Data Mining for Diabetes Readmission Prediction

Team Evolution

Yi Chun Chien, Xiayu Zeng, Hong Zhang, Yixi Zhang

Agenda

- Abstract
- Introduction
- Data Description
- Problem Statement
- Methodology
- Discussions and Results
- Conclusions

Abstract

Background: Alarmingly high risk of readmissions in the US

- Goal: discover factors contributing to hospitals readmissions
- Methods: C5.0 Decision Tree, Quest, Neural Network and Bayesian Network
- Results: emergency readmissions occur most frequently
- Conclusions: effective prediction on readmissions enables hospitals to identify and target patients at the highest risk

Introduction

- Topic: Diabetes Readmission Prediction
- What Is Readmission Rate
  - A hospitalization that occurs within 30 days after a discharge
- Why Is Readmission Important
  - Reduce cost of care and medical disputes
  - Improve patients' safety and health

Data Description

- Data source: The data is from the Center for Clinical and Translational Research, Virginia Commonwealth University

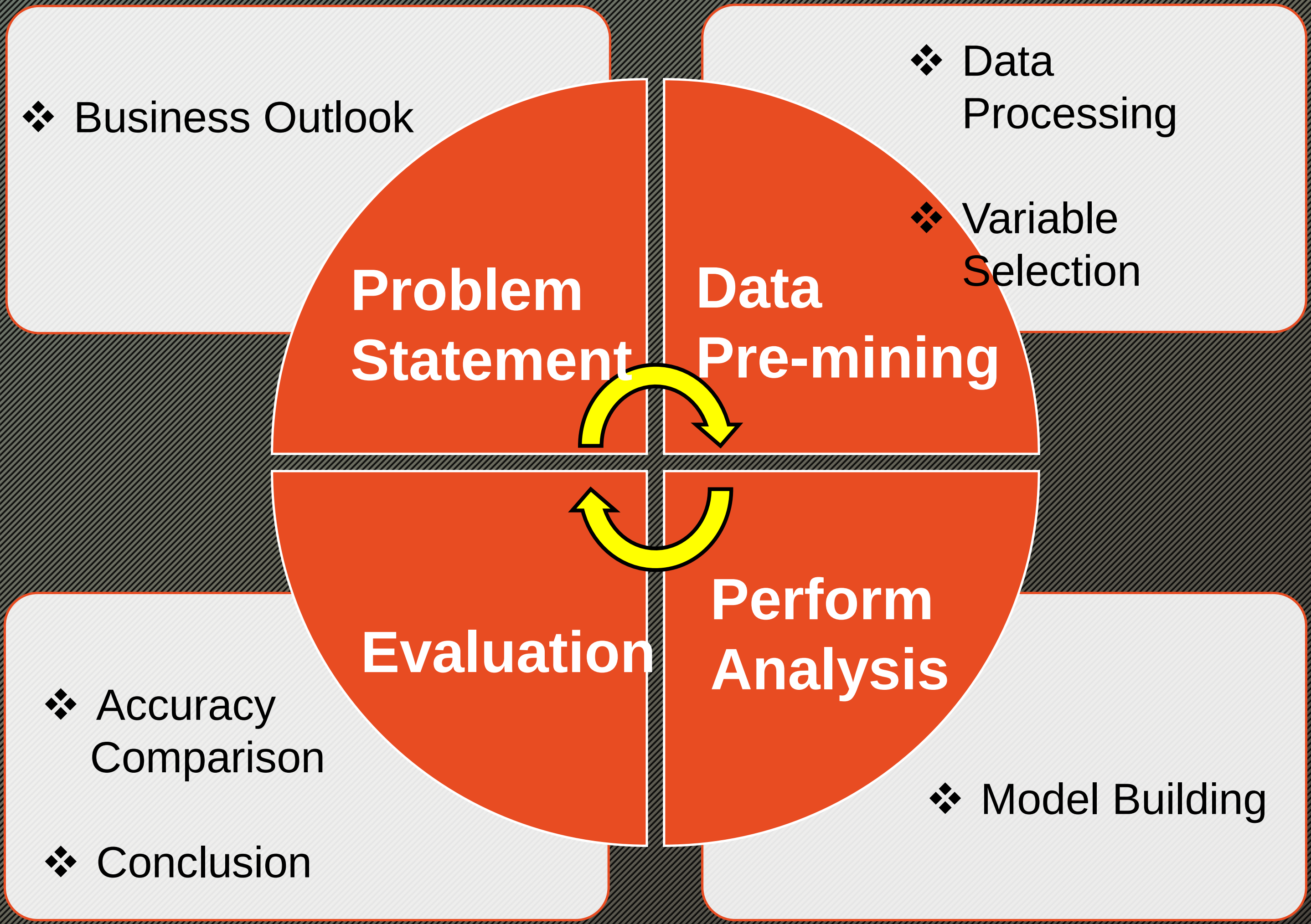
- The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes **101,766** instances and **55** features representing patient and hospital outcomes.

- Link: http://www.cioslab.vcu.edu/index.html

Main attributes :

Attribute	Description
Admission type	patient's admission type: emergency, urgent, elective, newborn, not available, etc.
Time in hospital	Integer number of days between admission and discharge
Number of lab procedures	Number of lab tests performed during the encounter
Number of outpatient visits	Number of outpatient visits of the patient in the year preceding the encounter
Number of emergency visits	Number of emergency visits of the patient in the year preceding the encounter
Number of inpatient visits	Number of inpatient visits of the patient in the year preceding the encounter
Number of diagnoses	Number of diagnoses entered to the system
A1c test result	The range of the result or if the test was not taken
Diabetes medications	Whether there was any diabetic medication prescribed or not
Readmitted	Days to inpatient readmission

Methodology



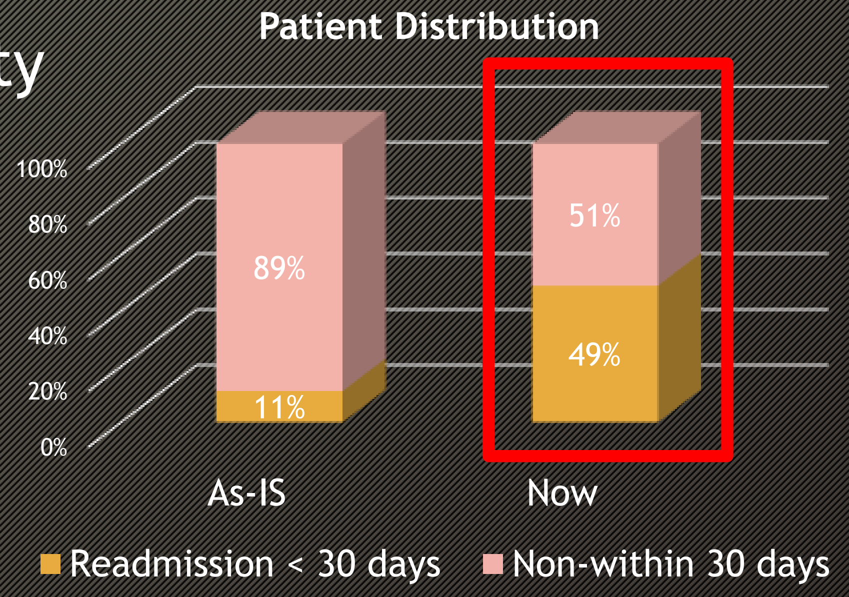
Problem Statement

- Identify the major factors that contribute to hospital readmissions
- Measure the influence of every attribute
- Compare accuracy of each model

Data Pre-mining

- Re-categorize Readmission group from 3 groups (<30, >30 days and No) to 2 groups (Readmission <30 days and Non-within 30 days)

- Review all variables’ distribution to ensure data quality
  - Make Readmission group’s sample ratio be 1:1
  - Review each variable’s relationship with Readmission
  - Review numeric variables’ correlation

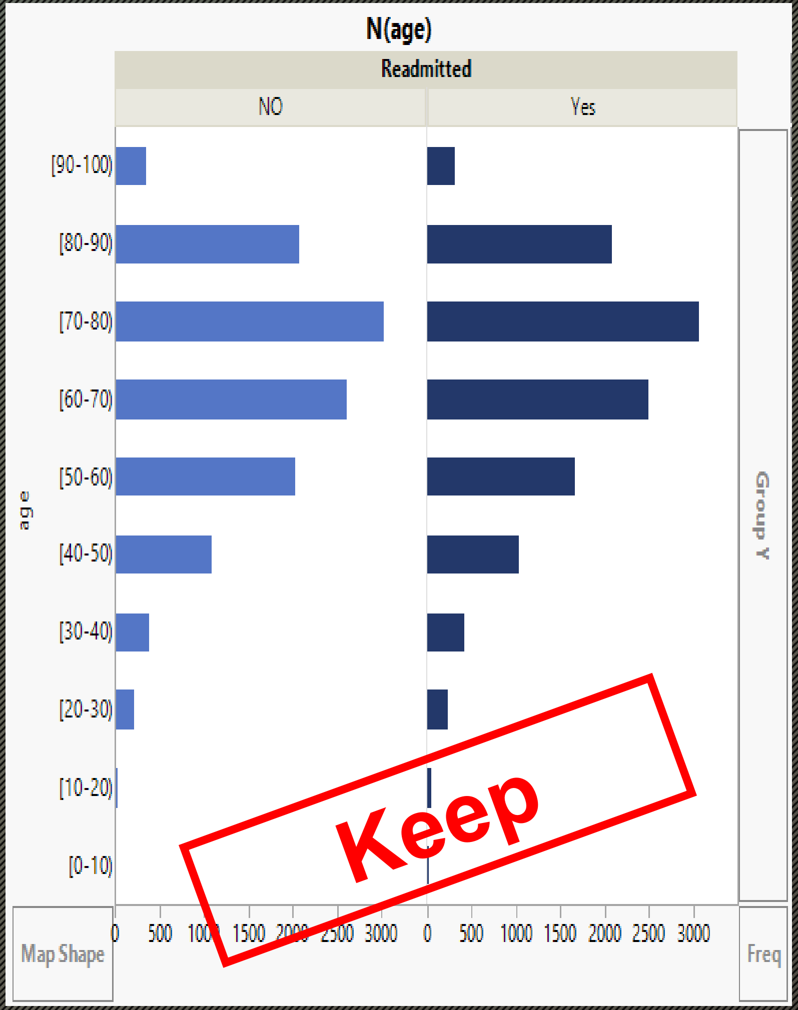


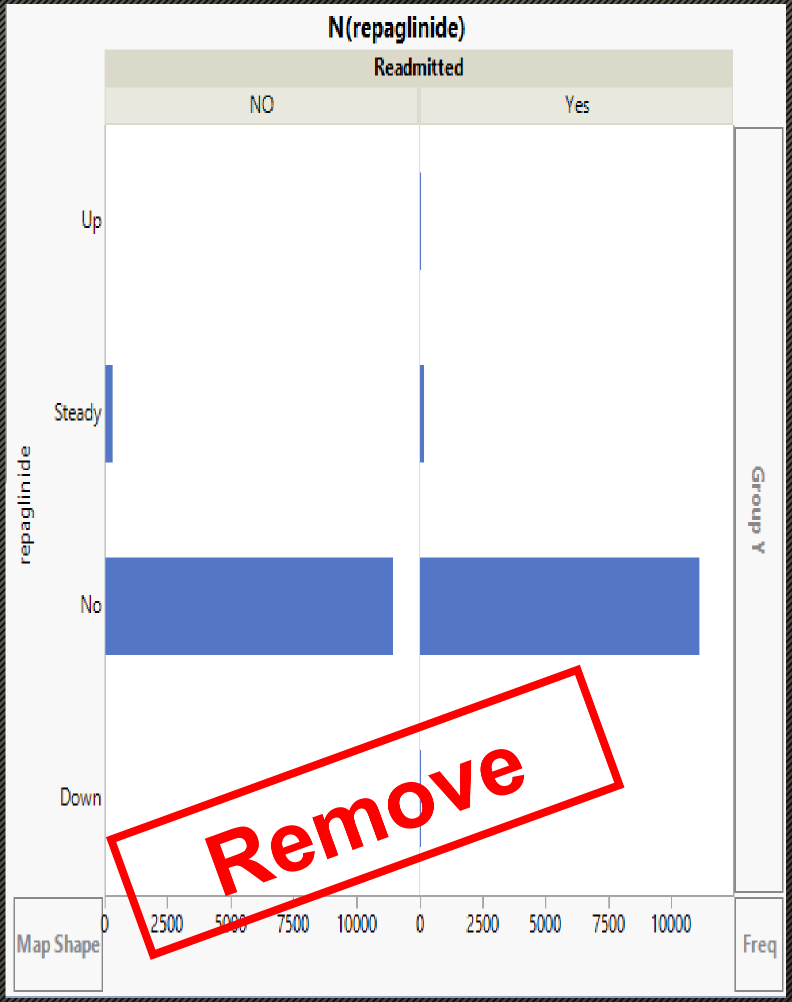
Variable's relationship with Readmission

- Remove irrelevant variables

  (EX: patient ID, payer code)
- Delete unknown value
- Eliminate variables only

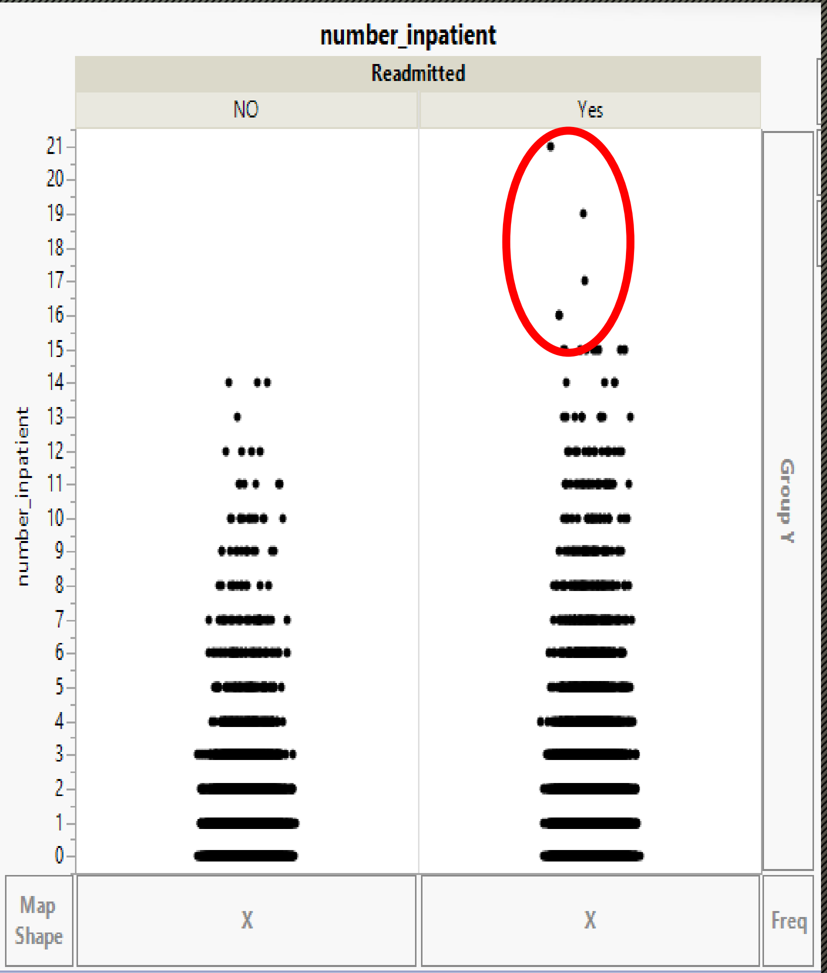
  has majority value

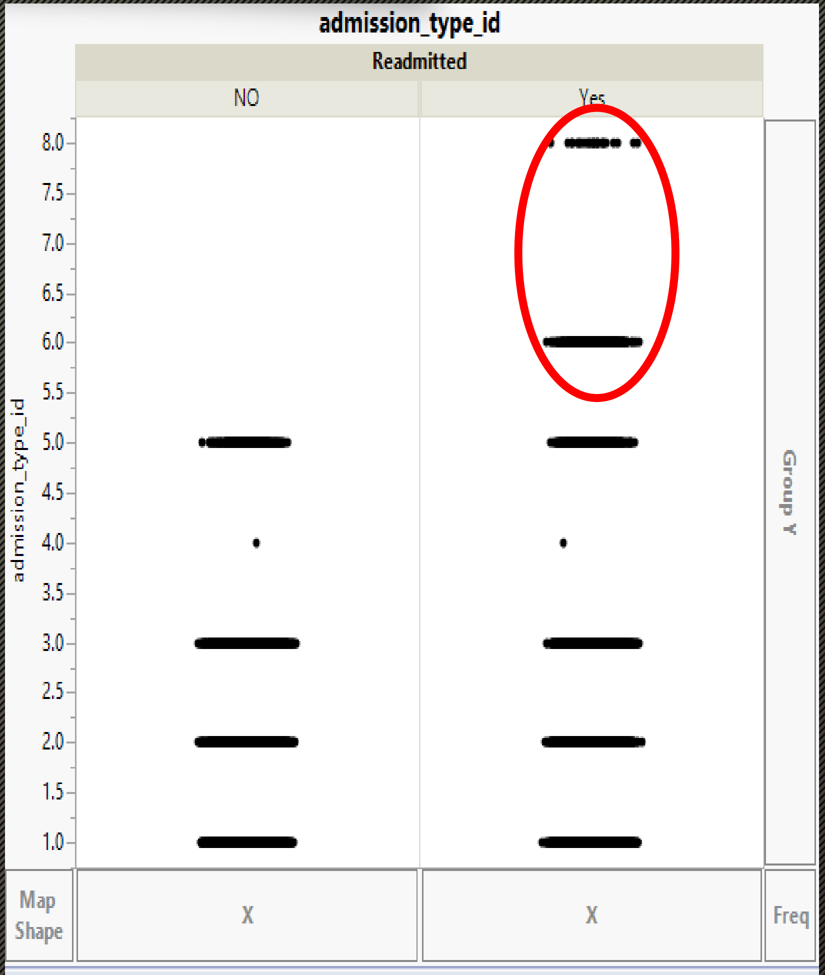




Variable's relationship with Readmission

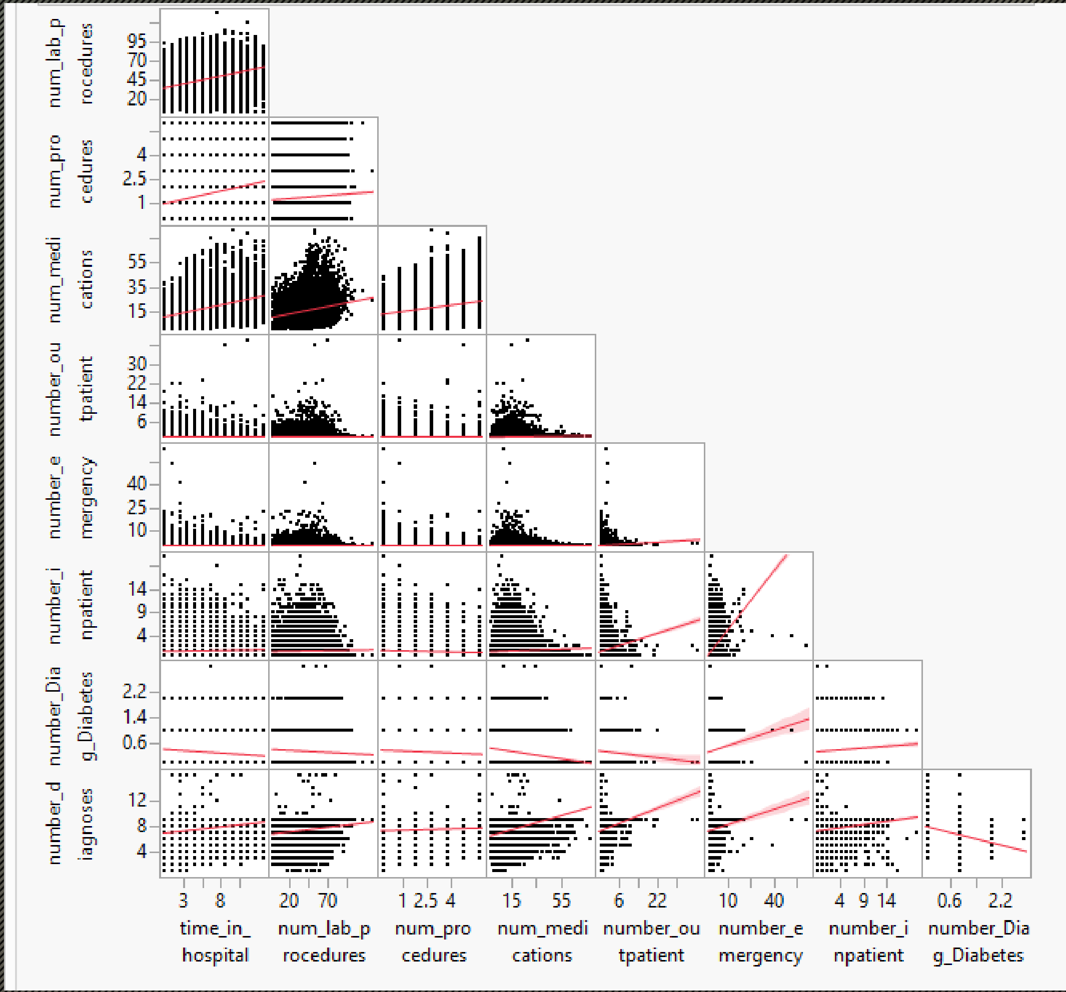
- Found some variables have different behavior between readmission group
  - Diagnoses Number
  - Admission Type ID
  - Inpatient Number





Review numeric variables' correlation

- No correlation between numeric variables' correlation
- Variables are independent



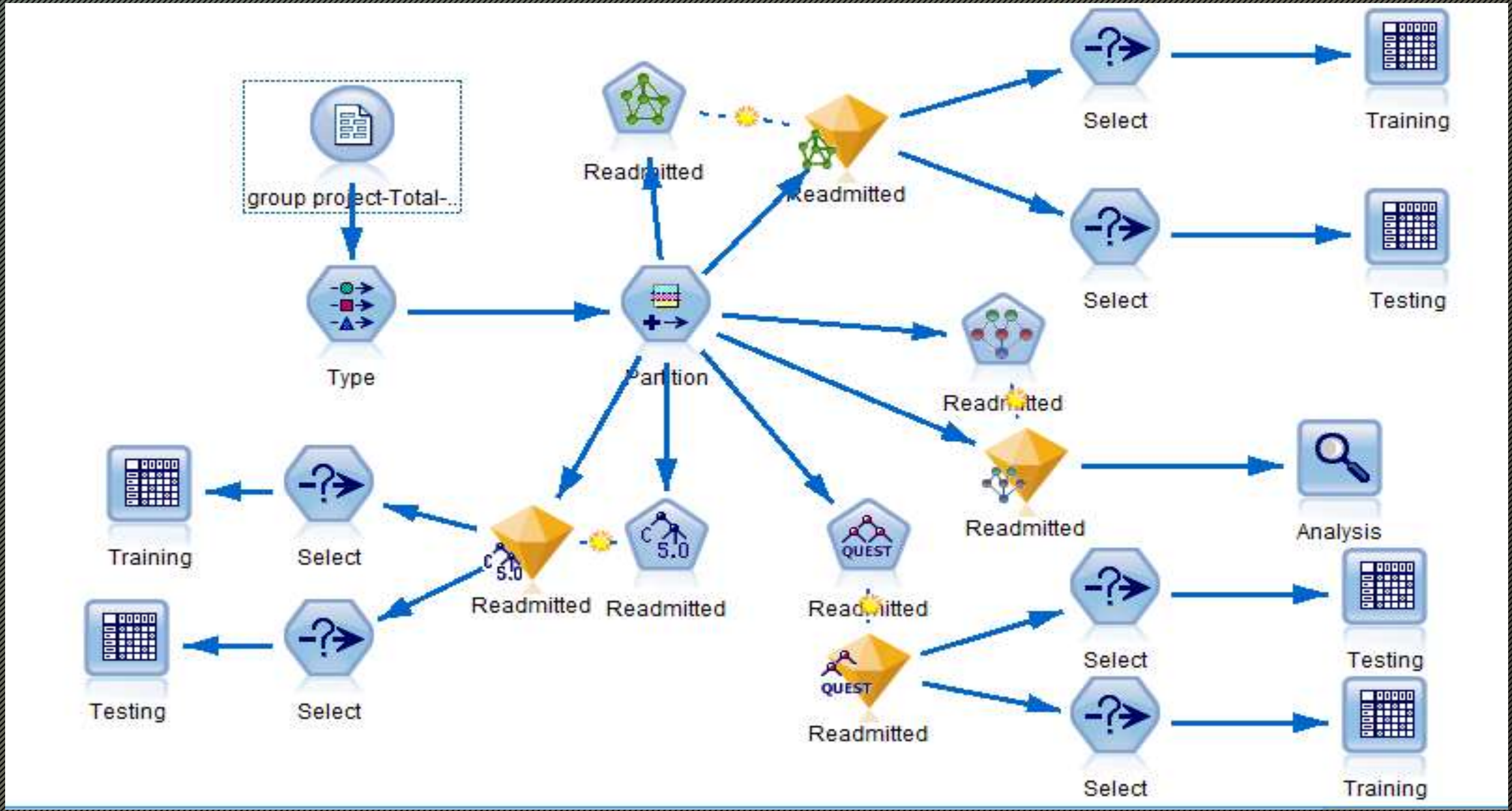
Perform Analysis

- Input variables: 14; Output variable: Readmission group
- Total sample size: 23,154
- Partition: 70% on Training; 30% on Testing
- Build models from (1) Decision Tree Analysis: C5.0 & QUEST

  (2) Apply Neural Network Methodology

  (3) Perform Bayes' Analysis

Model Stream



Neural Network - Comparison

Hidden layer = 5

Readmitted		$null$	NO	Yes
NO	Count	0	5909	2317
	Row %	0.000	71.833	28.167
Yes	Count	1	2698	5285
	Row %	0.013	33.793	66.195

$N-Readmitted

Readmitted		$null$	NO	Yes
NO	Count	2	2551	1018
	Row %	0.056	71.437	28.507
Yes	Count	1	1140	2232
	Row %	0.030	33.798	66.173

Keep

Observed	Predicted NO	Predicted Yes
NO	71.8%	28.2%
Yes	33.8%	66.2%

Row Percent
100.00
80.00
60.00
40.00
20.00
0.00

Hidden layer = 8

Readmitted		$null$	NO	Yes
NO	Count	0	5584	2642
	Row %	0.000	67.882	32.118
Yes	Count	1	2831	5152
	Row %	0.013	35.458	64.529

Readmitted		$null$	NO	Yes
NO	Count	2	2413	1156
	Row %	0.056	67.572	32.372
Yes	Count	1	1209	2163
	Row %	0.030	35.843	64.127

Remove

Observed	Predicted NO	Predicted Yes
NO	67.9%	32.1%
Yes	35.5%	64.5%

Row Percent
100.00
80.00
60.00
40.00
20.00
0.00

Neural Network

- Model Performance:
  - 66% accuracy in testing dataset

- **Predictor Importance:**
  Top 5:

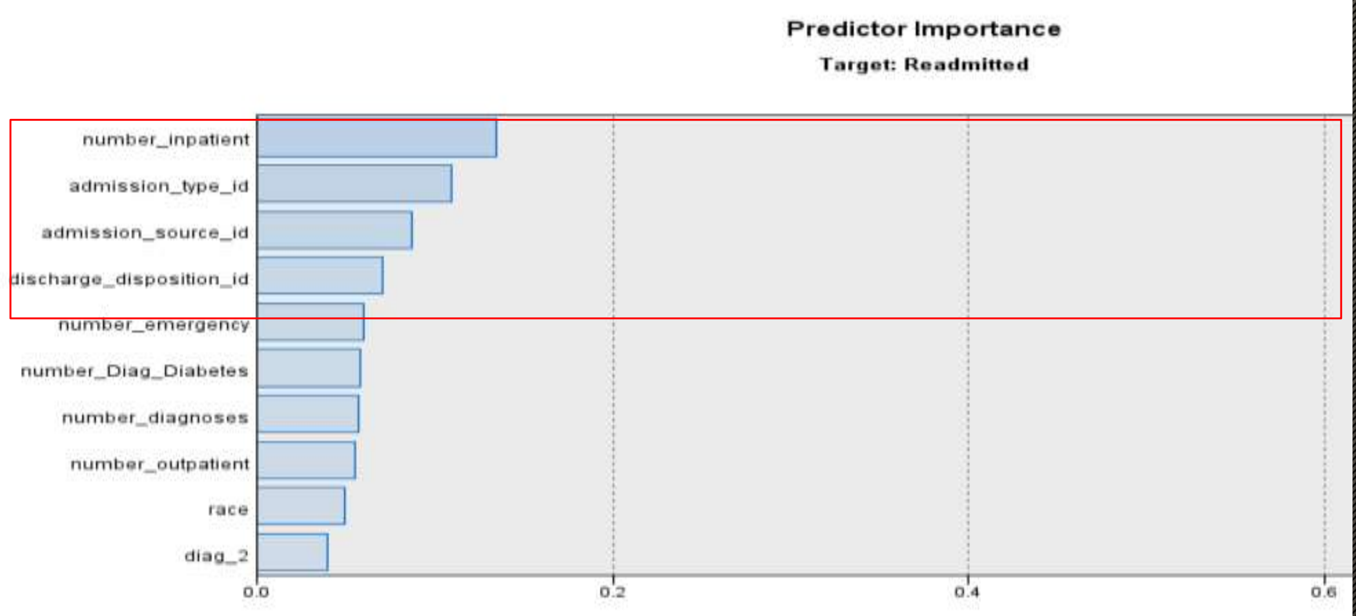
- Number of inpatient visits
- Admission type
- Admission source
- Discharge disposition
- Number of emergency visits

Training data

		$N-Readmitted		
Readmitted		$null$	NO	Yes
NO	Count	0	5909	2317
	Row %	0.000	71.833	28.167
Yes	Count	1	2698	5285
	Row %	0.013	33.793	66.195

Testing data

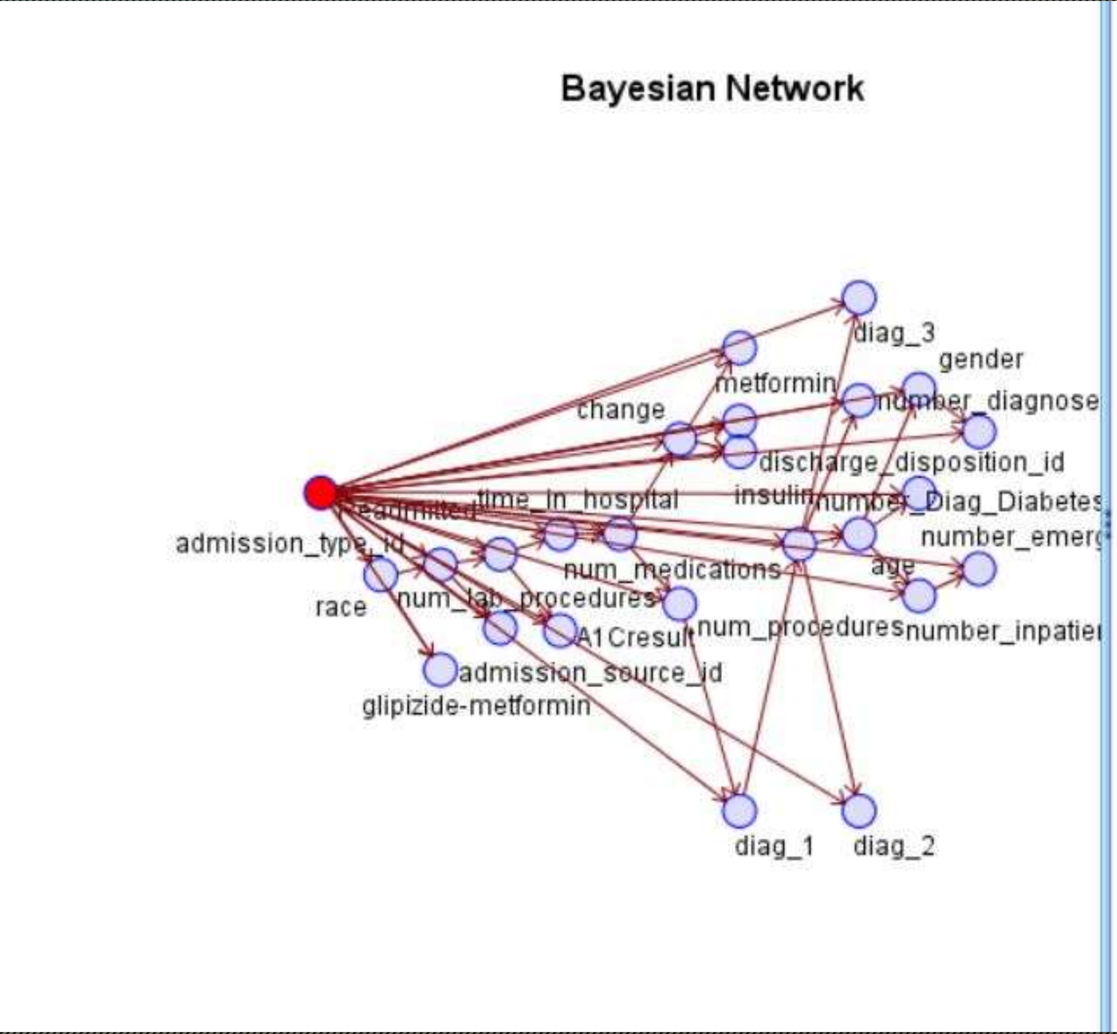
		$N-Readmitted		
Readmitted		$null$	NO	Yes
NO	Count	2	2551	1018
	Row %	0.056	71.437	28.507
Yes	Count	1	1140	2232
	Row %	0.030	33.798	66.173

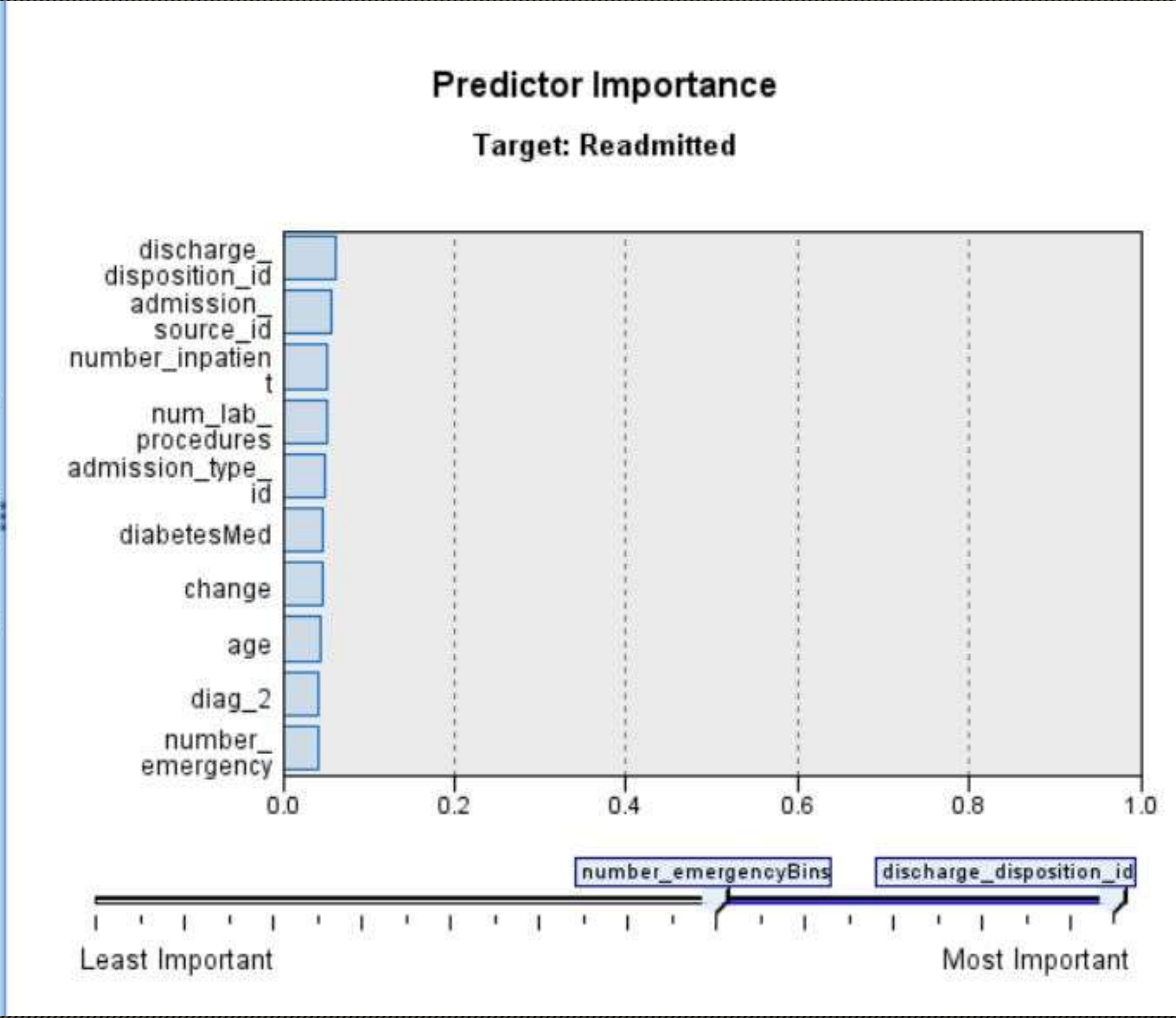


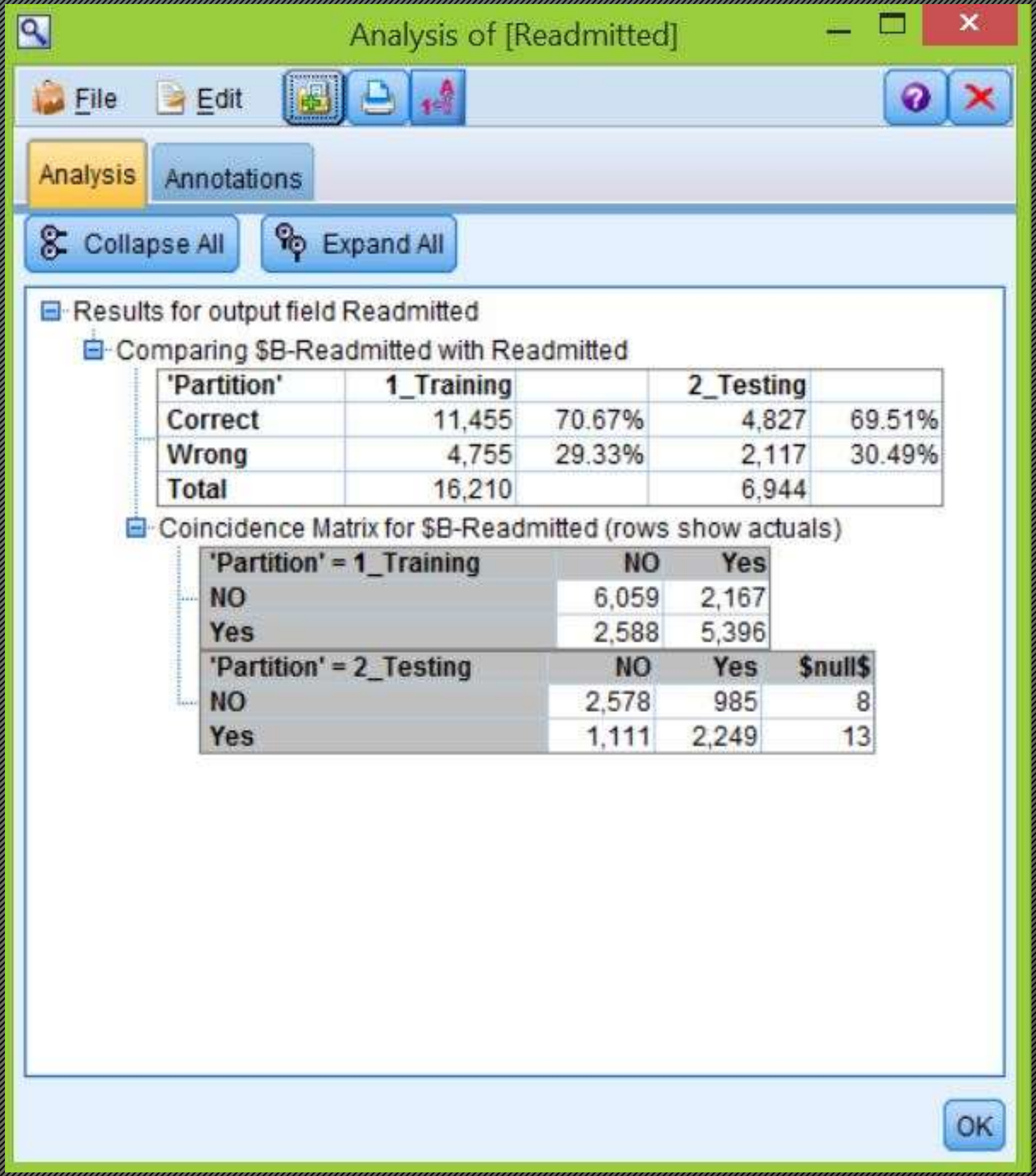
Why Use Bayesian Network

- Have probability for reference
- Allow variables’ dependency
- Well suited for categorical variables

Interpreting the Results







'Partition'	1_Training		2_Testing	
Correct	11,455	70.67%	4,827	69.51%
Wrong	4,755	29.33%	2,117	30.49%
Total	16,210		6,944	

'Partition' = 1_Training	NO	Yes
NO	6,059	2,167
Yes	2,588	5,396

'Partition' = 2_Testing	NO	Yes	$null$
NO	2,578	985	8
Yes	1,111	2,249	13

- Overall Model:

Training accuracy: 70.67%
Testing accuracy: 69.51%

- Model on “Yes” results:

Training accuracy: 71.35%
Testing accuracy: 69.54%

Decision Tree: C5.0

- Model Performance:
  - 65% accuracy in testing dataset

Training

File Edit Generate

Matrix Appearance Annotations

Readmitted		$C-Readmitted NO	Yes
NO	Count	6905	1321
	Row %	83.941	16.059
Yes	Count	1907	6077
	Row %	23.885	76.115

Testing

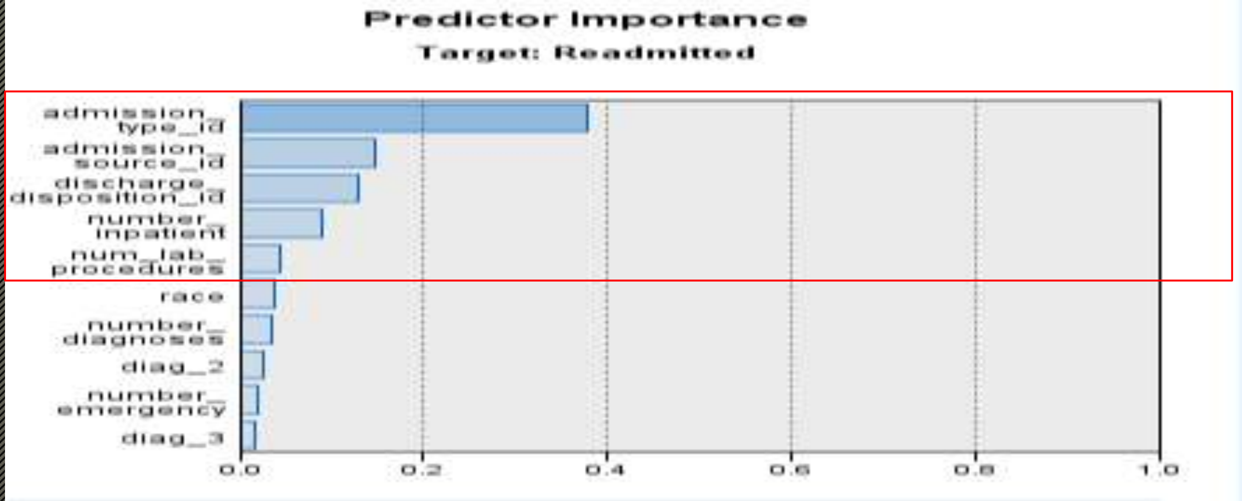
File Edit Generate

Matrix Appearance Annotations

Readmitted		$C-Readmitted NO	Yes
NO	Count	2594	977
	Row %	72.641	27.359
Yes	Count	1167	2206
	Row %	34.598	65.402

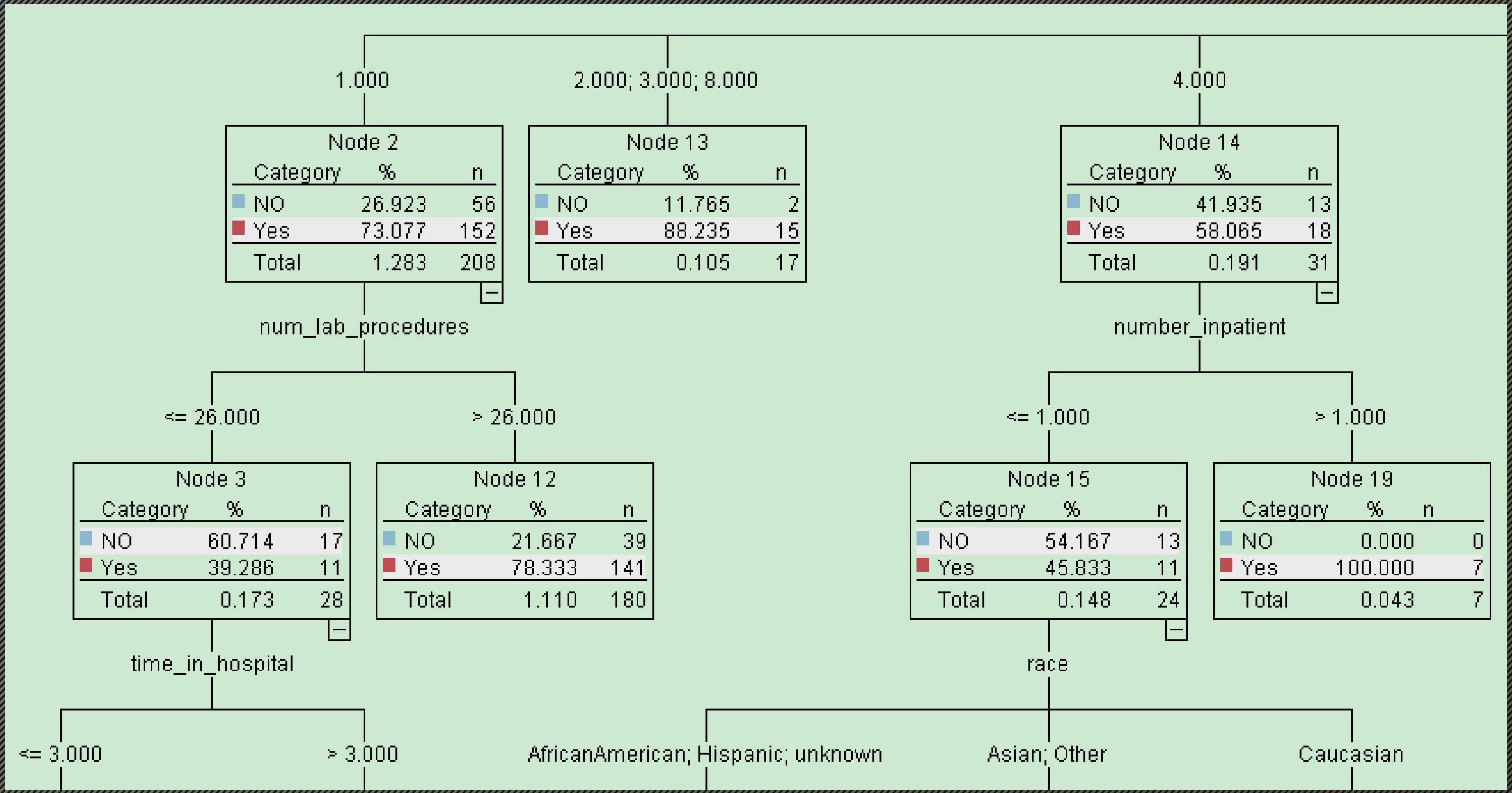
- Predictor Importance:

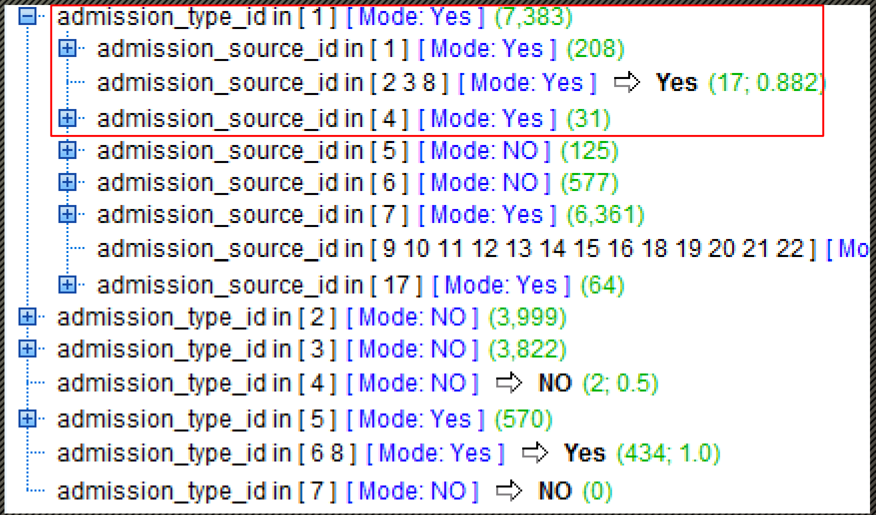
  Top 5:
  - Admission type
  - Admission source
  - Discharge disposition
  - Number of inpatient visits
  - Number of lab procedures



Decision Tree: C5.0

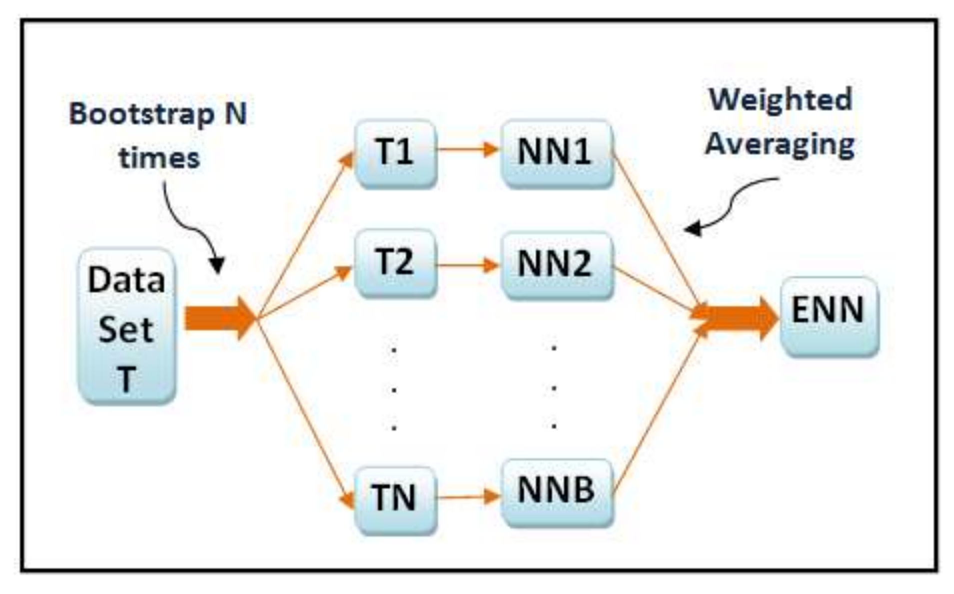
Decision tree (part)





Decision Tree: QUEST

- What's QUEST: *Quick, Unbiased and Efficient Statistical Tree*
- Advantage:
  - Easily handle categorical predictor variables with many categories
  - Use imputation to deal with missing values
  - It provides linear splits using Fisher's LDA method
- How to improve accuracy
  - Boosting: Randomly re-sample N dataset -> create N trees -> optimal by weighted average
  - Bagging: optimal by aggregated average



Decision Tree: QUEST

- Adopt Boosting method in QUEST
- Model Performance:
  - 72% accuracy in testing dataset

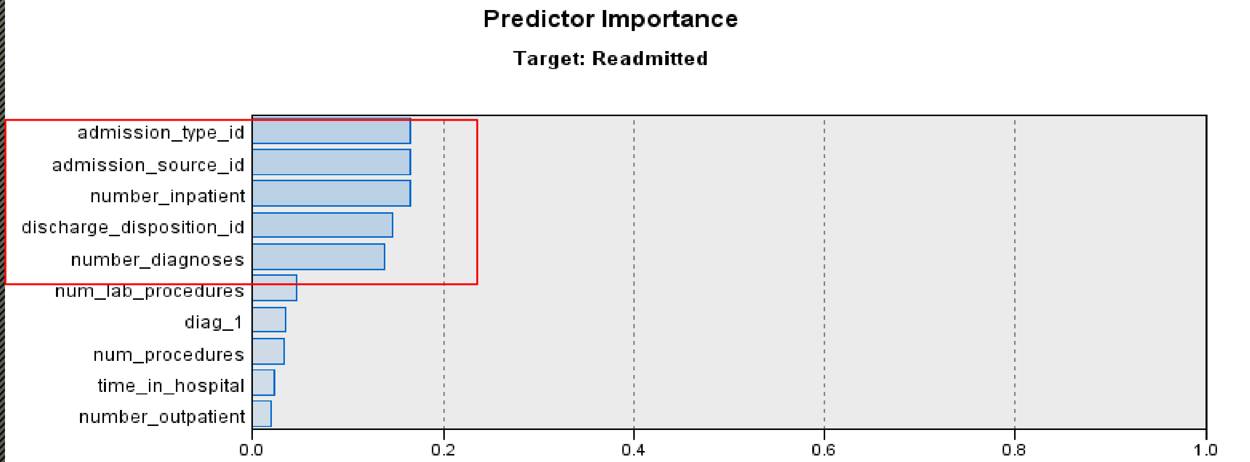
Training data

$R-Readmitted			
Readmitted		NO	Yes
NO	Count	5196	3030
	Row %	63.166	36.834
Yes	Count	2169	5815
	Row %	27.167	72.833

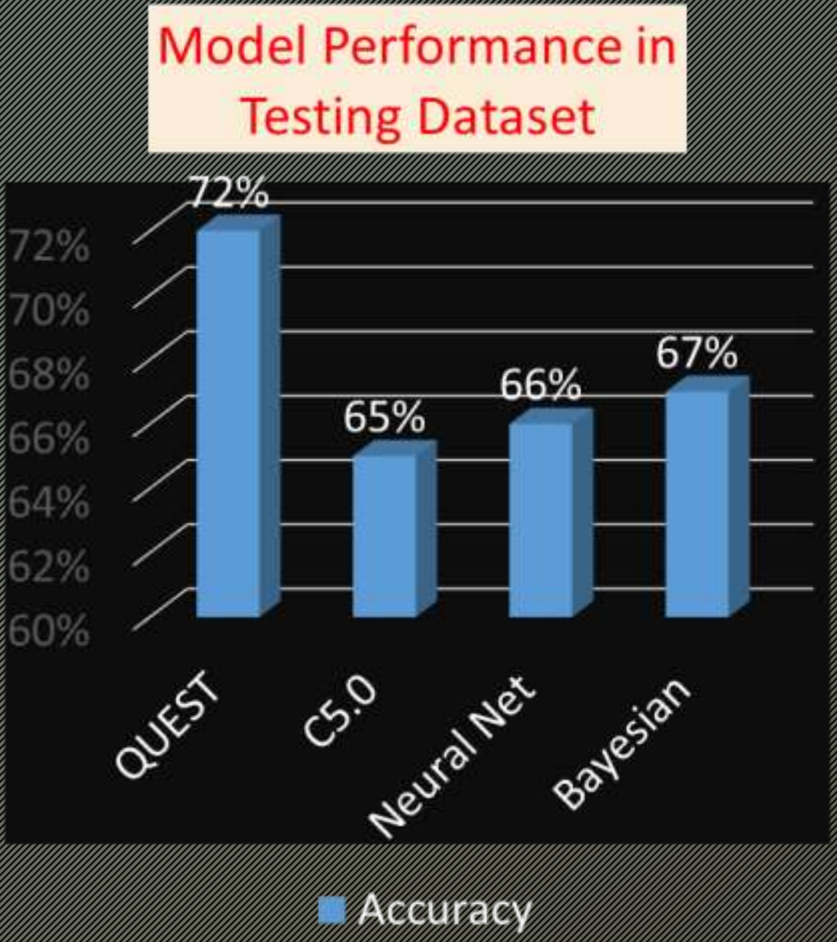
Testing data

$R-Readmitted			
Readmitted		NO	Yes
NO	Count	2258	1313
	Row %	63.232	36.768
Yes	Count	932	2441
	Row %	27.631	72.369

- Variable Importance:
  - Top 5: Admission Type,
    Admission Source,
    Number of inpatient visits
    Discharge Deposition,
    Number of diagnoses



Model Evaluation

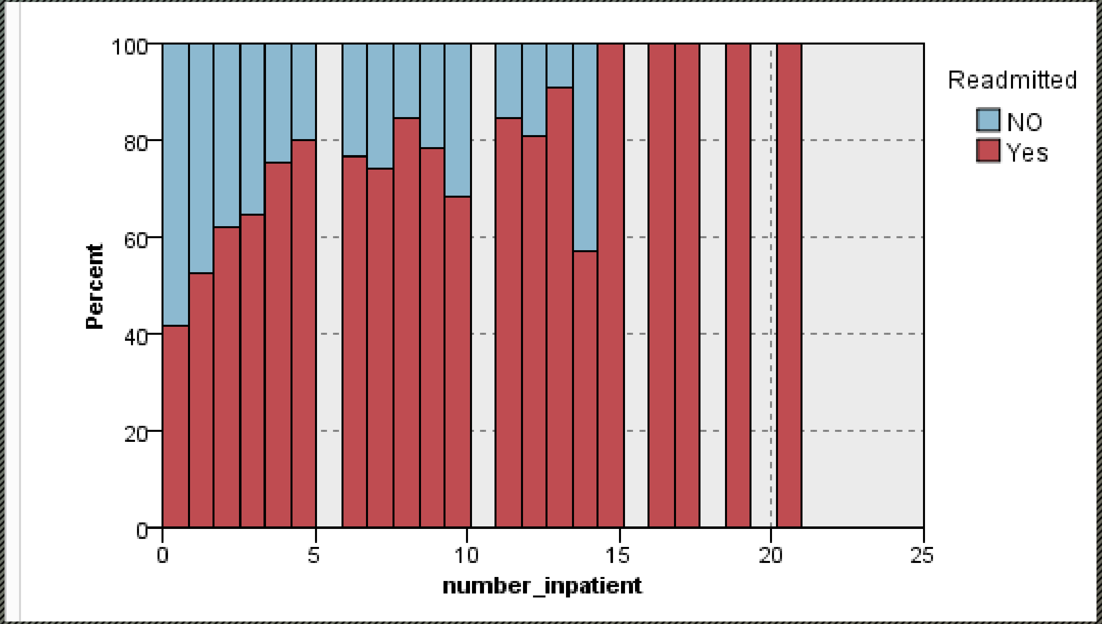


Top Important Variables:

QUEST	C5.0	Neural Net	Bayesian
Admission Type	Admission Type	Inpatient #	Discharge Disposition
Admission Source	Admission Source	Admission Type	Admission Source
Inpatient #	Discharge Disposition	Admission Source	Inpatient #
Discharge Disposition	Inpatient #	Discharge Disposition	Lab Procedures #
Diagonose #	Lab Procedures #	Emergency #	Admission Type

- QUEST has the best accuracy
- Top 4 variables: Admission Type, Admission Source, number of inpatient visits and Discharge disposition.

Evaluation of top 4 important attributes



NUMBER OF INPATIENT VISITS

(Number of inpatient visits of the patient in the year preceding the encounter)

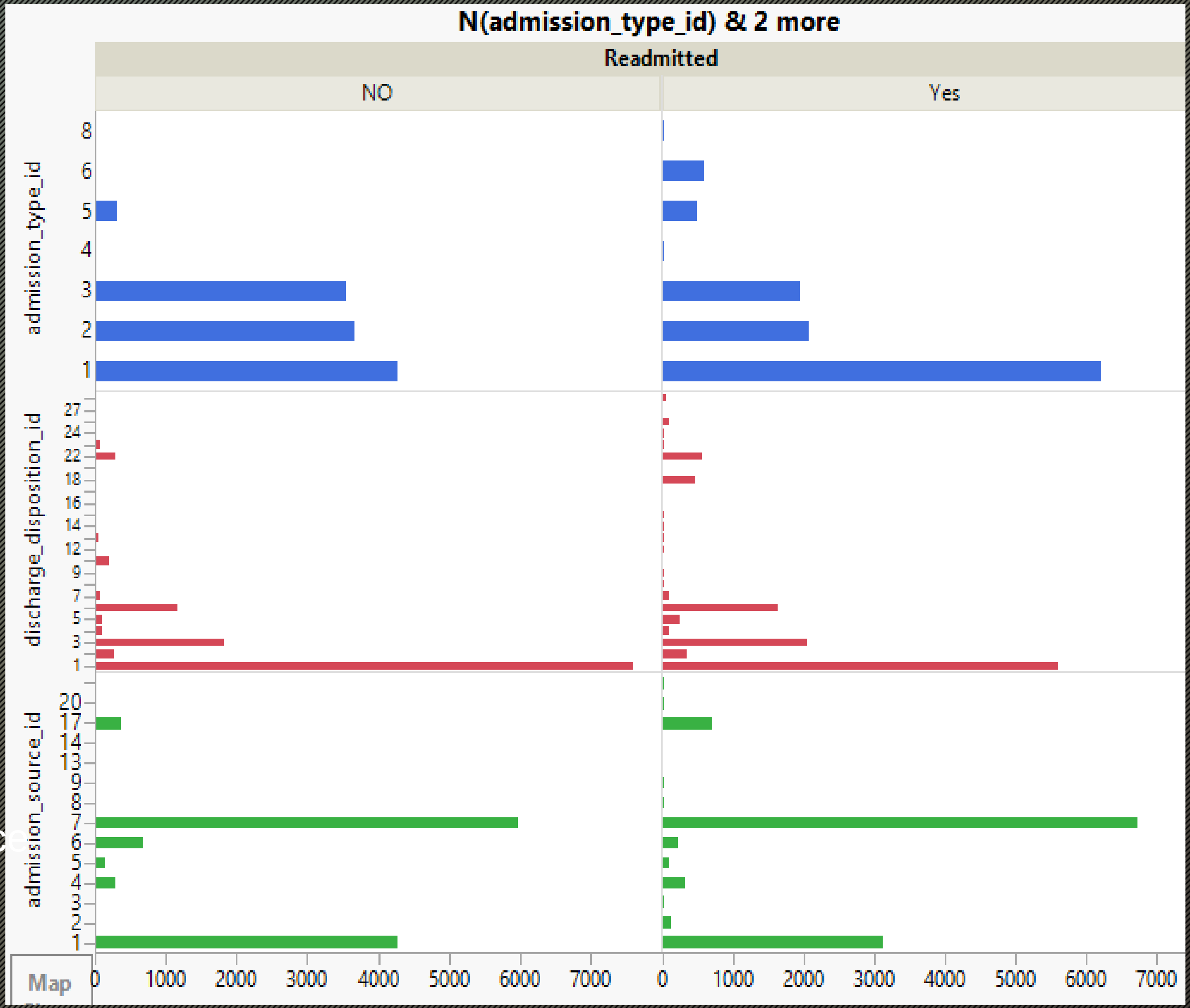
From the graph, it's shown that number of inpatient visits potentially increases the risk of readmission. For patients who stayed in a hospital over 15 times in a year, they would be 100% readmitted to the hospital within 30 days. Inpatient treatment facilities should better provide additional and more specialized medical care to reduce their readmission rate.

ADMISSION TYPE
(1.Emergency, 2.Urgent, 3.Elective, 4.Newborn, 5.Not Available, 6.NULL, 7.Trauma Center, 8.Not Mapped)

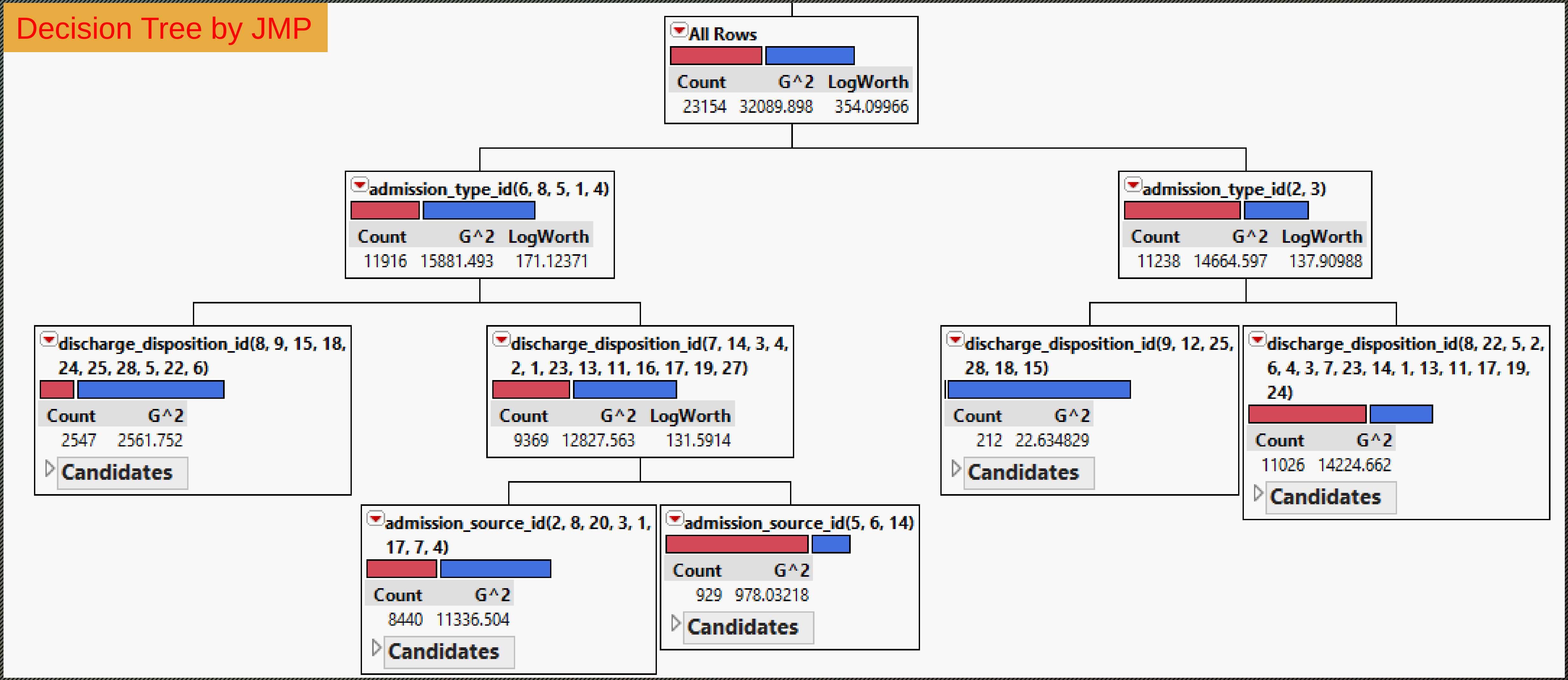
ADMISSION SOURCE

(physician referral, emergency room, and transfer from a hospital, etc.)

DISCHARGE DISPOSITION
(discharged to home, expired, and Hospice / home, etc.)



Decision tree(part)



Conclusion

- Data pre-mining is of upmost importance in improving the model accuracy (1%→72%);
- The readmission groups are related to admission source,admission type, discharge disposition and number of inpatient visits;
- The readmission groups do not solely depend on any single variable, but the interactions of related variables;
- Instead of tracking all 55 attributes, hospitals are suggested to focus on number of patient's inpatient visits, admission source, admission type, discharge disposition;
- Hospitals are advised to concern not only inpatient treatment but also continuing care after discharge.