



Problem

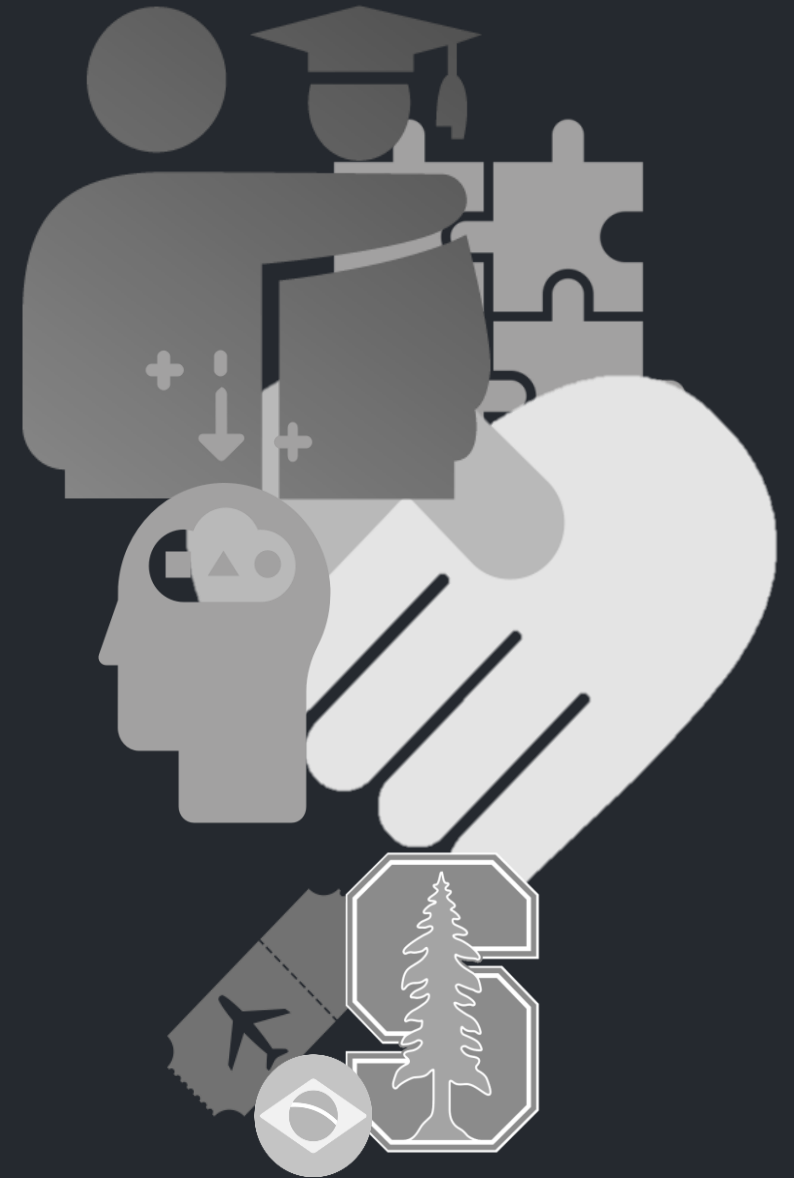
WiDS Datathon 2020

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

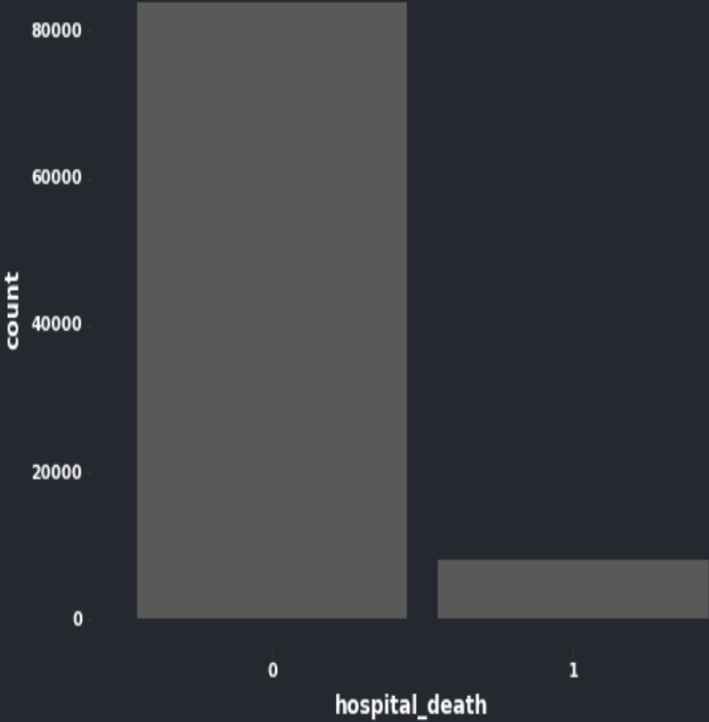
Column Categories

	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

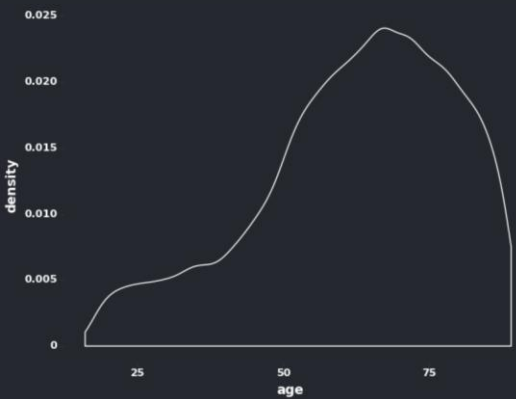
Shape

df.shape (91713, 186)

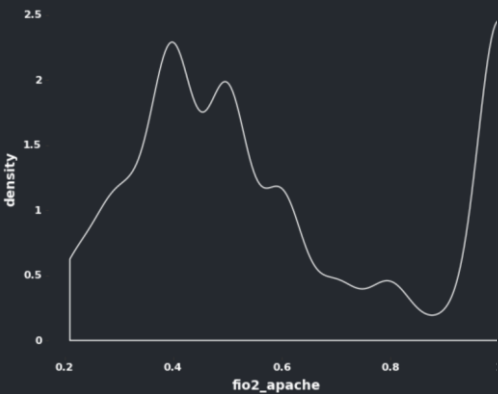
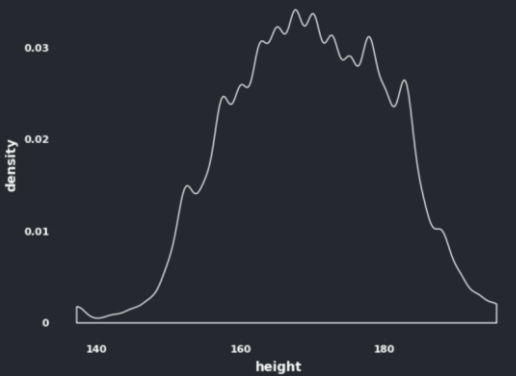
Target Variable



Messiness



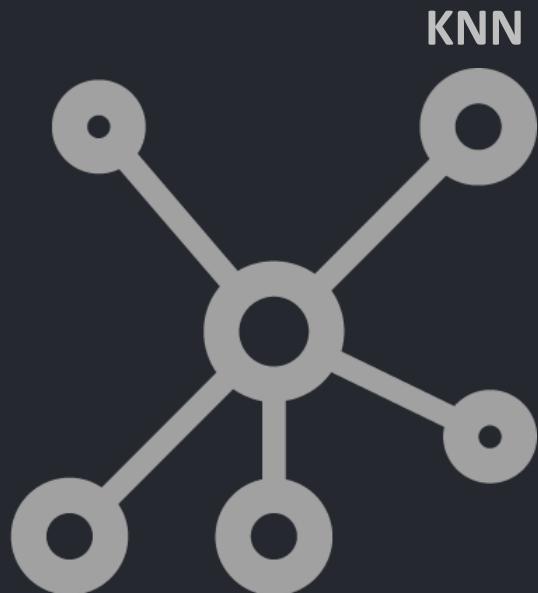
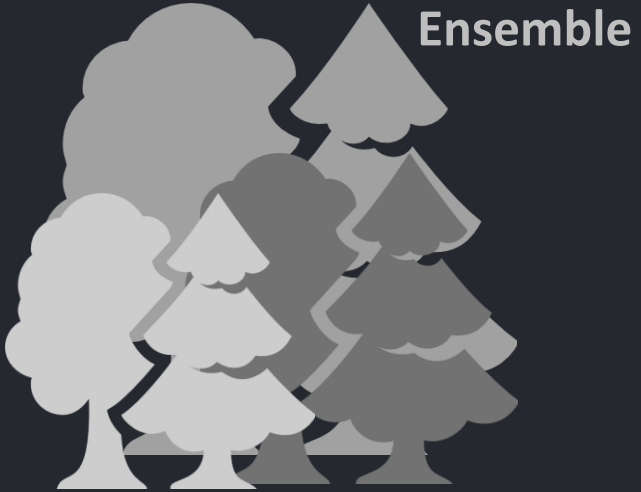
Skews



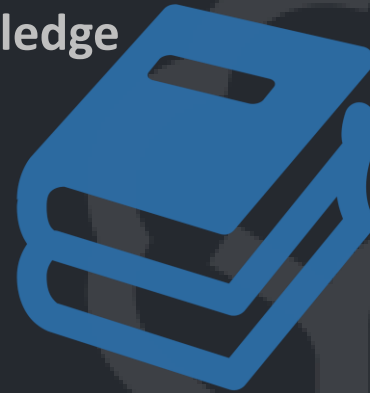
Missing Values



Models



Generalized
Knowledge



Experience/
Case-Based
Knowledge

Baseline

Missing values

Categorical
Columns

Modelling

Score

0.878
760/940

```
DEPENDENT_VAR = getDependentVariable()
catcols = getCategoricalColumns(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])
```

kaggle									
Search									
Overview	Data	Notebooks	Discussion	Leaderboard	Rules	Team	My Submissions		Late Submission
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

Knn Imputation

KDTree

Random Forest Imputation

Variance Based Drop
Column

PCA

Agglomerative Clustering

Random Forest Feature
Importance

Recursive Feature
Elimination

ANOVA

Random Forest

NNet

XGBoost , LGM

Stacking

Logistic Regression

682

▲16



0.89725

1

10d

1

▲1

Women PowerIL

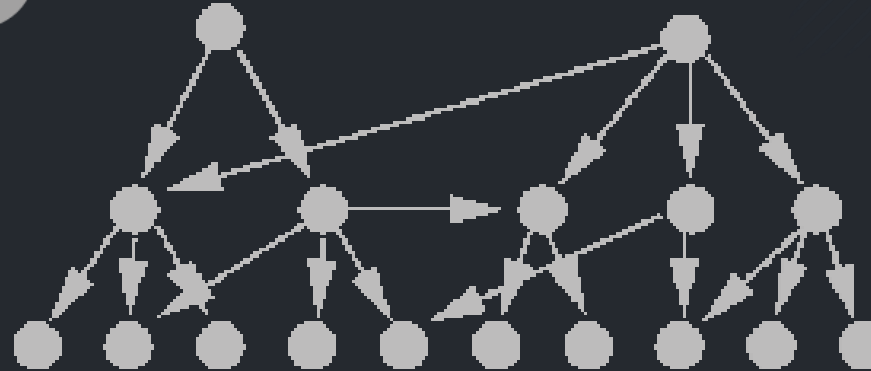
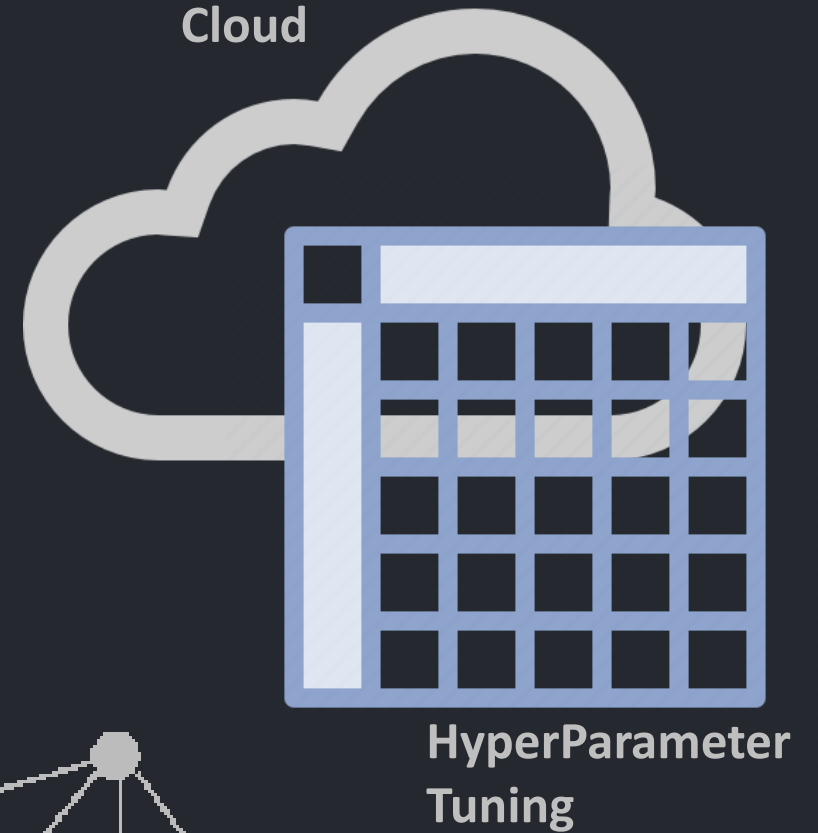
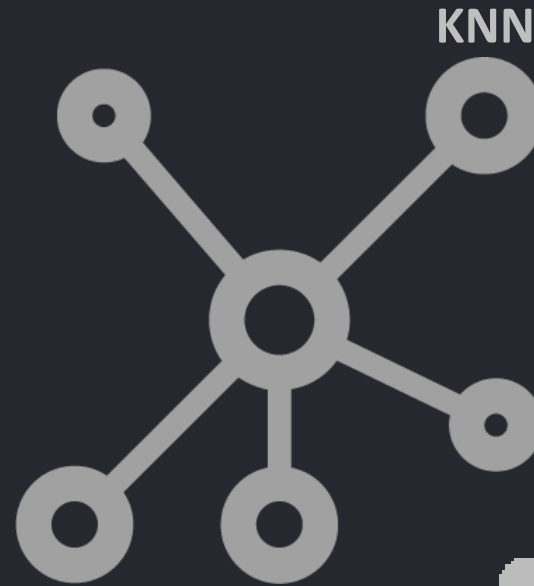
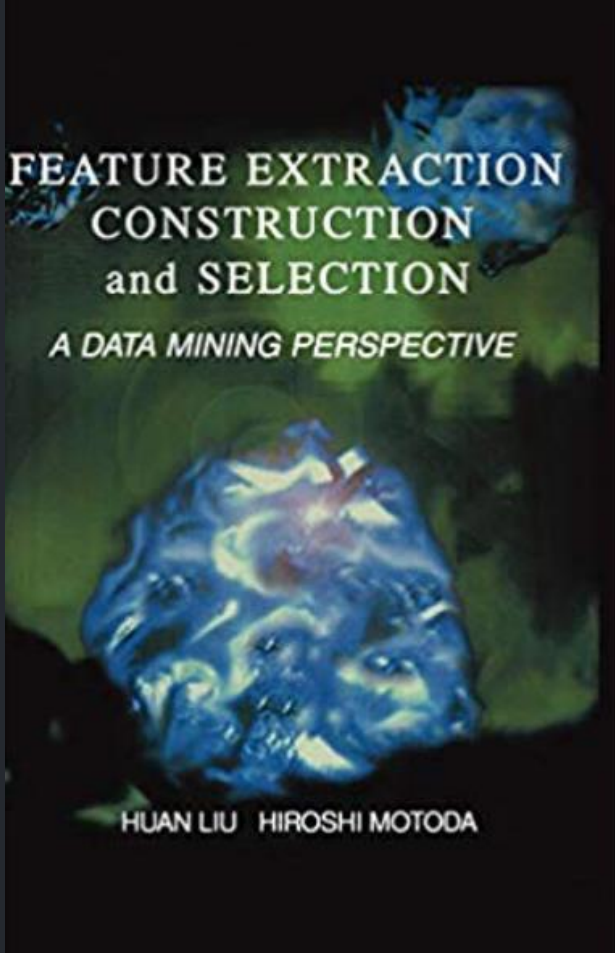


0.91497

205

10d

More to Come



Belief Networks