Supervised Learning Project

Predicting Sleep Disorders with Machine Learning

Sarah Goodyear

Supervised Learning Algorithms

- Learn from labeled training data to predict known outputs
- Goal: model dataset to generalize well on unseen data
- Common types:
  - Linear regression
  - Logistic regression
  - K-nearest-neighbors
  - Support vector machines
  - Decision trees
  - Random forests
  - Naïve bayes

Linear Regression

- Models the relationship between input and output variables with straight line equation
  - Minimizes mean squared error
- Regression
- Quantitative data
- Advantages: simply, interpretable, does well with linear relationships
- Disadvantages: sensitive to outliers, assumes linearity
- Scenario: predicting sales

Logistic Regression

- Models the probability of binary result using logistic function
- Classification
- Quantitative & qualitative data
- Advantages: computationally efficient, good with linearly separable classes
- Disadvantages: sensitive to correlated features
- Scenario: disease prediction

K-Nearest-Neighbors

- Classifies data points by measuring the Euclidean distance for k number of nearest neighbors
- Regression & classification
- Quantitative & qualitative data
- Advantages: simple intuitive, model complex decision boundaries
- Disadvantages: computationally expensive inference time, sensitive to feature & k selection
- Scenario: predicting car prices or diabetes

Support Vector Machines

- Identifies the optimal hyperplane to maximize the margin between classes
  - Kernel functions to model non-linear relationships
- Regression & classification
- Quantitative & qualitative (with transformations) data
- Advantages: resistant to overfitting with tuning, good for high dimensional data
- Disadvantages: computationally expensive for large datasets, requires parameter tuning
- Scenario: stock price prediction, image classification

Decision Tree

- Split data into subsets based on features using gini impurity, entropy, or mean squared error
- Regression & classification
- Quantitative & qualitative data
- Advantages: interpretable, handle missing data well, don’t need scaling/normalizing
- Disadvantages: prone to overfitting, sensitive to small changes in data
- Scenario: predicting house price, medical diagnosis

Random Forest

- Uses multiple decision trees trained on random subsets of data and features via bagging to aggregate predictions
- Regression & classification
- Quantitative & qualitative data
- Advantages: minimizes overfitting, good for large datasets, handles missing data well
- Disadvantages: computationally expensive, hard to interpret
- Scenario: predicting pollution levels, fraud detection

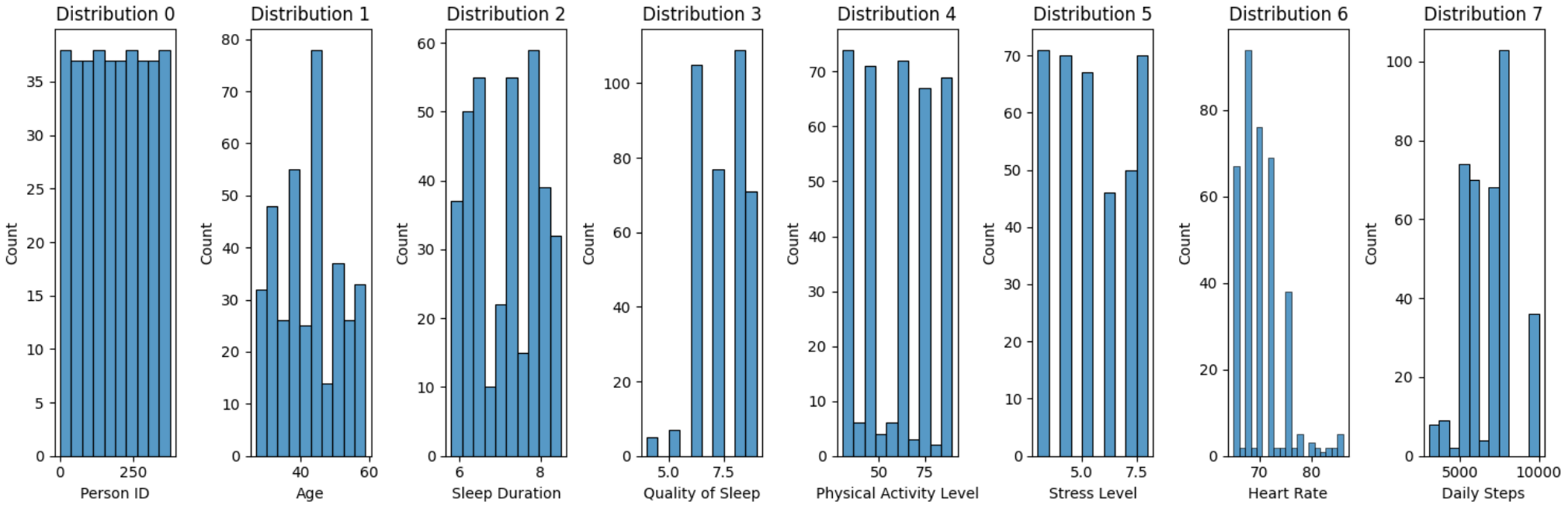
Naïve Bayes

- Family of classifiers based on Bayes' theorem that assumes features are independent
- Gaussian Naïve Bayes: performs classification tasks with data that follows a Gaussian distribution
  - Weather classification
- Multinomial Naïve Bayes: performs classification tasks with discrete count data
  - Natural language processing
- Bernoulli Naïve Bayes: performs classification tasks using binary categorical data by modeling features as binary occurrences
  - Binary text classification

Data Collection

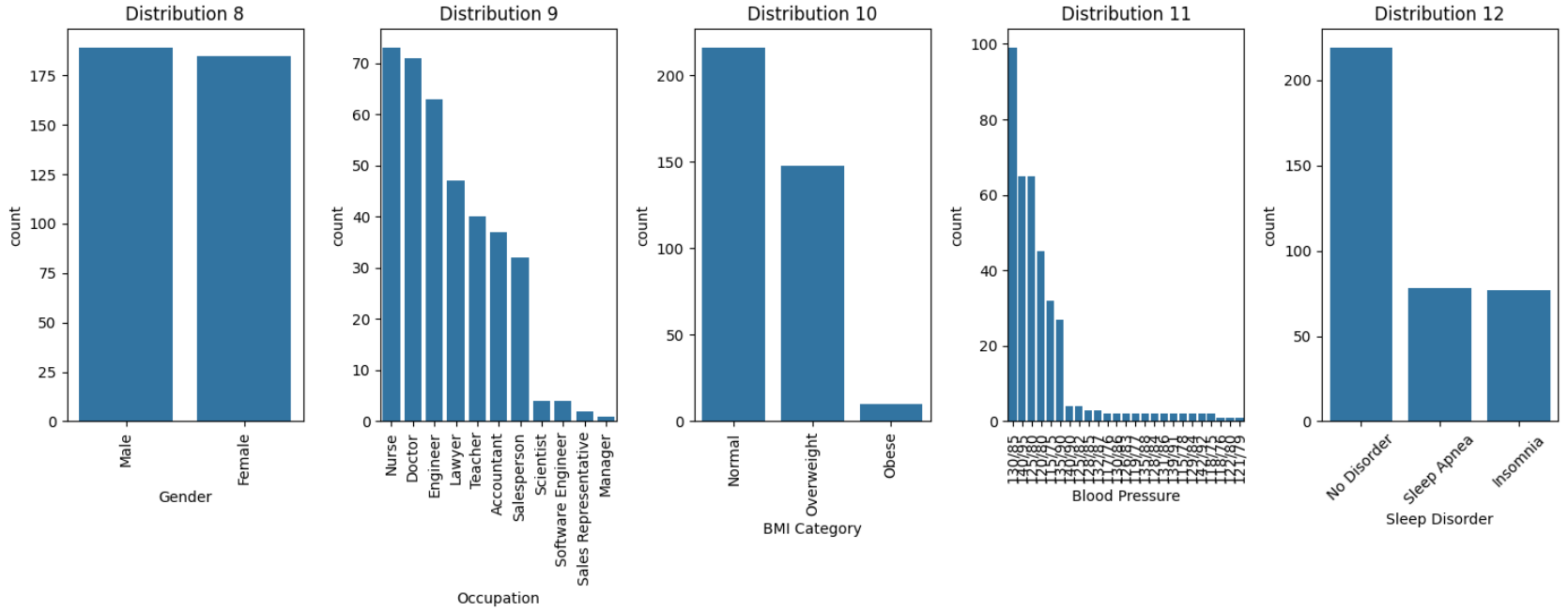
- “Sleep Health and Lifestyle Dataset” from Kaggle
- Uploaded on Feb. 25, 2024
- Synthetic data
- 374 samples
- 13 features: sleep pattern, lifestyle, and physiological characteristics

Data Preprocessing: Quantitative



- Person ID: arbitrary value for uniqueness

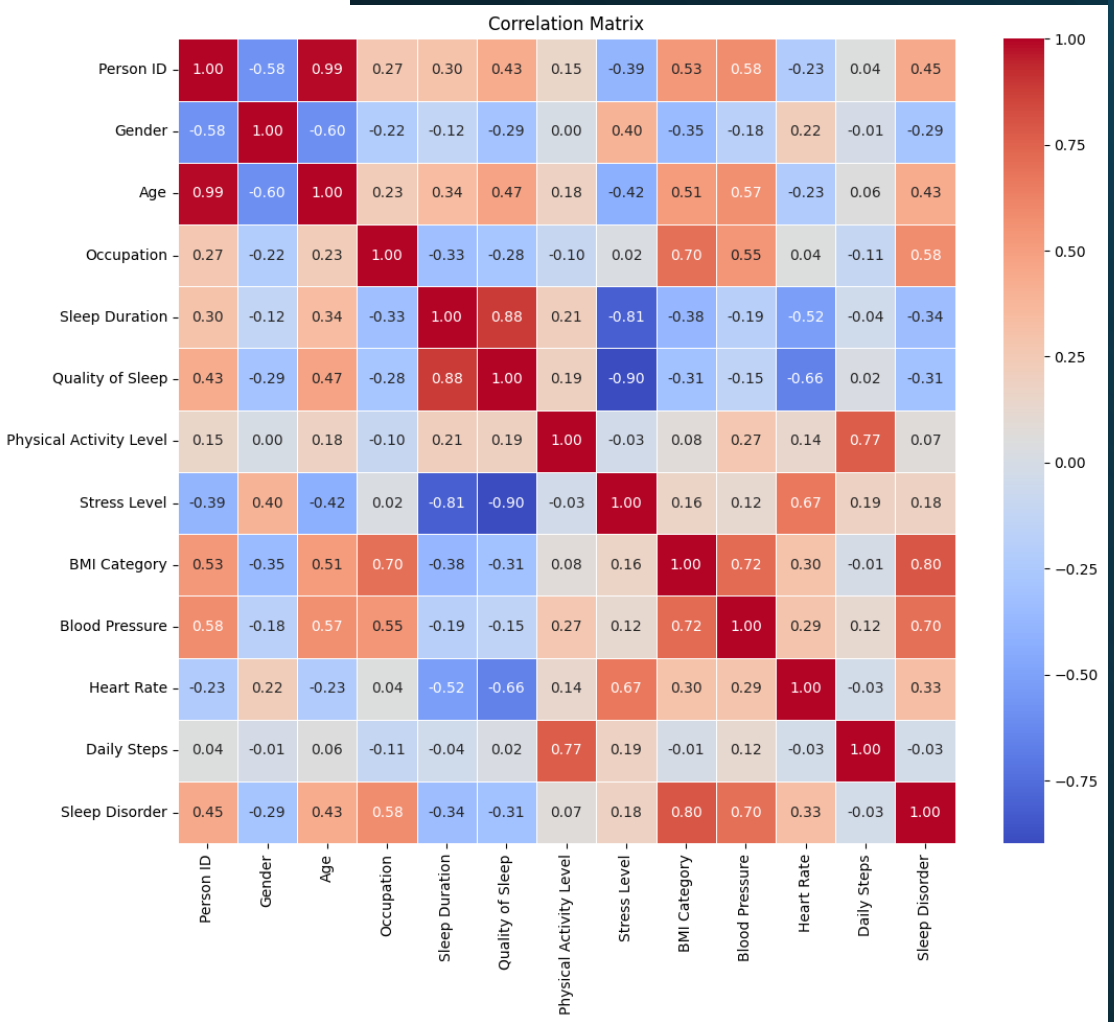
Data Preprocessing: Qualitative



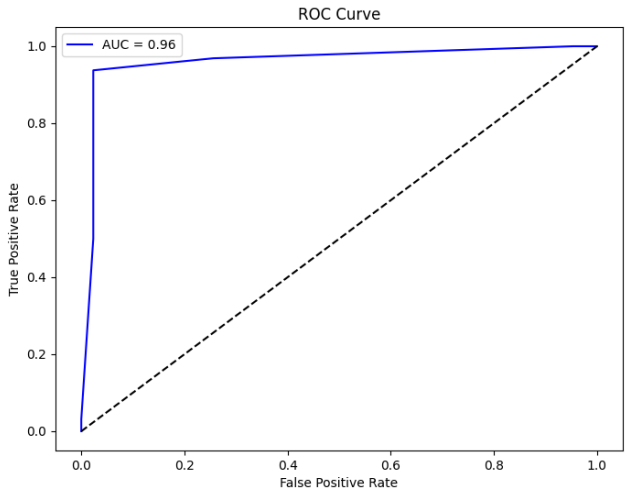
- BMI Category: Normal + Normal weight → Normal
- Sleep Disorder:
  - NaN → No Disorder
  - Sleep Apnea, Insomnia → Sleep Disorder

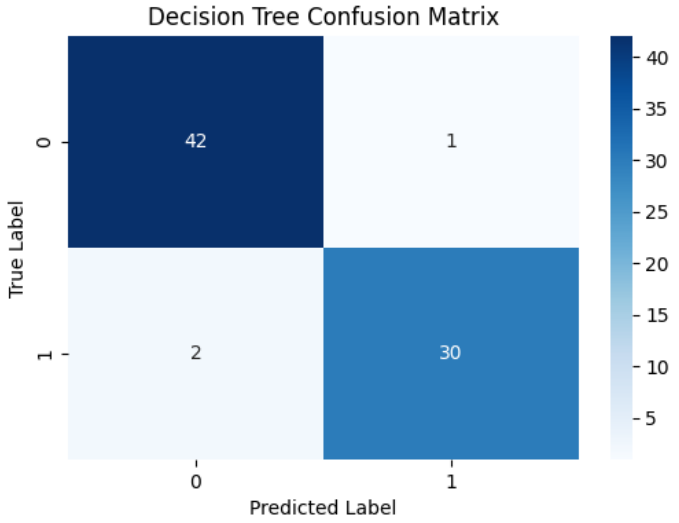
Feature Selection & Research Questions

- Target: Sleep Disorder
- Feature threshold: 0.4
  - Age
  - Occupation
  - BMI Category
  - Blood Pressure
- Can machine learning effectively predict the presence of a sleep disorder with physiological and lifestyle factors?
  - Which ML model?



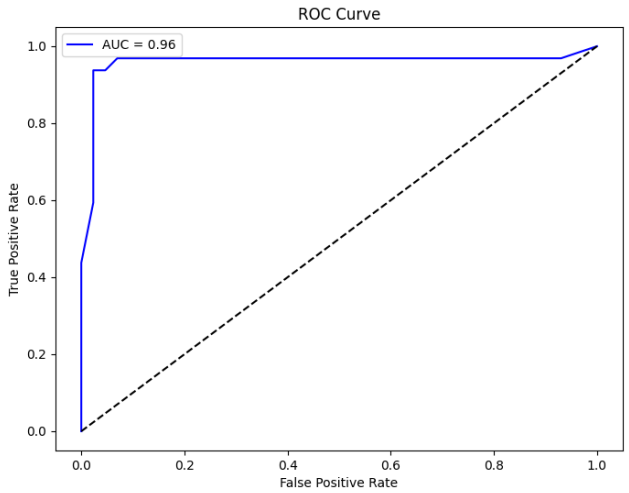
Decision Tree Application

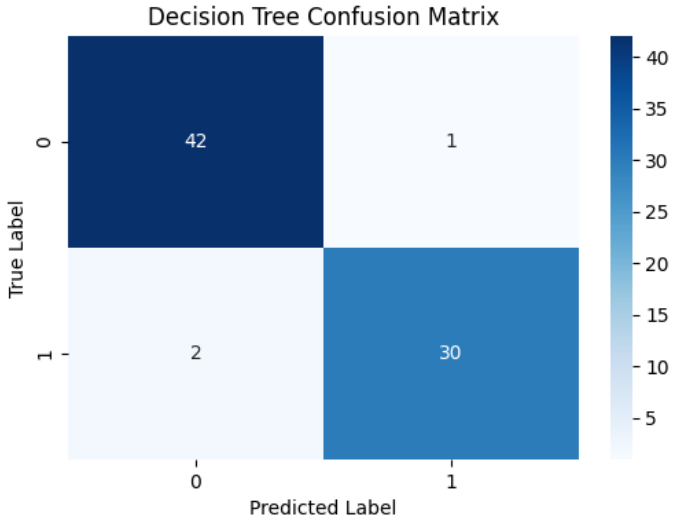




- Parameter optimization: maximum depth (3, 5, 10, none), minimum samples split (2, 5, 10), minimum samples leaf (1, 2, 4), criterion (gini, entropy)
- F1-score: 0.95

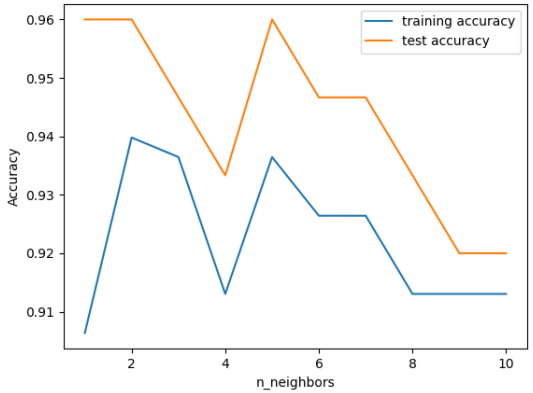
Random Forest Application

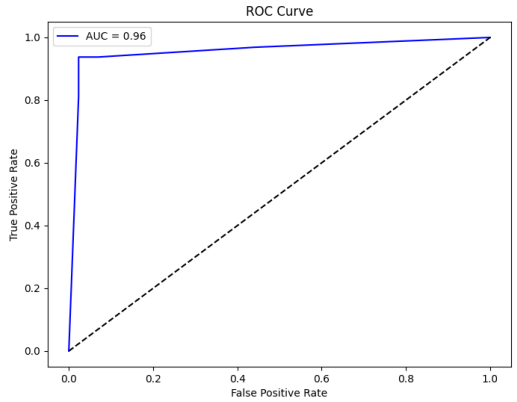


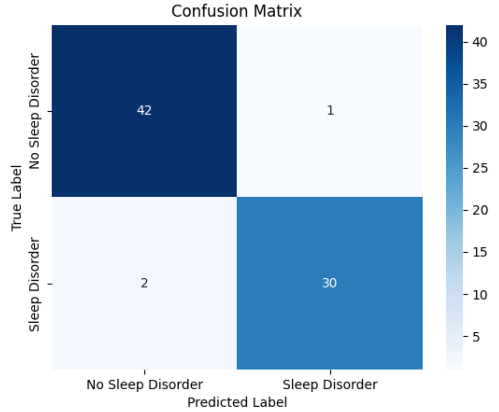


- Parameter optimization: n estimators (50, 100, 200), maximum depth (5, 10, 20, none), minimum samples split (2, 5, 10), minimum samples leaf (1, 2, 4), criterion (gini, entropy)
- F1-score = 0.95

kNN Application

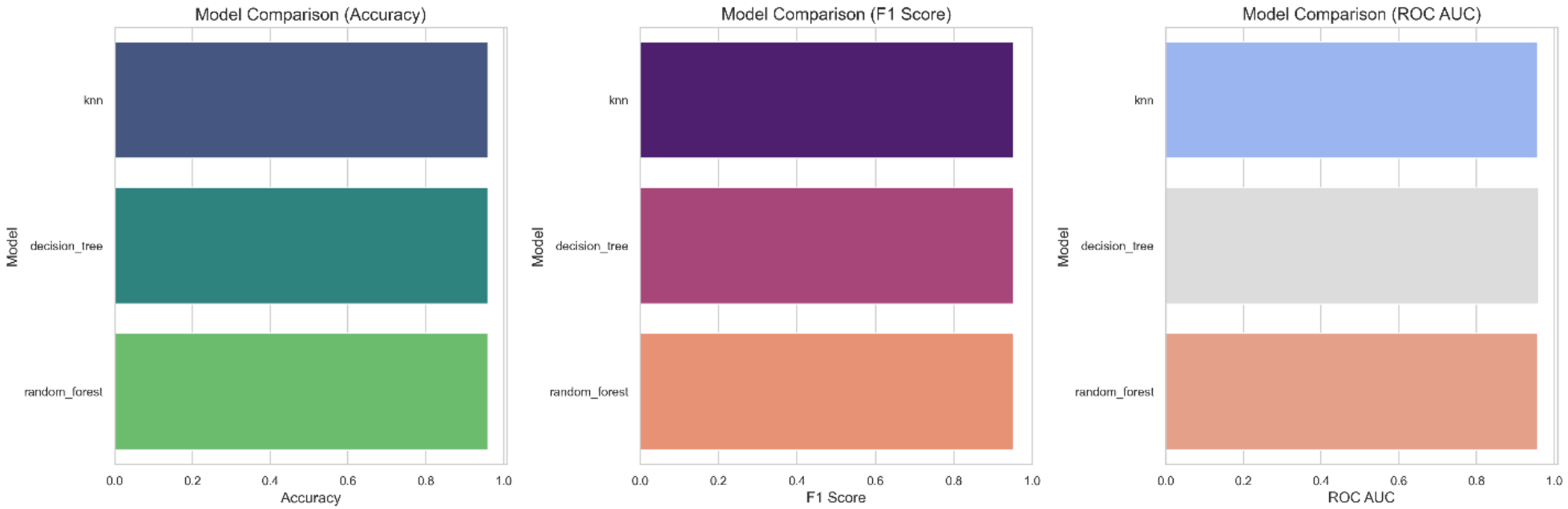






- Tested what k best maximized both training and testing accuracies on a range from k=1 to k=10
  - k = 5
- F1-score = 0.95

Results & Recommendations



	Accuracy	F1 Score	ROC AUC
kNN	0.96	0.952	0.957
Decision Tree	0.96	0.952	0.961
Random Forest	0.96	0.952	0.958