Chronic Kidney Disease Classification Using Machine Learning

Vignesh Nagarajan DeVos Graduate School, Northwood University MGT 665 - Solving Bus Problems W/ Machine Learning Dr. Mighty Itauma June 15th, 2025

Abstract

This study focusing on a classifying the chronic Kidney Disease using three supervised machine learning models logistic Regression K-Nearest Neighbors and Decision tree. This are the three used. The dataset sources which I took in the online platform UCI machine learning which is contains 400 patient records and included both the numerical values and categorical features after this did with the preprocessing and imputation of the missing value then the models were I trained and evaluated using a standard metrics such as a accuracy, precision, recall and F1 score. The results are shows that the all models are performed well with the decision tree which is slightly outperforming with others the study of demonstrates we had a missing dataset in the CKD but we cleaned and fixed it a properly and got a very good predictions.

Introduction

Chronic Kidney Disease (CKD) is a significant global health concern that requires early detection for effective treatment. Machine learning provides a valuable tool for developing predictive models based on patient data. This lab investigates the performance of three popular classifiers on a CKD dataset: Logistic Regression, k-NN, and Decision Tree. The dataset was chosen for its clinical relevance and structured attributes. The goal is to build accurate models, compare their performance, and discuss the handling of missing data. The CKD dataset used in this study was obtained from the UCI Machine Learning Repository (Rubini, 2015).

Literature Review

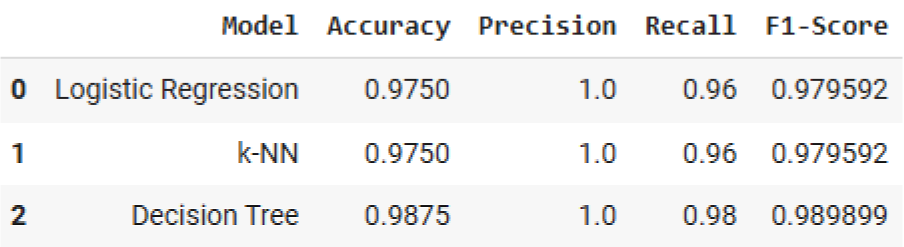
The numerous of the case studies which I have applied in the machine learning techniques for detection of a chronic kidney disease Tomar and Agarwal (2013) about the comprehensive of survey of data mining an approaches in a healthcare which is highlighting the classification as an a effective of the strategy for an disease prediction Krittanawong et al. (2017) they are demonstrating with the potential of an AI to improving the diagnosis of precisions in a cardiovascular and renal conditions in a context of a chronic kidney diseases models like Decision trees and logistic regression shown an promising the results due there interpretability and adaptability to be clinically data (Suresh & Guttag, 2021). This are findings to support the selection of this models are used in this to study and emphasize to importance of an explainability and be recall in a healthcare focused in a machine learning.

Methodology

The dataset which is containing the 25 features including a one target is class dependent variable data cleaning which is remove the extra spaces and to standardizing the labels and helps to prevent processing errors and be ensures the model interprets the data correctly. The missing numerical values were imputed using the median values and categorical values using the mode. The binary categorical variables were we encode into a 0 and 1 in this data was split into a 80% training and 20% testing sets this logistics regression and k-nearest neighbors are required to feature scaling with the standard scaler the decision tree did not require in the feature scaling because of the split the dataset based on a feature and thresholds instead of that calculating the distances or gradients unlike a logistic regression and K-nearest neighbors which is a sensitive to the scale of the inputs features decision tree to evaluate one feature at a time and be unaffected by a differences in a units or magnitude Python (Pedregosa et al., 2011).

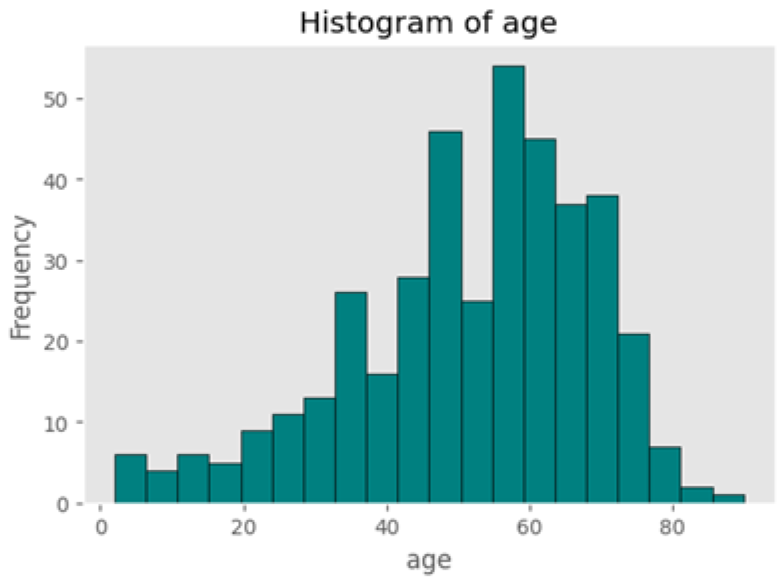
Results

The all three models are be achieved the high performance logistic regression and K-nearest neighbors are both the reached an a accuracy of 97.5% while a decision tree is achieved 98.75% in a precision was a 1 but in the decision tree has the highest recall of 0.98 and F1 -score is 0.989. this results are indicate in the excellent performance to identifying the chronic Kidney Disease cases.

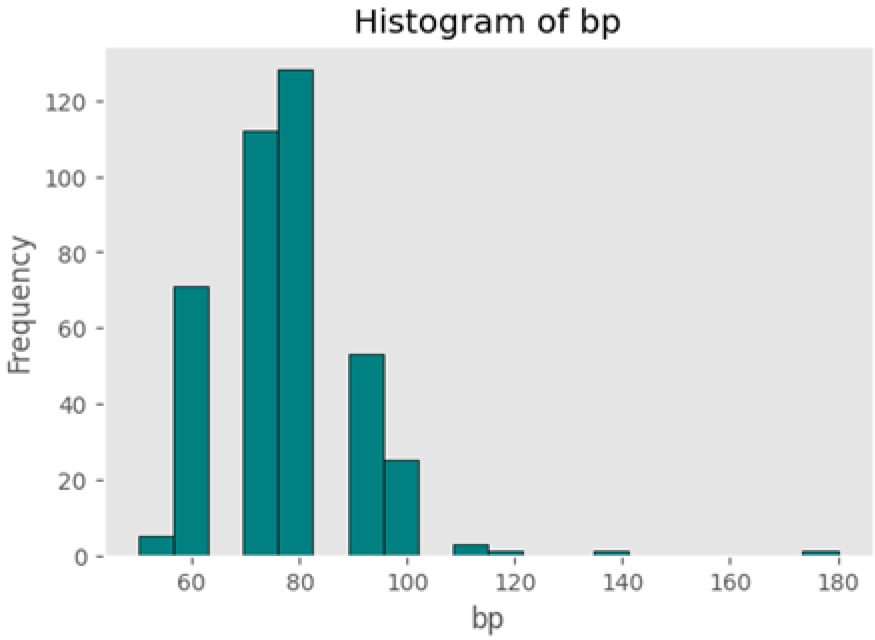


	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.9750	1.0	0.96	0.979592
1	k-NN	0.9750	1.0	0.96	0.979592
2	Decision Tree	0.9875	1.0	0.98	0.989899

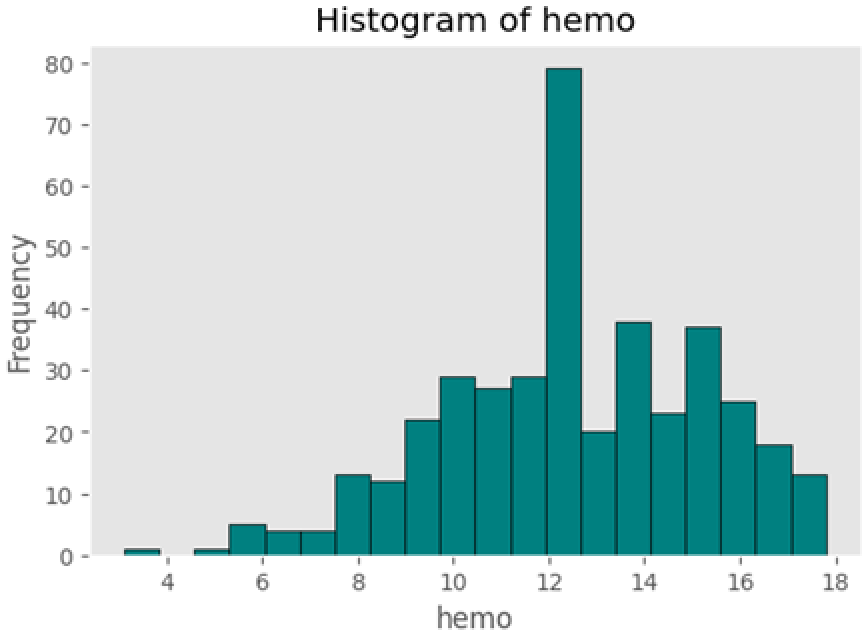
The exploratory data analysis is histograms were generated for key a features such as age, blood pressure, hemoglobin and serum creatinine. Figure 1 displays the distributions of a patients ages and is showing a higher of frequency of a chronic kidney disease in a older of individuals.



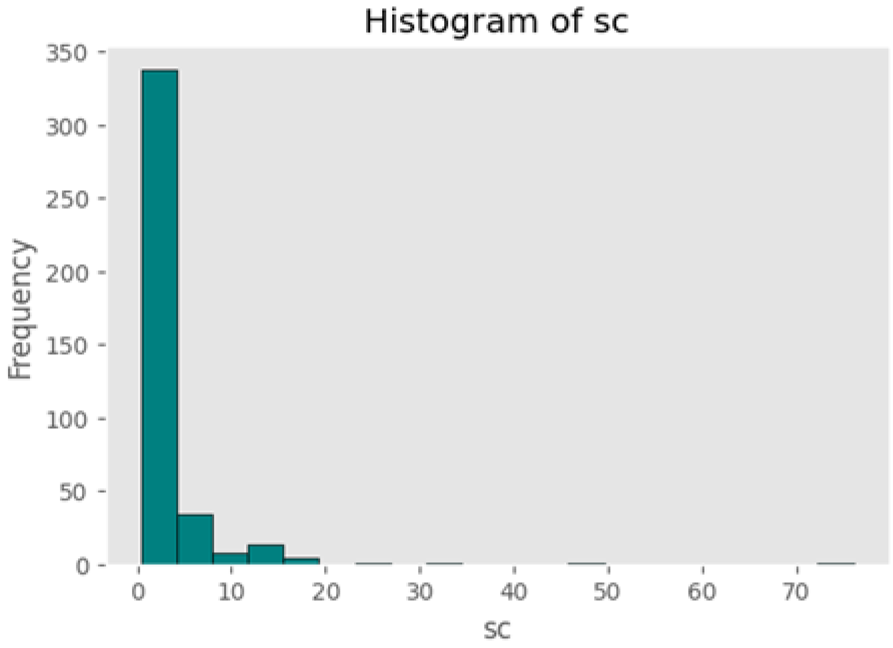
The figure 2 presents a distribution of the blood pressure among the patients most of the values are fall between a 70 and 90 mmHg which are within or a slightly above a normal of range with this noticeable of a portion of the patients have an elevated with the readings beyond a 90mmhg and a few of even exceed 120mmhg is indicating a potential hypertension of this finding suggest that is mildly elevated by a blood pressure is a common in a population and may can contribute to chronic Kidney diseases.



In the figure 3 which is histogram of hemoglobin are exhibits a left skewed negatively skewed distribution with a modal are peak of around 12 g/dl and a tail will extending a towards of the lower values this is a skewness that are indicates that is substantial of a proportion of a patients have a hemoglobin levels below this are a average of the suggesting that is presence of a distributional of asymmetry and anemia are related with the variability within a dataset.



The figure 4 histogram of the serum of creatinine is a strongly in a right skewed positively skewed with a sharp of concentration of a values of near the lower end of 0-5 mg/dl and a long of tail stretching with a towards of extreme values is greater than 70mg/dl this is distribution on reflects an a high degree of a kurtosis and also a positive of a outlier influence which is a characteristics of a clinical markers are affected by an progressive of diseases severity.



The predictive of accuracy of the models are confusion of matrices were we generated for an logistic regression K-nearest neighbors and decision tree classifiers these are the graphs figure each of the matrix are confirms that are the models are correctly will be classified ,most of the chronic kidney diseases and non-chronic kidney diseases cases with the only minor of the misclassifications this are classification are reports provided are detailed metrics the precision and recall scores were there are above of 95% for all models with the decision tree are showing the best of the balance of the high recall precisions. This is evaluation of the visuals support these numerical results are validating the effectiveness of an our preprocessing ang model are selection of strategies.

Logistic Regression Classification Report:

```
           precision    recall  f1-score   support

        0       0.94      1.00      0.97        30
        1       1.00      0.96      0.98        50

 accuracy                           0.97        80
```

macro avg 0.97 0.98 0.97 80 weighted avg 0.98 0.97 0.98 80

k-NN Classification Report:

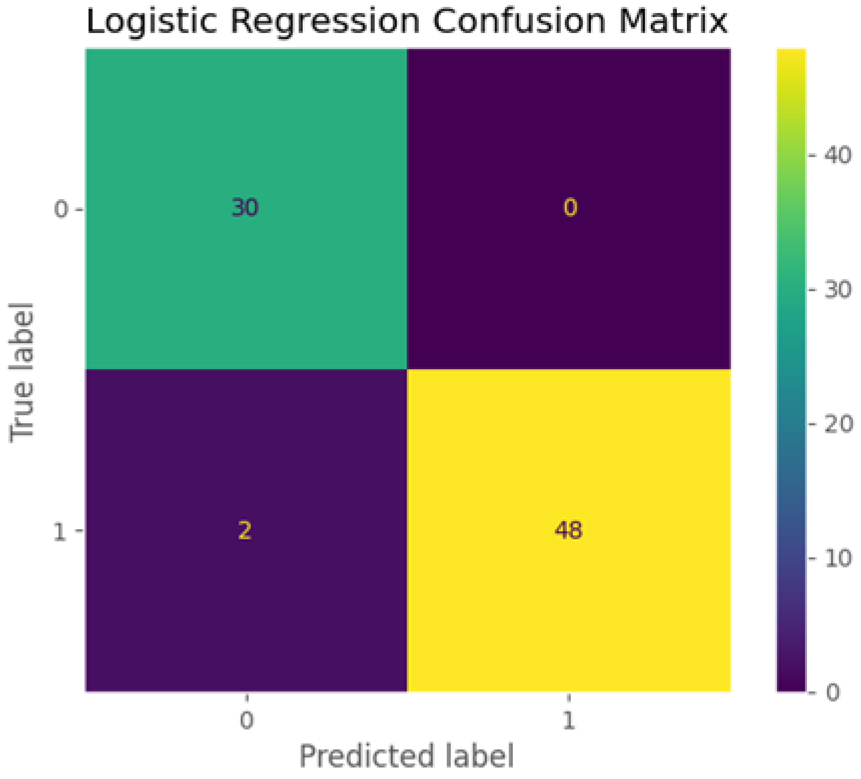
```
           precision    recall  f1-score   support

        0       0.94      1.00      0.97        30
        1       1.00      0.96      0.98        50

 accuracy                           0.97        80
```

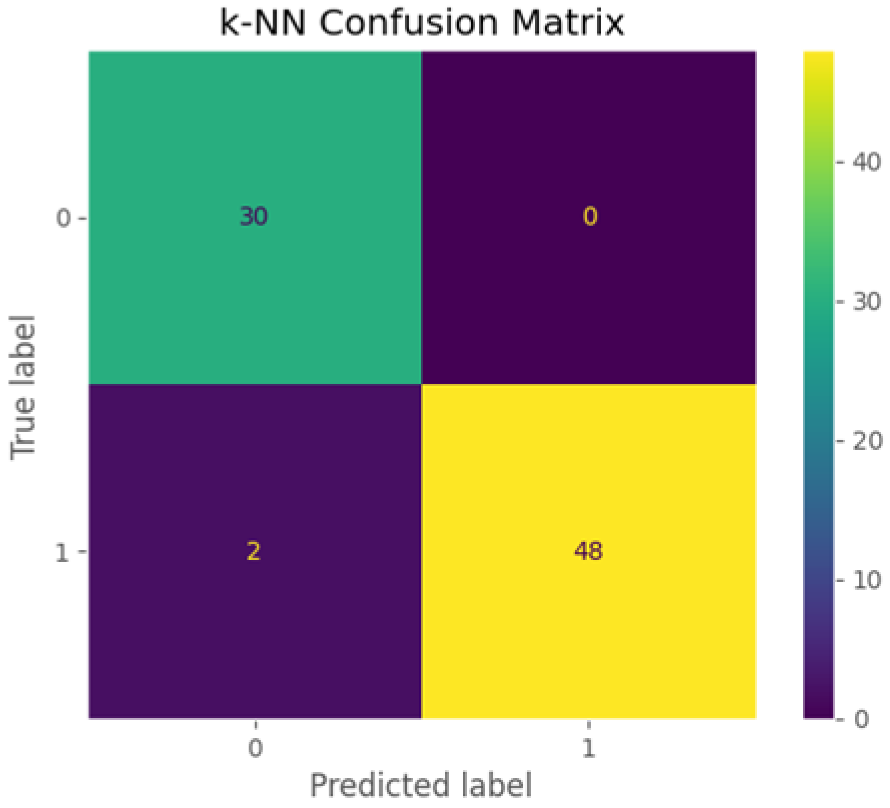
macro avg 0.97 0.98 0.97 80 weighted avg 0.98 0.97 0.98 80

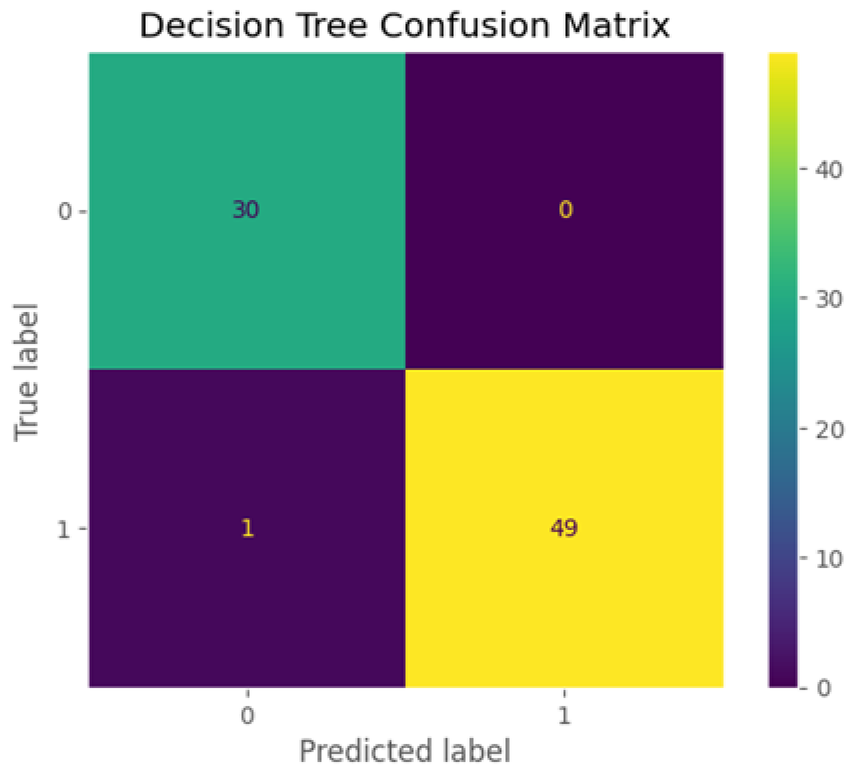
Decision Tree Classification Report:

```
           precision    recall  f1-score   support

        0       0.97      1.00      0.98        30
        1       1.00      0.98      0.99        50

 accuracy                           0.99        80
```

macro avg 0.98 0.99 0.99 80 weighted avg 0.99 0.99 0.99 80







Discussion

The three models are evaluated with the decision tree achieved the highest of the accuracy of 98.75% which is followed the closely by an a logistic regression and K-NN at 97.5% of a decision tree also we showed the a highest of F1 – score and recall are indicating its an strong of ability to be correctly classify a chronic kidney diseases cases while a logistics regression and a K-NN also we performed well be the non-linear of a nature of the decision of tree helped a capture a complex of patterns in the data.

The decision tree is a slightly to better performance can be a attributed to its be ability to be handle the both categorical and be continuous features without a requiring a scaling and to be create a hierarchical of split that is detect the subtle to be interactions between a features. Its is higher to recall means its correctly which is identified the more chronic kidney diseases cases which is a critical in a medical of applications of where we missing the positive of case are could have be serious of consequences this is elevated of F1 score reflects its is balances between to precision and recall in making it an a optimal of model in this study.

the overall of these are results are demonstrated that is with a proper of preprocessing and features of engineering even datasets with a missing values can be yield of highly accurate models this is consistency with a high precision across models are shows that is all models made a few false positives and we improved the recall of decision tree and shows a it's a strength in reducing a false negatives.

Conclusion

This is study with the demonstrates that is machine learning models are highly with the effective in a predicting a chronic kidney disease using a structed with the clinical data with a despite the presence of a missing a values in a dataset this are a application of a proper and preprocessing and including a median imputation for a numeric variables and a mode of imputation for a categorical of variables preserved of data quality and a model performance.

All the three of classification in a model are evaluated in this study of logistic regression K-NN and decision tree are achieved with the high precision and a accuracy of however this a decision tree models are provided are slightly superior results are especially in a recall and f1 score which are is critical for a minimizing a false of negatives in a medical of diagnistics.

The furthermore of decision trees are a inherently are interpretable and capable of this handling both a numerical and categorical features of without they are need for a scaling making them a well we suited for a real world clinical decision support and systems are overall this is study of reinforces the values of a machine learning in a healthcare and supports that is integration of a predictive models for a early of stage chronic kidney disease detection.

GitHub

File path : https://github.com/vigneshNagaraja/LAB-2-Assignment-.git/

References

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011).

Suresh, H., & Guttag, J. V. (2021). A framework for understanding unintended consequences of machine learning. Communications of the ACM, 64(3), 62–71

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 69(21), 2657–2664.

Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241–266.

```
#Step 1: Upload and Load Data

from google.colab import files
uploaded = files.upload()
```

```
import pandas as pd

df = pd.read_csv('ckd_cleaned_imputed.csv')
df.head()
```

Show hidden output

```
 #Step 2: Basic Cleaning (Whitespace + Categorical Encoding)

# Clean up any stray tabs or spaces
df = df.applymap(lambda x: x.strip() if isinstance(x, str) else x)

# Encode categories and target
binary_map = {
    'yes': 1, 'no': 0,
    'present': 1, 'notpresent': 0,
    'abnormal': 1, 'normal': 0,
    'good': 1, 'poor': 0,
    'ckd': 1, 'notckd': 0
}
df.replace(binary_map, inplace=True)
```

Show hidden output

```
 #Step 3: Exploratory Data Analysis (Histograms)

important_cols = ['age', 'bp', 'hemo', 'sc', 'class']

for col in important_cols:
    plt.figure(figsize=(6, 4))
    df[col].hist(bins=20, color='teal', edgecolor='black')
    plt.title(f'Histogram of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.grid(False)
    plt.show()
```

Show hidden output

```
#Step 4: Prepare Data for Modeling

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

```
X = df.drop("class", axis=1)
y = df["class"]

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42
)

# Scale features for LR and k-NN
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

```
 #Step 5: Train Models

from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier

# Initialize models
lr = LogisticRegression(max_iter=1000)
knn = KNeighborsClassifier(n_neighbors=5)
dt = DecisionTreeClassifier(random_state=42)

# Train models
lr.fit(X_train_scaled, y_train)
knn.fit(X_train_scaled, y_train)
dt.fit(X_train, y_train)  # Tree does not need scaling
```

Show hidden output

```
 #Step 6: Predict and Evaluate
from sklearn.metrics import classification_report, confusion_matrix, ConfusionMatrixDisplay

# Predictions
y_pred_lr = lr.predict(X_test_scaled)
y_pred_knn = knn.predict(X_test_scaled)
y_pred_dt = dt.predict(X_test)

# Evaluation Function
def evaluate_model(name, y_true, y_pred):
    print(f"\n{name} Classification Report:\n")
    print(classification_report(y_true, y_pred))
    ConfusionMatrixDisplay(confusion_matrix(y_true, y_pred)).plot()
    plt.title(f"{name} Confusion Matrix")
    plt.grid(False)
    plt.show()
```

```
evaluate_model("Logistic Regression", y_test, y_pred_lr)
evaluate_model("k-NN", y_test, y_pred_knn)
evaluate_model("Decision Tree", y_test, y_pred_dt)
```

Show hidden output

```
 #Step 7: Compare All Models

from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

def get_scores(y_true, y_pred, model):
    return {
        "Model": model,
        "Accuracy": accuracy_score(y_true, y_pred),
        "Precision": precision_score(y_true, y_pred),
        "Recall": recall_score(y_true, y_pred),
        "F1-Score": f1_score(y_true, y_pred)
    }

results = [
    get_scores(y_test, y_pred_lr, "Logistic Regression"),
    get_scores(y_test, y_pred_knn, "k-NN"),
    get_scores(y_test, y_pred_dt, "Decision Tree")
]

results_df = pd.DataFrame(results)
results_df
```

Show hidden output

Next steps: Generate code with `results_df` View recommended plots New interactive sheet

Could not connect to the reCAPTCHA service. Please check your internet connection and reload to get a reCAPTCHA challenge.