2023 Datathon

Saturday, October 28, 2023 1:27 PM

Data Traits

timeknown cost reflex sex blood bloodchem1 bloodchem2 temperature race heart psych1 glucose

age sleep dnr bloodchem5 pdeath meals pain primary psych4 disability administratorcost

psych2 dose psych3 bp bloodchem3 confidence bloodchem4 comorbidity totalcost breathing

urine diabetes income extraprimary bloodchem6 education psych5 psych6 information cancer death

Interpolate and Clean Up Data

Any missing values, values with incorrect data types, values that are zero when they shoudn't be, outlier data values, etc. need to be either removed (not preffered) or interpolated (preferred) so the data is more reliable.

Interpolating

We plan to take the median value to replace null/empty/zero values (that shouldn't be zero) rather than the mean as the median provides a more accurate "common" value because the mean is more susceptible to outliers.

Cleaning Data

There are some outliers in the data that yield an unrealistic scenario in which a patient lives/dies, i.e. a 200 yr. old patient living. This data will be labled extraneous and, thus, ignored. We may also replace these values with an interpolated value for that set depending on. after testing, which yields a higher accuracy.

Selecting Most Influential (towards Death) Data Traits

After the DataSet is (mostly) cleaned up, then, using recursive feature Elimination in the sklearn Library, we can run an algorithm ranking the most influential data sets (traits) that effect the deaths of the patients (increasing the accuracy).

Through trial & Error, we try varios combinations of the top 'n' number traits that we ranked above. By preprocessing and then training the Keras Sequential model on the bestcombination of traits to produce the higher accuracy and more consistent results.

Trial and Error Results

Before: 68.4%

Age: 69.7%
Blood: 52.3%
Reflex: 68.1%
BloodChem1: 68.0%
BloodChem2: 67.2%
Psych 1: 68.1%
Glucose: 66%

timeknown
cost
blood chem3
totalcost
pdeath
administrator cost
info
Psych6