Predictive Modeling: Heart Disease & Twitter Data

1. Introduction

This project applies scikit-learn to two datasets to develop predictive models:

- Heart Disease Dataset (Switzerland subset): Predict a person's age using two clinical features: blood pressure (`trestbps`) and cholesterol (`chol`). This is a regression task.
- Twitter Data: Predict tweet timestamps based only on geographic coordinates (latitude and longitude). This was an exploratory task to test if location reveals timing patterns — which we later realized was a flawed assumption.

We built Linear Regression and K-Nearest Neighbors (KNN) models for each dataset and compared our findings with prior research to better understand our results.

---

2. Dataset Exploration

2.1 Heart Disease Data

Motivation
The Heart Disease dataset is a well-known benchmark in medical machine learning research. We chose it because it allowed us to compare our model's performance to published studies.
Features Selected

- `trestbps` (resting blood pressure)
- `chol` (serum cholesterol)

Target Variable

- `age`

Preprocessing

- Removed rows with missing values.
- Data split: training (64%), validation (16%), test (20%). We started with an 80/20 train-test split, then further split the train set for validation.

Prior Work Comparison
Detrano et al. (1989) used logistic regression with 13 clinical features and achieved ~77% accuracy in predicting heart disease presence.
In contrast, we predicted age from just two features — making our task more difficult.

2.2 Twitter Data

Motivation
We wanted to explore whether geographic coordinates could be used to predict when a tweet was posted.
Features Selected

- `latitude`
- `longitude`

Target Variable

- `timestamp`

Key Challenges

- There is no strong direct link between a tweet's location and the time it was posted.
- This turned out to be a flawed problem formulation.

Prior Work Comparison
Jurgens (2015) showed that location is useful for predicting where a tweet comes from, not when it was posted.
This helped us understand that the problem itself was not well-suited for prediction.

---

3. Model Development with scikit-learn

3.1 Heart Disease Models

Linear Regression

- Best R^2 Score: 0.21
- High variability depending on the fold (some as low as -0.52)
- Result: Two features were insufficient for reliable prediction.

K-Nearest Neighbors (K=7)

- Best R^2 Score: 0.0286 (~2.86%)
- Barely better than random guessing.
- Reinforced the importance of feature selection.

3.2 Twitter Models

Linear Regression

- R^2 Score: ~0

- No meaningful pattern between location and tweet time.

K-Nearest Neighbors (K=5)

- R^2 Score: ~0
- KNN also failed to find useful patterns.

---

4. Hyperparameter Tuning and Model Comparison

We used `GridSearchCV` to tune the number of neighbors (`k`) for KNN models.

Dataset	Best k	Best R^2 Score
Heart Disease	7	0.0286
Twitter	5	0.00019

Even after tuning, performance remained poor — especially for the Twitter dataset — confirming that the feature selection and problem formulation were limiting factors.

---

5. Results and Literature Comparison

5.1 Heart Disease Data

- *Detrano et al. (1989)*: 77% accuracy using 13 features and classification (disease presence).
- Our work: R^2 up to 0.21 using only 2 features and regression (predicting age).

Key Lessons

- More features = better predictions.
- Classification tasks are better suited for this dataset than regression tasks.

5.2 Twitter Data

- Location does not meaningfully predict timestamp.
- *Jurgens (2015)* showed location data is useful for place prediction, not time prediction.

Key Lessons

- A bad problem setup cannot be fixed with better models.
- Data understanding is as important as model tuning.

---

6. Key Learnings

1. Feature Selection Matters
   Two features weren't enough for the heart dataset. Including more clinical variables could improve performance.
2. Problem Formulation is Crucial
   Predicting tweet time from location was not a meaningful task. The relationship between data and target must be logical.
3. Validation Helps Reveal Flaws
   Cross-validation showed early that our models weren't learning useful patterns. This guided how we interpreted results.

---

7. Conclusion

Even though our models did not perform well, this project was a valuable learning experience:

- Strong machine learning results depend on good features and well-posed problems.
- Scikit-learn provides powerful tools for model training, cross-validation, and hyperparameter tuning.
- Comparing our work to published research gave us insights into what worked and why.

📚 Works Cited

1. Detrano, Robert, et al. "International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease." *The American Journal of Cardiology*, 1989. UCI Repository
2. Jurgens, David. "That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships." *Journal of Computational Science*, 2015.