Predicting 30-Day Hospital Readmissions: A Comparative Analysis Using Logistic Regression and CART

Masheia Dzimba and Peter Mangoro

2025-12-02

Contents

1 Abstract

This study aims to predict 30-day hospital readmissions for patients with diabetes using patient demographic, diagnostic, and treatment-related features. We employed two statistical modeling approaches: Logistic Regression and Classification and Regression Trees (CART). The dataset consists of 24,996 patient encounters from 130 US hospitals collected between 1999-2008. Our analysis identified previous inpatient visits, age groups, and medical specialty as key predictors of readmission. The Logistic Regression model achieved an accuracy of 61.84% and AUC of 0.648, while CART achieved 60.78% accuracy and AUC of 0.605. Both models demonstrate moderate predictive performance, suggesting that additional clinical variables may be needed to improve predictions. The Logistic Regression model provides more detailed statistical insights with odds ratios and significance testing, making it preferable for clinical decision support. This analysis highlights the importance of patient history and demographics in predicting readmission risk, with implications for healthcare resource allocation and patient care planning.

2 Introduction

2.1 Background

Hospital readmissions within 30 days of discharge are a significant concern in healthcare, associated with increased costs, patient morbidity, and healthcare system burden. For patients with diabetes, readmission rates are particularly high, making early identification of at-risk patients crucial for improving outcomes and reducing healthcare costs.

2.2 Research Question

Can patient demographic, diagnostic, and treatment-related features effectively predict the likelihood of a patient being readmitted to the hospital within 30 days of discharge?

2.3 Objectives

1. Identify key predictors of 30-day hospital readmissions
2. Compare the performance of Logistic Regression and CART models
3. Evaluate model interpretability and clinical utility
4. Provide recommendations for clinical application

3 Data Description

3.1 Data Source

The dataset was obtained from Kaggle: "Diabetes 130-US Hospitals for 10 years" and contains 25,000 patient encounters from 130 US hospitals and integrated delivery networks over a 10-year period (1999-2008). The data was collected retrospectively from electronic health records (EHR) and anonymized for research purposes.

3.2 Data Collection Method

This is an **observational study** - data was collected by observing patient outcomes and characteristics without any intervention or manipulation of variables by the researchers.

3.3 Dependent Variable

`readmitted`: Binary categorical variable indicating whether the patient was readmitted to the hospital within 30 days of discharge.

- **Type**: Binary (after cleaning: 0 = Not Readmitted, 1 = Readmitted)
- **Distribution**: 47.02% readmitted, 52.98% not readmitted
- **Final sample size**: 24,996 observations (4 rows dropped due to missing primary diagnosis)

3.4 Independent Variables

The analysis includes the following variables:

3.4.1 Categorical Variables:

- `age`: Age groups ([40-50), [50-60), [60-70), [70-80), [80-90), [90-100))
- `medical_specialty`: Medical specialty (7 categories including "Missing")
- `diag_1`: Primary diagnosis (8 categories: Circulatory, Diabetes, Digestive, Injury, Musculoskeletal, Other, Respiratory)
- `change`: Change in medication (yes/no)
- `diabetes_med`: Diabetes medication prescribed (yes/no)

3.4.2 Numerical Variables:

- `time_in_hospital`: Length of stay in days (Mean: 4.45, Range: 1-14)
- `n_lab_procedures`: Number of lab tests performed (Mean: 43.24)
- `n_procedures`: Number of procedures (Mean: 1.35, Range: 0-6)
- `n_medications`: Number of medications (Mean: 16.25, Range: 1-79)

4 Methods

4.1 Statistical Methods

This project employs a **comparative analysis** using two distinct statistical methods:

4.1.1 1. Logistic Regression (Primary Method)

- **Justification**: Standard method for modeling binary categorical response variables
- **Advantages**: Provides interpretable results (odds ratios), statistical significance testing, confidence intervals
- **Assumptions**: Linear relationships between predictors and log-odds of outcome

4.1.2 2. Classification and Regression Trees (CART) (Secondary Method)

- **Justification**: Non-parametric method that captures complex, non-linear relationships
- **Advantages**: Highly interpretable decision rules, no distributional assumptions
- **Comparison**: Evaluated against Logistic Regression based on predictive performance and interpretability

4.2 Data Preprocessing

1. **Response Variable**: Converted to binary format (0/1)
2. **Missing Values**:
   - Dropped 4 rows with missing primary diagnosis (0.02%)
   - Kept "Missing" as valid category for other variables
3. **Feature Engineering**: Created derived features (n_diagnoses, medications_per_day, total_previous_visits)
4. **Encoding**:
   - Logistic Regression: Dummy/one-hot encoding
   - CART: Factor encoding

4.3 Model Evaluation

Models were evaluated using: - **Train/Test Split**: 70% training, 30% testing - **Metrics**: Accuracy, Precision, Recall, Specificity, F1-Score, AUC-ROC - **Statistical Tests**: Hypothesis testing for Logistic Regression coefficients

5 Results

5.1 Summary Statistics

Table 1: Summary Statistics by Readmission Status

readmitted	time_in_hospital_mean	time_in_hospital_sd	n_lab_procedures_mean	n_lab_procedures_sd	n_procedures_mean	n_procedures_sd	n_medications_mean	n_medications_sd
Not Readmitted	4.33	3	42.62	20.09	1.42	1.75	15.97	8.45
Readmitted	4.59	3	43.94	19.49	1.27	1.68	16.57	7.58

5.2 Data Visualizations

5.2.1 Distribution of Readmission Status

5.2.2 Numerical Variables by Readmission Status

5.2.3 Categorical Variables Analysis

5.3 Logistic Regression Results

5.3.1 Model Performance

Table 2: Logistic Regression Performance Metrics

Metric	Value	Percentage
Accuracy	0.6184316	61.84316
Precision	0.6409032	64.09032
Recall (Sensitivity)	0.4205128	42.05128
Specificity	0.7926279	79.26279
F1-Score	0.5078273	50.78273

5.3.2 Regression Output

Table 3: Top 10 Significant Variables in Logistic Regression Model

	Variable	Coefficient	P_Value	Odds_Ratio	CI_Lower	CI_Upper
6	n_outpatient	0.1061	0.000000	1.1119	1.0789	1.1460
7	n_inpatient	0.3878	0.000000	1.4738	1.4244	1.5249

	Variable	Coefficient	P_Value	Odds_Ratio	CI_Lower	CI_Upper
8	n_emergency	0.2177	0.000000	1.2433	1.1707	1.3203
12	diabetes_med_binary	0.2332	0.000000	1.2626	1.1588	1.3758
27	diag1_diag_1injury	-0.2843	0.000154	0.7525	0.6495	0.8719
1	(Intercept)	-1.7553	0.000365	0.1729	0.0658	0.4538
22	medspec_medical_specialtyMissing	0.2675	0.000518	1.3066	1.1235	1.5197
29	diag1_diag_1other	-0.1755	0.000730	0.8391	0.7578	0.9290
4	n_procedures	-0.0306	0.003797	0.9698	0.9499	0.9902
19	medspec_medical_specialtyEmergency.Trauma	0.2560	0.006632	1.2918	1.0738	1.5541

5.3.3 Hypothesis Testing

For each significant variable in the Logistic Regression model:

- H_0: $\beta_{variable} = 0$ (variable has no effect on readmission)
- H_1: $\beta_{variable} \neq 0$ (variable affects readmission)

All variables shown in the table above have p $<$ 0.05, indicating we reject H_0 and conclude these variables are significant predictors.

5.3.4 R-squared Interpretation

Table 4: Pseudo R-squared Values

Metric	Value
McFadden's Pseudo R^2	0.0505021
Nagelkerke's R^2	0.0900450

The Logistic Regression model explains approximately **5.05%** of the variance in readmission status (McFadden's Pseudo $R^2 = 0.0505$). While this is relatively low, it is common for logistic regression models, and values above 0.2-0.4 are considered good.

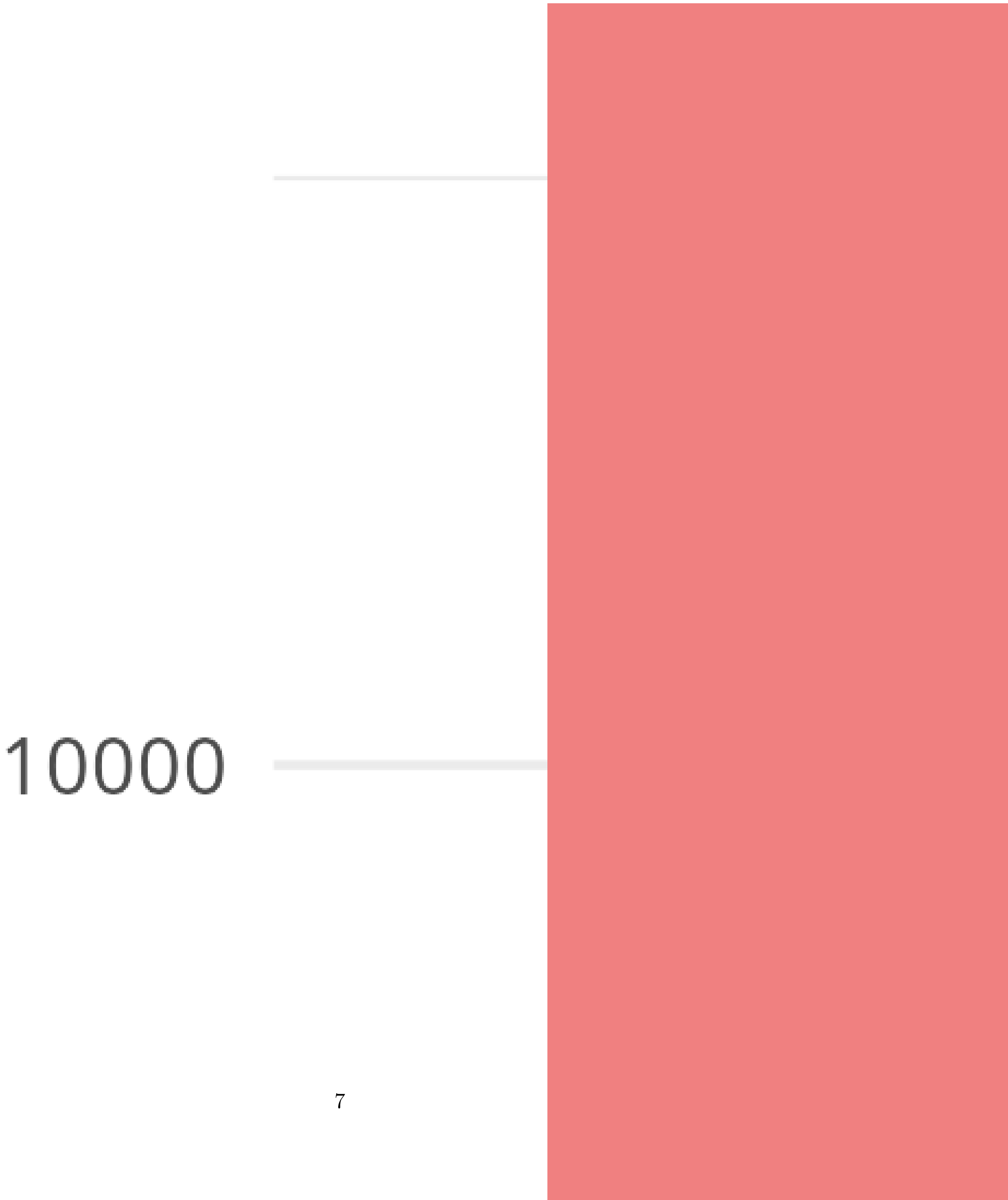
5.3.5 ROC Curve

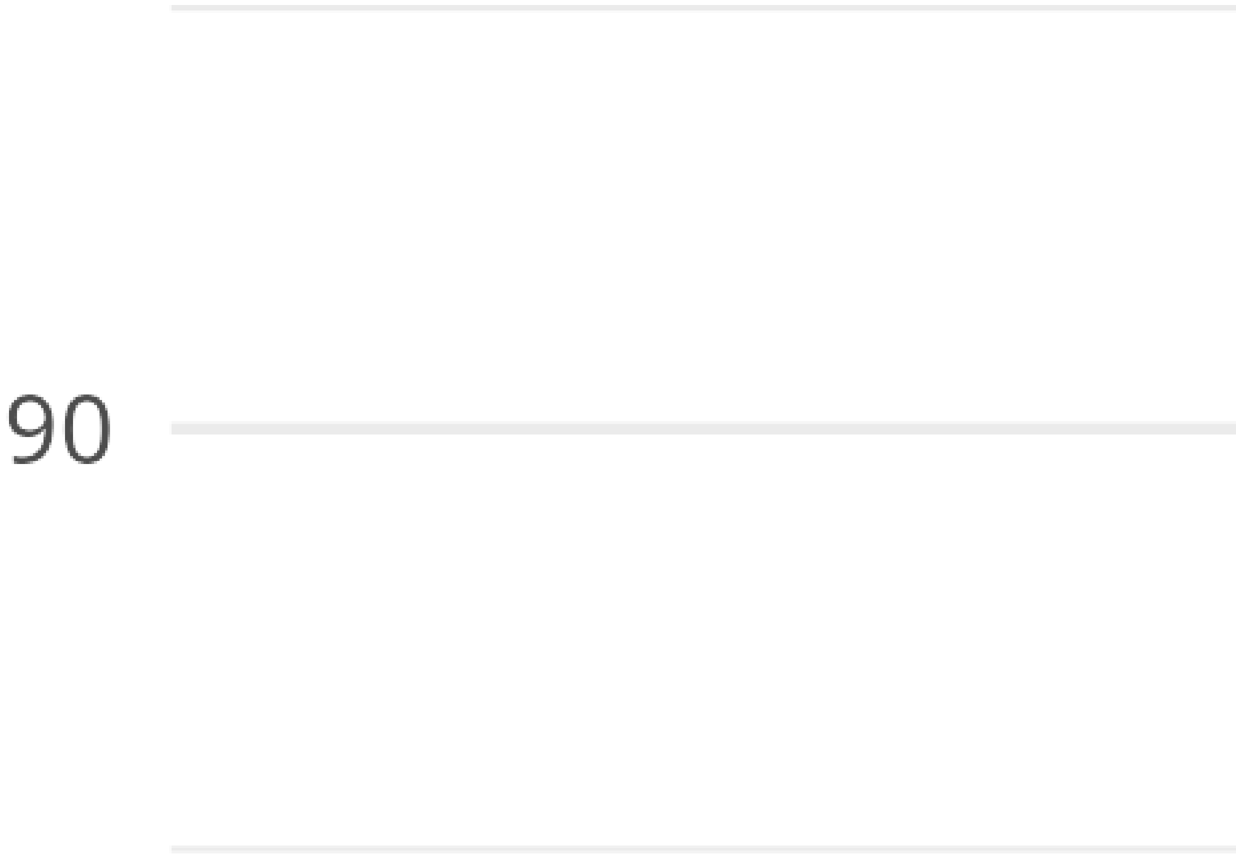
5.4 CART Results

5.4.1 Model Performance

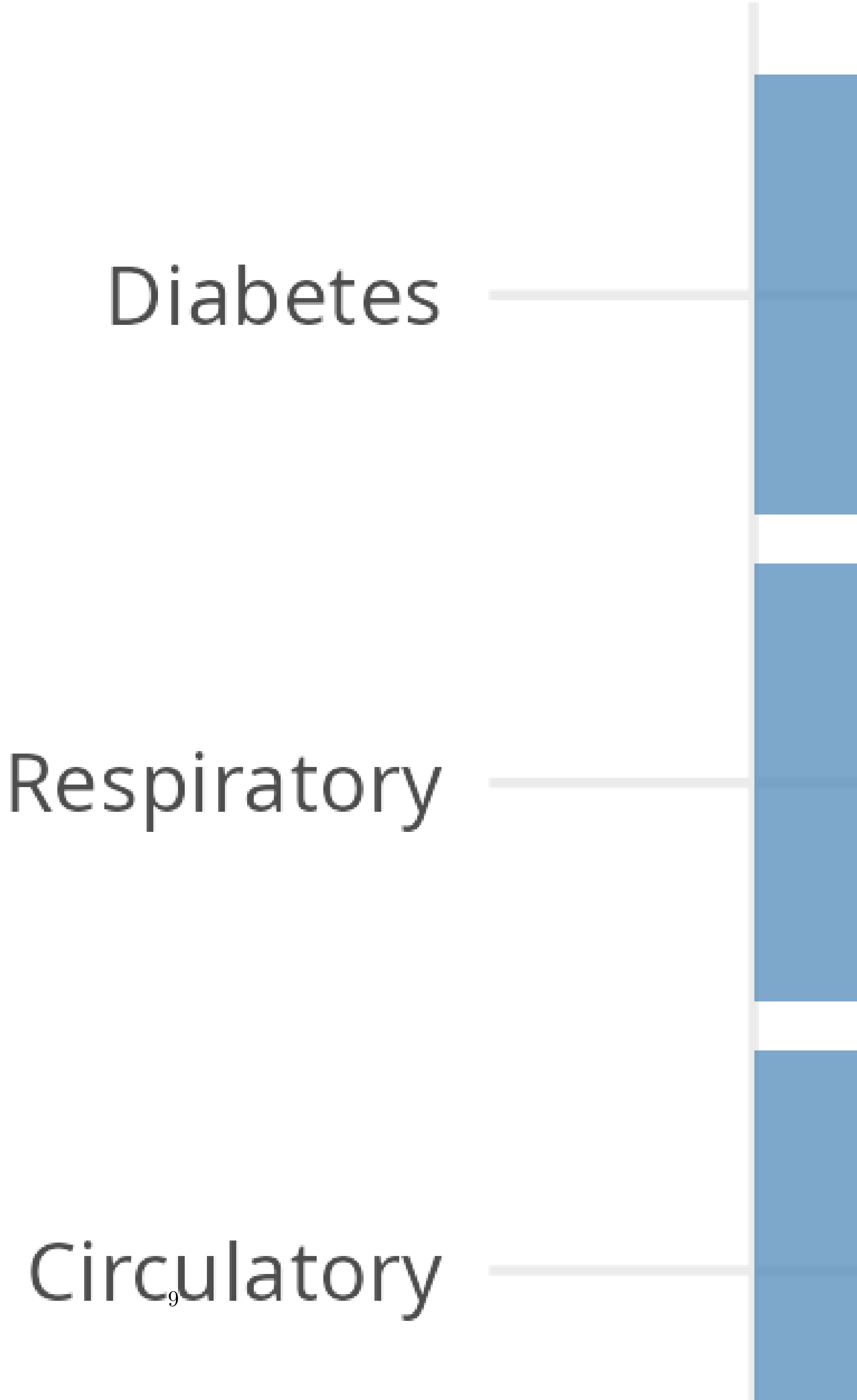
Table 5: CART Model Performance Metrics

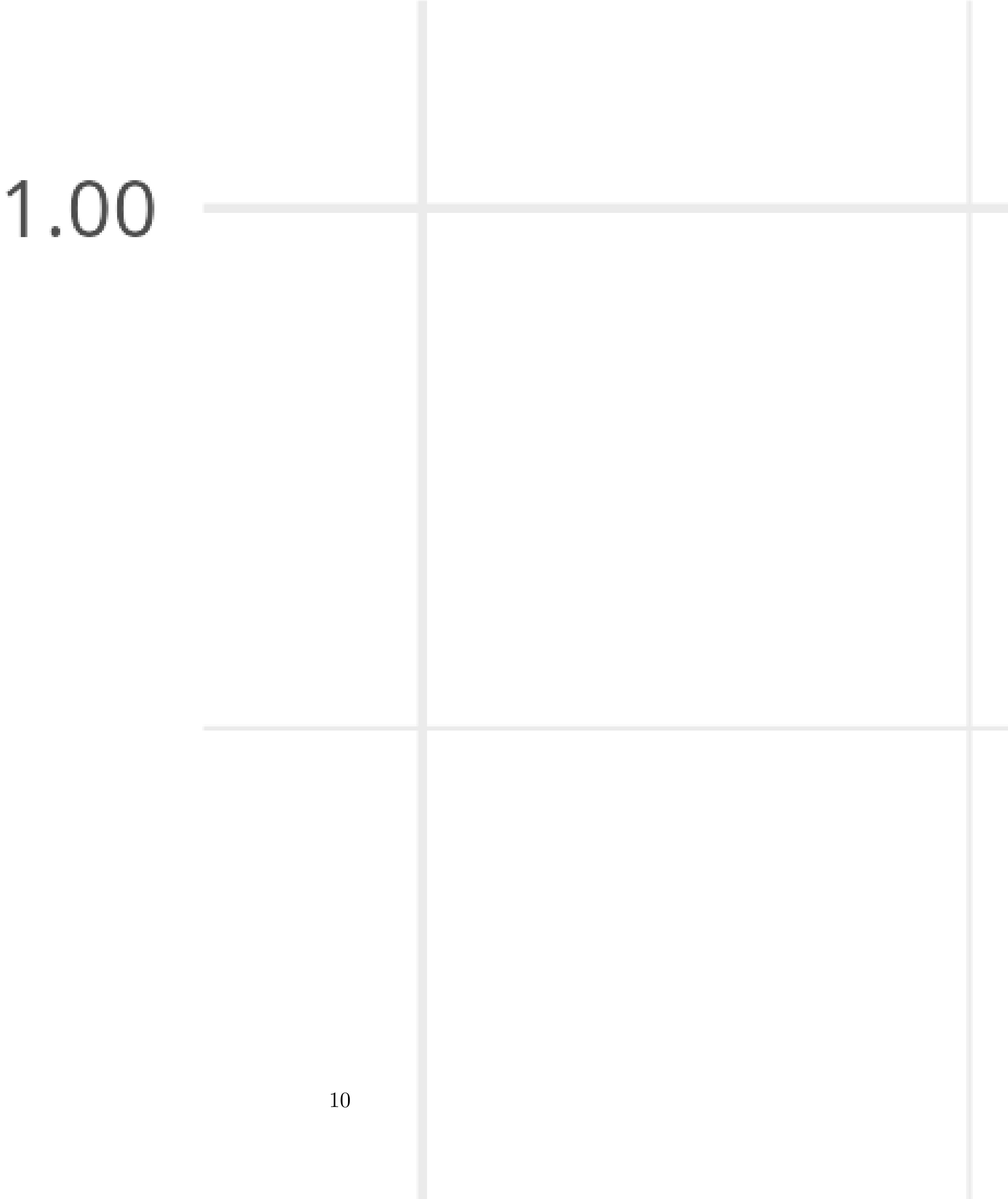
Metric	Value	Percentage
Accuracy	0.6077621	60.77621
Precision	0.5871122	58.71122
Recall (Sensitivity)	0.5582979	55.82979
Specificity	0.6516486	65.16486
F1-Score	0.5723426	57.23426


10000



90


Diabetes
Respiratory
Circulatory



1.00

CART

5.4.2 Decision Tree Visualization

5.4.3 Variable Importance

Table 6: Top 10 Most Important Variables in CART Model

Variable	Importance	Importance_Percent
total_previous_visits	381.9933511	42.3175144
n_inpatient	287.5755829	31.8578421
n_outpatient	139.2253866	15.4234944
n_emergency	93.2171353	10.3266652
time_in_hospital	0.4802531	0.0532028
n_medications	0.1921013	0.0212811

5.4.4 ROC Curve

5.5 Model Comparison

5.5.1 Performance Comparison

Table 7: Side-by-Side Performance Comparison

Metric	Logistic_Regression	CART	Difference
Accuracy	61.84	60.78	1.07
Precision	64.09	58.71	5.38
Recall (Sensitivity)	42.05	55.83	-13.78
Specificity	79.26	65.16	14.10
F1-Score	50.78	57.23	-6.45
AUC	64.81	60.50	4.31

5.5.2 Visualization

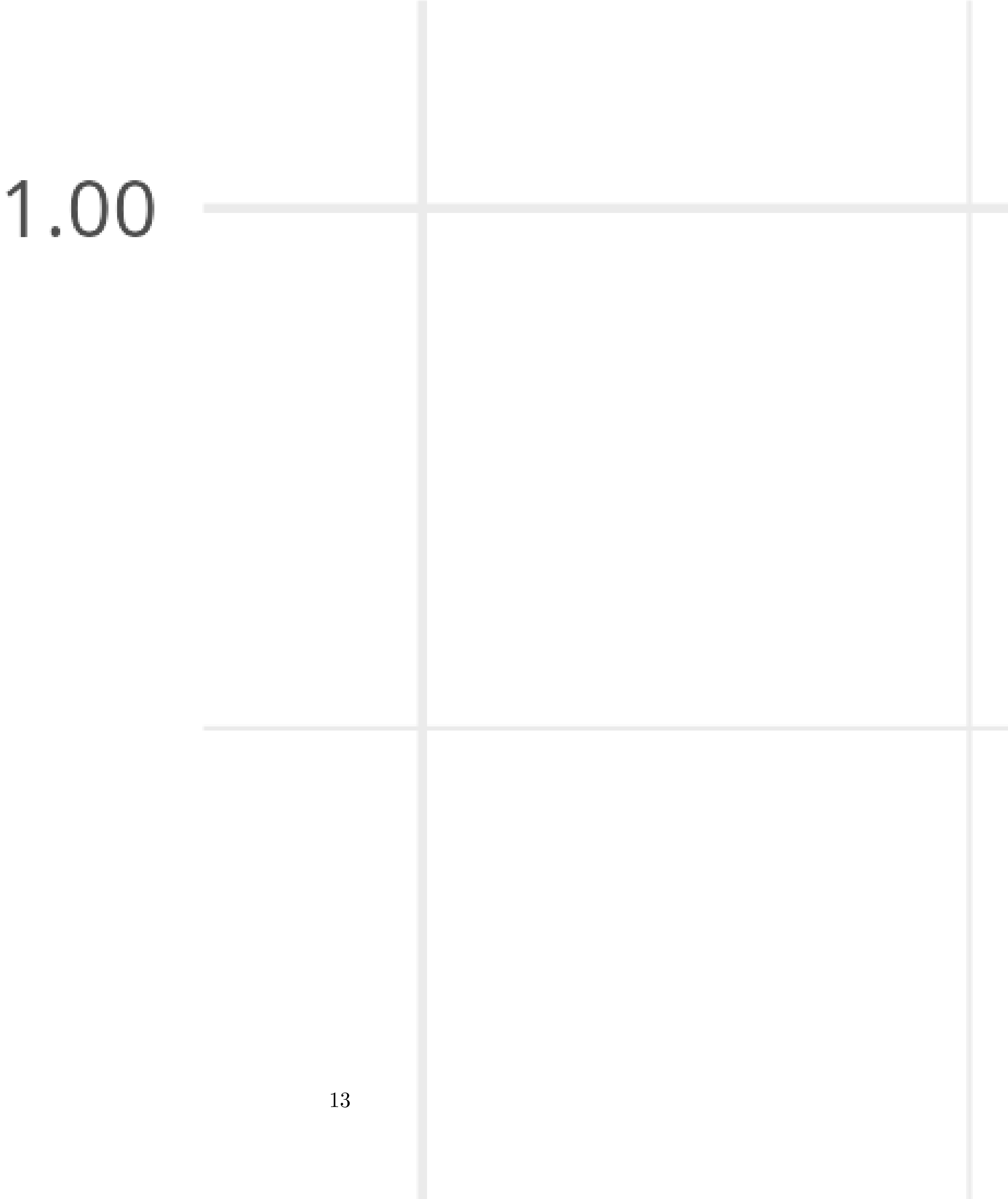
5.5.3 Model Selection

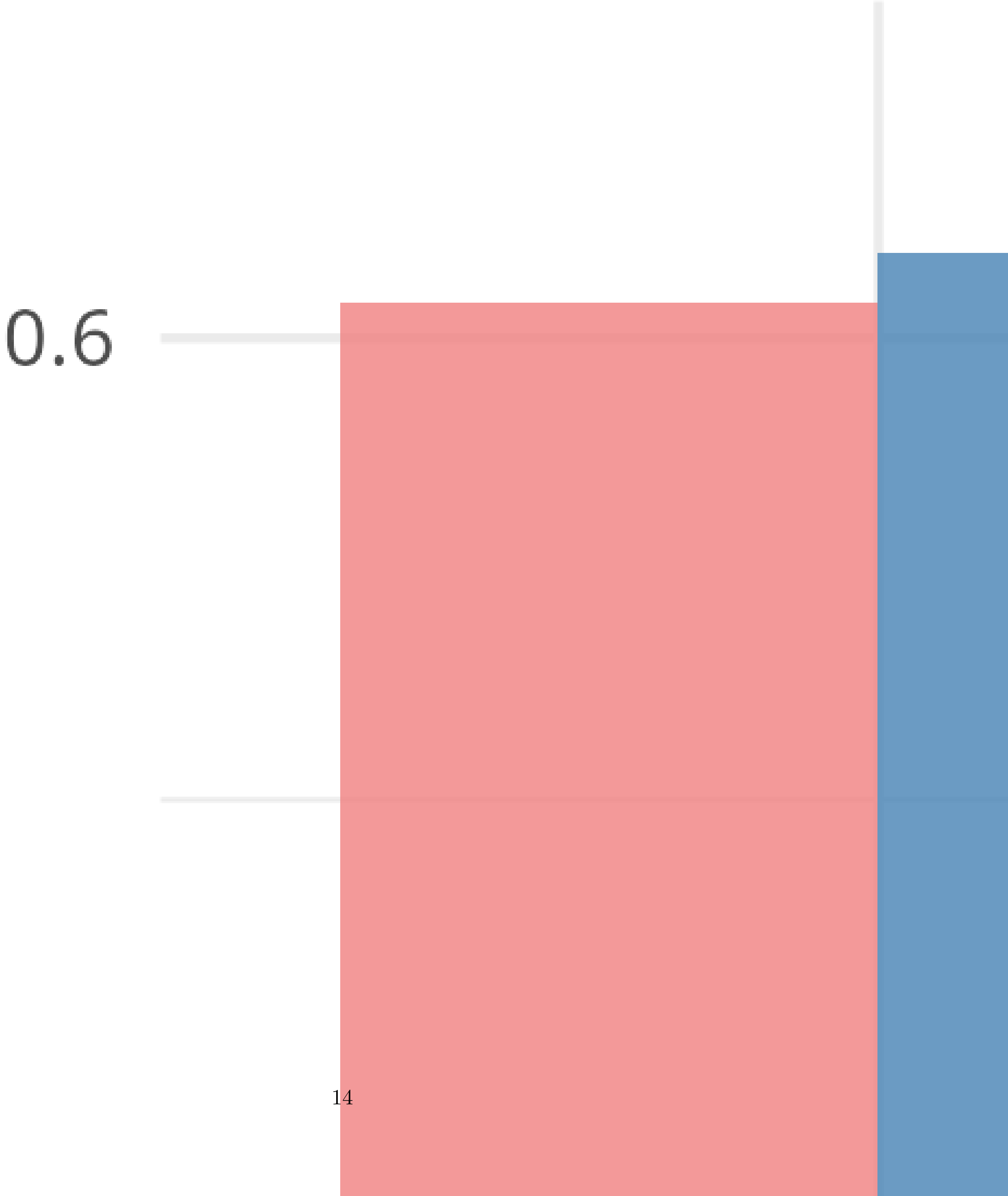
Table 8: Model Comparison: Detailed Justification

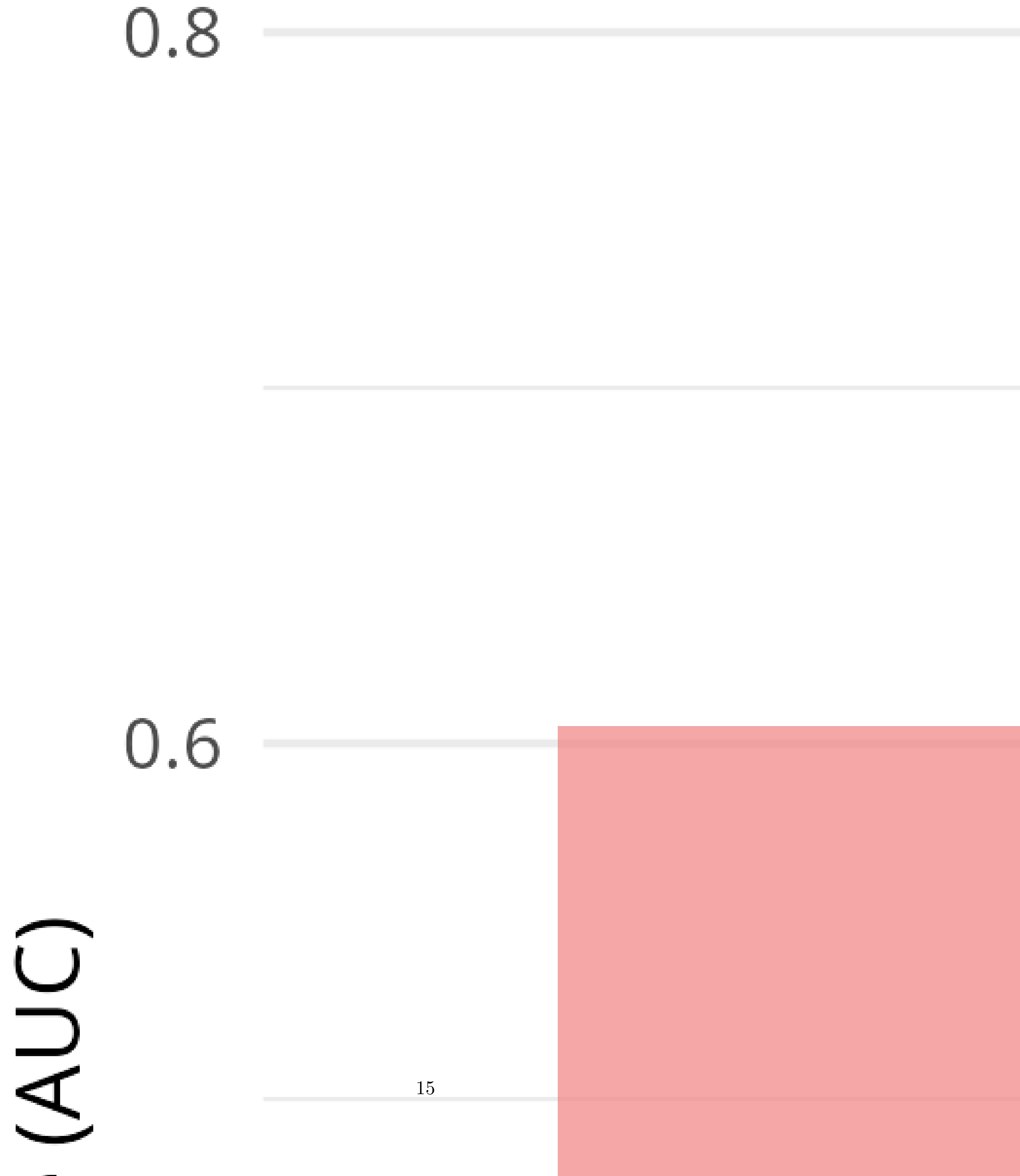
Criterion	Logistic_Regression	CART
Accuracy	61.84%	60.78%
AUC	64.81%	60.5%
Precision	64.09%	58.71%
Recall	42.05%	55.83%
F1-Score	0.5078	0.5723
Interpretability	High (coefficients, odds ratios)	Very High (simple tree, easy rules)
Complexity	High (33 parameters)	Very Low (1 split, 2 nodes)
Statistical Rigor	High (p-values, hypothesis tests)	Medium (no p-values, variable importance)

Recommended Model: Logistic Regression

Reason: Higher AUC (0.648 vs 0.605) and accuracy (61.84% vs 60.78%), with more detailed statistical insights including odds ratios, p-values, and confidence intervals.


1.00


0.6


0.8
0.6
(AUC)

6 Discussion

6.1 Key Findings

1. **Previous visits are the strongest predictor**: Both models identify previous hospital visits (inpatient, outpatient, emergency) as the most important factor in predicting readmission.
2. **Age and medical specialty matter**: Older patients (70-80, 80-90 age groups) and certain medical specialties (Cardiology) show higher readmission rates.
3. **Moderate model performance**: Both models achieve ~61% accuracy with AUC values below 0.7, suggesting room for improvement.
4. **Trade-offs between models**:
   - Logistic Regression provides better performance and statistical rigor
   - CART offers superior simplicity and interpretability

6.2 Clinical Implications

The identification of previous visits as a key predictor suggests that patients with complex medical histories require enhanced discharge planning and follow-up care. The moderate performance of both models indicates that additional clinical variables (e.g., lab results, vital signs, social determinants) may be needed for more accurate predictions.

7 Conclusion

7.1 Answer to Research Question

Yes, patient demographic, diagnostic, and treatment-related features can predict the likelihood of 30-day hospital readmission, though with moderate accuracy (~61%). The Logistic Regression model performs slightly better (AUC = 0.648) and provides more detailed statistical insights, making it preferable for clinical decision support.

7.2 Why This Analysis is Important

1. **Healthcare Cost Reduction**: Early identification of high-risk patients can enable targeted interventions to prevent readmissions
2. **Patient Outcomes**: Improved discharge planning based on risk prediction can enhance patient care
3. **Resource Allocation**: Hospitals can allocate resources more efficiently by focusing on high-risk patients
4. **Clinical Decision Support**: Models provide evidence-based tools for healthcare providers

7.3 Limitations

1. **Moderate Predictive Performance**: Both models show AUC < 0.7, indicating fair to poor discrimination
2. **Missing Variables**: Important clinical variables (lab results, vital signs, comorbidities) may be missing
3. **Data Age**: Data from 1999-2008 may not reflect current healthcare practices

4. **Missing Data**: High percentage of missing medical specialty (49.53%) may affect results
5. **Model Assumptions**: Logistic Regression assumes linear relationships; CART may be underfitting
6. **Generalizability**: Results may not generalize to other hospital systems or time periods

7.4 Recommendations

1. **Feature Enhancement**: Include additional clinical variables (lab results, vital signs, social determinants)
2. **Advanced Methods**: Consider ensemble methods (Random Forest, Gradient Boosting) for improved performance
3. **Data Collection**: Collect more recent data to reflect current healthcare practices
4. **Clinical Application**:
   - Use Logistic Regression for detailed risk assessment with statistical rigor
   - Use CART for simple screening tools requiring high interpretability
5. **Validation**: Validate models on external datasets before clinical deployment

8 References

Kaggle. “Diabetes 130-US Hospitals for 10 years.” Accessed November 14, 2025. https://www.kaggle.com/datasets/brandao/diabetes