12monkeys

Synthetic Health Data Hackathon 2020

REGION Rigshospitalet

Digital Hub Denmark

Erhvervsministeriet

biolib

Team

Kevin Albert

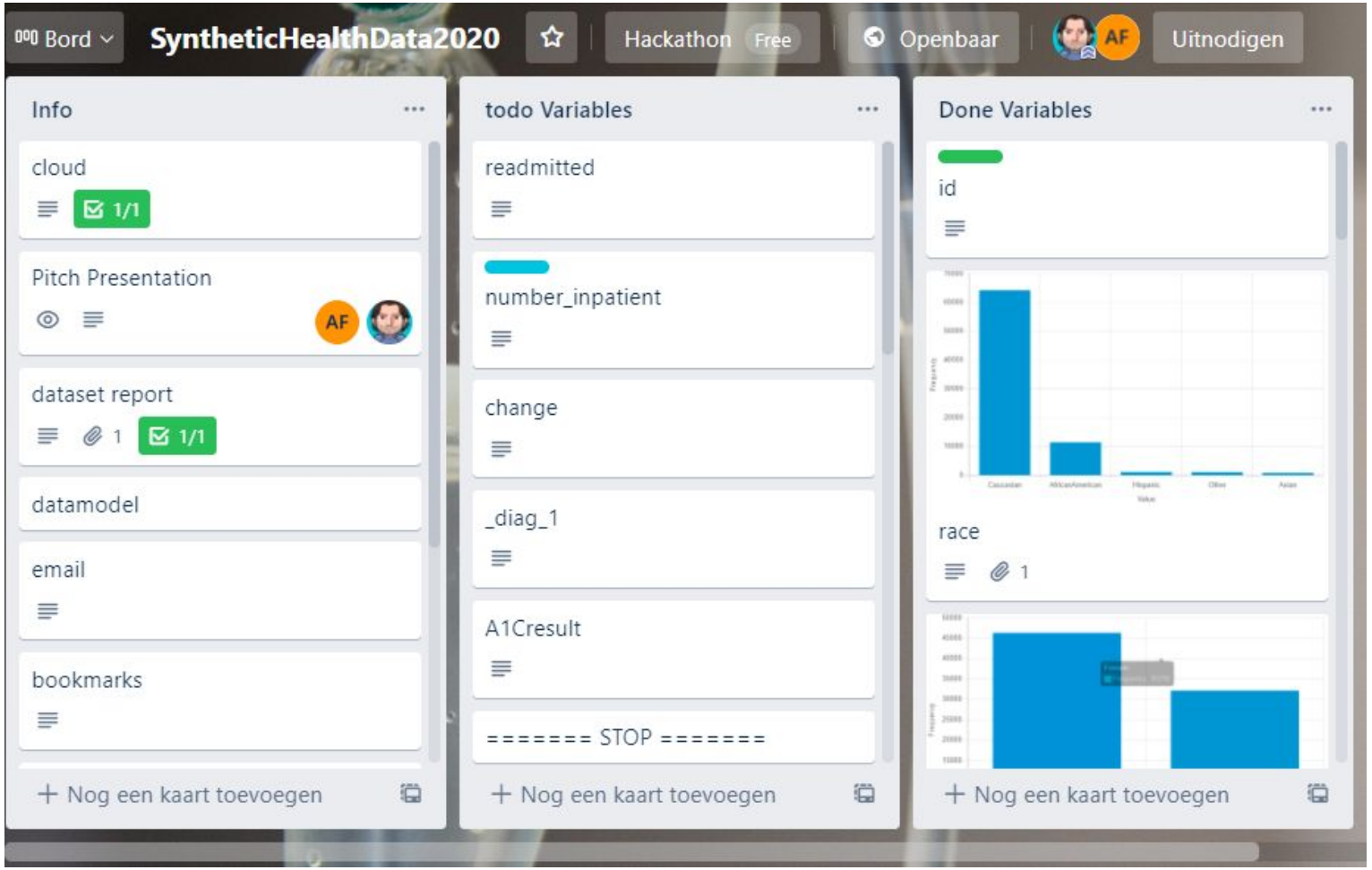
Andre Fontes

Track

- Diabetes - Bioinformatics - Data Set
- Diabetes - Machine Learning - Data Set

Tools

Status page

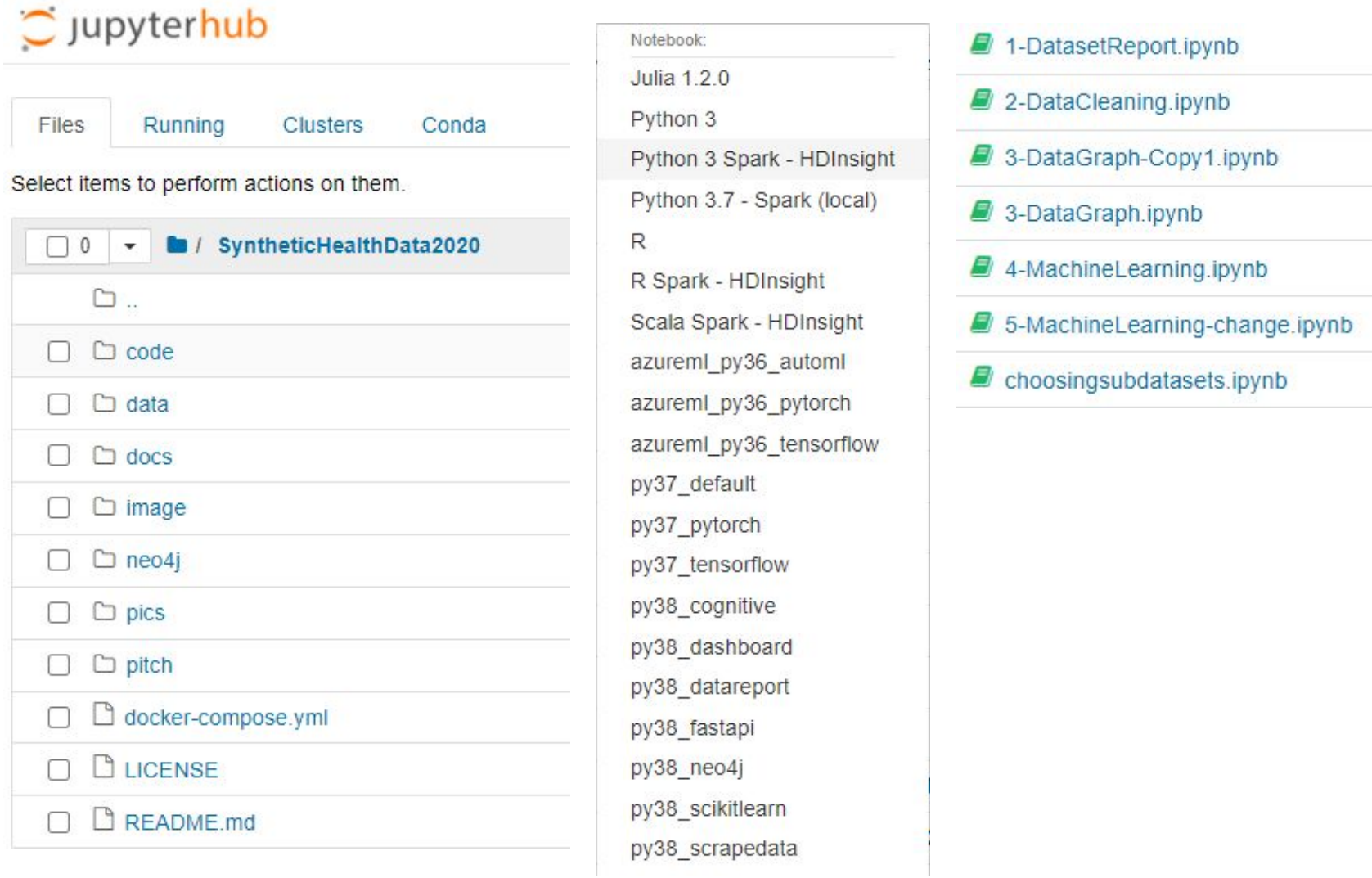


trello.com/b/hqd6UMGH

Cloud

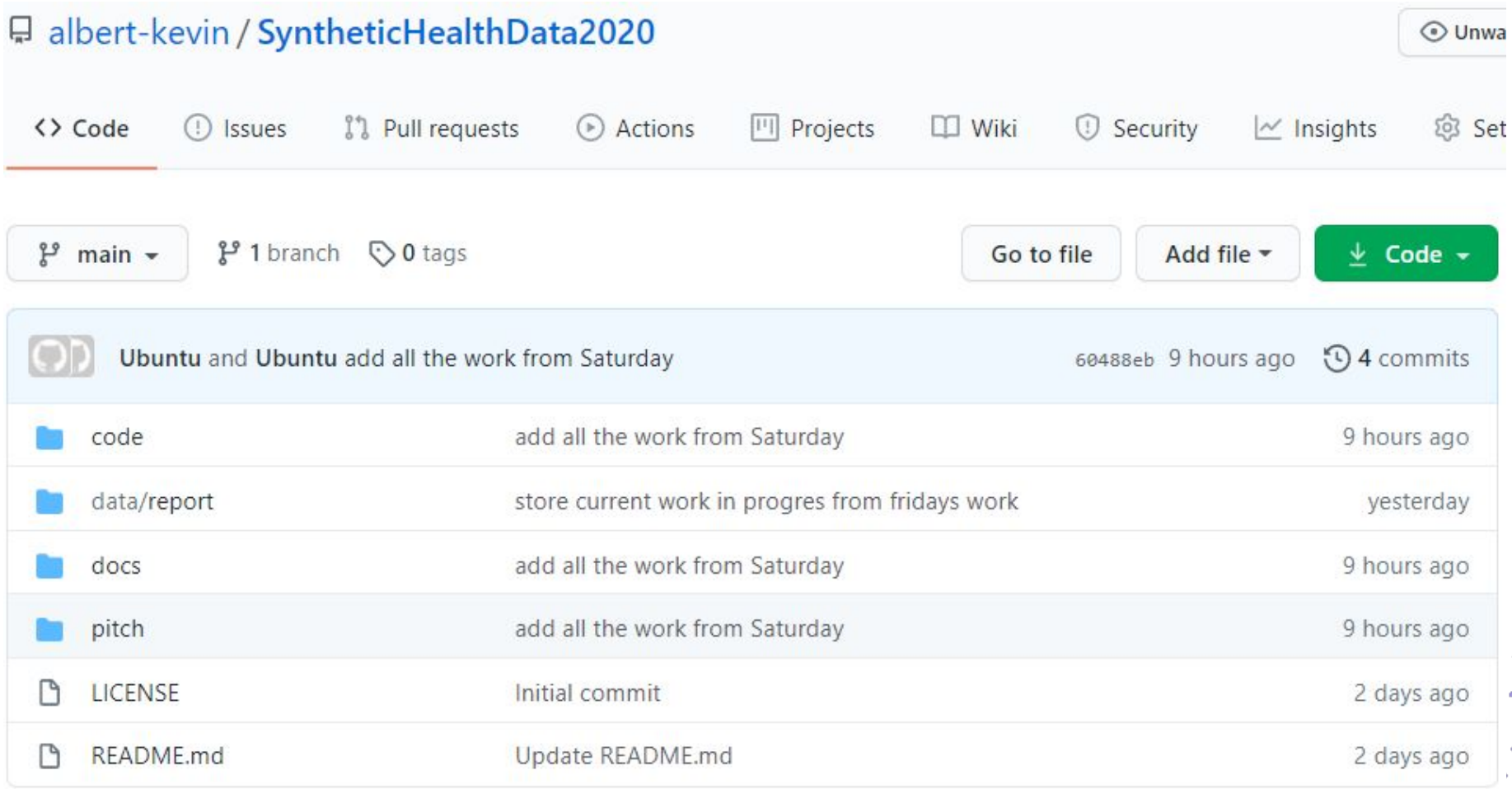
- anomalydetector27112020
- computervision27112020
- contentmoderator27112020
- customvisionPrediction27112020
- customvisionTraining27112020
- datalake27112020
- face27112020
- formrecognizer27112020
- functionApp27112020
- functionApp27112020
- languageunderstandingAuthoring27112020
- languageunderstandingPrediction27112020
- myDatabricks02
- myVM02
- myVM02_OsDisk_1_ba2
- myVM02NSG
- myVM02PublicIP
- myVM02VMNic
- machine_learning_workspace02
- machinelinsights73d0a6bb
- machinelkeyvaulte374e73a
- machinelstorage9af0d08f1
- myADF27112020

Development environment

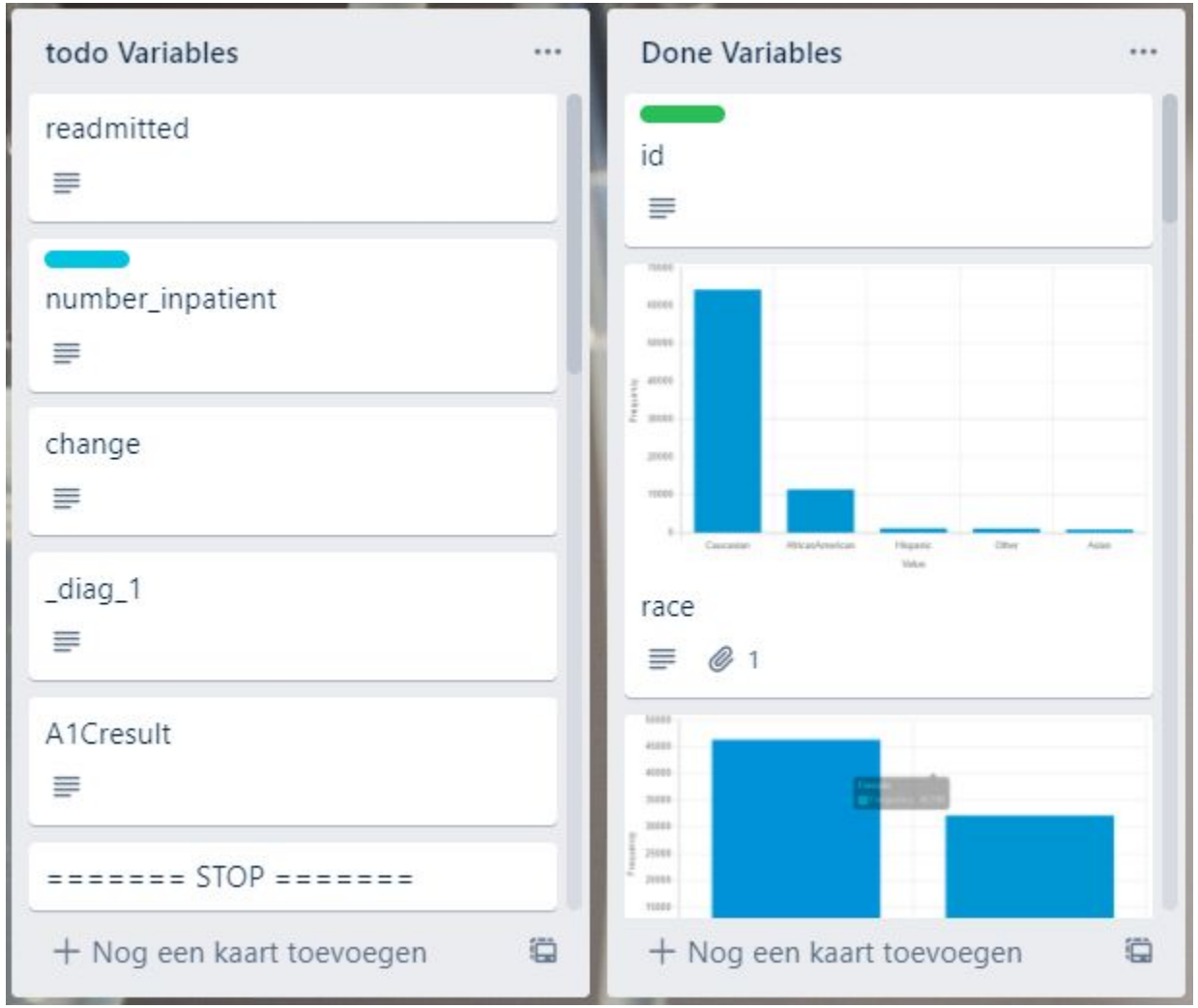


Git repo

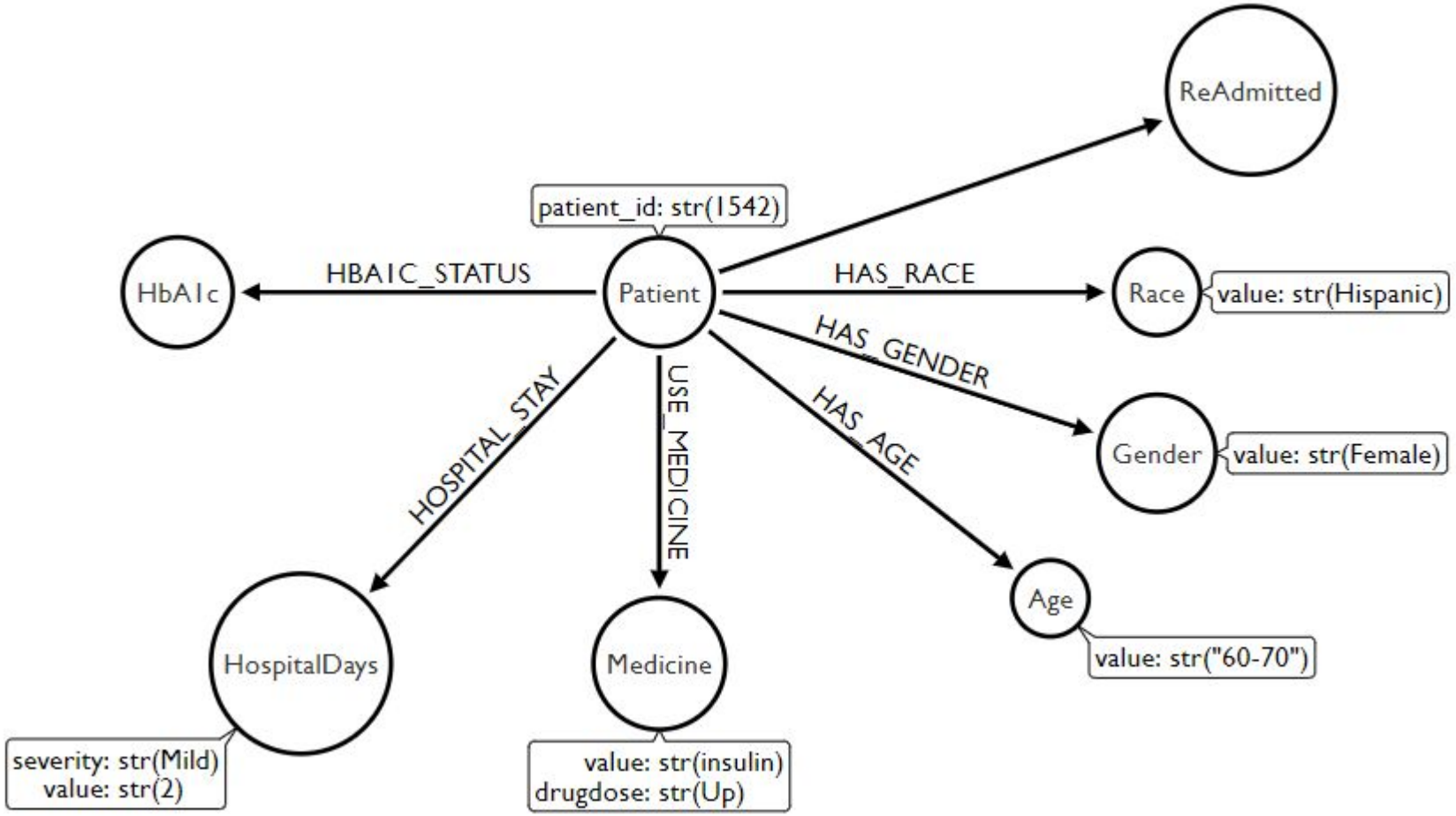
https://github.com/albert-kevin/**SyntheticHealthData2020**

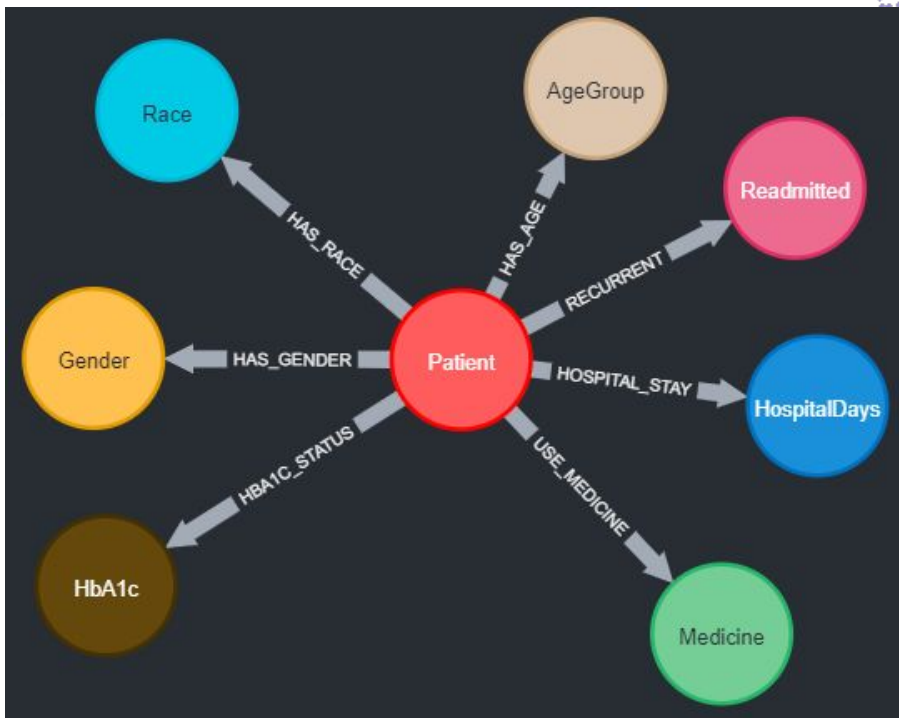


Datamodel


todo Variables
readmitted
number_inpatient
change
_diag_1
A1Cresult
======= STOP =======
+ Nog een kaart toevoegen
Done Variables
id
race
1
+ Nog een kaart toevoegen

Datamodel


ReAdmitted
patient_id: str(1542)
HbA1c
HBA1C_STATUS
Patient
HAS_RACE
Race
value: str(Hispanic)
HAS_GENDER
Gender
value: str(Female)
HOSPITAL_STAY
USE_MEDICINE
HAS_AGE
Age
value: str("60-70")
HospitalDays
severity: str(Mild)
value: str(2)
Medicine
value: str(insulin)
drugdose: str(Up)

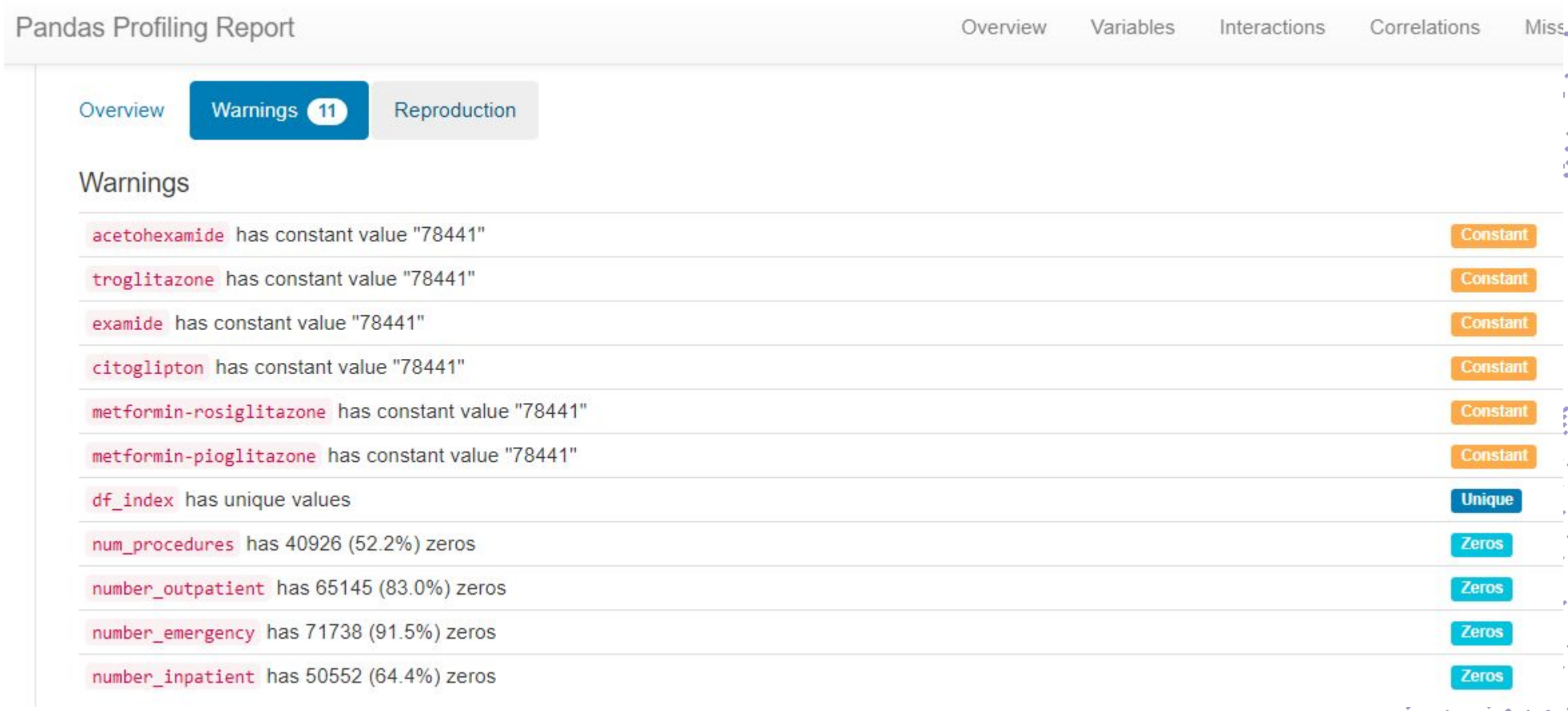

Race
AgeGroup
Readmitted
HAS_RACE
HAS_AGE
RECURRENT
Gender
HAS_GENDER
Patient
HOSPITAL_STAY
HospitalDays
HBA1C_STATUS
USE_MEDICINE
HbA1c
Medicine

interactive dataset dashboard

http://13.93.37.217:40000/dtale/main/1

42	tformin	glipizide-metformin	glimepiride-pioglitazone	metformin-rosiglitazone	metformin-pioglitazone	change
		No	No	No	No	No
		No	No	No	No	No
		No	No	No	No	Ch
		No	No	No	No	Ch
		No	No	No	No	Ch
		No	No	No	No	Ch
		No	No	No	No	Ch
		No	No	No	No	Ch
		No	No	No	No	Ch
		No	No	No	No	Ch
		No	No	No	No	No
		No	No	No	No	No
		No	No	No	No	Ch

static dataset report



Data cleaning

In [66]:
```
# creating new columns, binned in binary procedures or not
synthetic_df["num_procedures_bin"] = synthetic_df["num_procedures"].apply(lambda x: False if (x==0) else True)
real_df["num_procedures_bin"] = real_df["num_procedures"].apply(lambda x: False if (x==0) else True)
```

In [67]:
```
# created #meds / # days_in_hospital
synthetic_df["num_medications_perday"] = synthetic_df["num_medications"]/synthetic_df["time_in_hospital"]
real_df["num_medications_perday"] = real_df["num_medications"]/real_df["time_in_hospital"]
```

In [68]:
```
# synthetic_df["number_outpatient_perday"] = synthetic_df["number_outpatient"]/synthetic_df["time_in_hospital"]
# synthetic_df["number_inpatient_perday"] = synthetic_df["number_inpatient"]/synthetic_df["time_in_hospital"]
```

In [69]:
```
# created a binned version - severity lvl
bin_labels_4 = ['Normal', 'Mild', 'Moderate', 'Severe']
synthetic_df['time_in_hospital_severitylvl'] = pd.qcut(synthetic_df['time_in_hospital'], q=[0, 0.25, 0.5, 0.75, 1], labels=bin_la
real_df['time_in_hospital_severitylvl'] = pd.qcut(real_df['time_in_hospital'], q=[0, 0.25, 0.5, 0.75, 1], labels=bin_labels_4)
```

In [70]:
```
# synthetic_df["number_emergency"].apply(lambda x: "normalPatient" if (x==0) else x)
```

In [71]:
```
# remove the [] and () from age with regex
synthetic_df["age"] = synthetic_df["age"].str.replace(r'[^\-\w ]', '')
real_df["age"] = real_df["age"].str.replace(r'[^\-\w ]', '')
```

Machine Learning

- Multilabel Classification
- target “readmitted”
- primary metric “AUC_weighted”
- cross validation 5 (full dataset)
- no DeepLearning used
- autoML
- auto feature engineering
- run on cloud VM (4 core, 32GB) can scale to cluster 20 nodes
- cloud infrastructure (compute, storage, VM)
- synthetic dataset Version 1

Machine Learning

ITERATION	PIPELINE	DURATION	METRIC	BEST
0	MaxAbsScaler LightGBM	0:00:41	0.8949	0.8949
1	MaxAbsScaler XGBoostClassifier	0:00:56	0.8776	0.8949
2	MaxAbsScaler RandomForest	0:00:21	0.8334	0.8949
3	MaxAbsScaler SGD	0:00:21	0.8822	0.8949
4	MaxAbsScaler SGD	0:00:23	0.8799	0.8949
5	MaxAbsScaler ExtremeRandomTrees	0:00:28	0.8120	0.8949
6	MaxAbsScaler ExtremeRandomTrees	0:00:25	0.7914	0.8949
7	MaxAbsScaler SGD	0:00:24	0.8828	0.8949
8	MaxAbsScaler RandomForest	0:00:27	0.8197	0.8949
9	MaxAbsScaler SGD	0:00:26	0.8834	0.8949
10	MaxAbsScaler RandomForest	0:00:23	0.8102	0.8949
11	MaxAbsScaler SGD	0:00:28	0.8818	0.8949
12	MaxAbsScaler SGD	0:00:22	0.8820	0.8949
13	MaxAbsScaler RandomForest	0:00:23	0.8033	0.8949
14	StandardScalerWrapper ExtremeRandomTrees	0:00:26	0.7867	0.8949
15	SparseNormalizer ExtremeRandomTrees	0:00:25	0.8209	0.8949
16	SparseNormalizer ExtremeRandomTrees	0:00:26	0.8013	0.8949
17	MaxAbsScaler SGD	0:00:24	0.8833	0.8949

Machine Learning

Synthetic data

AUC_weighted

0.8957

log_loss

0.5496

accuracy

0.7771

f1_score_weighted

0.7736

Real data

AUC_weighted

0.6533

log_loss

0.8977

accuracy

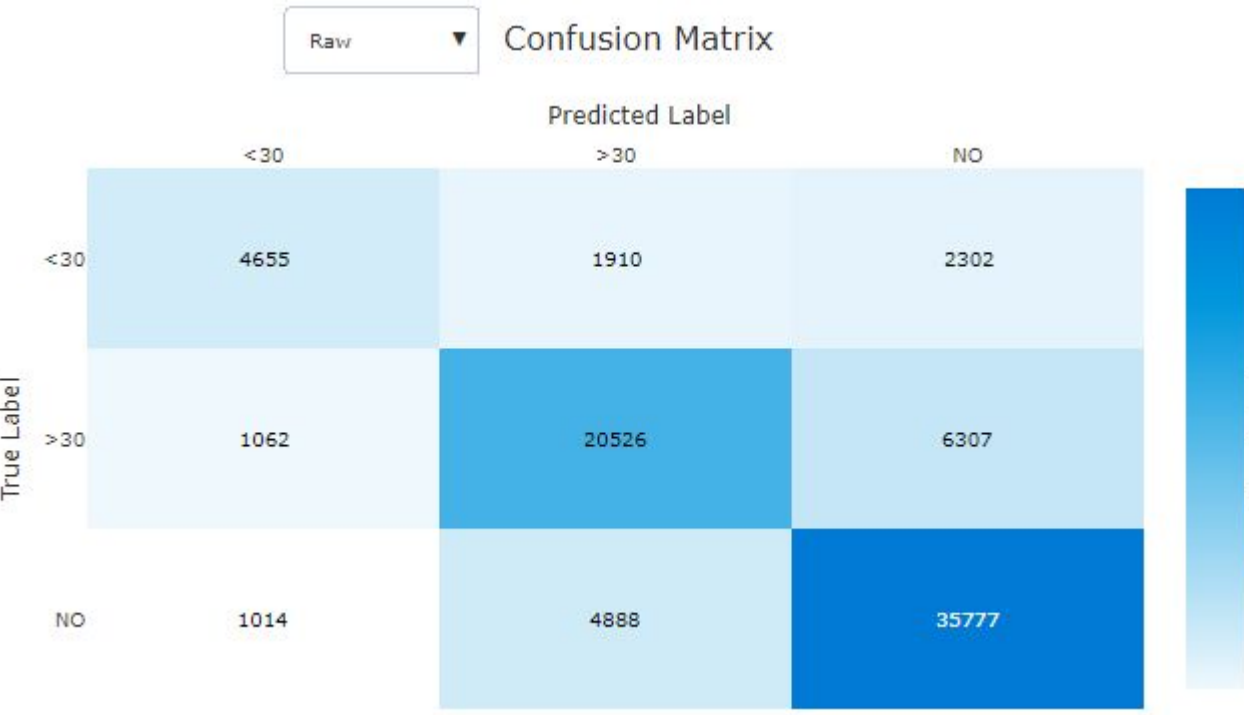
0.5755

f1_score_weighted

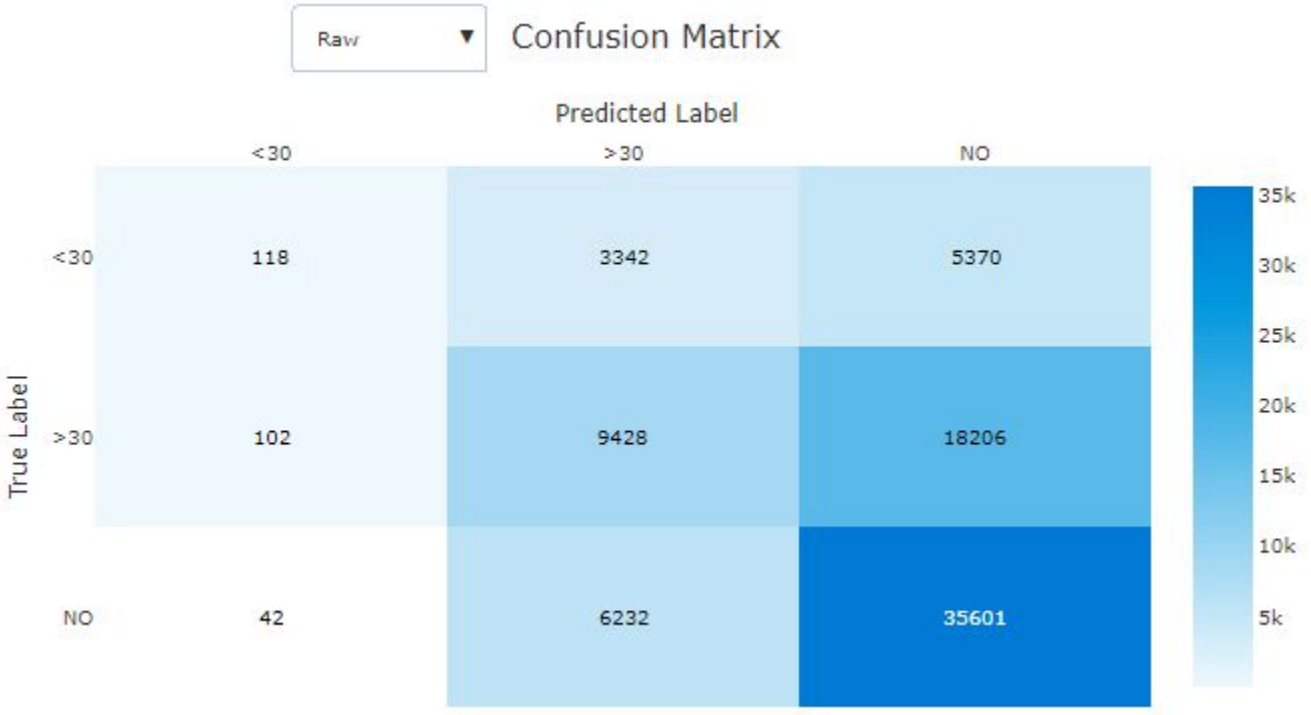
0.5217

Machine Learning

Synthetic data

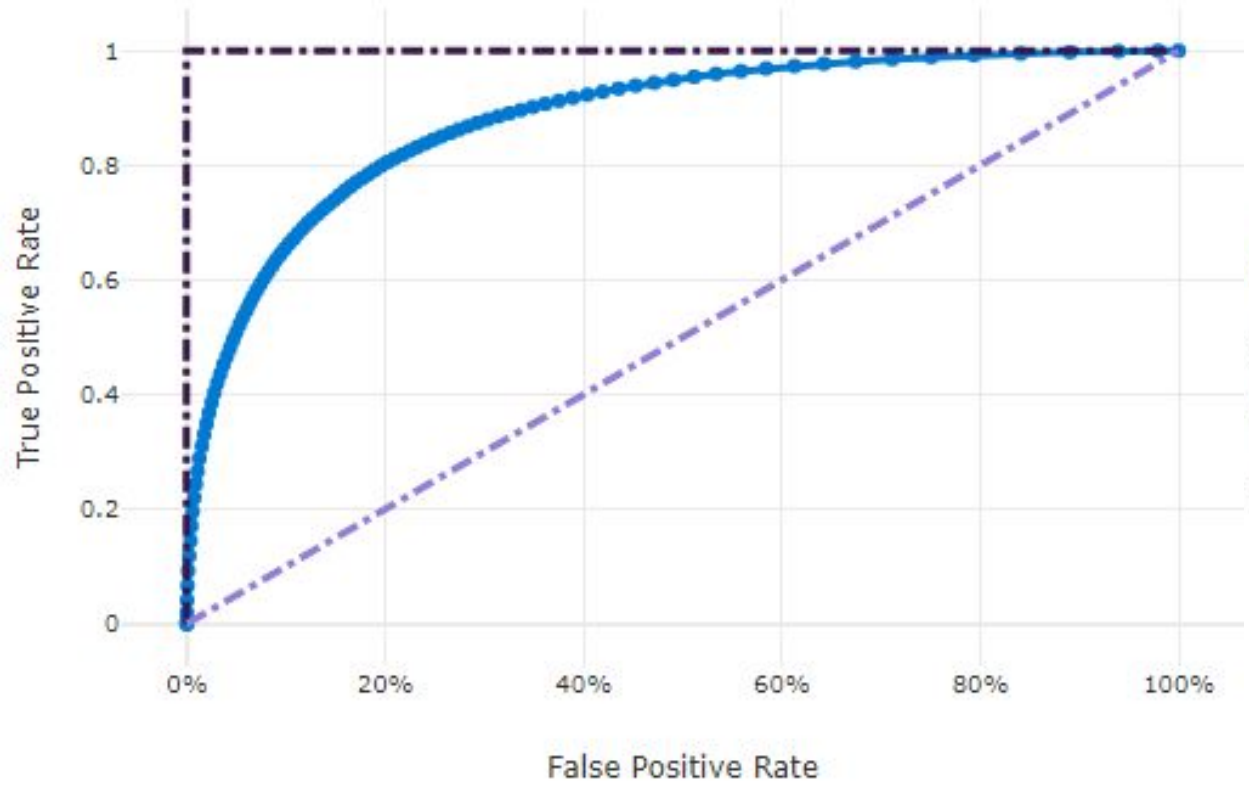


Real data



ROC

Synthetic data



Real data

Feature Importance

Synthetic data

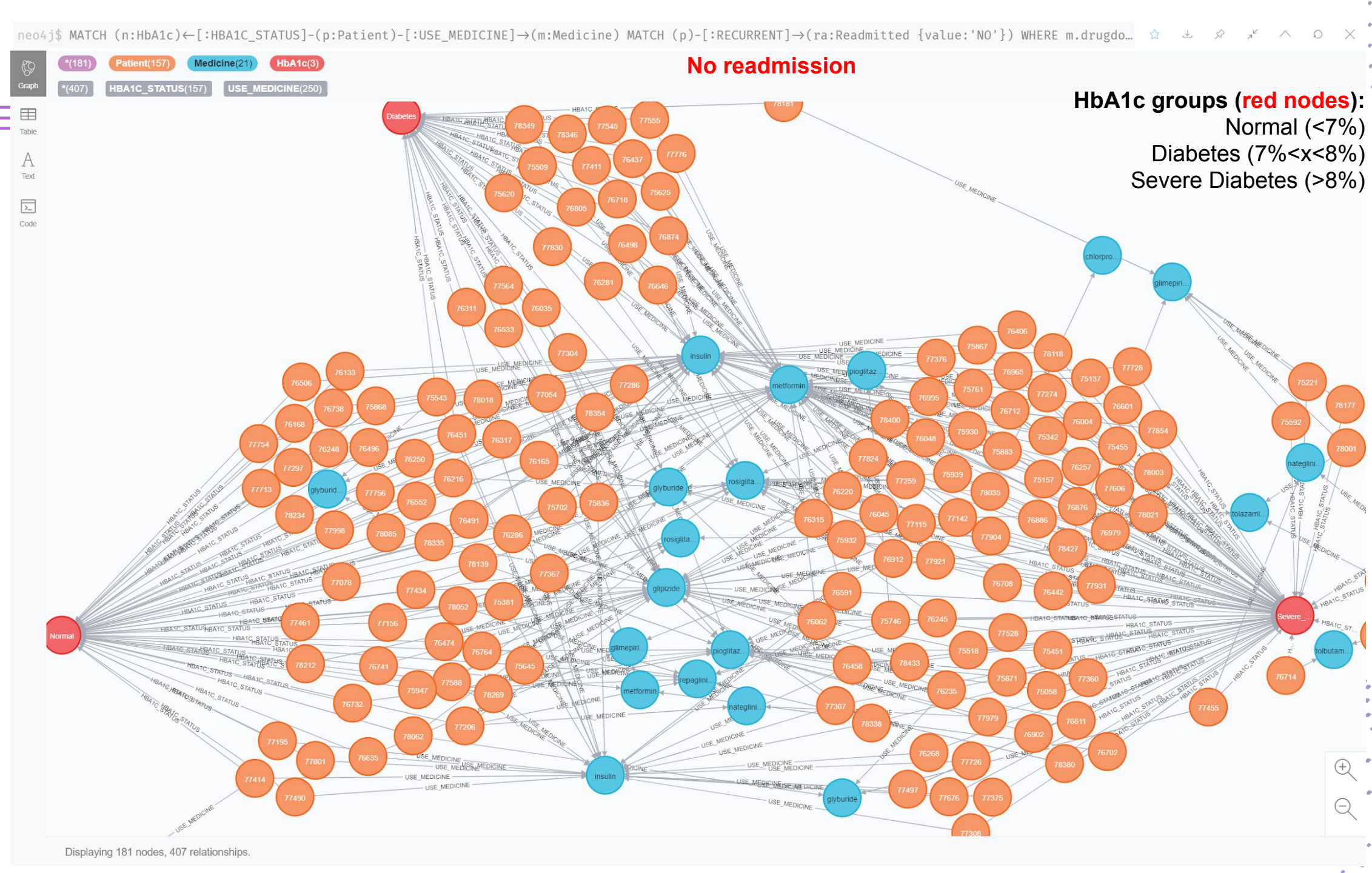
	modelFeatureImportance_name	modelFeatureImportance_value	modelFeatureImportance_relativeWeight
0	number_inpatient_CharGramCountVectorizer_0	0.80	0.16
1	change_ModeCatImputer_LabelEncoder	0.39	0.08
2	_diag_1_CharGramCountVectorizer_Diseases of the circulatory system	0.27	0.06
3	time_in_hospital_severitylvl_CharGramCountVectorizer_Normal	0.20	0.04
4	_diag_3_CharGramCountVectorizer_Diabetes mellitus	0.19	0.04
5	age_CharGramCountVectorizer_80-90	0.18	0.04
6	_diag_2_CharGramCountVectorizer_Diabetes mellitus	0.14	0.03
7	number_diagnoses_CharGramCountVectorizer_9	0.14	0.03
8	_diag_1_CharGramCountVectorizer_Diseases of the respiratory system	0.13	0.03
9	num_medications_MeanImputer	0.12	0.02
10	number_inpatient_CharGramCountVectorizer_2	0.11	0.02

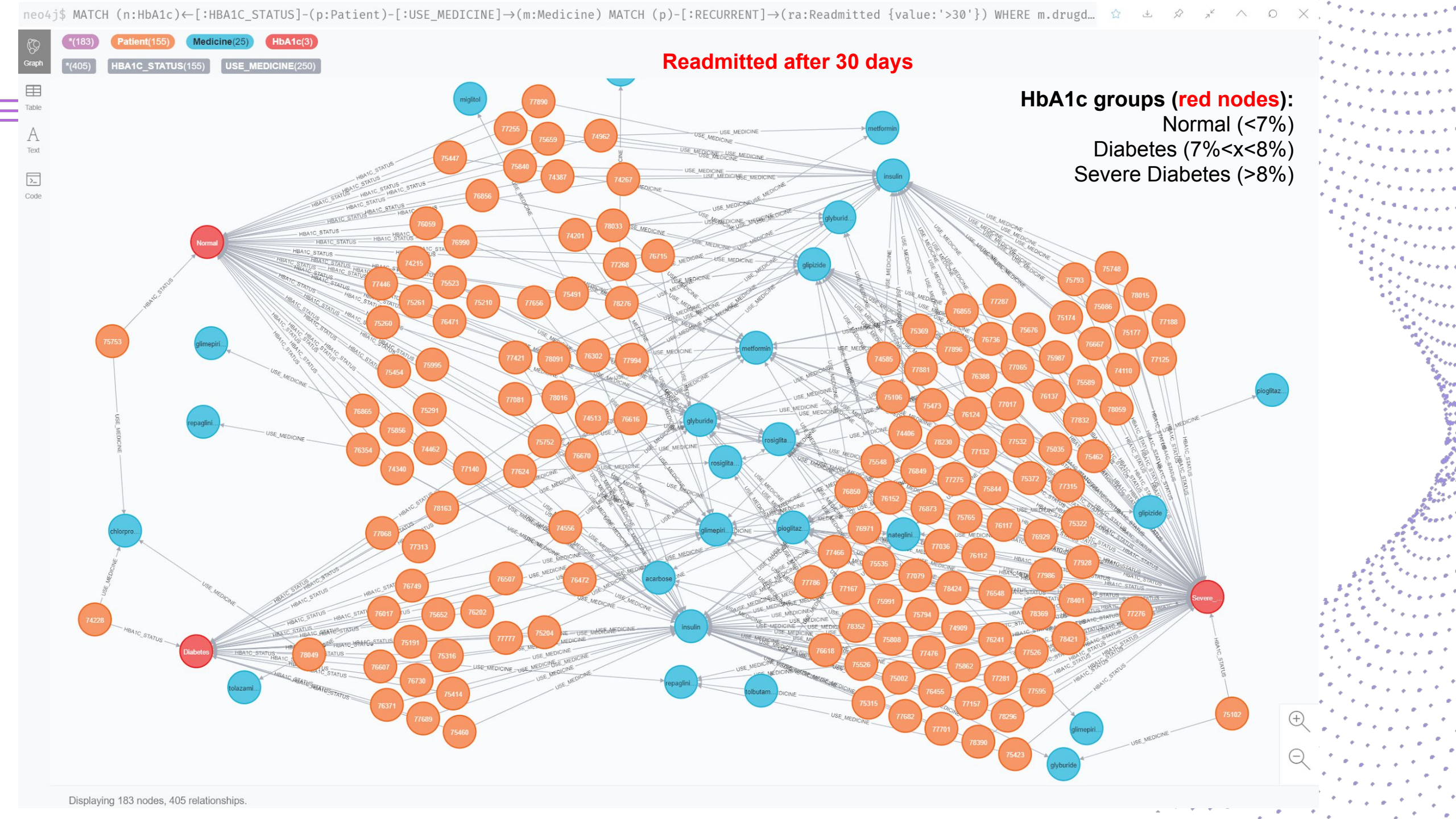
Feature Importance

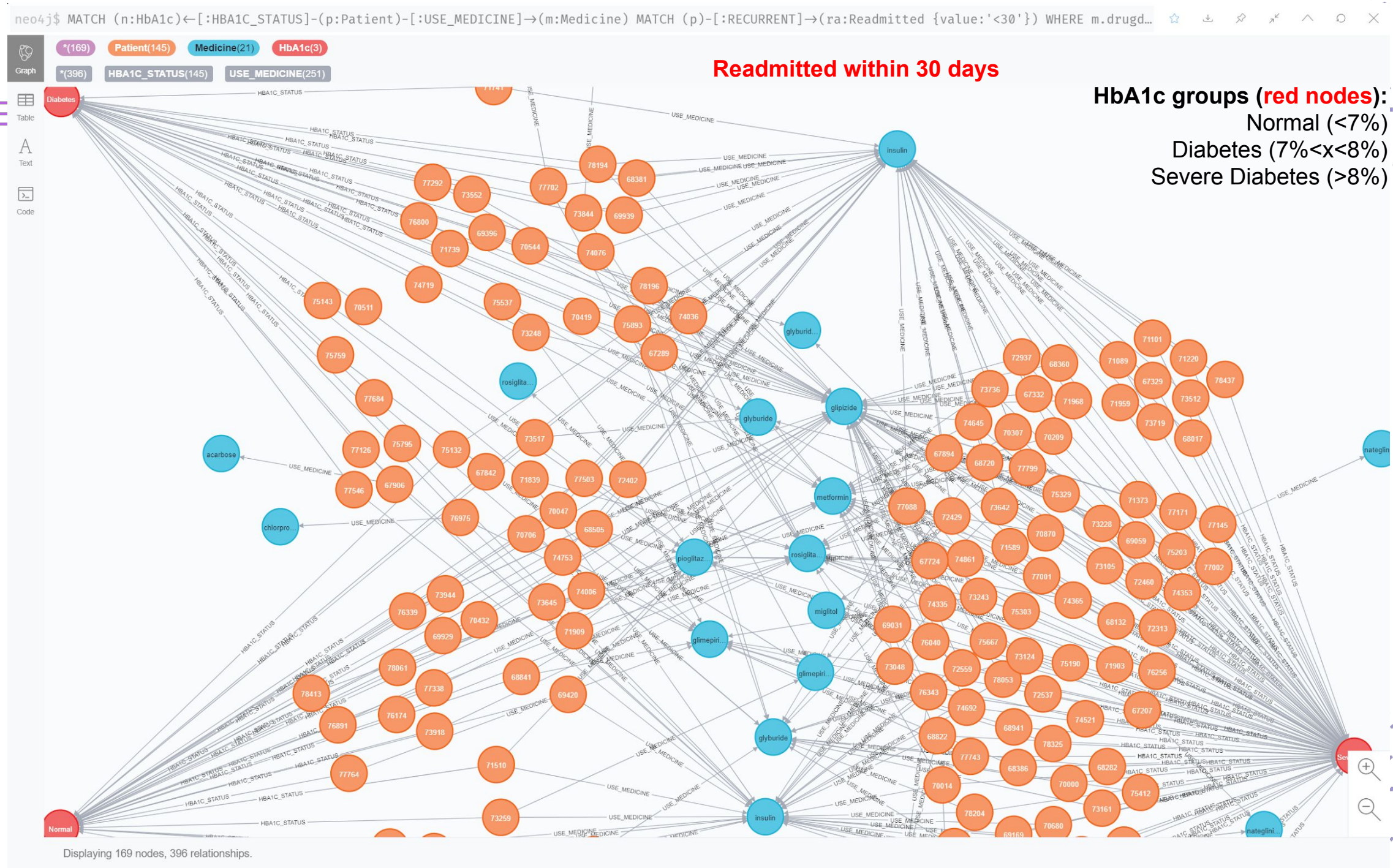
Real data

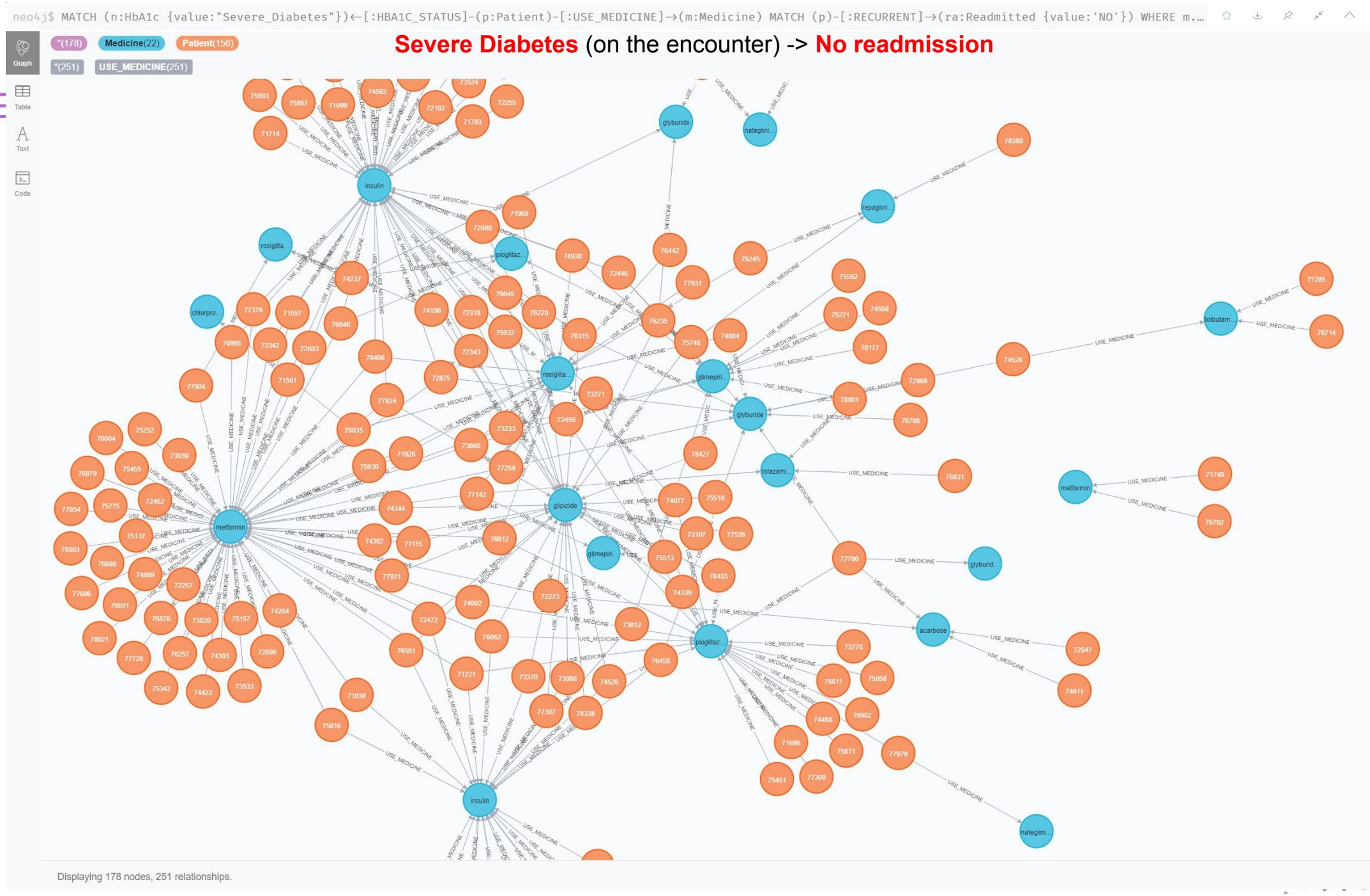
	modelFeatureImportance_name	modelFeatureImportance_value	modelFeatureImportance_relativeWeight
0	_diag_1_CharGramCountVectorizer_Diseases of the nervous system	3.79	0.15
1	age_CharGramCountVectorizer_50-60	2.97	0.12
2	num_medications_perday_MeanImputer	2.63	0.10
3	number_inpatient_CharGramCountVectorizer_11	2.14	0.08
4	number_inpatient_CharGramCountVectorizer_0	1.95	0.08
5	time_in_hospital_CharGramCountVectorizer_9	1.41	0.05
6	num_lab_procedures_MeanImputer	1.36	0.05
7	num_medications_MeanImputer	0.90	0.04
8	repaglinide_CharGramCountVectorizer_Steady	0.85	0.03
9	_diag_2_CharGramCountVectorizer_Neoplasms	0.63	0.02
10	_diag_1_CharGramCountVectorizer_Endocrine, nutritional, and metabolic diseases and immunity diso...	0.55	0.02

Some Interesting Findings


neo4j$ MATCH (n:HbA1c)←[:HBA1C_STATUS]-(p:Patient)-[:USE_MEDICINE]→(m:Medicine) MATCH (p)-[:RECURRENT]→(ra:Readmitted {value:'NO'}) WHERE m.drugdo…
Graph
Table
Text
Code
*(181)
Patient(157)
Medicine(21)
HbA1c(3)
*(407)
HBA1C_STATUS(157)
USE_MEDICINE(250)
No readmission
HbA1c groups (red nodes):
Normal (<7%)
Diabetes (7%<x<8%)
Severe Diabetes (>8%)
Diabetes
Normal
Severe_...
insulin
metformin
glyburide
glipizide
rosiglita...
pioglitaz...
glimepiri...
repaglini...
nateglini...
chlorpro...
tolazami...
tolbutam...
USE_MEDICINE
HBA1C_STATUS
Displaying 181 nodes, 407 relationships.


neo4j$ MATCH (n:HbA1c)←[:HBA1C_STATUS]-(p:Patient)-[:USE_MEDICINE]→(m:Medicine) MATCH (p)-[:RECURRENT]→(ra:Readmitted {value:'>30'}) WHERE m.drugd…
Graph
Table
Text
Code
*(183)
Patient(155)
Medicine(25)
HbA1c(3)
*(405)
HBA1C_STATUS(155)
USE_MEDICINE(250)
Readmitted after 30 days
HbA1c groups (red nodes):
Normal (<7%)
Diabetes (7%<x<8%)
Severe Diabetes (>8%)
Normal
Diabetes
Severe_...
miglitol
metformin
insulin
glyburid...
glipizide
glyburide
rosiglita...
pioglitaz...
glimepiri...
repaglini...
nateglini...
acarbose
tolbutam...
tolazami...
chlorpro...
Displaying 183 nodes, 405 relationships.


neo4j$ MATCH (n:HbA1c)←[:HBA1C_STATUS]-(p:Patient)-[:USE_MEDICINE]→(m:Medicine) MATCH (p)-[:RECURRENT]→(ra:Readmitted {value:'<30'}) WHERE m.drugd…
Graph
Table
Text
Code
*(169)
Patient(145)
Medicine(21)
HbA1c(3)
*(396)
HBA1C_STATUS(145)
USE_MEDICINE(251)
Readmitted within 30 days
HbA1c groups (red nodes):
Normal (<7%)
Diabetes (7%<x<8%)
Severe Diabetes (>8%)
Diabetes
Normal
insulin
glyburide
glipizide
metformin
rosiglita...
pioglitaz...
miglitol
glimepiri...
acarbose
chlorpro...
nateglini...
HBA1C_STATUS
USE_MEDICINE
Displaying 169 nodes, 396 relationships.


neo4j$ MATCH (n:HbA1c {value:"Severe_Diabetes"})←[:HBA1C_STATUS]-(p:Patient)-[:USE_MEDICINE]→(m:Medicine) MATCH (p)-[:RECURRENT]→(ra:Readmitted {value:'NO'}) WHE
Graph
Table
Text
Code
*(178)
Medicine(22)
Patient(156)
*(251)
USE_MEDICINE(251)
Severe Diabetes (on the encounter) -> No readmission
insulin
metformin
glipizide
glyburide
rosiglita...
pioglitaz...
glimepiri...
repaglini...
nateglini...
tolbutam...
tolazami...
acarbose
chlorpro...
metformin
glyburid...
USE_MEDICINE
Displaying 178 nodes, 251 relationships.

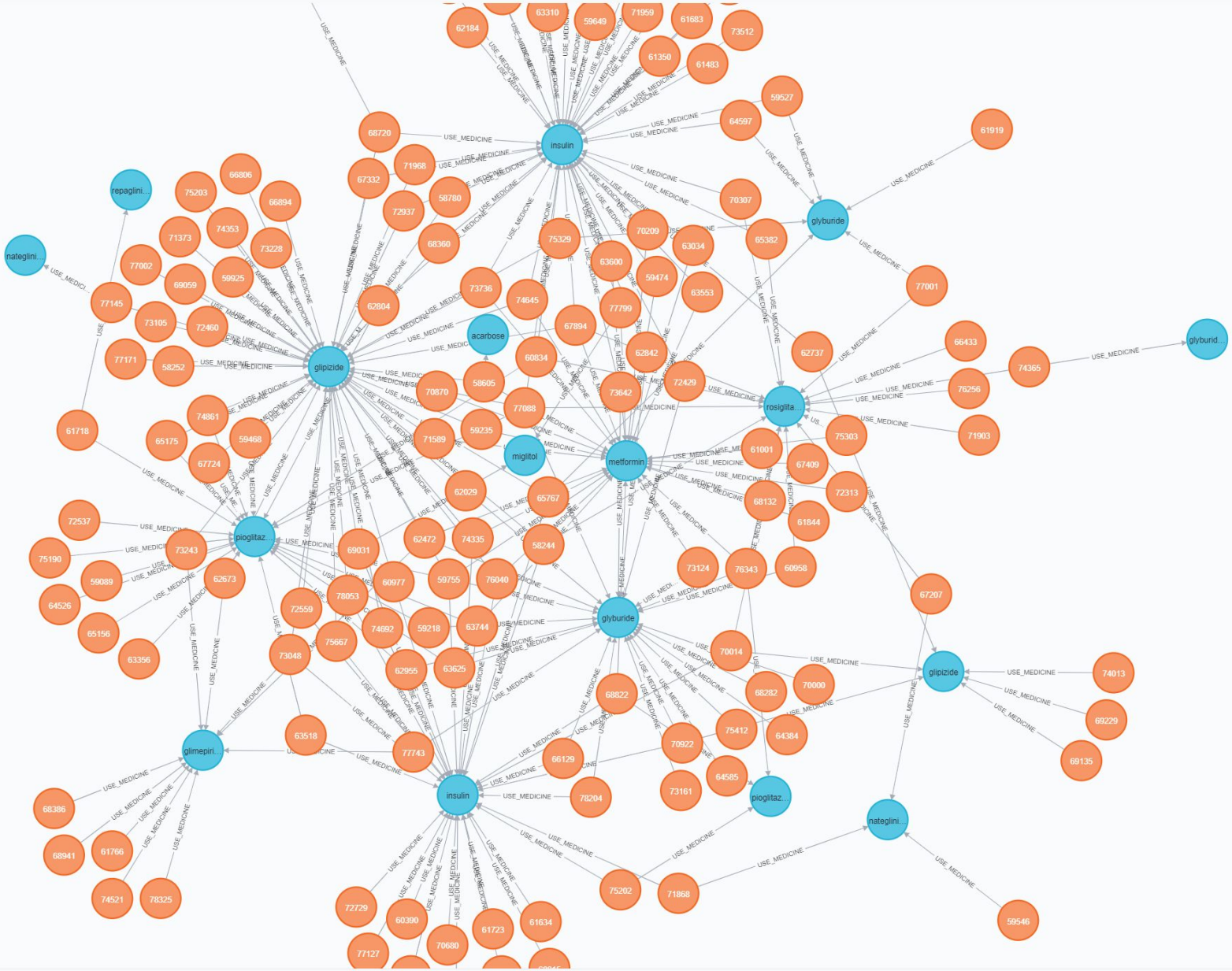
```
neo4j$ MATCH (n:HbA1c {value:"Severe_Diabetes"})←[:HBA1C_STATUS]-(p:Patient)-[:USE_MEDICINE]→(m:Medicine) MATCH (p)-[:RECURRENT]→(ra:Readmitted {value:'<30'}) WHERE m…
```

Graph | Table | Text | Code

*(166) Medicine(19) Patient(147)

*(251) USE_MEDICINE(251)

Severe Diabetes (on the encounter) -> Admission in less than 30 days



Displaying 166 nodes, 251 relationships.