PAD project

Alexey Rasskazov

Topic

AI can assist in various situations. We can implement simple statistical analysis and achieve great results.

Heart Disease

- Cleveland
- Hungarian
- Switzerland
- Long Beach VA
- Statlog (Heart) Data Set.

This heart disease dataset is curated by combining five popular heart disease datasets that were previously available independently but not combined before.

Collecting and finding data

Heart Disease Dataset

The dataset is sourced from Kaggle.

Dowloading dataset

```
import kaggle

kaggle.api.authenticate()
kaggle.api.dataset_download_files(
    'mexwell/heart-disease-dataset',
    path='data',
    unzip=True
)
```

`Dataset URL: https://www.kaggle.com/datasets/mexwell/heart-disease-dataset`

The dataset is automatically downloaded to the local data folder. Ensure you set your Kaggle username and authentication key.

Cleaning data and handling missing values

Let's explore the dataset for the most obvious errors.

Visualize incorrect rows

```
coloms_to_remove = data.loc[
    (data['resting bp s']<10) | (data['ST slope']==0)
]
coloms_to_remove[['resting bp s', 'ST slope']]
```

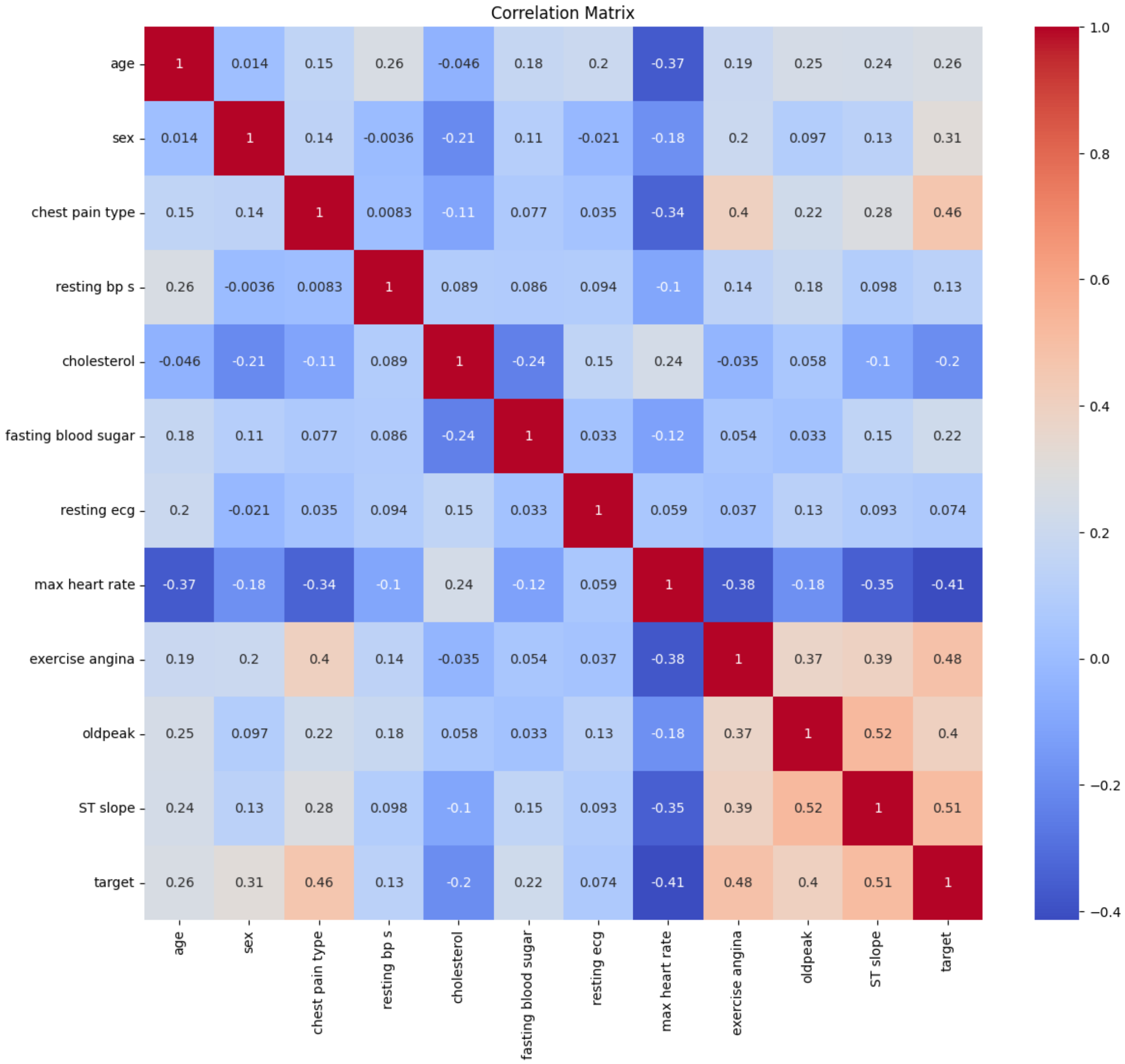
	resting bp s	ST slope
450	0	2
517	150	0

Two rows are clearly impossible for human beings. The first has a heart rate of zero beats per second, and the second has a slope of the peak exercise ST segment of zero, which is not possible according to the documentation.

Data analysis – attribute dependencies

This section focuses on attribute dependencies.

Heatmap



The main attributes describing variance in the data are chest pain type, exercise-induced angina, and the slope of the peak exercise ST segment.

Creating a dashboard

```
python3 dashboard.py
```

A dashboard is created using Plotly, showing the distribution of variables.

*Evaluation

Problem type: classification

Let's move to the evaluation part.

Split data

```
from sklearn.model_selection import train_test_split

X = data.drop('target', axis=1)
y = data['target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

The data was split into training and testing datasets.

Data preparation

```
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer


imputer = SimpleImputer(strategy='mean')
X_train = imputer.fit_transform(X_train)
X_test = imputer.transform(X_test)
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

The cleaned data was prepared for evaluation.

Training data

```python
from sklearn.metrics import accuracy_score, classification_report
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.naive_bayes import GaussianNB

models = {
    'Logistic Regression': LogisticRegression(max_iter=10000),
    'Decision Tree': DecisionTreeClassifier(),
    'Random Forest': RandomForestClassifier(),
    'SVM': SVC(),
    'KNN': KNeighborsClassifier(),
    'Gradient Boosting': GradientBoostingClassifier(),
    'Naive Bayes': GaussianNB(),
}

for name, model in models.items():
    print(f"Training {name}...")
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    print(f"{name} Accuracy: {accuracy:.2f}")
    print(classification_report(y_test, y_pred))
```

```
Training Logistic Regression...
Logistic Regression Accuracy: 0.85
              precision    recall  f1-score   support

           0       0.84      0.84      0.84       112
           1       0.86      0.86      0.86       126

    accuracy                           0.85       238
   macro avg       0.85      0.85      0.85       238
weighted avg       0.85      0.85      0.85       238

Training Decision Tree...
Decision Tree Accuracy: 0.87
              precision    recall  f1-score   support
           0       0.83      0.90      0.87       112
           1       0.91      0.84      0.87       126

    accuracy                           0.87       238
   macro avg       0.87      0.87      0.87       238
weighted avg       0.87      0.87      0.87       238

Training Random Forest...
Random Forest Accuracy: 0.93
              precision    recall  f1-score   support

           0       0.91      0.94      0.93       112
           1       0.94      0.92      0.93       126

    accuracy                           0.93       238
   macro avg       0.93      0.93      0.93       238
weighted avg       0.93      0.93      0.93       238

Training SVM...
SVM Accuracy: 0.87
              precision    recall  f1-score   support

           0       0.88      0.83      0.85       112
           1       0.86      0.90      0.88       126

    accuracy                           0.87       238
   macro avg       0.87      0.86      0.86       238
weighted avg       0.87      0.87      0.87       238

Training KNN...
KNN Accuracy: 0.86
              precision    recall  f1-score   support

           0       0.84      0.87      0.85       112
           1       0.88      0.85      0.86       126

    accuracy                           0.86       238
   macro avg       0.86      0.86      0.86       238
weighted avg       0.86      0.86      0.86       238

Training Gradient Boosting...
Gradient Boosting Accuracy: 0.91
```

```
              precision    recall  f1-score   support

           0       0.91      0.90      0.91       112
           1       0.91      0.92      0.92       126

    accuracy                           0.91       238
   macro avg       0.91      0.91      0.91       238
weighted avg       0.91      0.91      0.91       238

Training Naive Bayes...
Naive Bayes Accuracy: 0.88
              precision    recall  f1-score   support

           0       0.85      0.89      0.87       112
           1       0.90      0.87      0.88       126

    accuracy                           0.88       238
   macro avg       0.88      0.88      0.88       238
weighted avg       0.88      0.88      0.88       238
```

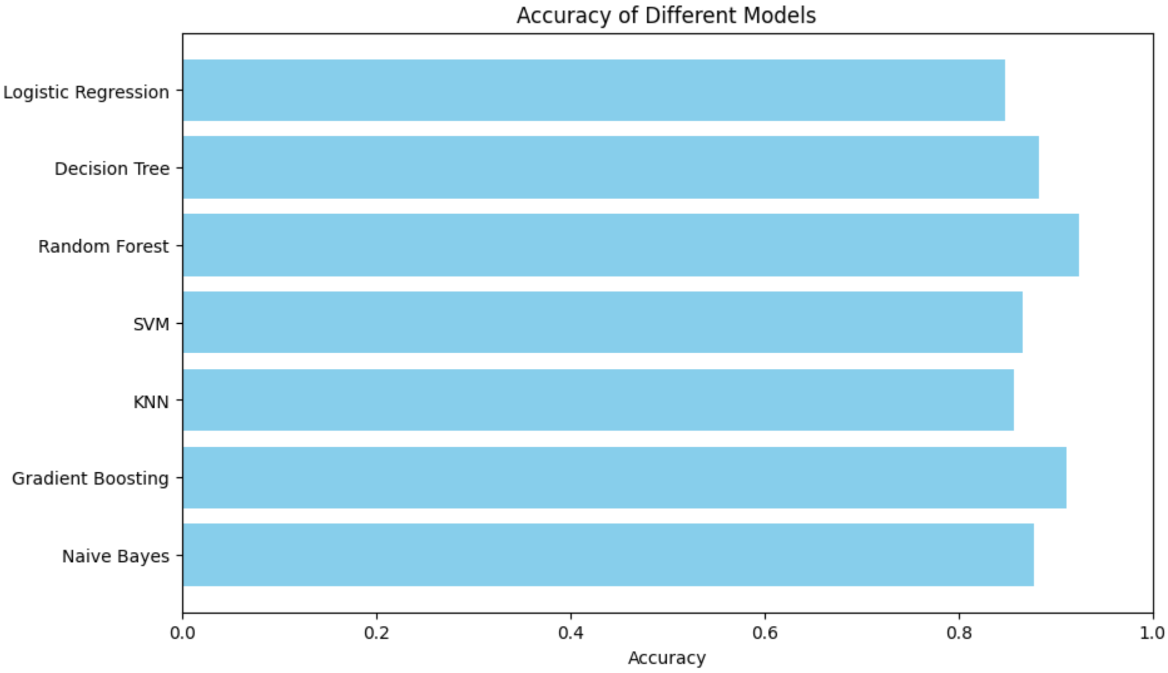
The data was trained on various models.

Stat analysis

```
Training Logistic Regression...
Logistic Regression Accuracy: 0.85
              precision    recall  f1-score   support

           0       0.84      0.84      0.84       112
           1       0.86      0.86      0.86       126

    accuracy                           0.85       238
   macro avg       0.85      0.85      0.85       238
weighted avg       0.85      0.85      0.85       238

Training Decision Tree...
Decision Tree Accuracy: 0.87
              precision    recall  f1-score   support

           0       0.83      0.90      0.87       112
           1       0.91      0.84      0.87       126
    accuracy                           0.87       238
   macro avg       0.87      0.87      0.87       238
weighted avg       0.87      0.87      0.87       238

Training Random Forest...
Random Forest Accuracy: 0.93
              precision    recall  f1-score   support

           0       0.91      0.94      0.93       112
           1       0.94      0.92      0.93       126

    accuracy                           0.93       238
   macro avg       0.93      0.93      0.93       238
weighted avg       0.93      0.93      0.93       238

Training SVM...
SVM Accuracy: 0.87
              precision    recall  f1-score   support

           0       0.88      0.83      0.85       112
           1       0.86      0.90      0.88       126

    accuracy                           0.87       238
   macro avg       0.87      0.86      0.86       238
weighted avg       0.87      0.87      0.87       238

Training KNN...
KNN Accuracy: 0.86
              precision    recall  f1-score   support

           0       0.84      0.87      0.85       112
           1       0.88      0.85      0.86       126

    accuracy                           0.86       238
   macro avg       0.86      0.86      0.86       238
weighted avg       0.86      0.86      0.86       238

Training Gradient Boosting...
Gradient Boosting Accuracy: 0.91
              precision    recall  f1-score   support

           0       0.91      0.90      0.91       112
           1       0.91      0.92      0.92       126
    accuracy                           0.91       238
   macro avg       0.91      0.91      0.91       238
weighted avg       0.91      0.91      0.91       238

Training Naive Bayes...
Naive Bayes Accuracy: 0.88
              precision    recall  f1-score   support

           0       0.85      0.89      0.87       112
           1       0.90      0.87      0.88       126

    accuracy                           0.88       238
   macro avg       0.88      0.88      0.88       238
weighted avg       0.88      0.88      0.88       238
```

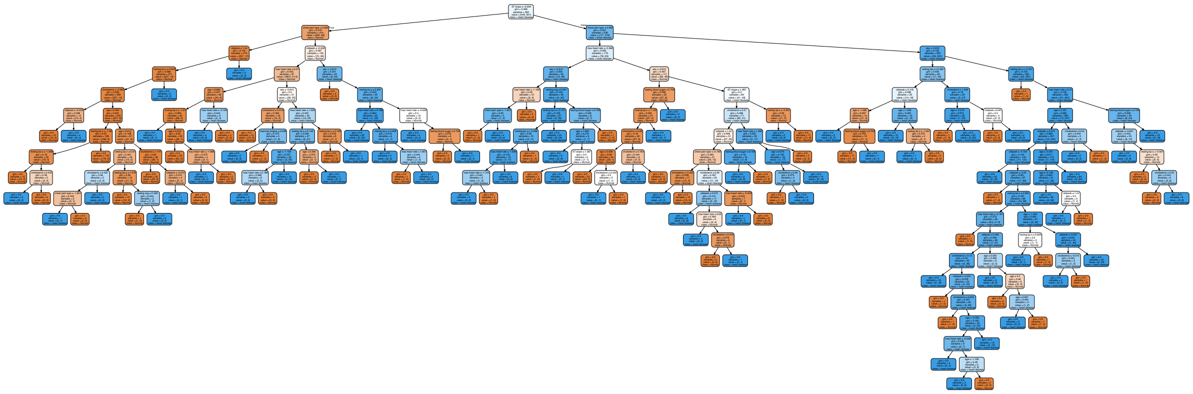
Stat analysis is dane with use of precision, recall, f1-score

Accuracy Comparison



The Random Forest algorithm shows the best results.

Building a tree



Here is the decision tree itself.