Predicting Work Absenteeism

ENGR 121 Final Project
San Jose State University

December 2, 2020

Oscar Alonso
Jeremy Buban
Joanna Rashid

PROBLEM STATEMENT

- Contributing factors to absenteeism
- Increase organizational efficiency
- Reduce absenteeism as measured in hours

OBJECTIVES OF THIS PROJECT

1. Visualize the features associated with workplace absenteeism
2. Prepare the data for to make suitable for various machine learning algorithms
3. Develop and refine several machine learning models to predict employee absenteeism
4. Evaluate model accuracy and interpret results to understand variables that are most strongly associated with work absence.

DATA SOURCE

- São Paolo courier company.
- 21 features
  - São Paolo courier company.
  - 740 observations of employee absences

METHODS (Data Preparation)

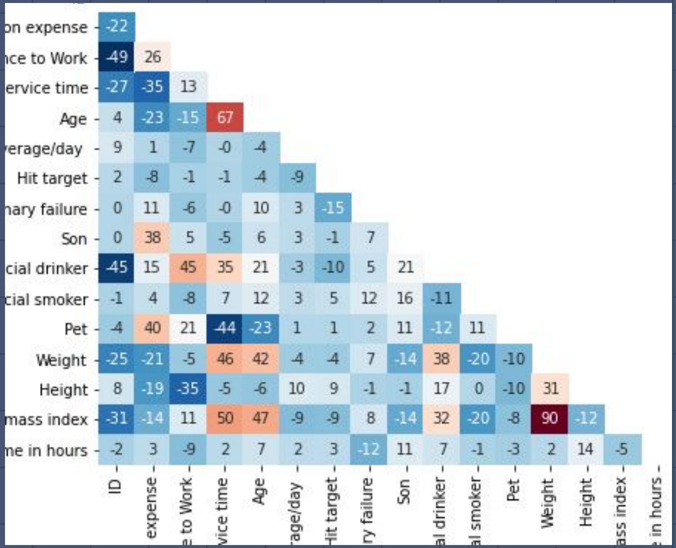
Data Cleaning

This data set was prepared by UCI and is relatively clean. There were no missing values, However, some values needed to be converted to 'bool' or categorical types for use with classifier algorithms.

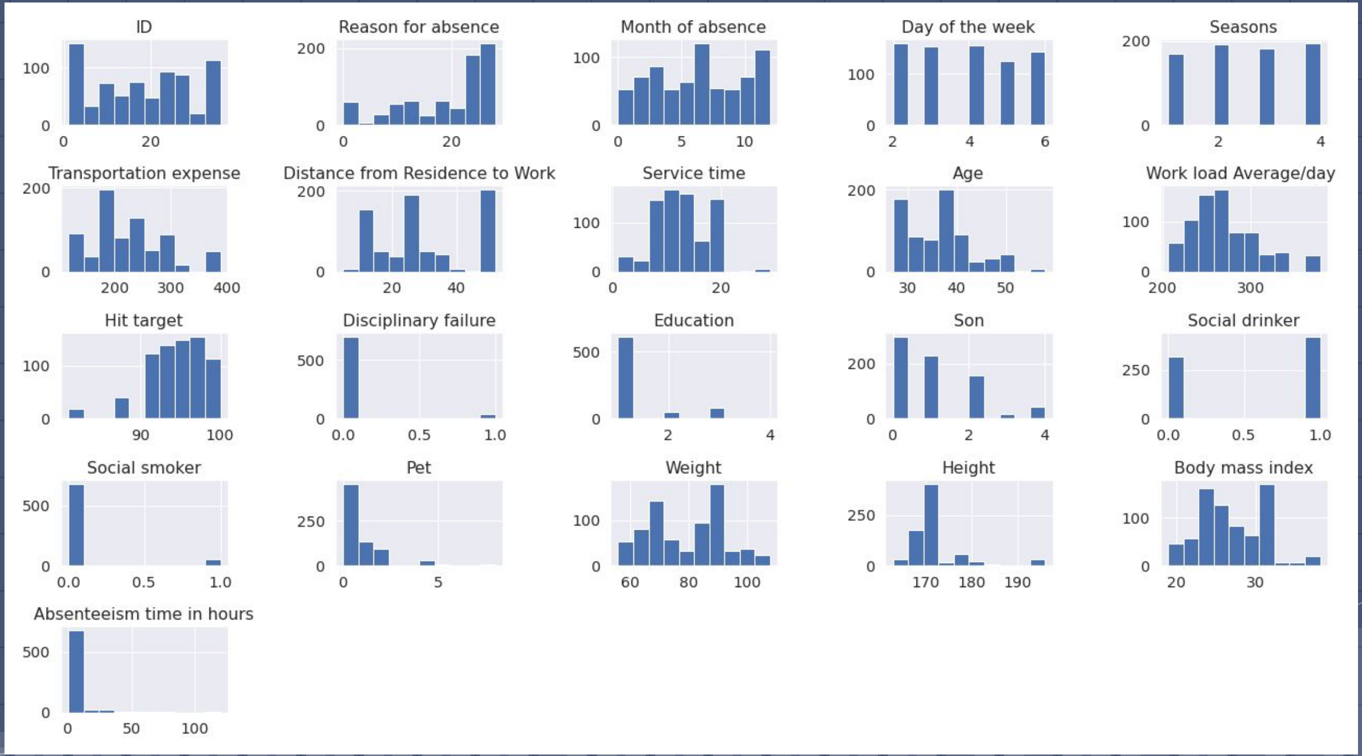
Feature Sorting

Sets were used to group features into numerical, categorical, or explanatory variable groups. This was useful for calling groups of features better suited to classification algorithms, and others better suited to regression algorithms.

Feature Reduction



METHODS (Visualization)



MODELING TOOLS

Hierarchical Clustering

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram).

This tool can be found in the sklearn library.

Linear Regression

This module allows estimation by ordinary least squares (OLS) regression of all non-categorical features on the target variable, absenteeism time.

This tool can be found in the statsmodels library.

Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers the features to predict absenteeism time accurately and control over-fitting.

This tool can be found in the sklearn library.

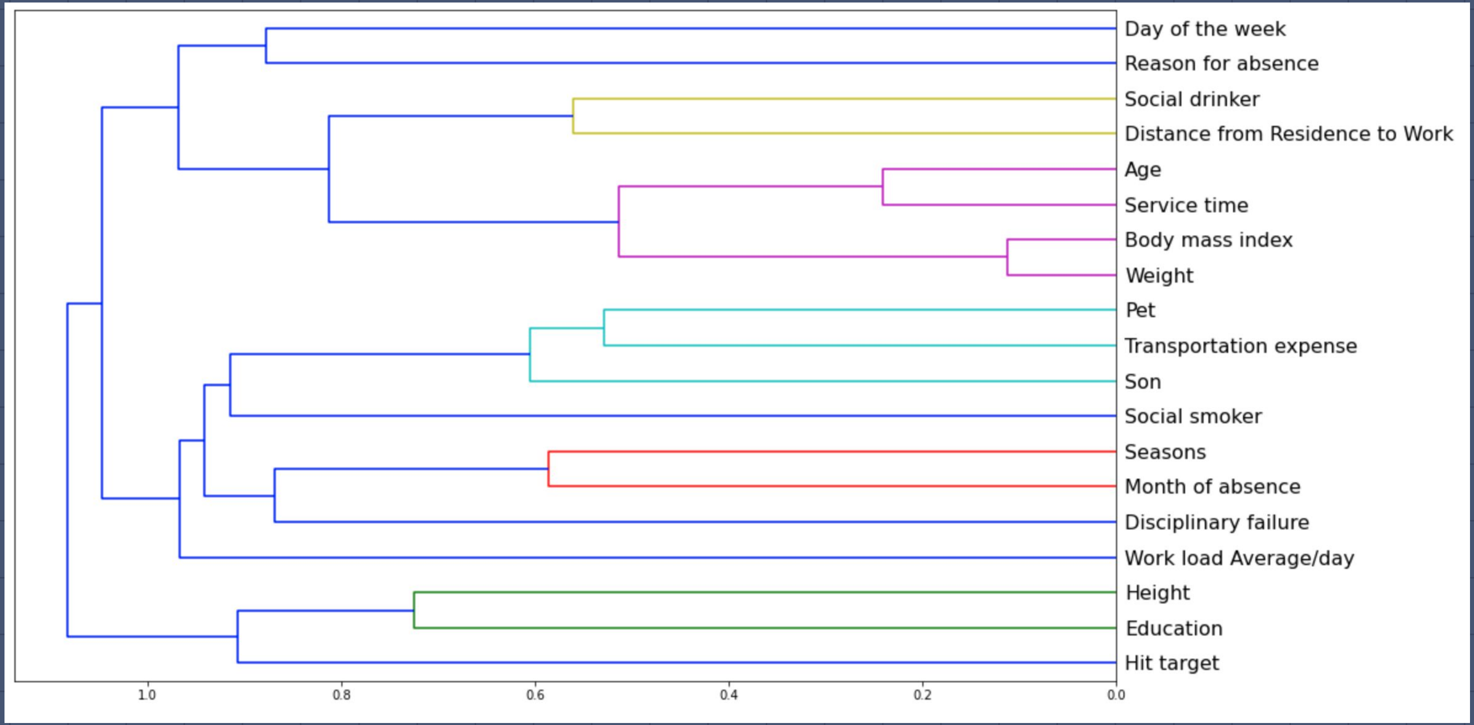
Decision Tree

Decision Trees are a supervised learning method used for classification and regression. This algorithm was is used to predict the target variable by learning simple decision rules inferred from the data features.

This tool can be found in the sklearn library.

RESULTS

Hierarchical Clustering



RESULTS OLS REGRESSION

OLS Regression Results

Dep. Variable:	Absenteeism time in hours	**R-squared (uncentered):**	0.265
Model:	OLS	**Adj. R-squared (uncentered):**	0.252
Method:	Least Squares	**F-statistic:**	20.17
Date:	Wed, 02 Dec 2020	**Prob (F-statistic):**	7.46e-41
Time:	06:27:42	**Log-Likelihood:**	-2940.7
No. Observations:	740	**AIC:**	5907.
Df Residuals:	727	**BIC:**	5967.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Transportation expense	0.0089	0.009	1.019	0.309	-0.008	0.026
Distance from Residence to Work	-0.1170	0.042	-2.794	0.005	-0.199	-0.035
Age	0.1799	0.093	1.940	0.053	-0.002	0.362
Work load Average/day	-0.0025	0.012	-0.200	0.841	-0.027	0.022
Hit target	-0.0852	0.117	-0.731	0.465	-0.314	0.144
Son	0.8264	0.510	1.621	0.105	-0.174	1.827
Pet	0.0335	0.423	0.079	0.937	-0.798	0.865
Height	0.1160	0.065	1.780	0.076	-0.012	0.244
Body mass index	-0.3466	0.143	-2.424	0.016	-0.627	-0.066
Social drinker	2.2830	1.372	1.663	0.097	-0.412	4.977
Social smoker	-1.7576	1.984	-0.886	0.376	-5.653	2.138
Disciplinary failure	-8.7828	2.186	-4.018	0.000	-13.074	-4.492
Education	-1.2734	0.852	-1.494	0.136	-2.947	0.400

OLS Regression Results

Dep. Variable:	Absenteeism time in hours	**R-squared (uncentered):**	0.233
Model:	OLS	**Adj. R-squared (uncentered):**	0.230
Method:	Least Squares	**F-statistic:**	74.77
Date:	Wed, 02 Dec 2020	**Prob (F-statistic):**	3.16e-42
Time:	09:01:51	**Log-Likelihood:**	-2956.4
No. Observations:	740	**AIC:**	5919.
Df Residuals:	737	**BIC:**	5933.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Distance from Residence to Work	-0.0841	0.032	-2.639	0.008	-0.147	-0.022
Son	1.4095	0.442	3.191	0.001	0.542	2.277
Height	0.0466	0.007	7.152	0.000	0.034	0.059

Omnibus:	832.459	**Durbin-Watson:**	1.991
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	46553.444
Skew:	5.560	**Prob(JB):**	0.00
Kurtosis:	40.232	**Cond. No.**	159.

RESULTS Decision Tree

```
Accuracy: 0.44324324324324327
-------------------------
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12
1	0.00	0.00	0.00	24
2	0.35	0.65	0.46	40
3	0.20	0.24	0.22	25
4	0.00	0.00	0.00	16
5	0.00	0.00	0.00	2
8	0.55	0.72	0.62	53
16	0.00	0.00	0.00	3
24	0.00	0.00	0.00	5
32	0.00	0.00	0.00	1
40	0.00	0.00	0.00	2
80	0.00	0.00	0.00	2
accuracy			0.44	185
macro avg	0.18	0.22	0.19	185
weighted avg	0.33	0.44	0.37	185

Decision Tree

- This model was selected because it is useful for both categorical features and those that do not have linear relationships
- Expected higher accuracy given the poor linear model
- Resulting accuracy of 0.44 (Low)
- Improve Random Forest with categorical variables

RESULTS

Random Forest (Continuous Target Variable)

Accuracy: 0.4648648648648649
----

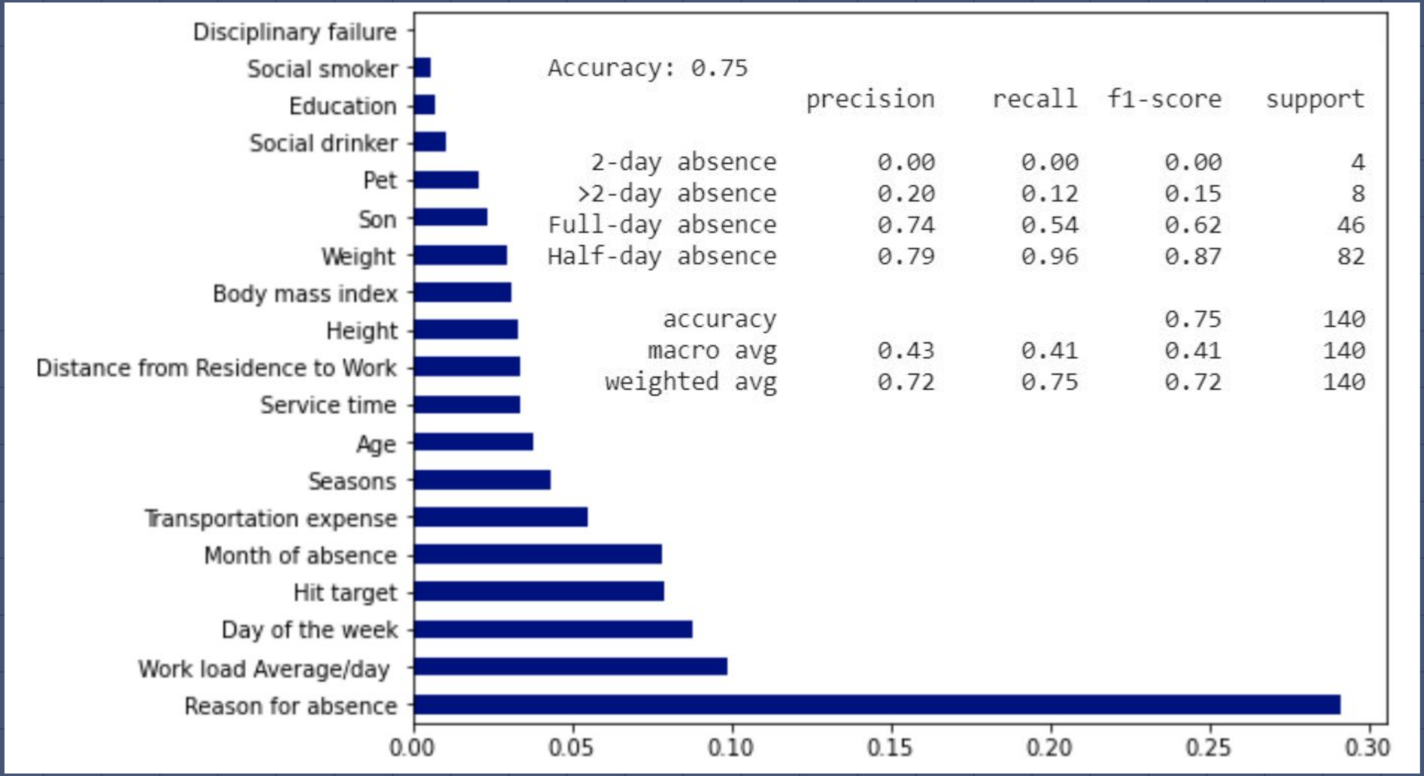
	precision	recall	f1-score	support
0	1.00	1.00	1.00	12
1	0.25	0.21	0.23	24
2	0.42	0.40	0.41	40
3	0.39	0.36	0.37	25
4	0.30	0.19	0.23	16
5	0.00	0.00	0.00	2
8	0.55	0.77	0.65	53
16	0.00	0.00	0.00	3
24	0.00	0.00	0.00	5
32	0.00	0.00	0.00	1
40	0.00	0.00	0.00	2
64	0.00	0.00	0.00	0
80	0.00	0.00	0.00	2
accuracy			0.46	185
macro avg	0.22	0.23	0.22	185
weighted avg	0.43	0.46	0.44	185

Random Forest

- Similar to Decision Tree, good with nonlinear categorical variables
- Expected higher accuracy given the poor linear model
- Resulting accuracy of 0.46 (Low)
- Improve Random Forest with categorical variables

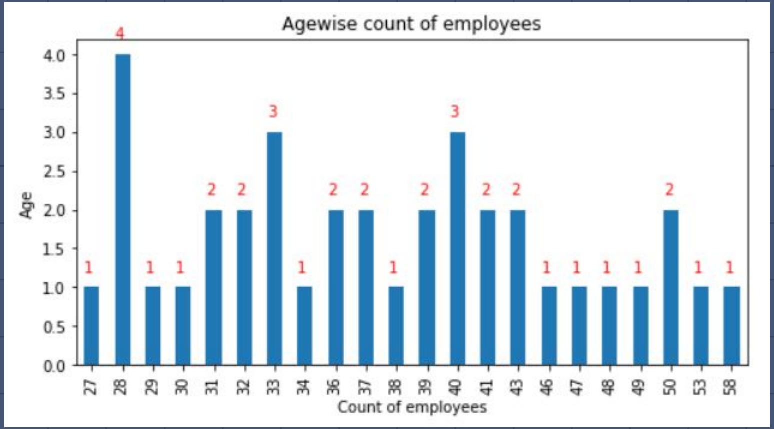
RESULTS

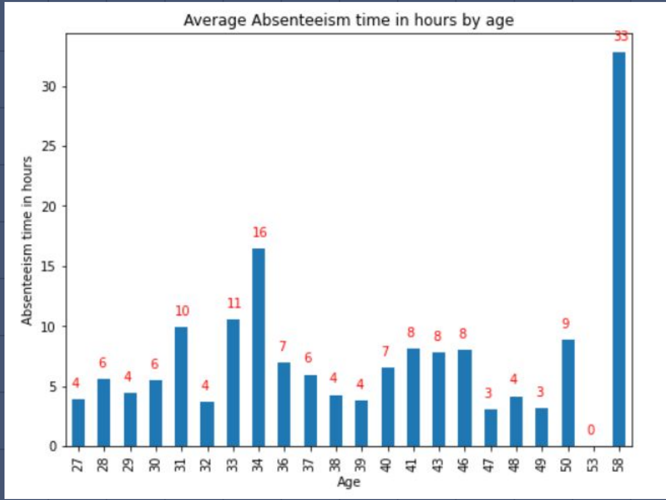
Random Forest Classification (Categorical Target Variable)

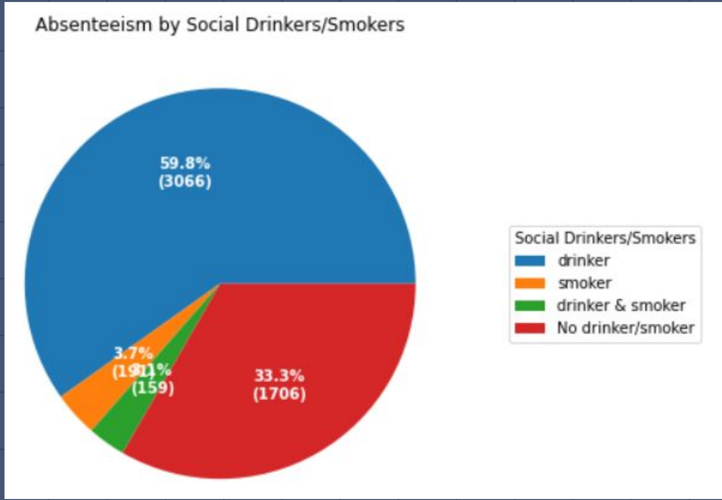


Analysis

- Important Features/Patterns
- Related Variables



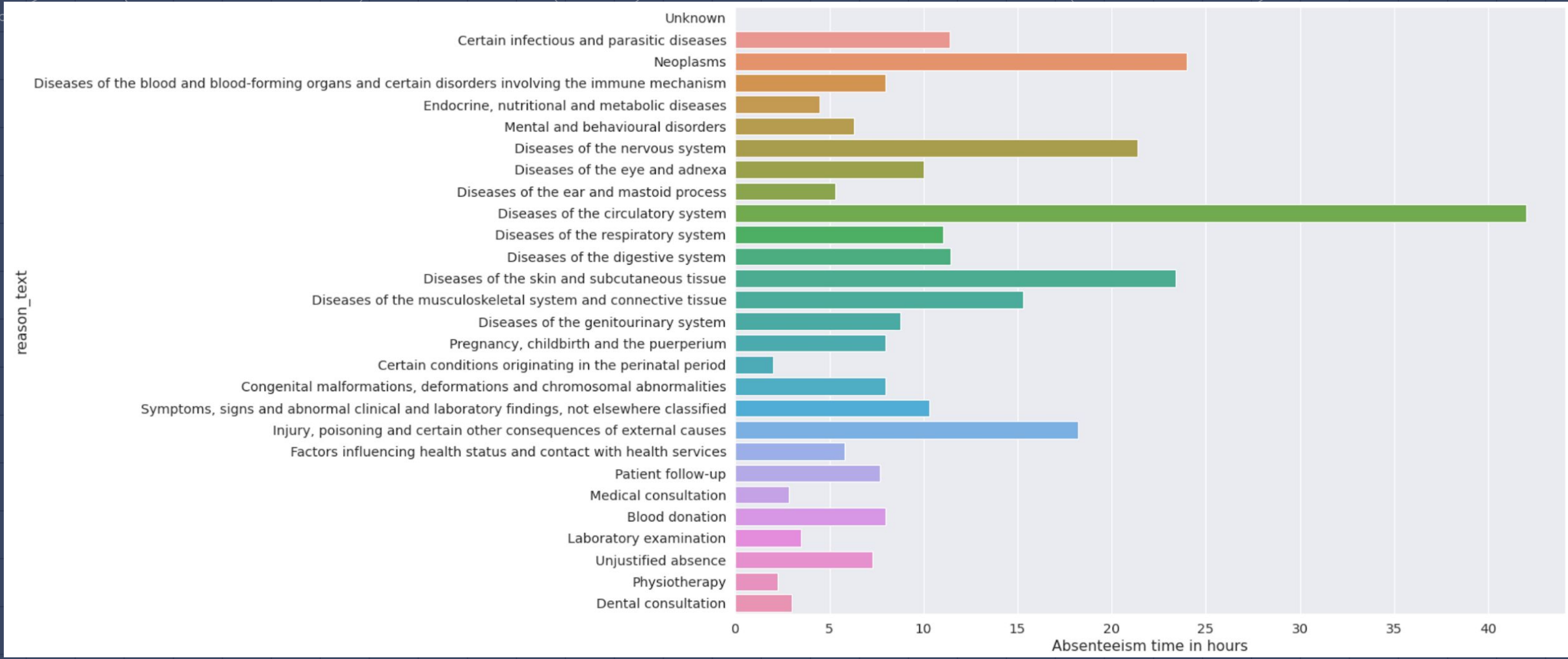




MODEL COMPARISON

	ACCURACY
Linear Regression	0.265066
Decision Tree	0.443243
Random Forest (continuous)	0.448649
Random Forest (categorical)	0.757143

The Random Forest model tells us that "Reason for absence" is the most important feature in the model. Some data visualization on this feature can help to clarify the cause of absenteeism.



CONCLUSION

- Variables contributing to absenteeism
- Interesting findings
- Continuous vs Categorical in Random Forest
- Why Linear Regression doesn't work as well
- Further Improvements