Lynne Wang

Heart Attack Prediction

Objective

The objective is to use features in a dataset to train classification models to predict whether a person is at risk of developing a heart attack. The dataset utilized in this study comes from Kaggle and can be accessed at: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.

Executive Summary

Different classification models are built using input features including, age, sex, chest pain type (cp), resting blood pressure (trtbps), cholesterol level (chol), fasting blood pressure (fbs), resting electrocardiogram results (restecg), maximum heart rate achieved (thalachh), exercise induced angina (exng), ST depression induced by exercise relative to rest (oldpeak), slope of the peak exercise ST segment (slp), number of major vessels colored by fluoroscopy (caa), and thalassemia (thall). Logistic regression, KNN, decision tree, random forest, and ensemble algorithms were used to train these models. Grid searches are conducted to adjust hyperparameters to improve the models. Grid search is able improve some models (e.g., KNN and random forest models), but not others (e.g., logistic regression, decision tree, SVC, and ensemble models). Out of all the models, the ensemble model is the best performer, achieving an accuracy rate of 90%.

Understanding Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trtbps    303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalachh  303 non-null    int64
 8   exng      303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slp       303 non-null    int64
 11  caa       303 non-null    int64
 12  thall     303 non-null    int64
 13  output    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

0. Age (age): Age is a critical risk factor in heart disease. Understanding the age distribution of a population can help tailor health services and insurance policies.

1. Sex (sex): 0 for female and 1 for male.

2. Chest Pain Type (cp):

- Value 1: typical angina
- Value 2: atypical angina
- Value 3: non-anginal pain
- Value 4: asymptomatic

3. Resting Blood Pressure (trtbps): High blood pressure is a major risk factor for heart disease. Monitoring and managing blood pressure is key for preventative health services.

4. Cholesterol Level (chol): Cholestoral in mg/dl fetched via BMI sensor. High cholesterol is another significant risk factor, informing both pharmaceutical and lifestyle intervention programs.

5. Fasting Blood Sugar (fbs): fasting blood sugar > 120 mg/dl: 1 = true; 0 = false. High fasting blood sugar levels can indicate diabetes, which is closely linked to heart health. This information could be used to integrate diabetes management and heart disease prevention.

6. Resting Electrocardiogram Results (restecg): This can show heart rhythm and function irregularities, informing immediate medical interventions and monitoring strategies.

- Value 0: normal
- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

7. Maximum Heart Rate Achieved (thalachh): This metric can be used in fitness and health monitoring, potentially guiding personalized exercise programs.

8. Exercise Induced Angina (exng): 1 = yes; 0 = no. Occurrence of angina during exercise is a significant indicator of coronary artery disease. This could inform emergency response services and patient education on activity limits.

9. ST Depression Induced by Exercise Relative to Rest (oldpeak): An important predictor of coronary artery disease used in diagnostic processes.

10. Slope of the Peak Exercise ST Segment (slp): This can indicate the severity of ischemic heart disease and guide treatment protocols.

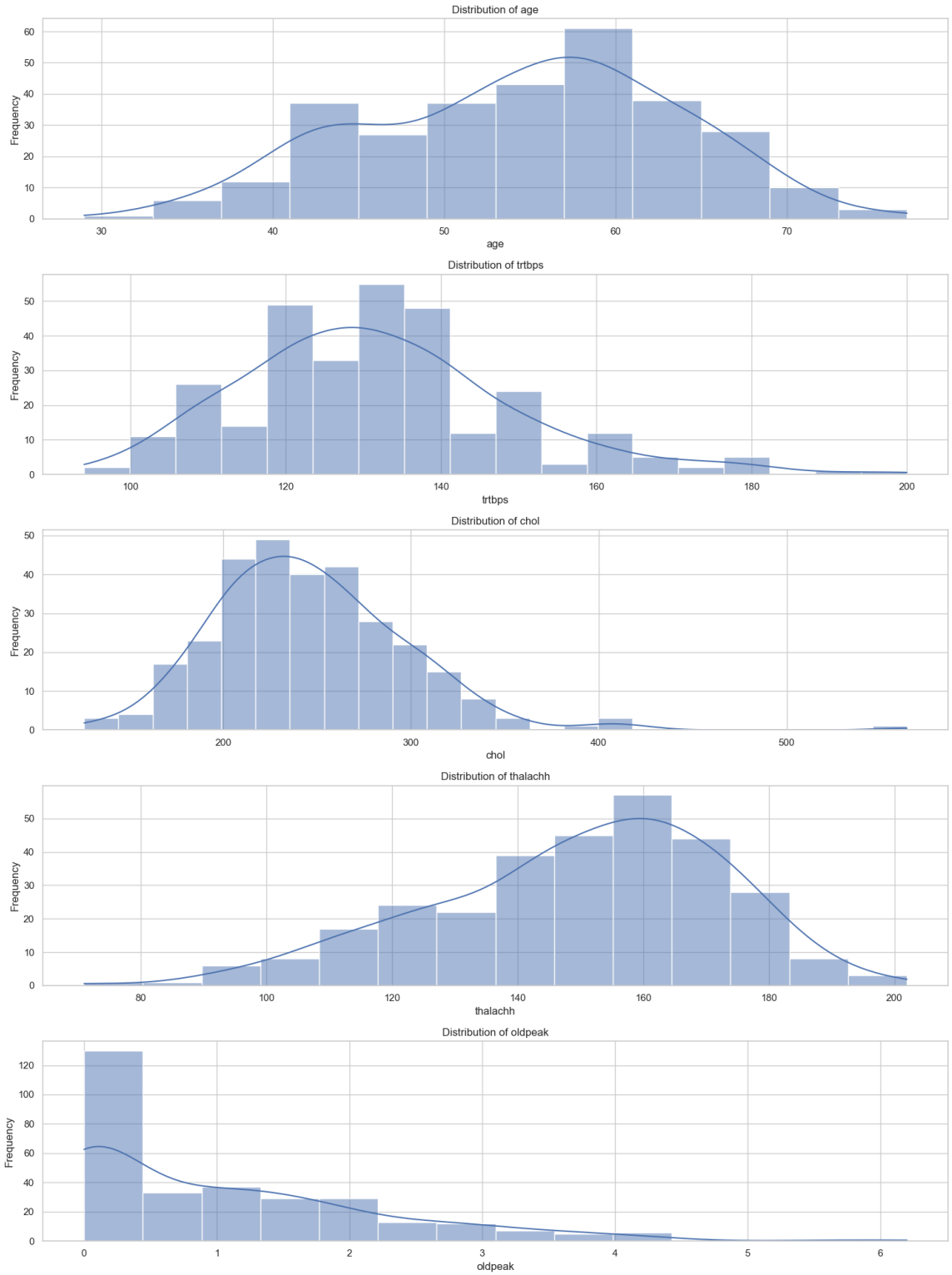
11. Number of Major Vessels Colored by Fluoroscopy (caa): 0-3. Reflects the extent of coronary artery blockage, which is crucial for surgical planning and risk assessment.

12. Thalassemia (thall): A blood disorder that can affect heart health. Understanding its prevalence can help tailor specific health services.

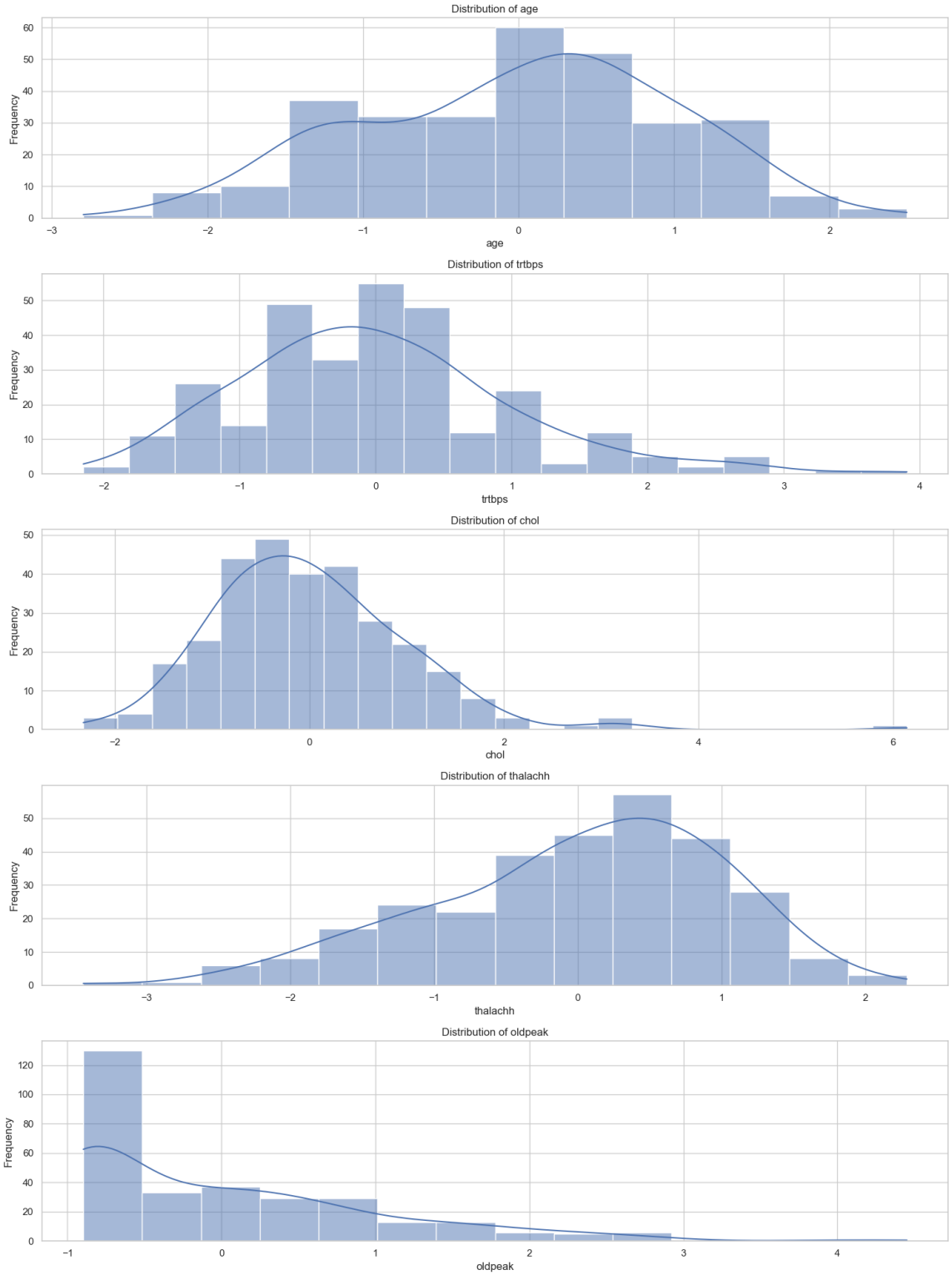
13. Output (output): 0= less chance of heart attack 1= more chance of heart attack

Visualizing Data

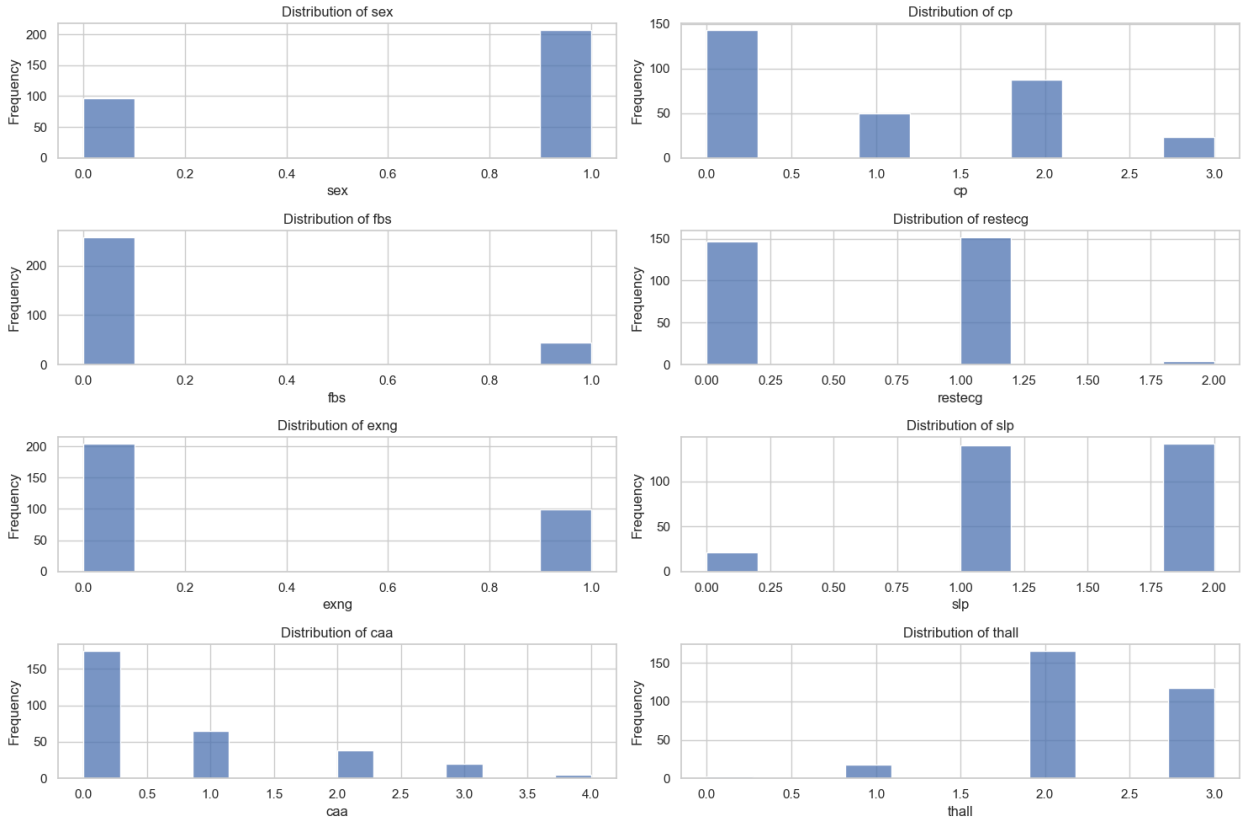
Visualizing Original Numeric Data



Visualizing Scaled Numeric Data

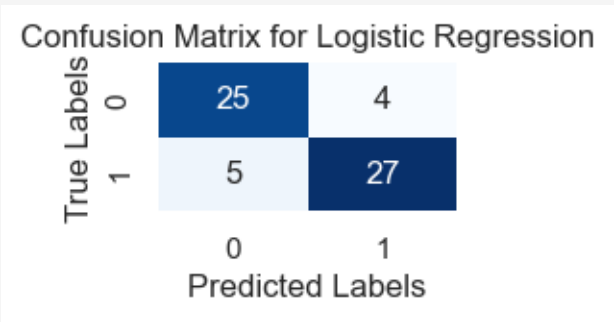


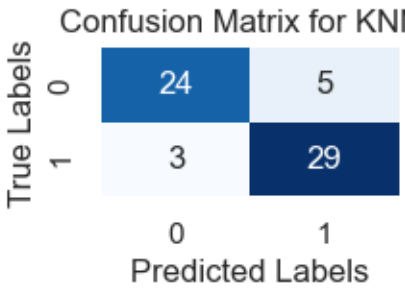
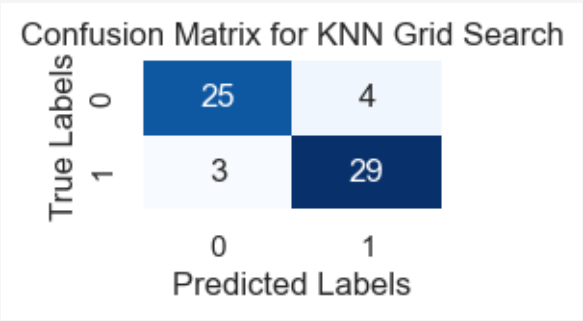
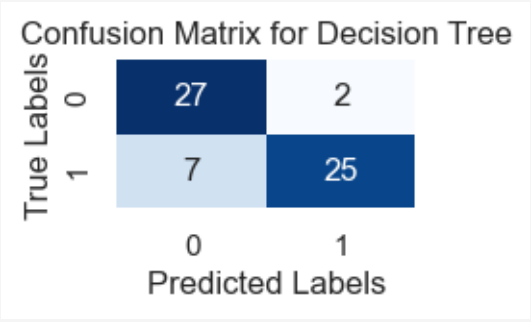
Visualizing Categorical Data

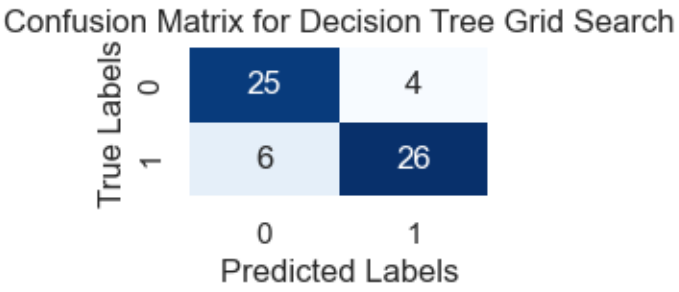
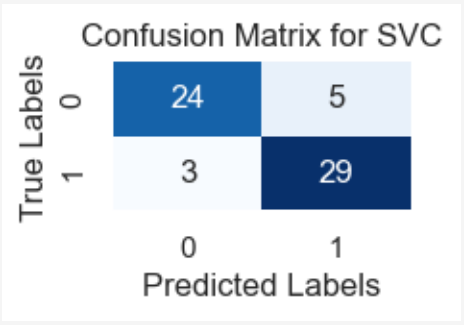
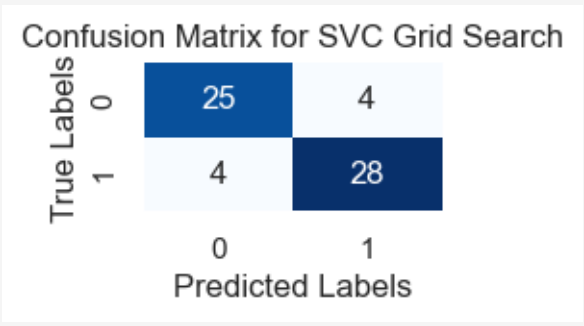
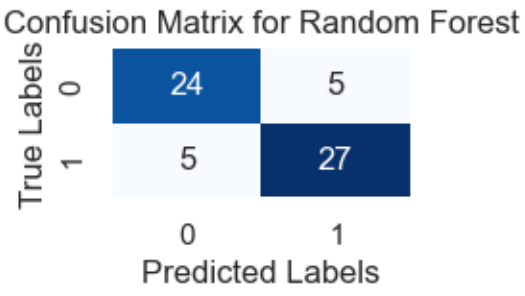


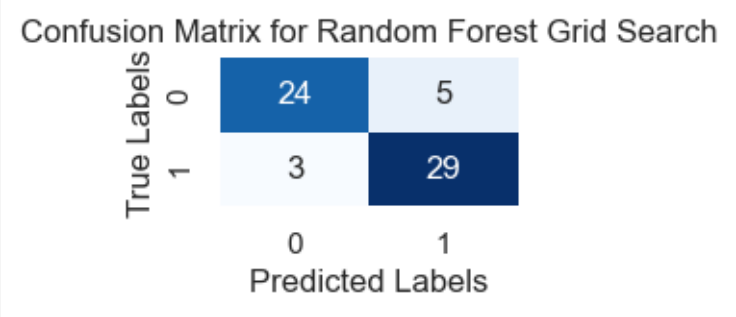
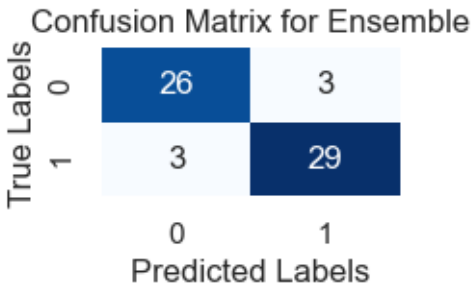
Model Accuracies

Logistic regression, KNN, decision tree, SVC, random forest, and ensemble algorithms are used to train classification models. Their accuracies are shown below. Among all the models, the ensemble model has the highest test accuracy of 90%. Interestingly, grid search did not improve the ensemble model.

Model	Train Accuracy	Test Accuracy	Confusing Matrix
Logistic Regression	0.863636	0.852459	

Model	Train Accuracy	Test Accuracy	Confusing Matrix
Logistic Regression Grid Search	0.863636	0.852459	
KNN5	0.880165	0.868852	
KNN Grid Search (N = 16)	1.000000	0.885246	
Decision Tree	1.000000	0.852459	

Model	Train Accuracy	Test Accuracy	Confusing Matrix
Decision Tree Grid Search	0.847107	0.836066	
SVC	0.888430	0.868852	
SVC Grid Search	0.863636	0.868852	
Random Forest	1.000000	0.836066	

Model	Train Accuracy	Test Accuracy	Confusing Matrix
Random Forest Grid Search	0.925620	0.868852	
Ensemble	0.933884	0.901639	
Ensemble Grid Search	0.900826	0.885246