Symptom Smart

Predictive Disease Diagnosis Using Machine Learning

Presented by: James Harris

Program: University of Houston Downtown – MSDA

Date: May 2nd, 2025

Physician Burnout

A long term stress reaction

- Emotional exhaustion
- Depersonalization
- Decreased feeling of personal achievement

Burnout Factors

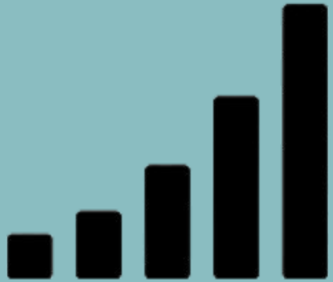
Image sourced from the Agency for Healthcare Research and Quality

Impetus of Project

- Provides the ability to spend more qualitative time with patients
- Predictive models can support evidence-based decisions
- Early diagnosis can significantly improve treatment outcomes

Project Overview

Develop a model that accurately classifies diseases



Improving diagnostic speed and accuracy

Excel, R, Ripper Rule Learner, Decision Trees & Random Forest Classifiers

Dataset Overview

- Data was sourced **Kaggle** – "*Disease Prediction from Symptoms*"
- 4920 entries
  - 131 symptoms
  - 41 diseases
- Each row represents a patient

Disease	Symptom_1	Symptom_2	Symptom_3
Fungal infection	itching	nodal_skin_eruptions	dischromic _patches
Fungal infection	itching	skin_rash	dischromic _patches
Fungal infection	itching	skin_rash	nodal_skin_eruptions
Fungal infection	itching	skin_rash	nodal_skin_eruptions
Allergy	continuous_sneezing	shivering	chills
Allergy	shivering	chills	watering_from_eyes
Allergy	continuous_sneezing	chills	watering_from_eyes
Allergy	continuous_sneezing	shivering	watering_from_eyes

Data Cleaning & Model Preparation

With Excel:

- Removed unnecessary spaces and creating consistency
- Corrected spelling issues
- Created dummy variables

With R:

- Exploratory data analysis
- Created factors from text
- Created training and test dataset
- Test models with mock patient

Exploratory Data Analysis

Exploratory Data Analysis (cont.)

Chicken pox	Chronic cholestasis
120	120
Common Cold	Dengue
120	120
Diabetes	Dimorphic hemmorhoids(piles)
120	120
Drug Reaction	Fungal infection
120	120
Gastroenteritis	GERD
120	120
Heart attack	hepatitis A
120	120
Hepatitis B	Hepatitis C
120	120
Hepatitis D	Hepatitis E
120	120

fatigue	vomiting	high_fever
1933	1915	1363
loss_of_appetite	nausea	headache
1153	1147	1135
abdominal_pain	yellowish_skin	yellowing_of_eyes
1032	913	817
chills		
798		

Data Preprocessing & Model Splitting

- Converted symptom text into binary features (1,0)
- Partitioning used to preserve proportions
  - 75/25 split

Disease	abdominal_pain	abnormal_menstruation	acidity
Alcoholic hepatitis	1	0	0
Alcoholic hepatitis	1	0	0
Alcoholic hepatitis	1	0	0
Alcoholic hepatitis	1	0	0
Alcoholic hepatitis	1	0	0
Alcoholic hepatitis	1	0	0
Alcoholic hepatitis	1	0	0
Alcoholic hepatitis	1	0	0
Alcoholic hepatitis	1	0	0
Alcoholic hepatitis	1	0	0
Alcoholic hepatitis	1	0	0

Model Training - Ripper Rule Learner

```
Number of Rules : 68
```

```
=== Summary ===

Correctly Classified Instances        3674               99.5664 %
Incorrectly Classified Instances        16                0.4336 %
Kappa statistic                          0.9956
Mean absolute error                      0.0004
Root mean squared error                  0.0142
Relative absolute error                  0.8524 %
Root relative squared error              9.2326 %
Total Number of Instances             3690
```

Model Training – Decision Tree

```
Call:
C5.0.default(x = symptom_train[-1], y = symptom_train$Disease)

Classification Tree
Number of samples: 3690
Number of predictors: 131

Tree size: 74
```

Model Training – Random Forest

```
Call:
 randomForest(formula = Disease ~ ., data = symptom_train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 11

        OOB estimate of  error rate: 0%
```

Attribute Usage

- Useful for feature selection
- Most discriminative feature is "**abnormal_menstruation**"
- Helps explainability

```
Attribute usage:

100.00% abnormal_menstruation
 95.12% muscle_pain
 85.47% malaise
 75.69% excessive_hunger
 68.92% yellowing_of_eyes
 68.59% chills
 63.77% family_history
 51.95% muscle_weakness
 49.65% high_fever
 49.57% yellow_crust_ooze
 47.40% fatigue
```

Models Performance with Test Data

Metric	Ripper Rule Learner	Decision Tree	Random Forest
Accuracy	99.2%	99.6%	100%
Precision	96% or higher	96% or higher	100%
Recall (Sensitivity)	90% or higher	93% or higher	100%
Specificity	99% or higher	99% or higher	100%
F1 Score	94% or higher	96% or higher	100%
Kappa	99.2%	99.6%	100%

Mock Patient

- **Input:** [‘abdominal_pain’, ‘chills’, ‘continuous_sneezing’, ‘fever’, ‘shivering’, ‘watering_from_eyes’, ‘cough’]
- Created new data frame and set active features

```
# Turn on the symptoms the patient has
mock_case$abdominal_pain        <- 1
mock_case$chills                <- 1
mock_case$fever                 <- 1
mock_case$continuous_sneezing   <- 1
mock_case$shivering             <- 1
mock_case$watering_from_eyes    <- 1
mock_case$cough                 <- 1
```

Model Predictions

Ripper Learner

```
> cat("Predicted disease:", ripper_pred_name, "\n")
Predicted disease: Chronic cholestasis
> cat("Confidence:", round(ripper_confidence * 100, 4), "%\n")
Confidence: 26 %
```

Decision Tree

```
> cat("Predicted Disease:", pred_name, "\n")
Predicted Disease: Allergy
> cat("Model Confidence:", round(confidence * 100, 2), "%\n")
Model Confidence: 98.84 %
```

Random Forest

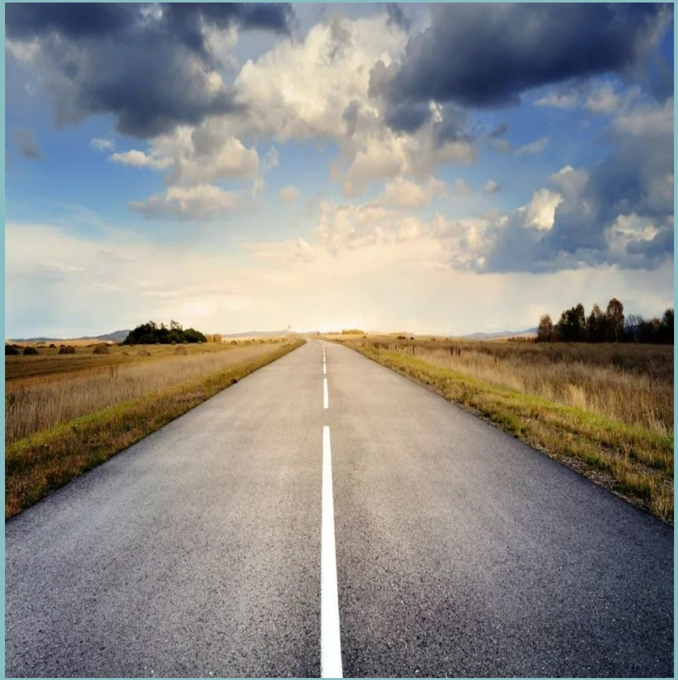
```
> cat("Predicted disease:", rf_pred_name, "\n")
Predicted disease: Allergy
> cat("Confidence:", round(rf_confidence * 100, 4), "%\n")
Confidence: 82 %
```

Key Insights and Limitations

- Equal distribution of diseases
- Lack of clinical context
- Equal weight of symptoms
- Single disease assumption
- We are not the same

Looking Ahead

- Integration of real-world clinical data
- The potential to aggregate trends globally
- Doctors can now spend more quality time with patients
- Ability to incorporate other ML techniques



Thank You