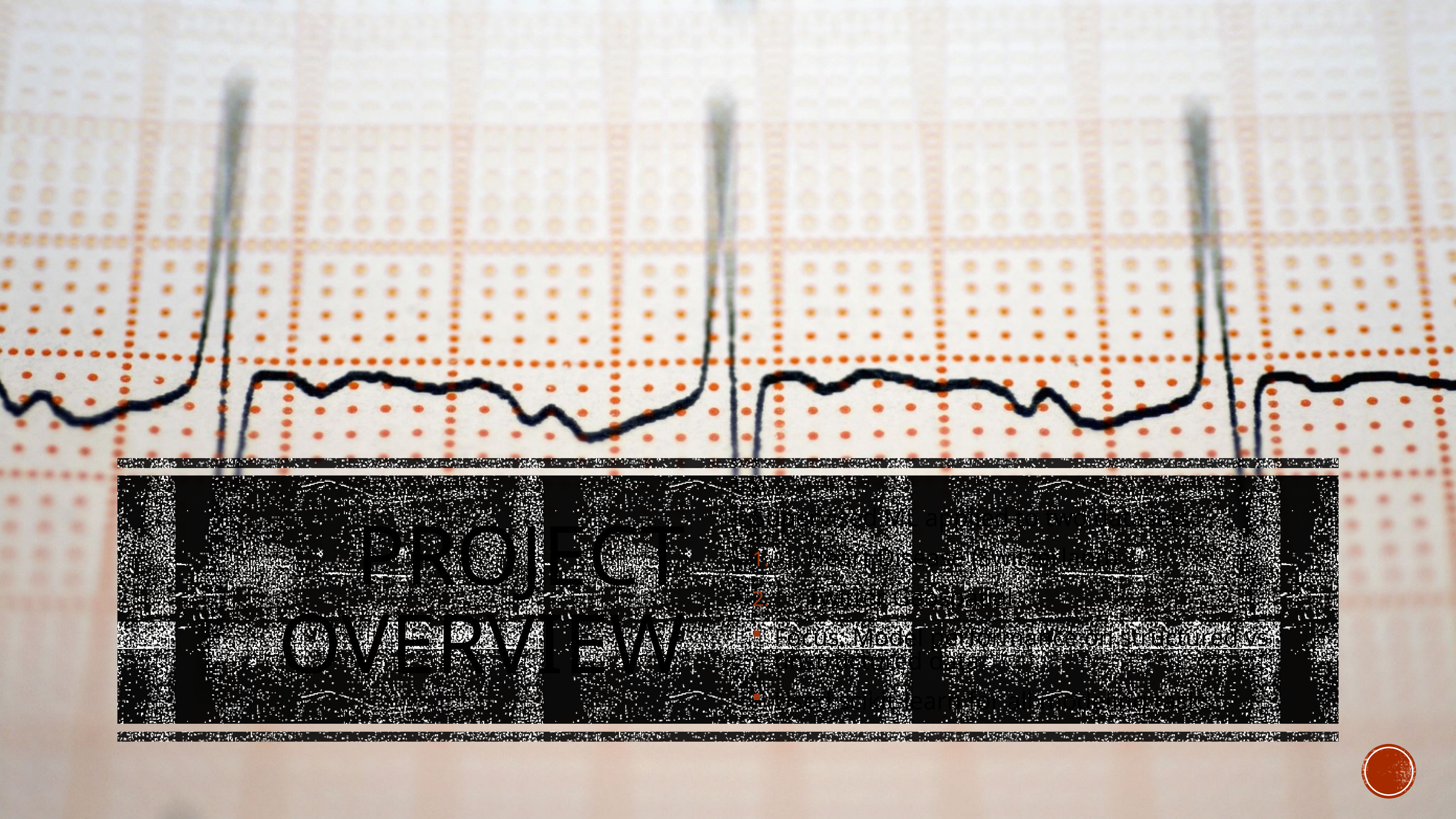
APPLIED MACHINE LEARNING WITH SCIKIT-LEARN

CSCI 164 - PROJECT

Team Members: Joshua Martinez, Anushka Patwa, Surya Gona



Project Overview

Supervised ML applied to two datasets:

1. Heart Disease (Switzerland)
2. Twitter Geospatial

- Focus: Model performance on structured vs unstructured data
- Used scikit-learn for all modeling tasks

DATASET 1 – HEART DISEASE

- Structured medical dataset (Switzerland subset)
- Binary target: Heart Disease (yes/no)
- Preprocessing: missing values, encoding, normalization

DATASET 2 – TWITTER GEOSPATIAL

- Real-world, noisy, geospatial tweet metadata
- Required text + location preprocessing
- Challenge: unstructured and imbalanced data

MODEL SELECTION

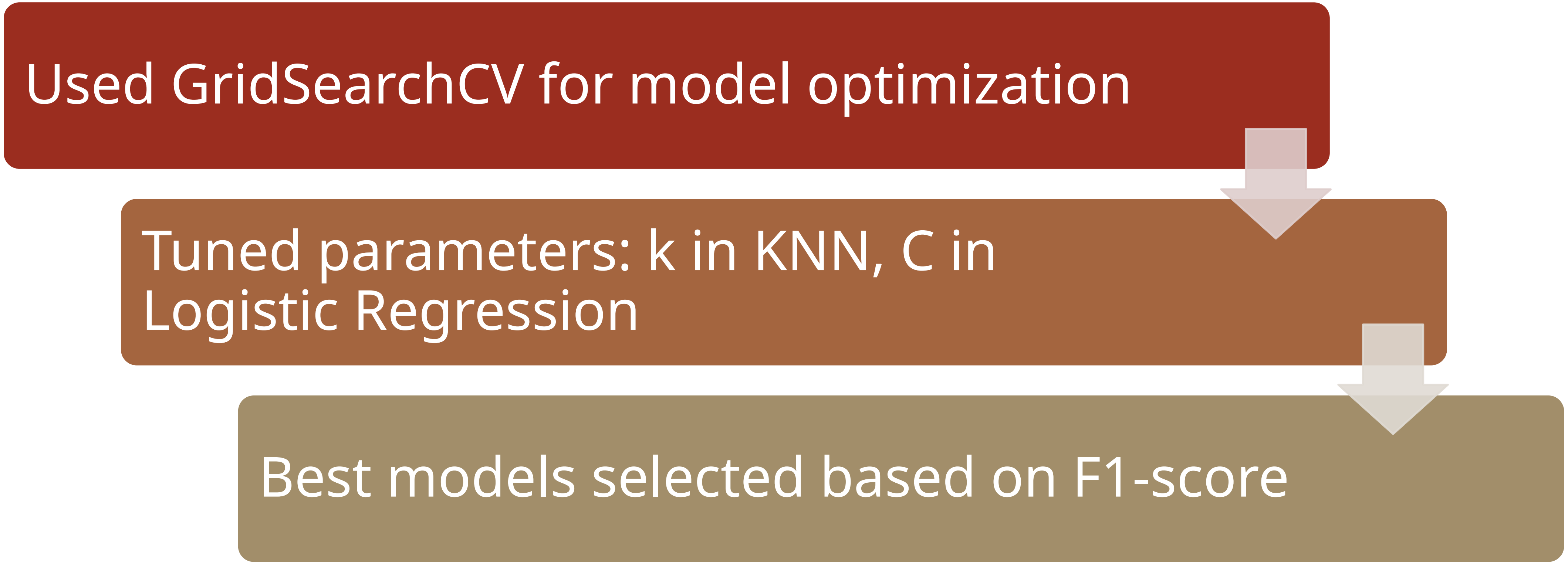
Logistic Regression: good for linearly separable classes

k-NN: works well with small datasets.

EVALUATION METRICS

- Accuracy, Precision, and F1 score are used for interpretation
- Prediction scatter plots are plotted to compare models
- Results are related as the F1 score balances precision and recall to show overall performance.

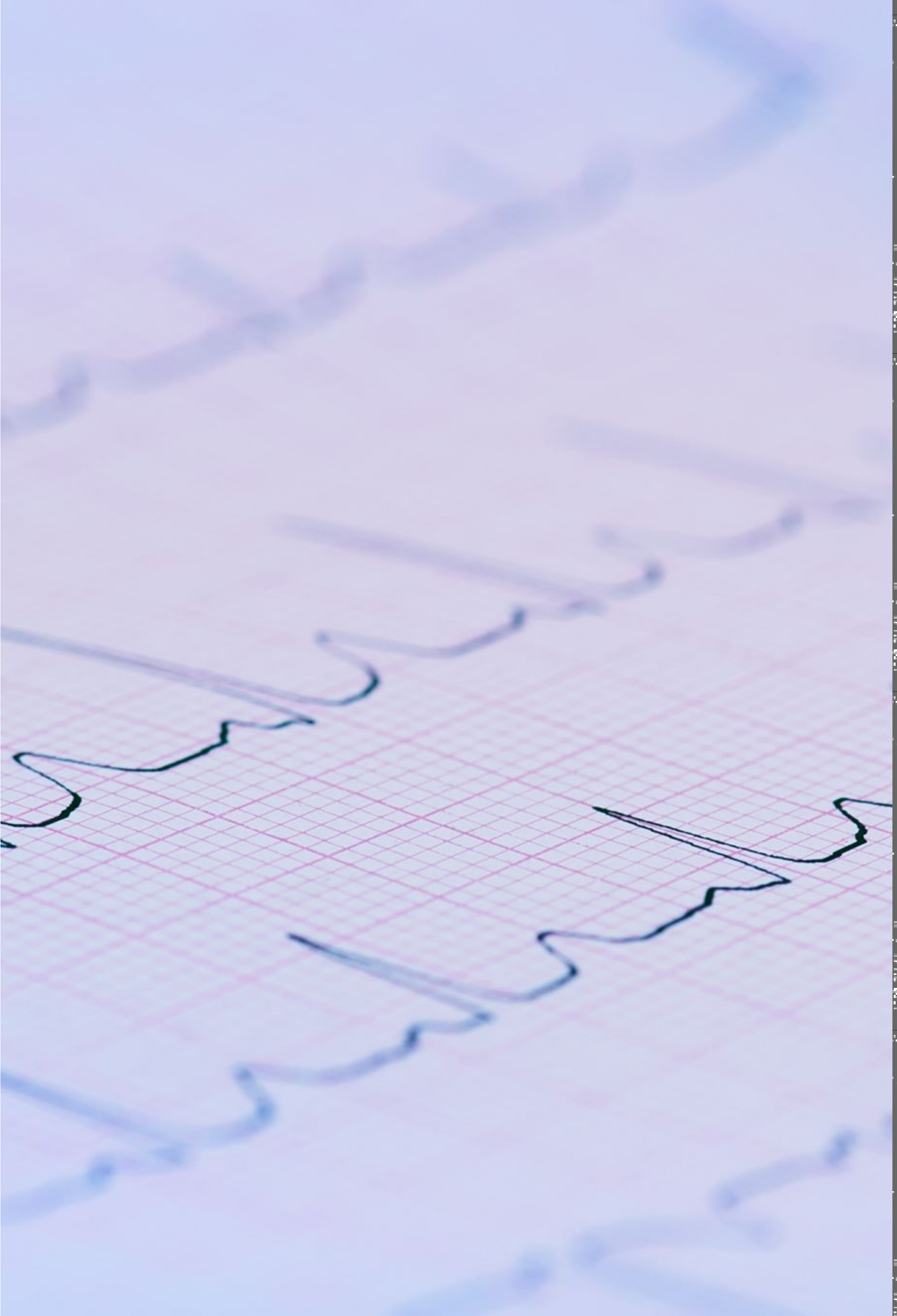
HYPERPARAMETER TUNING



Used GridSearchCV for model optimization

Tuned parameters: k in KNN, C in Logistic Regression

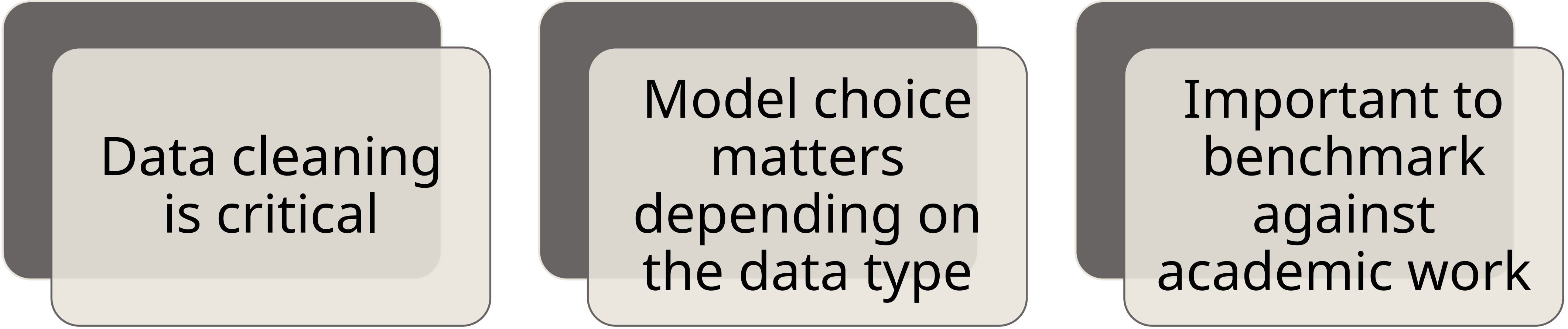
Best models selected based on F1-score



RESULTS

- Heart Disease: results aligned with published ML work
- Twitter: results are lower due to noisy data
- LR performed best on Heart and k-NN better on Twitter

KEY TAKEAWAYS



Data cleaning is critical

Model choice matters depending on the data type

Important to benchmark against academic work

REFERENCES

- UCI Machine Learning Repository
- scikit-learn documentation
- Prior published papers on ML in health and geospatial data