participants
Severity
intensive
patient
final
identifying
Stanford
environment
analytics
individuals
inspire
countries
supporters
platform
faculty
field
team
hours
tutorials
Discussion
invite
resources
half
Care
Lab
model
will
create
Science
aims
care
open
share
community
must
data
code
livestream
students
teams
sample
tried
learning
join
Harvard
Initiative
States
Brought
Brazil
health
Data
used
across
non-profit
Illness
February
throughout
Open
others
GOSSIS
roles
anyone
University
machine
thanks
spanning
via
tools
science
Big
industry
posted
least
preferred
WiDS
Sri
Hub
female
women
time
led
ends
uses
predictions
academic
government
collaboration
Forum
closes
Score
datathon
Winners
new
using
West
various
worldwide
include
release
can
sponsors
series
skills
announced
Source
Toward
never
growing
build
ICU
reaching
provided
Special
Datathon
Unit
Acknowledgements
initiative
winners
social
MIT
get
help
contest
Global
tutorials
worked
hone
webinars
learn
Kaggle
well
Privacy
first
determined
completed
New
Committee
Toronto
veterans
Tutorials
year's
Innovation
started
organizations
advance
participate
March
impact
connect
supportive
visits
focuses
leaderboard
Women
Participants
survival
Overview
enthusiasts

Problem

WiDS Datathon 2020 focused on

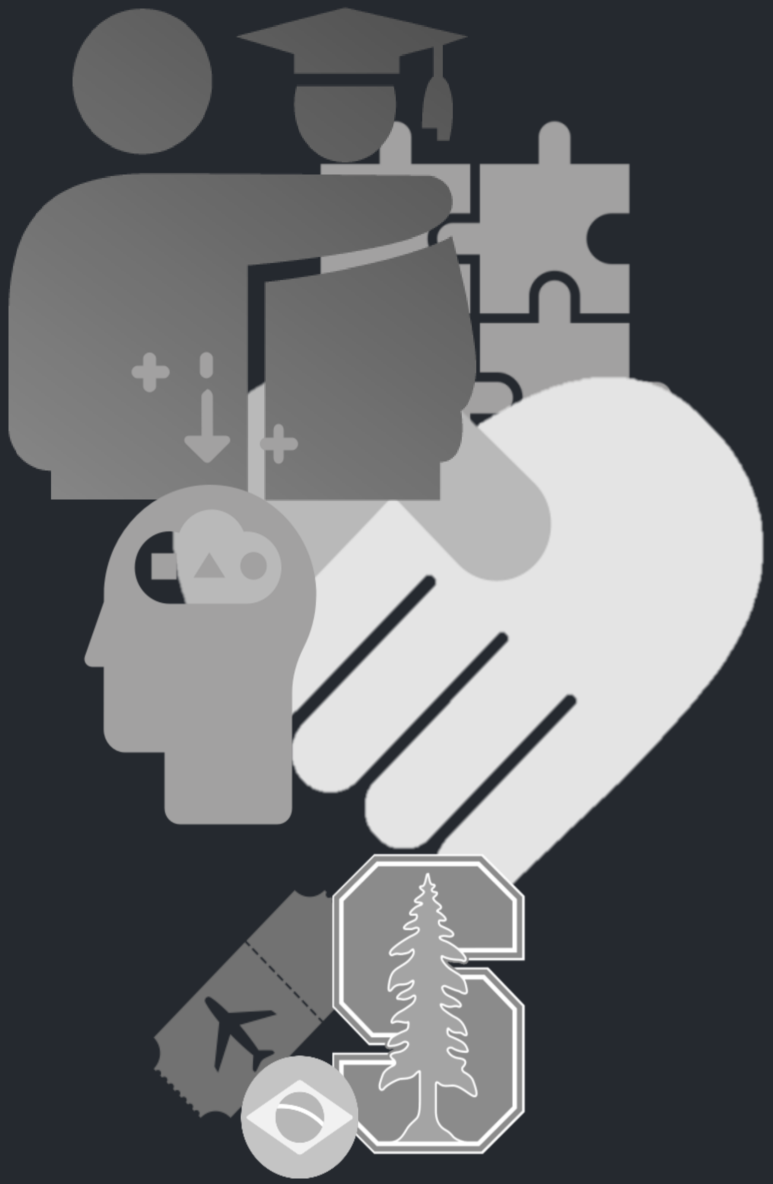
patient health through data from

MIT’s GOSSIS (Global Open Source Severity of Illness Score) initiative.

The challenge was to **create a model** that uses data from

the first 24 hours of intensive care to **predict**

patient survival.



Data

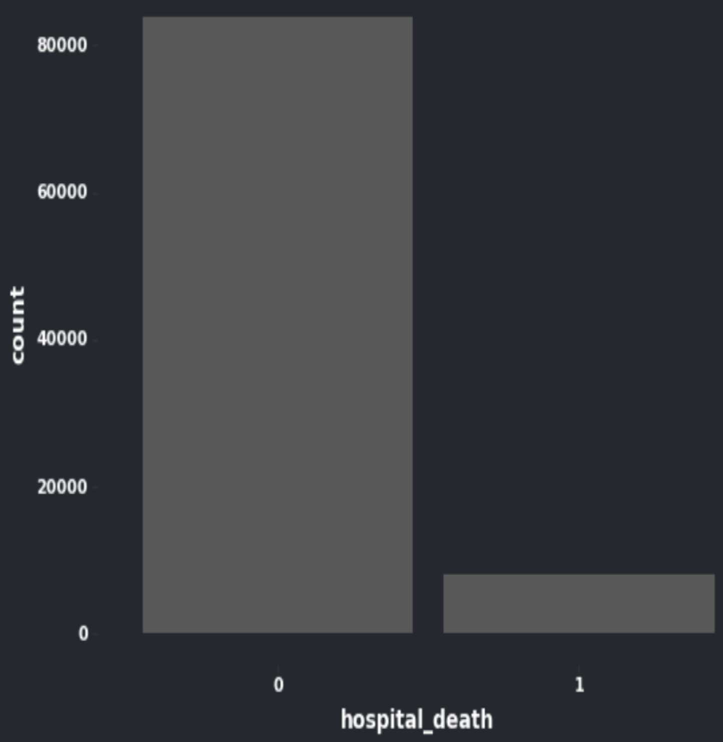
Shape

```
df.shape  (91713, 186)
```

Column Categories

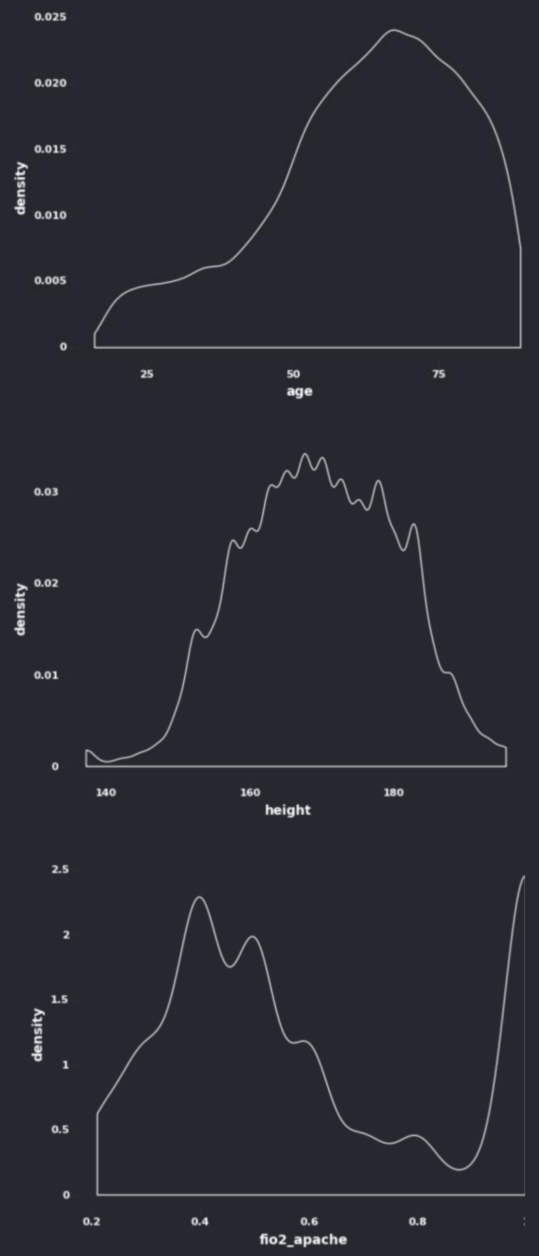
	Variable Name
Category	
APACHE comorbidity	8
APACHE covariate	28
APACHE grouping	2
APACHE prediction	2
GOSSIS example prediction	1
demographic	16
identifier	3
labs	60
labs blood gas	16
vitals	52

Target Variable

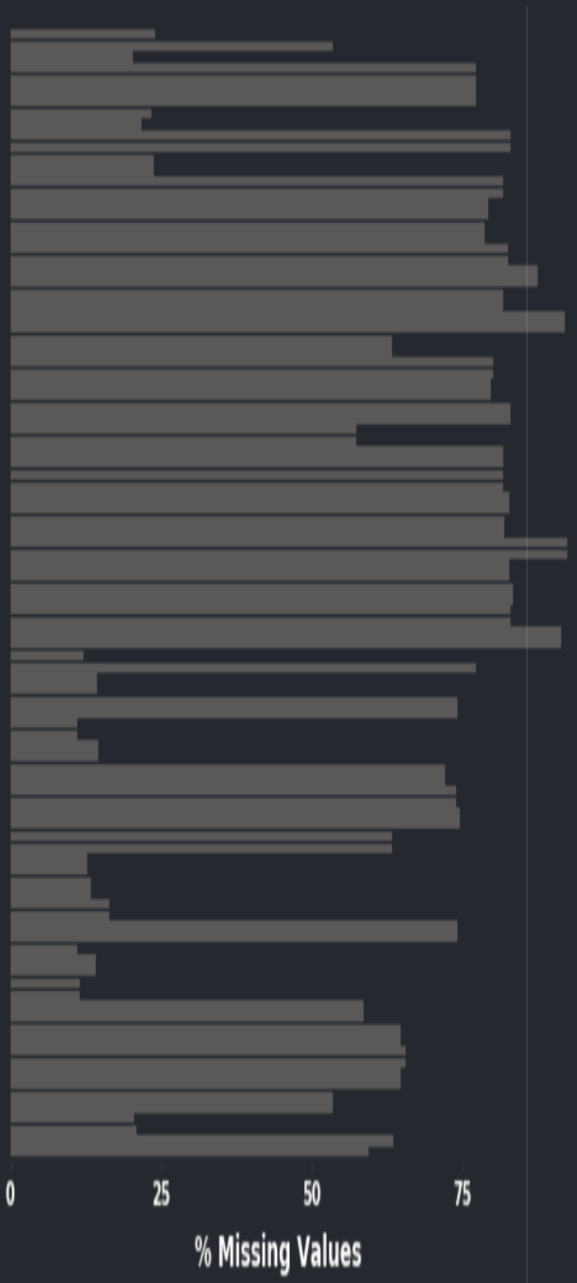


Messiness

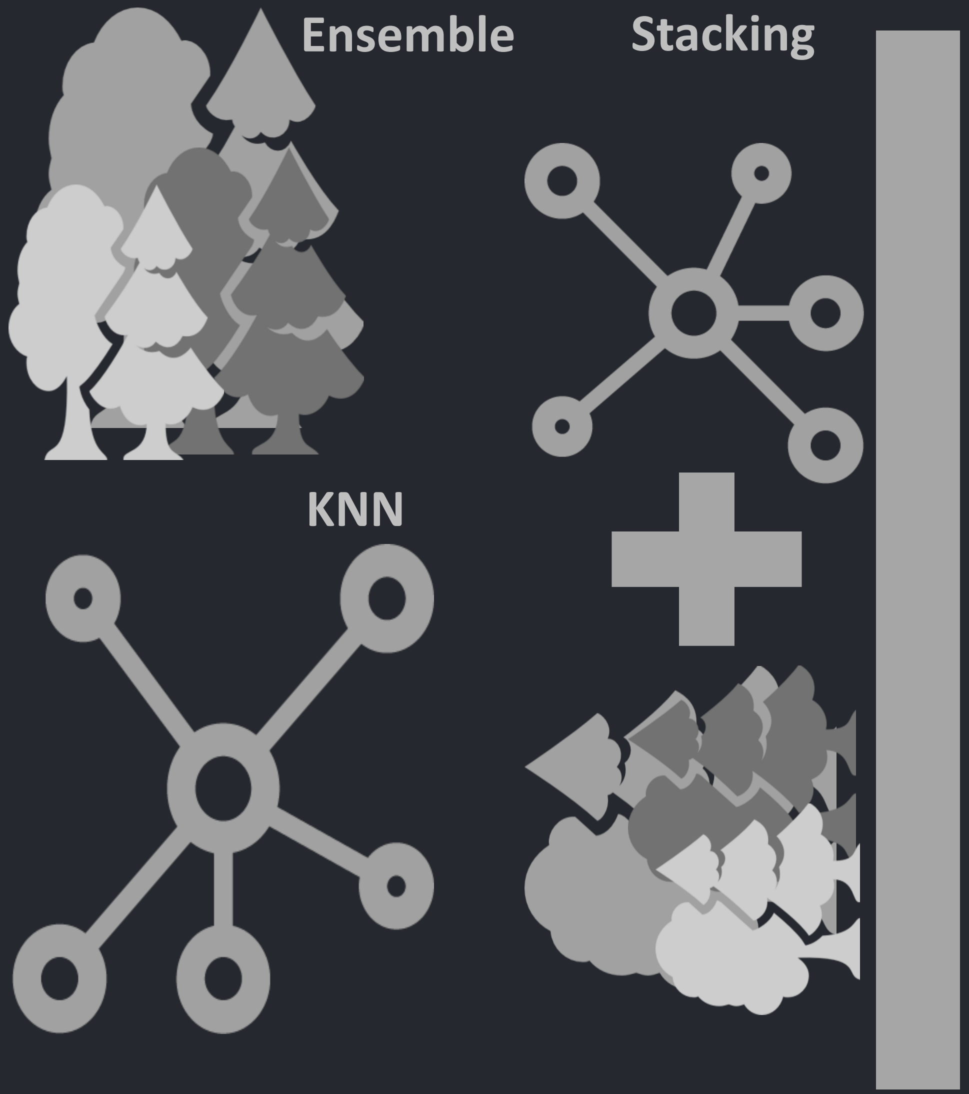
Skews



Missing Values



Models



Ensemble

Stacking

KNN

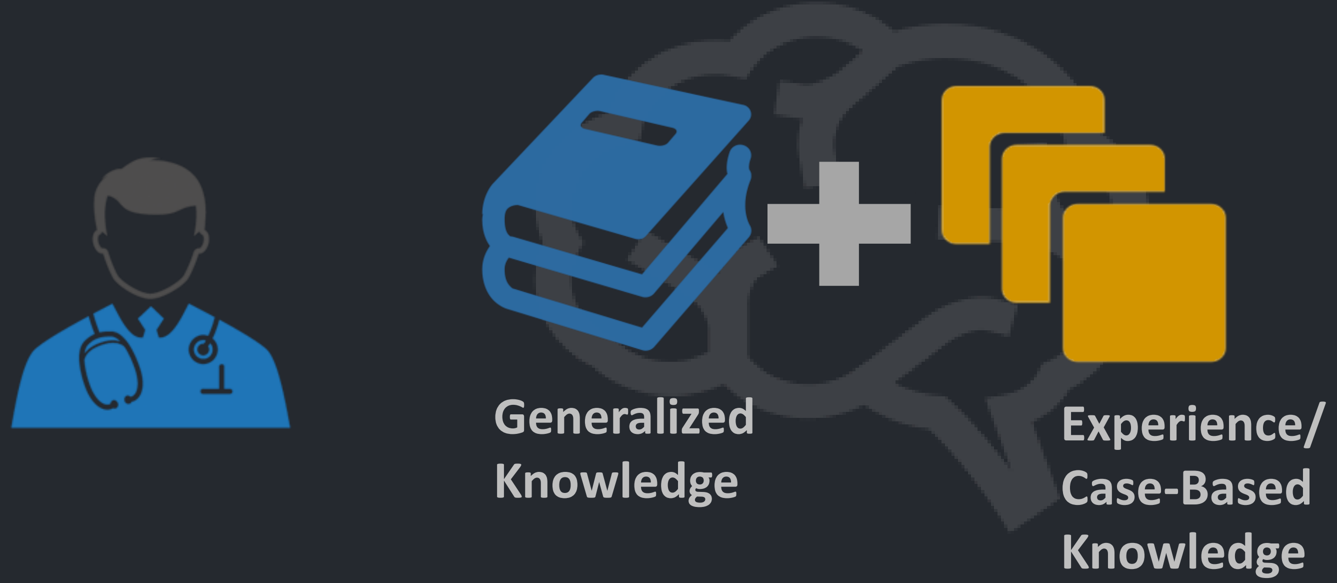
Research Questions

- What features influence the survival of the patients the most ?

- How does age affect patient survival ? Are older patients at higher risk irrespective of the condition?

- Is there any strong correlation between disease/Condition and the number of fatalities ?

- Is there a correlation between disease/Condition and patient admissions and readmissions?



Generalized Knowledge

Experience/ Case-Based Knowledge

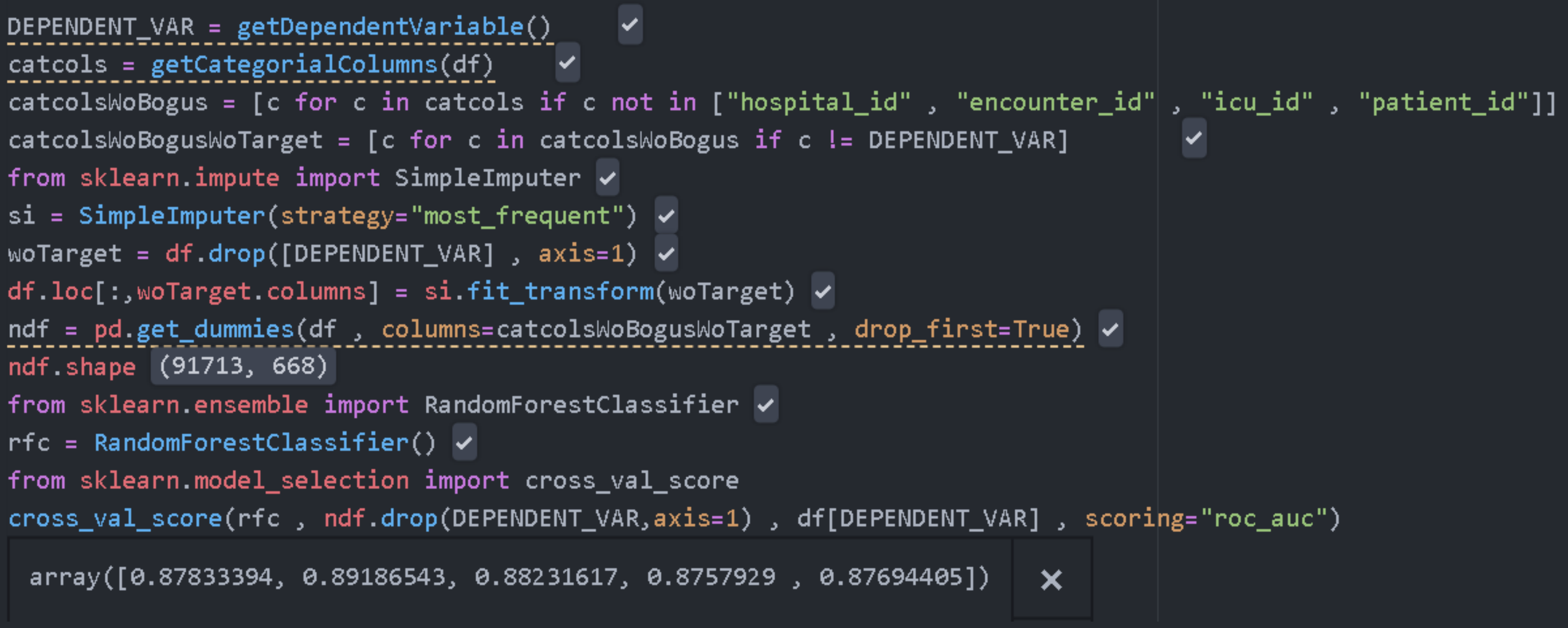
Baseline

Missing values

Categorical Columns

Modelling

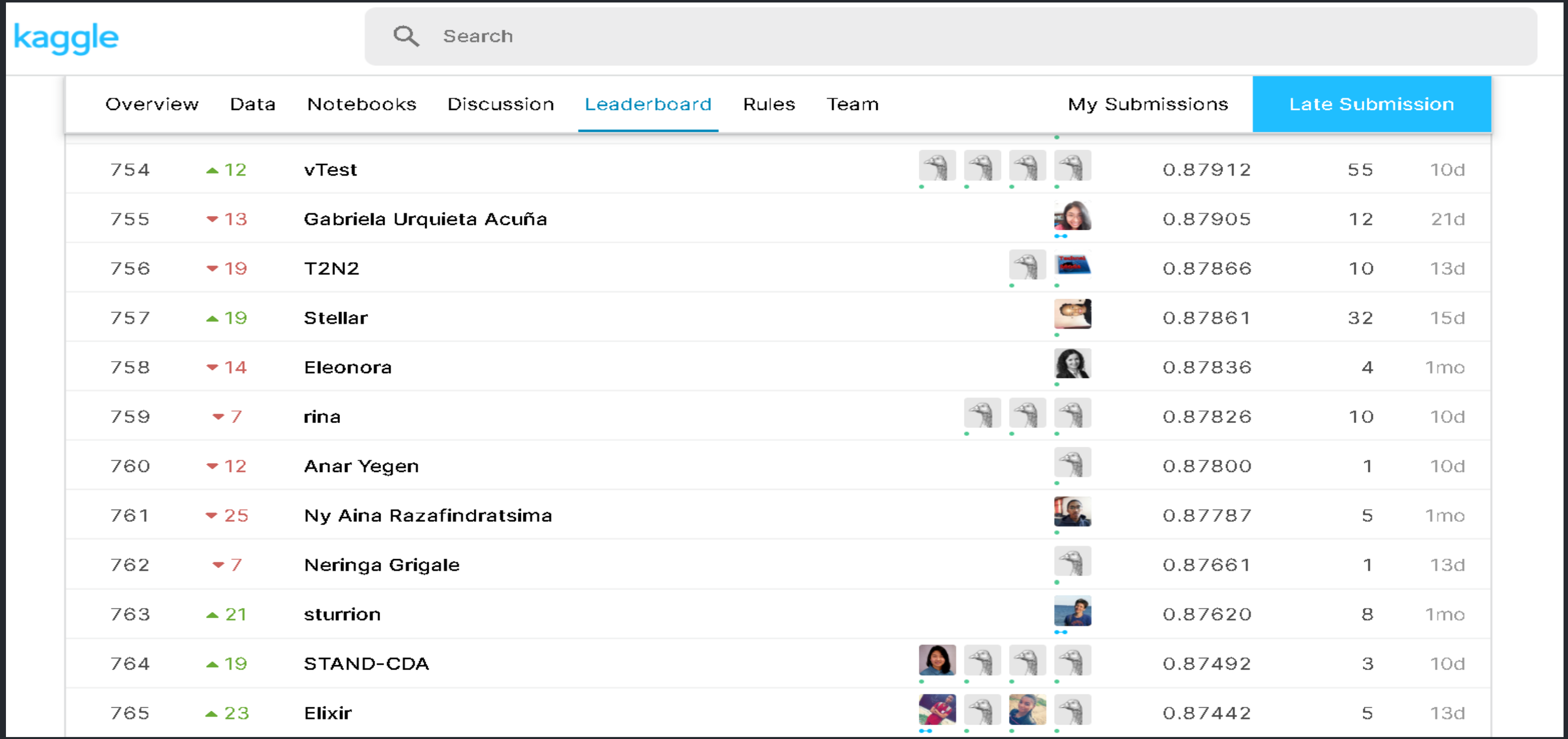
Score



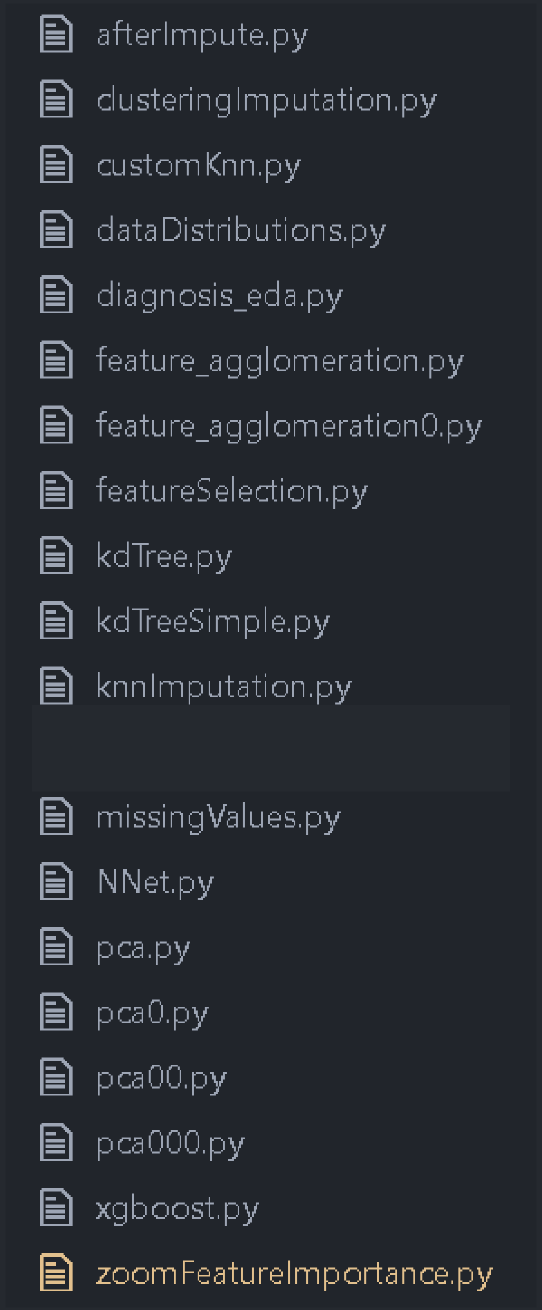
```
DEPENDENT_VAR = getDependentVariable() ✓
catcols = getCategorialColumns(df) ✓
catcolsWoBogus = [c for c in catcols if c not in ["hospital_id" , "encounter_id" , "icu_id" , "patient_id"]]
catcolsWoBogusWoTarget = [c for c in catcolsWoBogus if c != DEPENDENT_VAR] ✓
from sklearn.impute import SimpleImputer ✓
si = SimpleImputer(strategy="most_frequent") ✓
woTarget = df.drop([DEPENDENT_VAR] , axis=1) ✓
df.loc[:,woTarget.columns] = si.fit_transform(woTarget) ✓
ndf = pd.get_dummies(df , columns=catcolsWoBogusWoTarget , drop_first=True) ✓
ndf.shape (91713, 668)
from sklearn.ensemble import RandomForestClassifier ✓
rfc = RandomForestClassifier() ✓
from sklearn.model_selection import cross_val_score
cross_val_score(rfc , ndf.drop(DEPENDENT_VAR,axis=1) , df[DEPENDENT_VAR] , scoring="roc_auc")

array([0.87833394, 0.89186543, 0.88231617, 0.8757929 , 0.87694405]) ✕
```

0.878

760/940

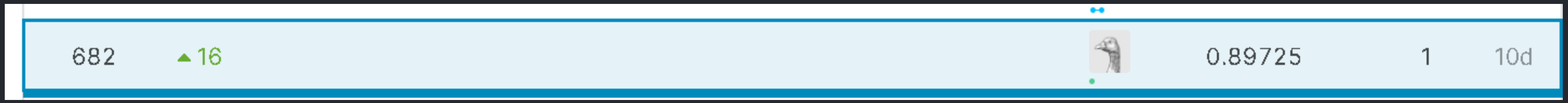


754	▲ 12	vTest	0.87912	55	10d
755	▼ 13	Gabriela Urquieta Acuña	0.87905	12	21d
756	▼ 19	T2N2	0.87866	10	13d
757	▲ 19	Stellar	0.87861	32	15d
758	▼ 14	Eleonora	0.87836	4	1mo
759	▼ 7	rina	0.87826	10	10d
760	▼ 12	Anar Yegen	0.87800	1	10d
761	▼ 25	Ny Aina Razafindratsima	0.87787	5	1mo
762	▼ 7	Neringa Grigale	0.87661	1	13d
763	▲ 21	sturrion	0.87620	8	1mo
764	▲ 19	STAND-CDA	0.87492	3	10d
765	▲ 23	Elixir	0.87442	5	13d

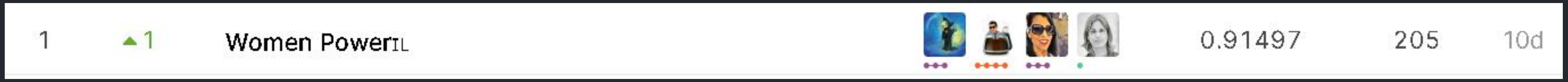


afterImpute.py
clusteringImputation.py
customKnn.py
dataDistributions.py
diagnosis_eda.py
feature_agglomeration.py
feature_agglomeration0.py
featureSelection.py
kdTree.py
kdTreeSimple.py
knnImputation.py
missingValues.py
NNet.py
pca.py
pca0.py
pca00.py
pca000.py
xgboost.py
zoomFeatureImportance.py

Missing values	Feature Engineering	Modelling
Simple Imputation	PCA	Random Forest
Knn Imputation	Agglomerative Clustering	NNet
KDTree	Random Forest Feature Importance	XGBoost , LGM
Random Forest Imputation	Recursive Feature Elimination	Stacking
Variance Based Drop Column	ANOVA	Logistic Regression

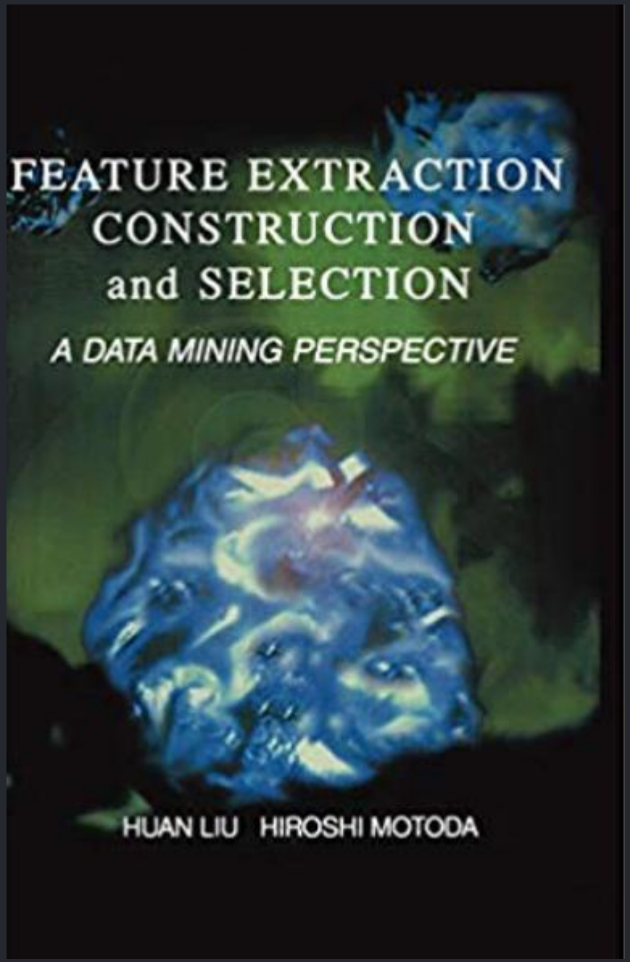


682 ▲16 0.89725 1 10d



1 ▲1 Women PowerIL 0.91497 205 10d

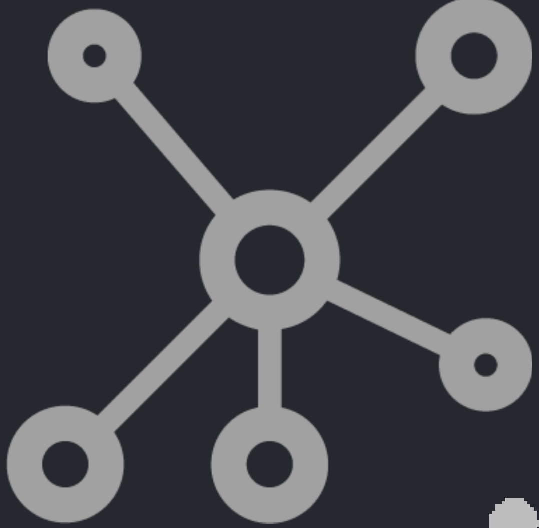
More to Come



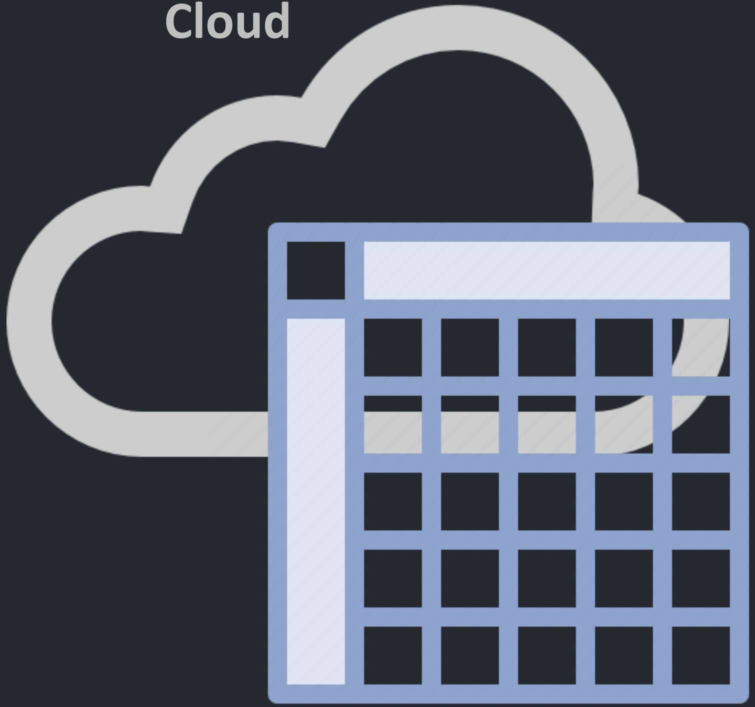
FEATURE EXTRACTION CONSTRUCTION and SELECTION

A DATA MINING PERSPECTIVE

HUAN LIU HIROSHI MOTODA

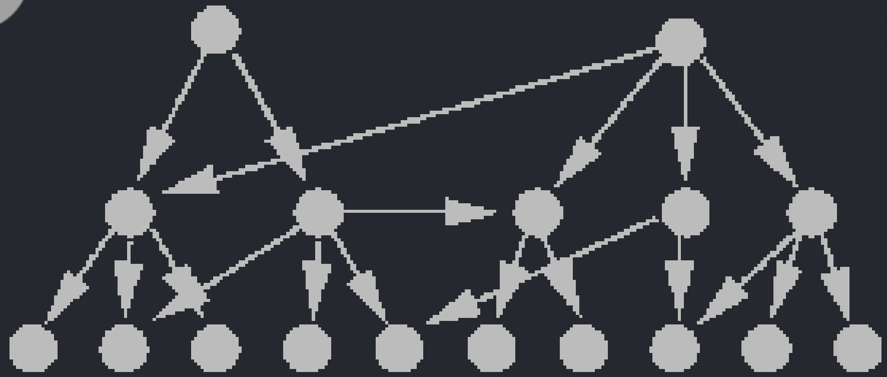


KNN



Cloud

HyperParameter Tuning



Belief Networks