Damon Lepine, Andrew Woloszyn 469 Final Project

Introduction: The problem we are trying to solve is identifying diseases given a person's symptoms. If solved, we will be able to more accurately predict diseases given fewer symptoms from a patient.

Formulation: The data mining task it can be formulated to is a mix of both Classification and Recommendation. The input we are taking in is a list of records which contains true/false flags for which symptoms they produce, the name of the disease, and the count of the disease in the dataset. Our expected output is a decision tree that will lead one to a disease given all of the appropriate symptoms from the patient along with the accuracy in which are trained model can predict a disease.

Datasets: The dataset we used was acquired from the following page: http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html
It contains a listing of diseases, the symptoms they present, and the number of patients who were diagnosed with each disease over the course of one year at a hospital. To format this data into something we could train our classifier on, we made each column of our dataset represent a symptom, with each row representing a patient. The data in each row is an array of 1s and 0s (1 means the symptom presents with the disease, and 0 means it does not). In the final column, we have the disease that the patient was diagnosed with. We add a row for each patient, according to the number of times each disease was diagnosed. Finally, since every disease diagnosis has the exact same set of symptoms, we slightly randomize the data for more realistic training. For every symptom that a patient has, there is a 5% chance that the patient does not report it (1 gets flipped to 0), and for every symptom that the disease doesn't have, there is a 1% chance the patient reports it (0 gets flipped to 1). This is to simulate the difficulty in diagnosing patients with irrelevant symptoms or symptoms that go unnoticed.

The dataset contains:

- 134 diseases
- 402 unique symptoms
- 33,724 patient records

Algorithm: We applied the Decision Tree algorithm to this dataset.

Experiments: We randomly selected 80% of the data to train the Decision Tree classifier, then used the remaining 20% of the data to test the predictions that it makes. As a result, we obtained a model with 0.9 F1 score on testing. While this is not accurate enough to perfectly diagnose people without doctor intervention, it is accurate enough to

give doctors a starting point to look at when diagnosing somebody. If we create a system where patients can list all their symptoms and get a preliminary diagnosis, they can bring that to their doctor and save time, ultimately leading to a speedier diagnosis and potentially a higher chance of recovery.

Full testing results:

	precision	recall	f1-score	support
Alzheimer's_disease	0.81	0.62	0.70	21
Pneumocystis_carinii_pneumonia	0.84	0.76	0.80	21
accident_cerebrovascular	0.93	0.98	0.95	165
acquired_immuno-deficiency_syndrome^HIV^hiv_infections	0.90	0.90	0.90	73
adenocarcinoma	0.97	0.83	0.90	36
adhesion	0.86	0.43	0.57	14
affect_labile	0.31	0.44	0.36	9
anemia	0.98	0.98	0.98	115
anxiety_state	0.86	0.95	0.90	83
aphasia	0.58	0.64	0.61	11
arthritis	0.75	0.86	0.80	28
asthma	0.95	0.98	0.97	166
bacteremia	0.96	0.88	0.92	26
benign_prostatic_hypertrophy	0.76	0.76	0.76	25
bipolar_disorder	0.90	0.96	0.93	47
bronchitis	0.96	0.96	0.96	48
candidiasis^oral_candidiasis	0.83	0.91	0.87	22
carcinoma	0.90	0.95	0.92	58
cardiomyopathy	0.98	0.92	0.95	65
cellulitis	0.89	0.95	0.92	60
cholecystitis	0.73	0.73	0.73	11
cholelithiasis^biliary_calculus	0.58	0.78	0.67	18
chronic_alcoholic_intoxication	0.67	0.67	0.67	9
chronic_kidney_failure	0.91	0.98	0.95	63
chronic_obstructive_airway_disease	0.97	0.96	0.97	109
cirrhosis	0.87	0.85	0.86	46
colitis	0.81	0.61	0.69	28
confusion	0.98	0.95	0.96	86
coronary_arteriosclerosis^coronary_heart_disease	0.97	0.98	0.98	257
decubitus_ulcer	0.21	0.50	0.30	6
deep_vein_thrombosis	0.92	0.86	0.89	69
degenerative_polyarthritis	0.90	0.95	0.92	96
deglutition_disorder	0.95	0.83	0.88	23
dehydration	0.95	0.97	0.96	61
delirium	0.75	0.60	0.67	15
delusion	0.71	0.83	0.77	6
dementia	0.91	0.94	0.92	102
dependence	1.00	0.93	0.96	14
depression_mental^depressive_disorder	0.95	0.97	0.96	265
diabetes	0.99	0.96	0.97	279
diverticulitis	0.96	1.00	0.98	27
diverticulosis	0.94	0.79	0.86	19
edema_pulmonary	1.00	0.88	0.94	34
effusion_pericardial^pericardial_effusion_body_substance	0.86	1.00	0.92	12
embolism_pulmonary	0.81	0.87	0.84	45
emphysema_pulmonary	1.00	0.73	0.84	11
encephalopathy	0.52	0.60	0.56	25
endocarditis	0.44	0.50	0.47	14
epilepsy	0.91	0.96	0.94	75
exanthema	0.81	0.85	0.83	41
failure_heart	0.91	0.94	0.92	31

failure_heart_congestive	0.91	0.96	0.94	198
failure_kidney	0.77	0.61	0.68	33
fibroid_tumor	0.95	0.86	0.90	21
gastritis	0.86	0.86	0.86	21
gastroenteritis	0.78	0.67	0.72	21
gastroesophageal_reflux_disease	0.86	0.93	0.89	58
glaucoma	0.63	0.71	0.67	17
gout	0.65	0.80	0.71	25
hemiparesis	0.75	0.70	0.72	30
hemorrhoids	0.45	0.50	0.48	10
hepatitis	0.78	0.67	0.72	27
hepatitis_B	0.80	0.91	0.85	22
hepatitis_C	0.84	0.91	0.87	53
hernia	0.96	0.77	0.85	30
hernia_hiatal	0.46	0.55	0.50	11
hyperbilirubinemia	0.41	0.44	0.42	16
hypercholesterolemia	0.96	0.97	0.97	110
hyperglycemia	0.88	0.71	0.79	21
hyperlipidemia	0.90	0.83	0.86	52
hypertension_pulmonary	0.79	0.79	0.79	29
hypertensive_disease	0.99	0.99	0.99	662
hypoglycemia	0.74	0.74	0.74	23
hypothyroidism	0.99	0.89	0.94	74
ileus	0.33	0.47	0.39	15
incontinence	0.75	0.93	0.83	29
infection	0.98	0.97	0.97	129
infection_urinary_tract	0.97	0.99	0.98	142
influenza	0.95	0.95	0.95	20
insufficiency_renal	0.96	0.93	0.95	88
ischemia	0.94	0.92	0.93	48
ketoacidosis_diabetic_	0.52	1.00	0.69	12
kidney_disease	0.94	0.80	0.86	20
kidney_failure_acute	0.95	0.83	0.89	47
lymphatic_diseases	0.81	0.92	0.86	38
lymphoma	0.62	0.71	0.67	21
malignant_neoplasm_of_breast^carcinoma_breast	0.94	0.89	0.92	37
malignant_neoplasm_of_lung^carcinoma_of_lung	0.59	0.67	0.62	15
malignant_neoplasm_of_prostate^carcinoma_prostate	0.87	0.83	0.85	24
malignant_neoplasms	0.60	0.56	0.58	16
malignant_neoplasms^primary_malignant_neoplasm	0.98	0.84	0.90	94
malignant_tumor_of_colon^carcinoma_colon	0.73	0.80	0.76	20
manic_disorder	0.75	0.55	0.63	22
melanoma	0.47	0.64	0.55	14
migraine_disorders	1.00	0.80	0.89	15
mitral_valve_insufficiency	0.77	0.89	0.83	38
myocardial_infarction	0.98	0.97	0.98	141
neoplasm	0.88	0.86	0.87	71
neoplasm_metastasis	0.67	0.60	0.63	20
neuropathy	0.95	0.95	0.95	22
neutropenia	0.86	0.60	0.71	10
obesity	0.83	0.91	0.87	55
obesity_morbid	0.94	0.89	0.91	18
osteomyelitis	0.87	0.74	0.80	27
osteoporosis	0.89	0.94	0.92	36
overload_fluid	1.00	1.00	1.00	25
pancreatitis	0.86	0.76	0.81	25
pancytopenia	0.53	0.80	0.64	10
paranoia	0.87	0.75	0.81	36
parkinson_disease	0.83	0.79	0.81	19
paroxysmal_dyspnea	0.93	0.83	0.88	30
peripheral_vascular_disease	0.88	0.91	0.90	33
personality_disorder	0.74	0.78	0.76	18

pneumonia	0.98	0.94	0.96	187
pneumonia_aspiration	0.67	0.82	0.73	22
pneumothorax	0.91	0.59	0.71	17
primary_carcinoma_of_the_liver_cells	0.76	0.59	0.67	22
psychotic_disorder	0.88	0.84	0.86	45
pyelonephritis	0.82	0.93	0.87	15
respiratory_failure	0.88	0.62	0.73	24
schizophrenia	0.88	0.92	0.90	25
septicemia^systemic_infection^sepsis_(invertebrate)	0.90	0.85	0.88	66
sickle_cell_anemia	0.96	0.90	0.93	29
spasm_bronchial	0.75	0.64	0.69	14
stenosis_aortic_valve	0.76	0.81	0.79	32
suicide_attempt	0.95	0.95	0.95	20
tachycardia_sinus	0.89	0.89	0.89	9
thrombocytopaenia	0.93	0.68	0.79	19
thrombus	0.57	0.57	0.57	14
tonic-clonic_epilepsy^tonic-clonic_seizures	1.00	0.65	0.79	20
transient_ischemic_attack	0.97	0.97	0.97	37
tricuspid_valve_insufficiency	0.80	0.89	0.84	18
ulcer_peptic	0.86	0.89	0.88	28
upper_respiratory_infection	0.93	0.93	0.93	28
accuracy			0.90	6745
macro avg	0.83	0.82	0.82	6745
weighted avg	0.91	0.90	0.90	6745