Name.**
- **Prepare a short introduction slide** about your team (team members, countries, something interesting about you).

Project Task

Each team must prepare **a short presentation** about the following **Machine Learning algorithms**:

- **k-Means Clustering**
- **Naive Bayes**
- **Principal Component Analysis (PCA)**
- **Decision Tree**
- **k-Nearest Neighbors (k-NN)**

Your presentation should briefly cover for each algorithm:

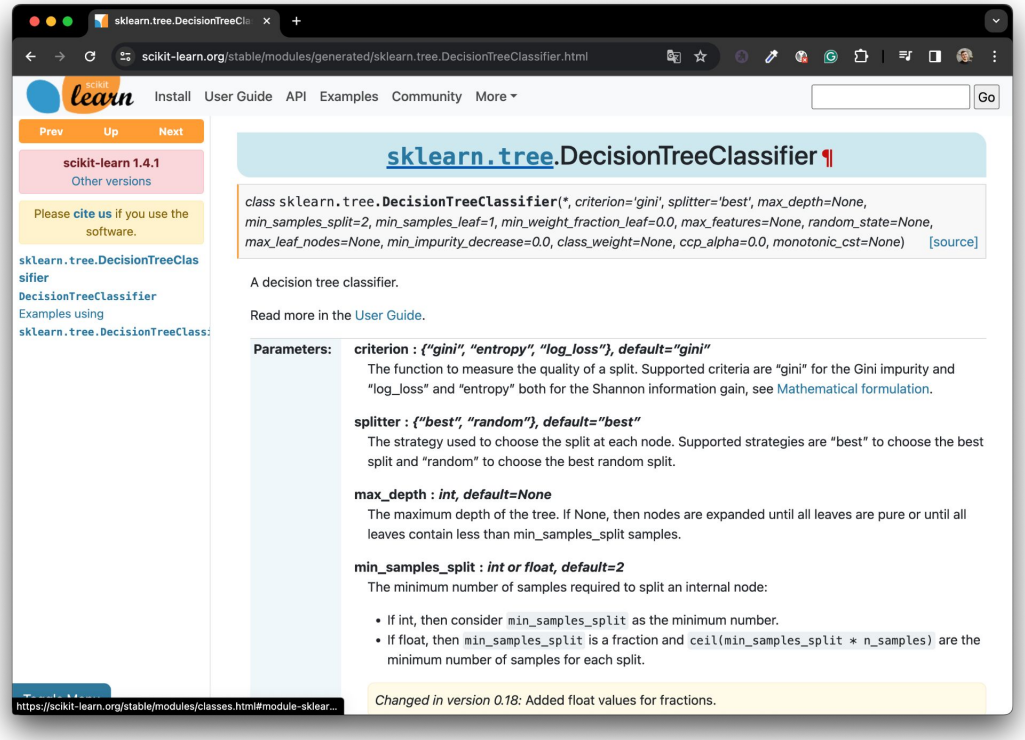
- ❏ Definition and main idea
- ❏ How it works (basic steps)
- ❏ Typical use cases
- ❏ Advantages and disadvantages

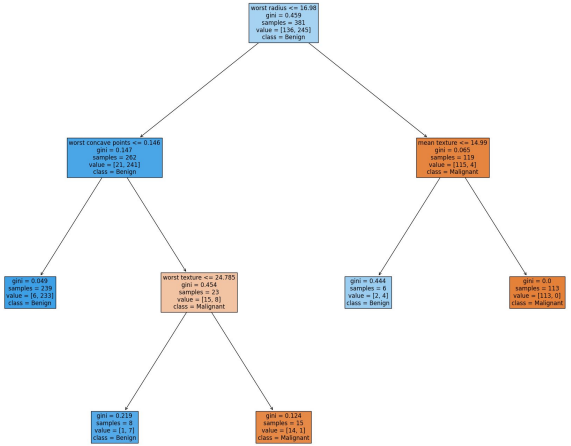
UCI Machine Learning Repository: The UCI repository hosts a wide range of datasets suitable for machine learning tasks, including datasets related to medicine. You can explore their collection and find datasets that fit your requirements. Here's the link: UCI Machine Learning Repository

Kaggle Datasets: Kaggle is a platform for data science competitions, and it also hosts a large collection of datasets contributed by the community. You can search for medical datasets on Kaggle and find ones suitable for building decision tree models. Here's the link: Kaggle Datasets

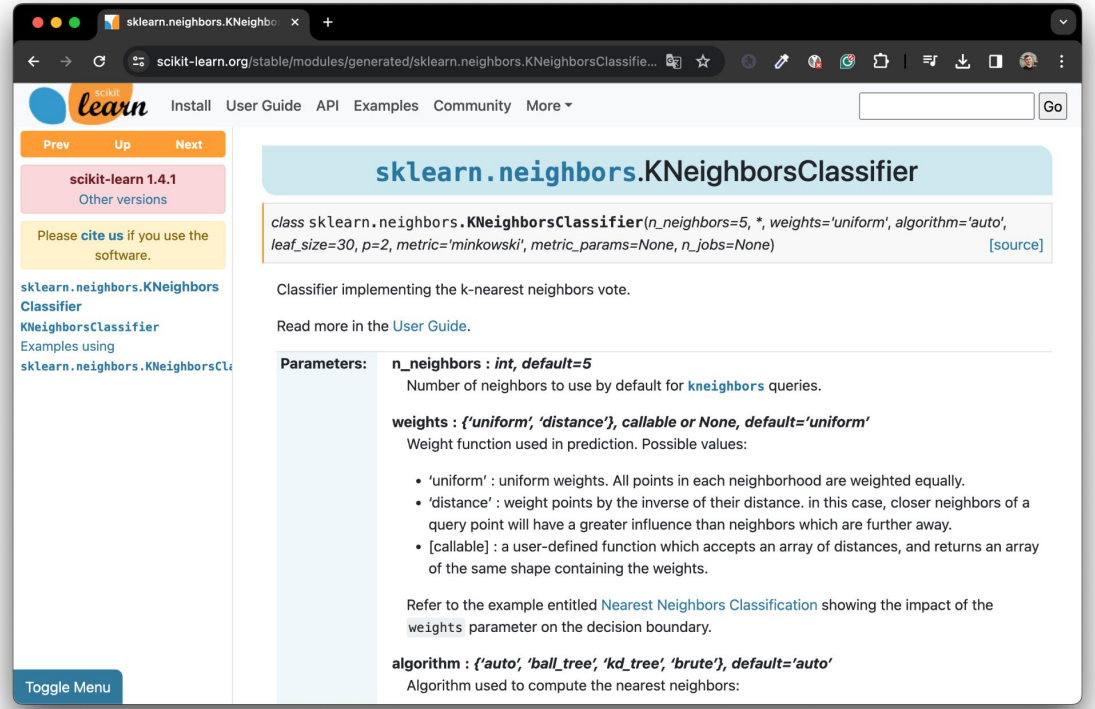
OpenML: OpenML is an online platform where you can find datasets, machine learning tasks, and experiments. It hosts a variety of datasets from different domains, including medicine. You can explore their collection and find datasets suitable for your project. Here's the link: OpenML

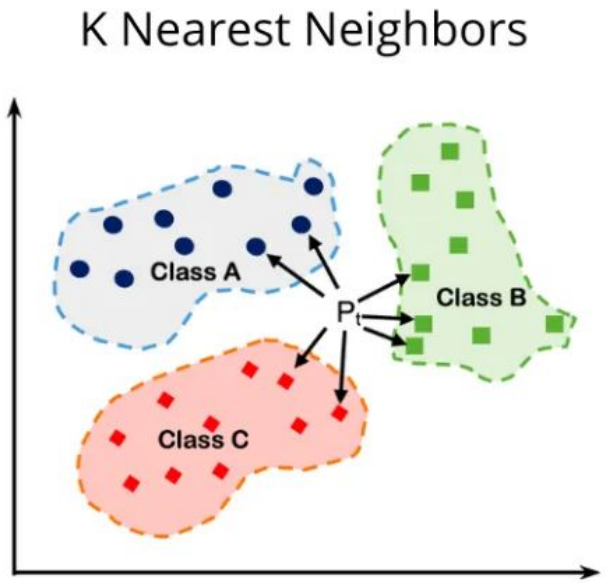
Decision Tree





K-Nearest Neighbors (KNN)





Submission Requirements

You must submit **three components**:

1. **Presentation** — will be delivered live on **January 8, 2026**, during the computer lab.

2. **Google Colab Notebook (.ipynb file)** — submit a file with your slides or additional explanations if needed.

   Project should use medical diagnosis data or security anomalies. Usage of minimum 2 machine learning classifier is required. Please add text description above code, explain how you process data. Test different data divisions.

3. **PDF Report** — submit a short written report (maximum 2-3 pages) summarizing the key points of your presentation.

Important Rules

All team members **must participate** during the final presentation.

Deadline for submission: Before the computer lab on January 8th, 2026.

Submission format: Send your files (IPYNB and KEYNOTE PDF, REPORT PDF) via the platform/email (teacher will provide further details).

Late submissions may lower the final grade!