Università degli Studi di Milano
Data Science and Economics (LM-91)

Early-Stage Diabetes Risk Prediction:

which physiological symptoms to watch out for?

Shihab Hamati

Nov 3, 2022

Objective

Diabetes is major and growing global problem

- Chronic illness which is the result of the body **not correctly producing or utilizing insulin**, the blood glucose regulating hormone

- About **1 out of every 11 to 12 adults** suffer from it

- Direct cause of death of **1.5 million** + **half a million** indirect kidney disease death + **one-fifth** of cardiovascular deaths

- 3% increase in diabetes related deaths (2000-19), **higher in lower-middle-income countries, at 13%**

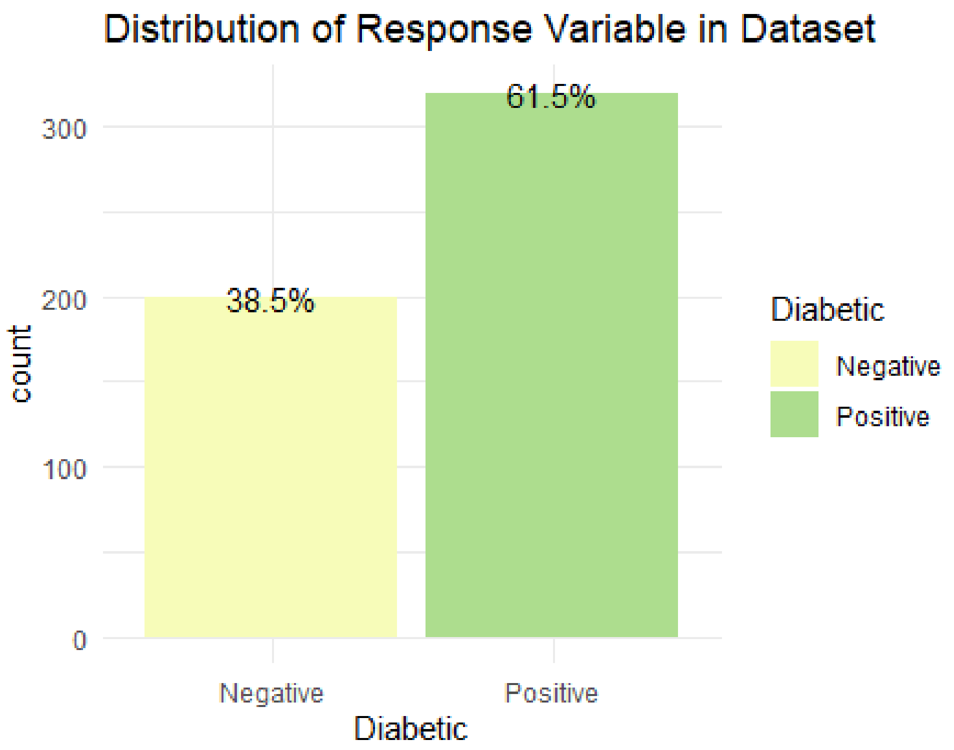
The goal is to predict the diabetic status

- 520 records and 17 features:
  - 1 numeric feature
  - 15 binary categorical features
  - 1 binary categorical response variable

- 61.5% of the records are diabetic
  (hospitals have disproportionate access to ill patients)

- This can be corrected for during production, e.g. in logistic regression:

$$\hat{\beta}_0^* = \hat{\beta}_0 + log\frac{\pi}{1-\pi} - log\frac{\tilde{\pi}}{1-\tilde{\pi}}$$

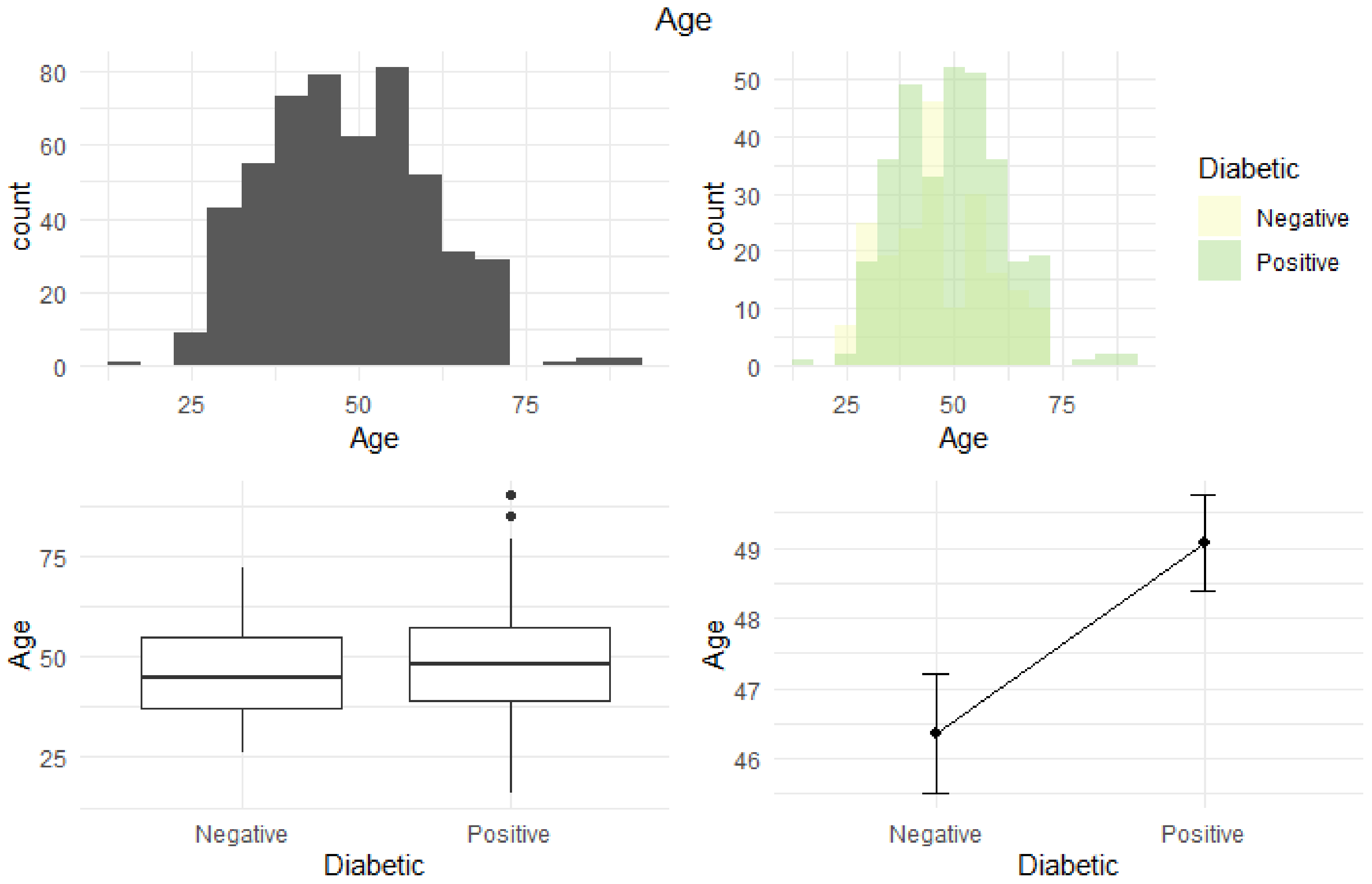


Exploratory Data Analysis

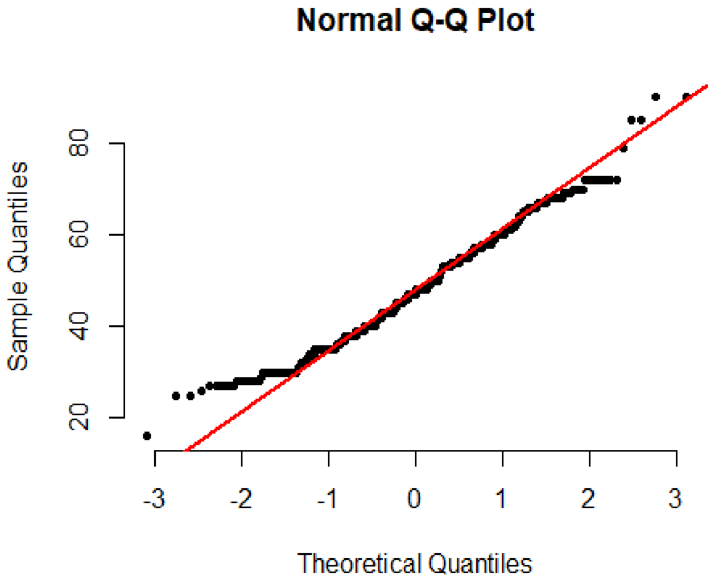
Dataset contains physiological symptoms

- **Overall Characteristic Traits:** Age, Gender, Sudden Weight Loss
- **Commonly Named Symptoms:** Weakness, Muscle Stiffness, Visual Blurring, Itching, Irritability, Delayed Healing
- **Medically Named Symptoms:**
  - Polyuria: excessive urination either in frequency or volume
  - Polydipsia: excessive thirst
  - Polyphagia: excessive eating
  - Paresis: muscular weakness, partial in this case
  - Alopecia: bodily hair loss
  - Genital Thrush: a fungal infection in the genitals

Diabetics are approximately 2.5 years older

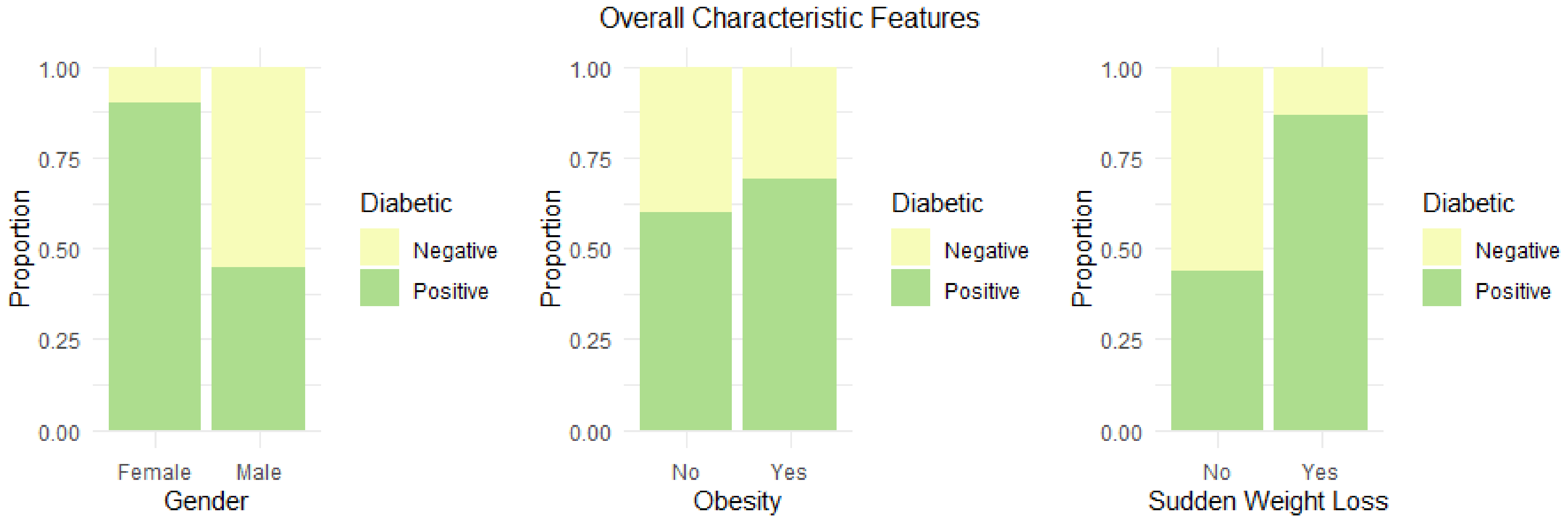


- ANOVA test indicates the mean age in both groups is **significantly different**
- The age variable does not pass the Shapiro test for normality, but is **acceptably bell-shaped**

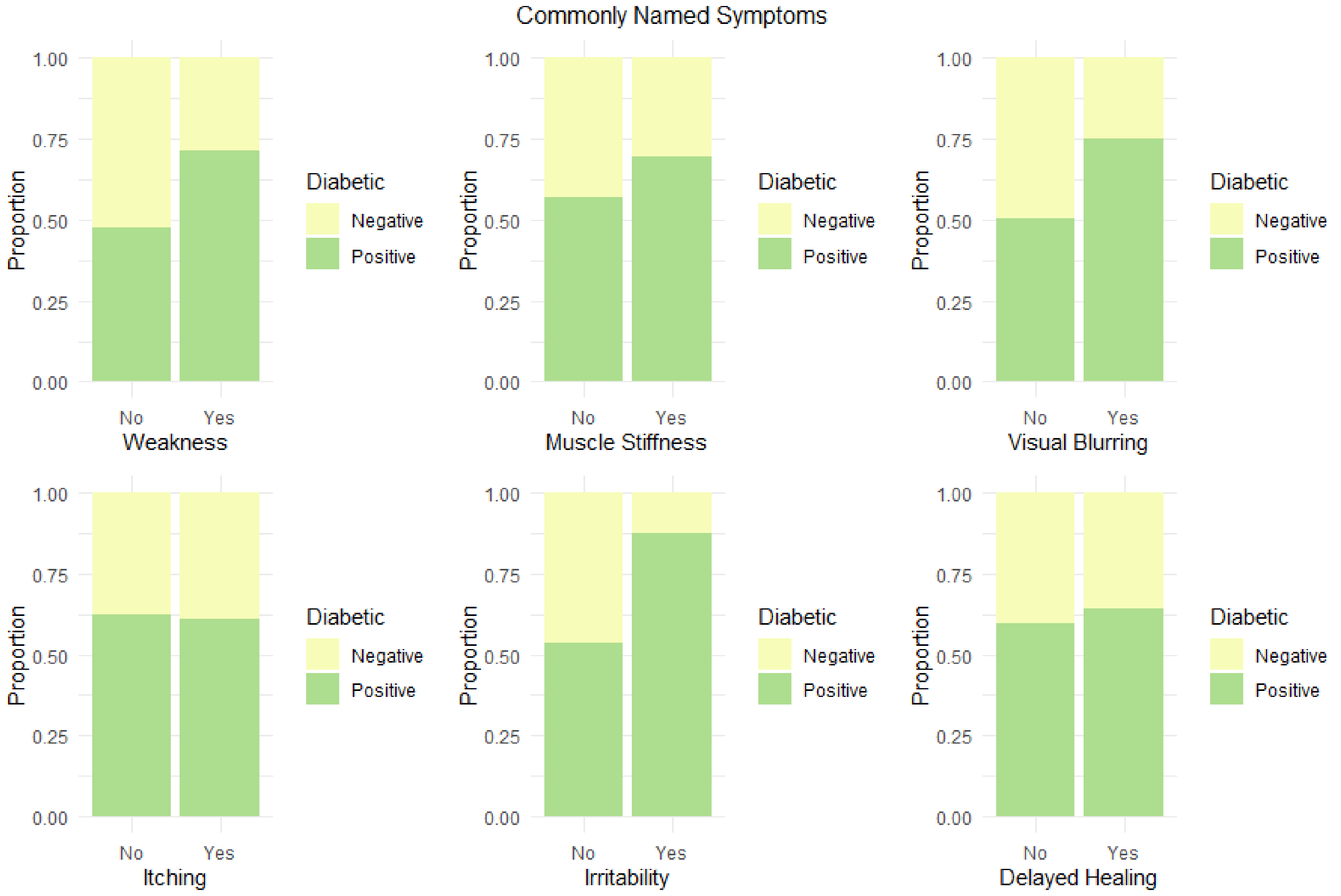


Obesity and sudden weight loss in diabetics

- In line with common medical knowledge, diabetics are **more likely to be obese**
- Also, **sudden and unintentional weight loss is more common** in diabetics

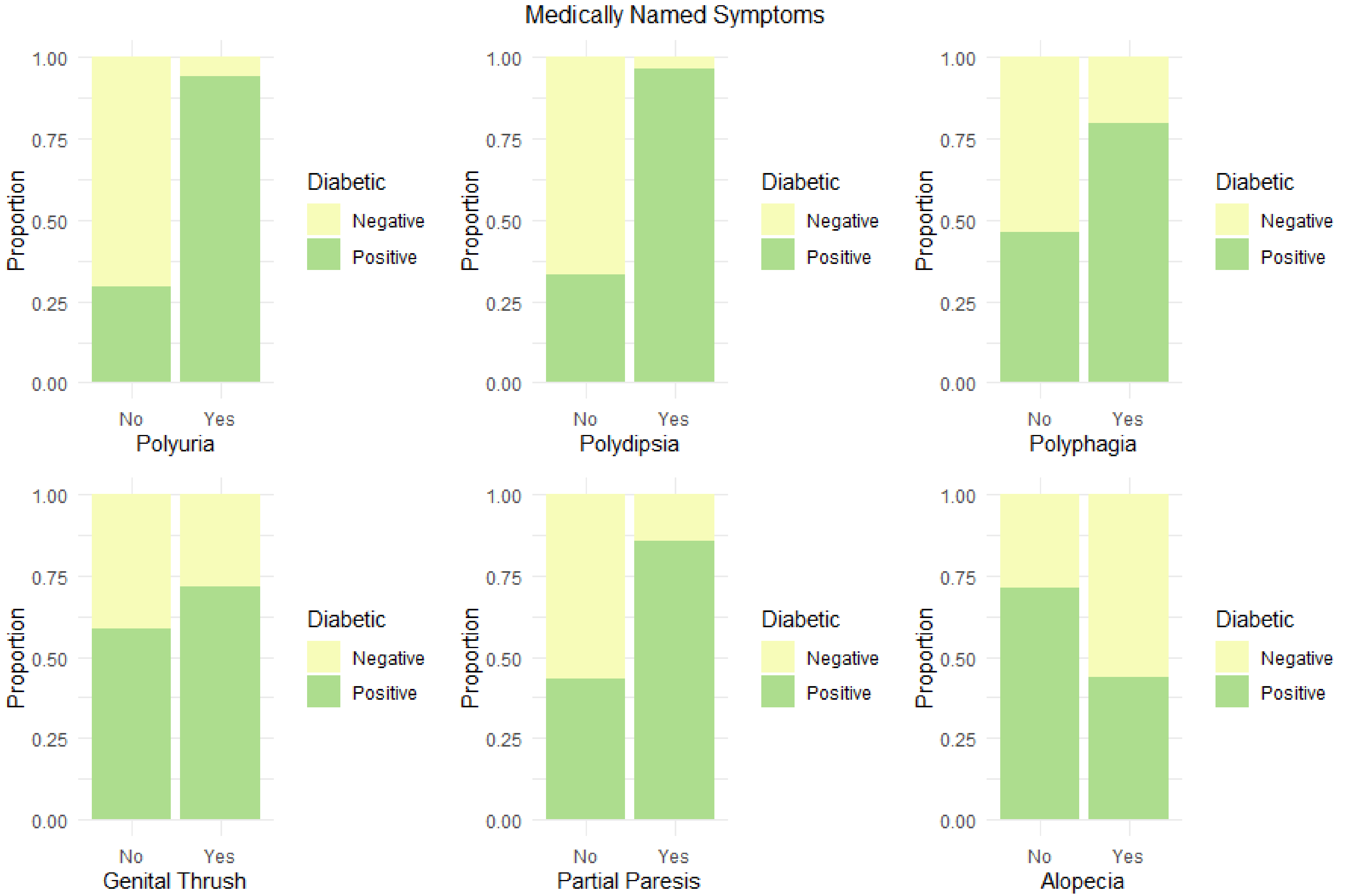


Weakness, stiffness, blurring, and irritability are more present in diabetics



- Diabetic patients are **more likely** to present signs of weakness, muscle stiffness, visual blurring, and irritability
- Itching and delayed healing **do not appear to be significantly different** across the two groups

Excessive urination and thirst are the most different symptom across classes

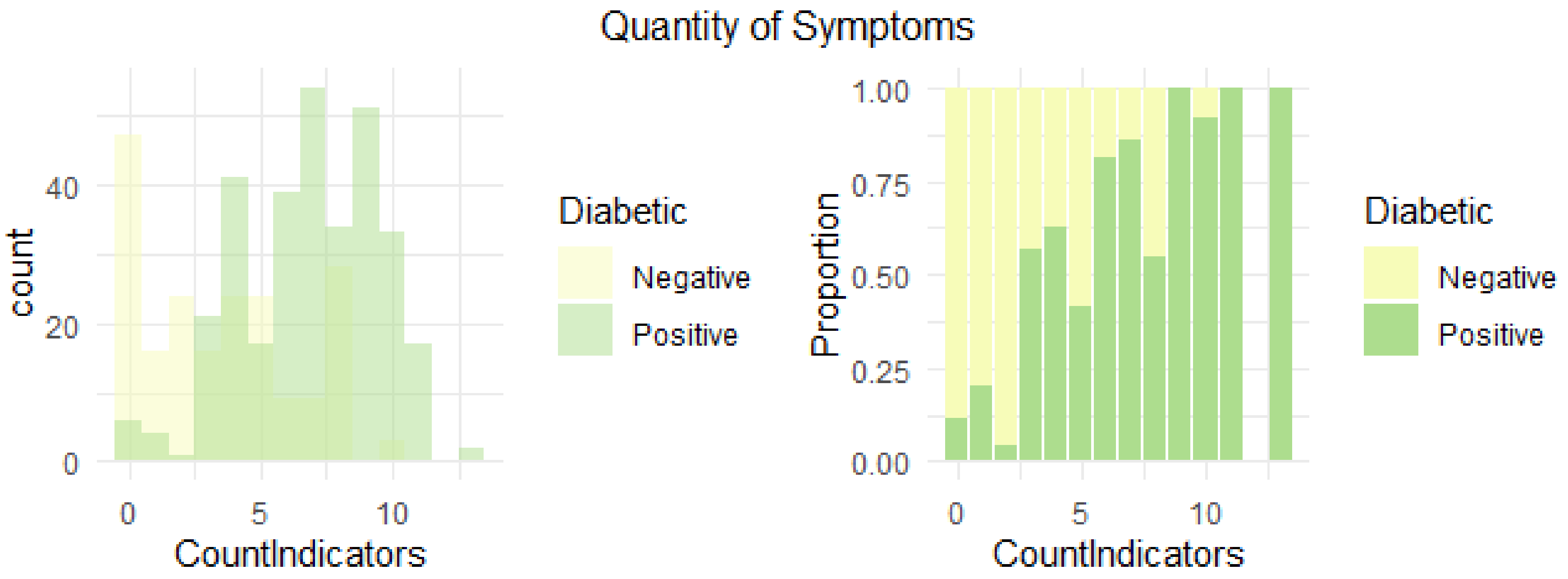


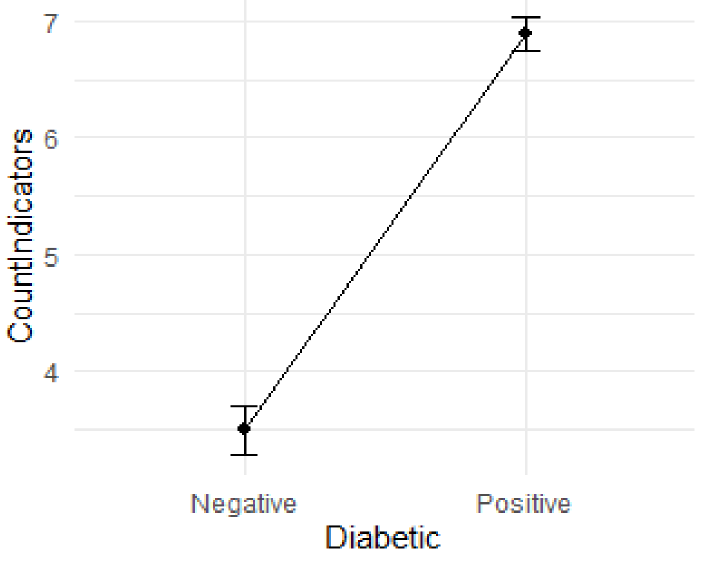
- Excessive urination, thirst, or hunger, as well as partial muscular weakness or genital thrush are **more prevalent** in diabetics
- On the other hand, loss of hair seems to be **less prevalent** in diabetics

Diabetics tend to present more symptoms

- A **new feature was created by simply counting** how many symptoms each patient reported in the questionnaire

- Diabetics tend to **report twice as many symptoms** than non-diabetics

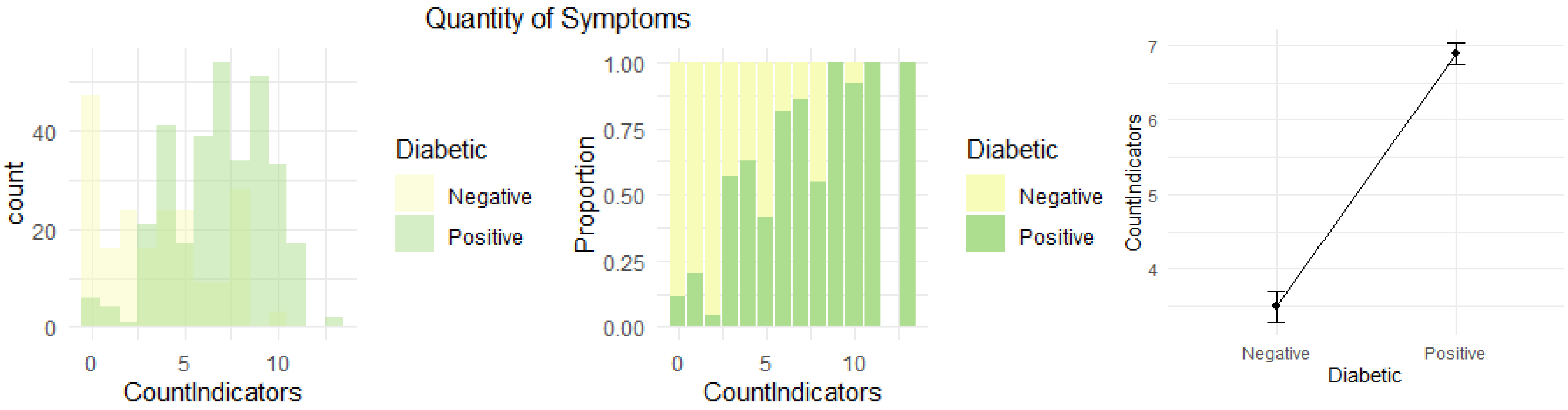




Diabetics tend to present more symptoms

- A **new feature was created by simply counting** how many symptoms each patient reported in the questionnaire

- Diabetics tend to **report twice as many symptoms** than non-diabetics



Statistical Models

1 Logistic Regression

- Dataset is split **70% as training and 30% as test**

- Logistic regression takes as an input multiple features and **outputs a value between 0 and 1.**

- This value denoted p(X) is the conditional probability that the response variable is "positive" given X:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}}$$

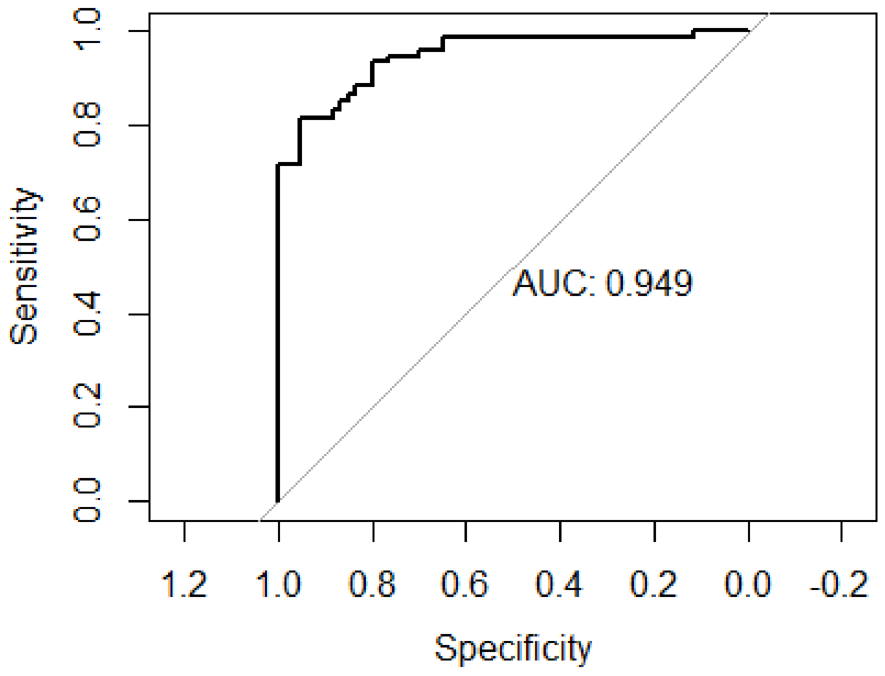
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.36265	1.36587	1.730	0.08367	.
Age	-0.06586	0.03447	-1.911	0.05606	.
GenderMale	-4.04206	0.77206	-5.235	1.65e-07	***
PolyuriaYes	**5.16819**	1.24070	4.166	3.11e-05	***
PolydipsiaYes	**6.82621**	1.48174	4.607	4.09e-06	***
sudden.weight.lossYes	0.78488	0.72712	1.079	0.28039	
weaknessYes	2.39592	0.75820	3.160	0.00158	**
PolyphagiaYes	0.99165	0.70983	1.397	0.16241	
Genital.thrushYes	2.49791	0.80174	3.116	0.00184	**
visual.blurringYes	1.91706	0.94529	2.028	0.04256	*
ItchingYes	-3.84983	0.97118	-3.964	7.37e-05	***
IrritabilityYes	0.66696	0.96321	0.692	0.48866	
delayed.healingYes	0.62316	0.81488	0.765	0.44443	
partial.paresisYes	2.37005	0.79030	2.999	0.00271	**
muscle.stiffnessYes	-1.55000	0.90375	-1.715	0.08633	.
AlopeciaYes	-0.67015	0.90816	-0.738	0.46056	
ObesityYes	0.16864	0.91390	0.185	0.85360	

1 Logistic Regression

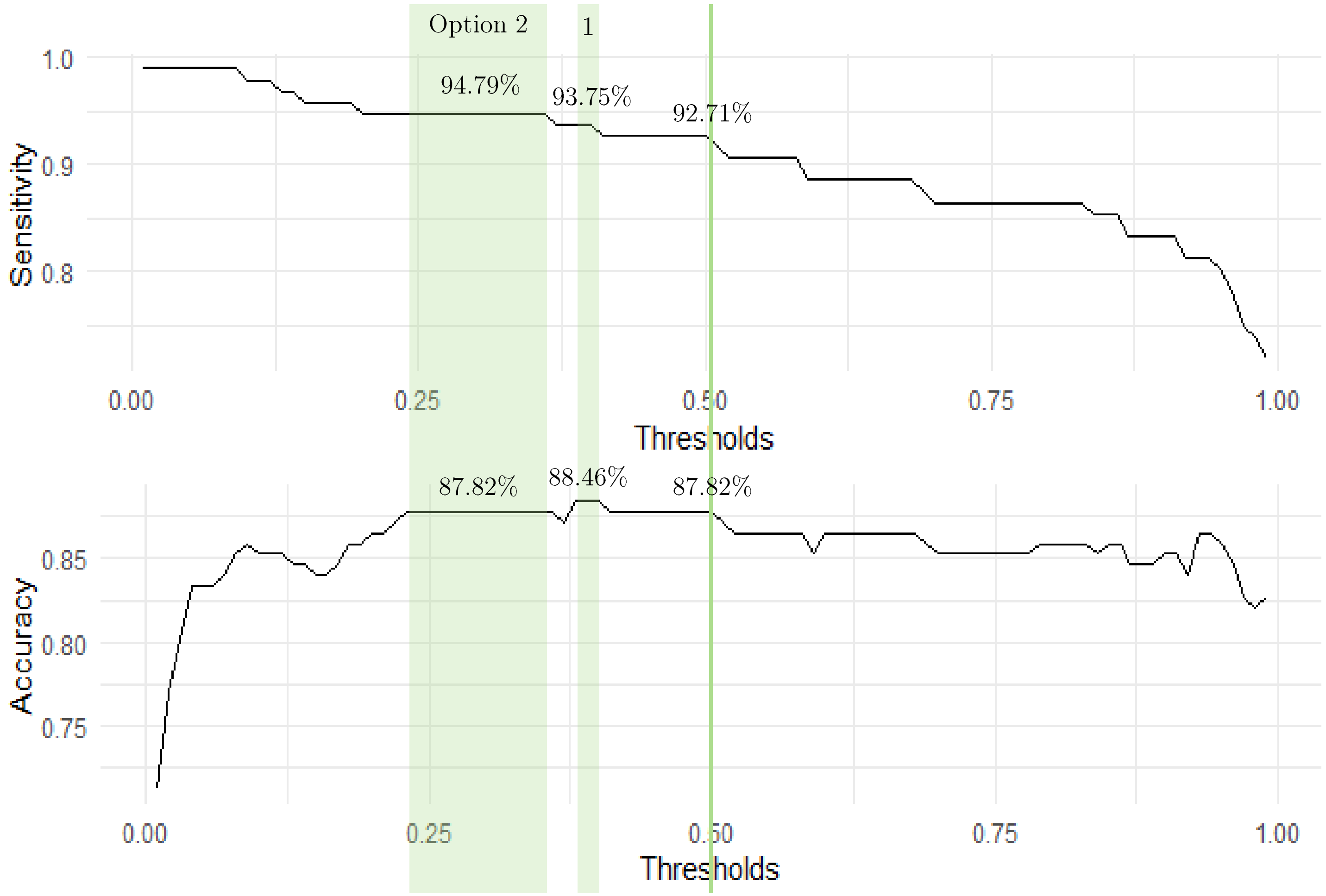
- The baseline LR model **achieves a good test performance** with an accuracy of 87.82% and a sensitivity of 92.71%

Training Dataset			Test Dataset		
	Reference			Reference	
Prediction	Negative	Positive	Prediction	Negative	Positive
Negative	**130**	**11**	Negative	**48**	**7**
Positive	**10**	**213**	Positive	**12**	**89**
Accuracy : **0.9423**			Accuracy : **0.8782**		
Kappa : 0.8783			Kappa : 0.7386		
Sensitivity : **0.9509**			Sensitivity : **0.9271**		
Specificity : 0.9286			Specificity : 0.8000		
Balanced Accuracy : 0.9397			Balanced Accuracy : 0.8635		



1 Optimizing the Threshold

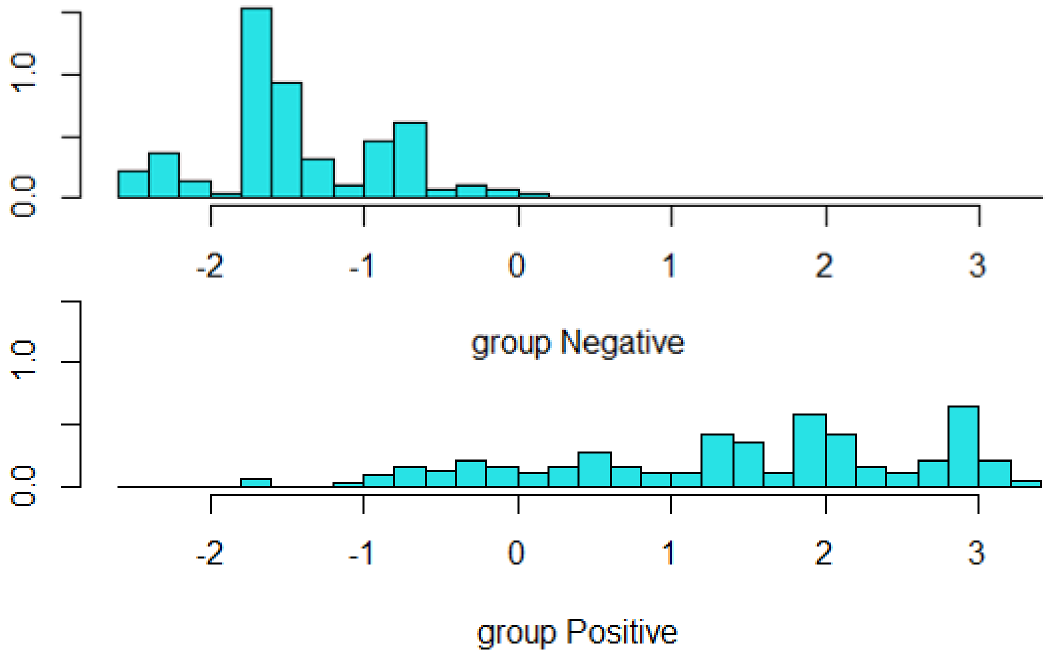
- 100 thresholds (t) were assessed, from 0 to 1 at 0.01 intervals
- **Option 1:** Maximize Accuracy (t=0.38-0.4)
- **Option 2:** Conservatively Increasing Sensitivity (t=0.23-0.36)



2 Linear Discriminant Analysis (LDA)

- DA **models the distribution each feature of each class separately**, and then uses Bayes' theorem **to flip things around** and obtain Pr(Y|X)

- There is a **good separation** (discrimination) between the output classes, yielding a good test accuracy (88.46%)



Training Dataset

Reference		
Prediction	Negative	Positive
Negative	**138**	**31**
Positive	**2**	**193**

Accuracy :	**0.9093**
Kappa :	0.8156
Sensitivity :	**0.8616**
Specificity :	0.9857
Balanced Accuracy :	0.9237

Test Dataset

Reference		
Prediction	Negative	Positive
Negative	**53**	**11**
Positive	**7**	**85**

Accuracy :	**0.8846**
Kappa :	0.7593
Sensitivity :	**0.8854**
Specificity :	0.8833
Balanced Accuracy :	0.8844

3 Decision Tree

- Tree based models are **simple and interpretable**: it segments the prediction space into several simple regions. At each step, the variable and threshold yielding the best separation is chosen

- At the final level, the **leaf assigns the majority class** to the data points as a prediction

- This model achieve an accuracy **higher than the previous two** (89.74%)

Training Dataset

Prediction	Reference: Negative	Reference: Positive
Negative	132	15
Positive	8	209

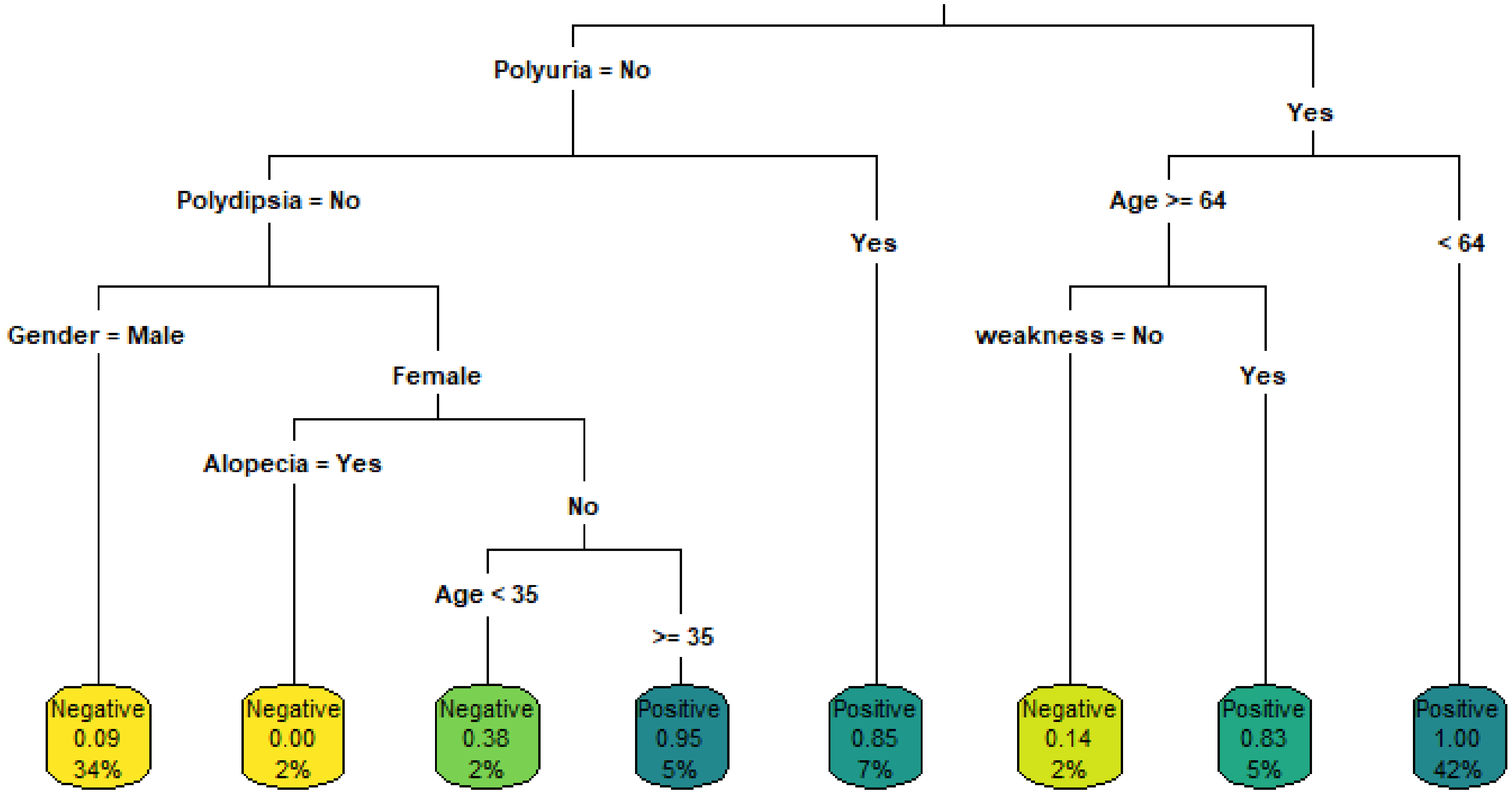
Accuracy	**0.9368**
Kappa	0.8678
Sensitivity	**0.9330**
Specificity	0.9429
Balanced Accuracy	0.9379

Test Dataset

Prediction	Reference: Negative	Reference: Positive
Negative	52	8
Positive	8	88

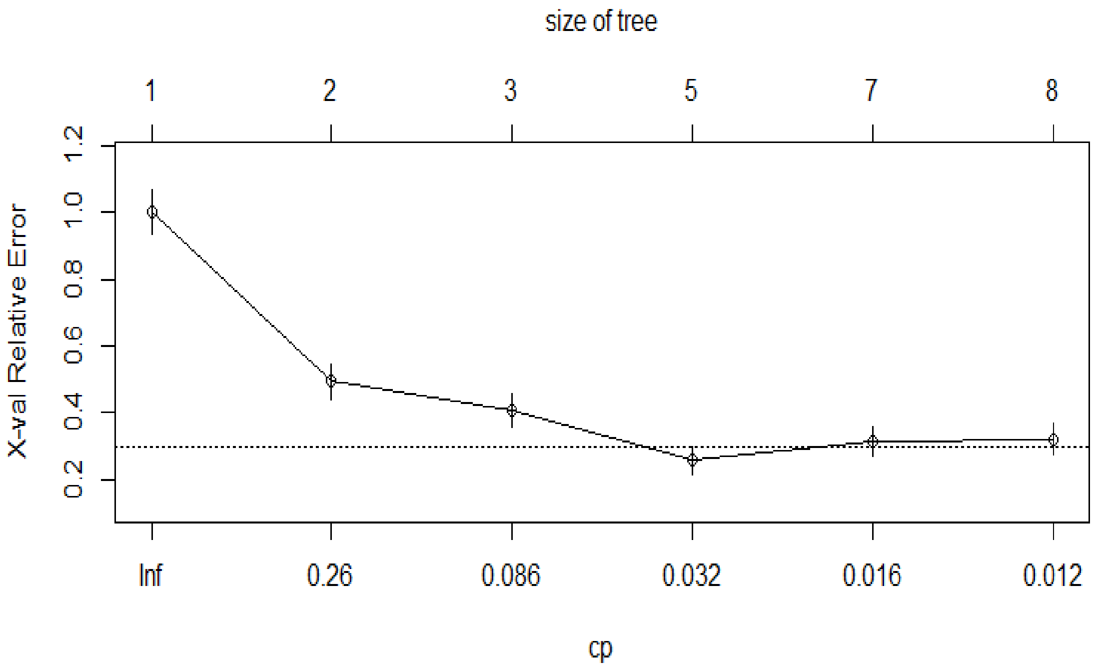
Accuracy	**0.8974**
Kappa	0.7833
Sensitivity	**0.9167**
Specificity	0.8667
Balanced Accuracy	0.8917

3 Decision Tree

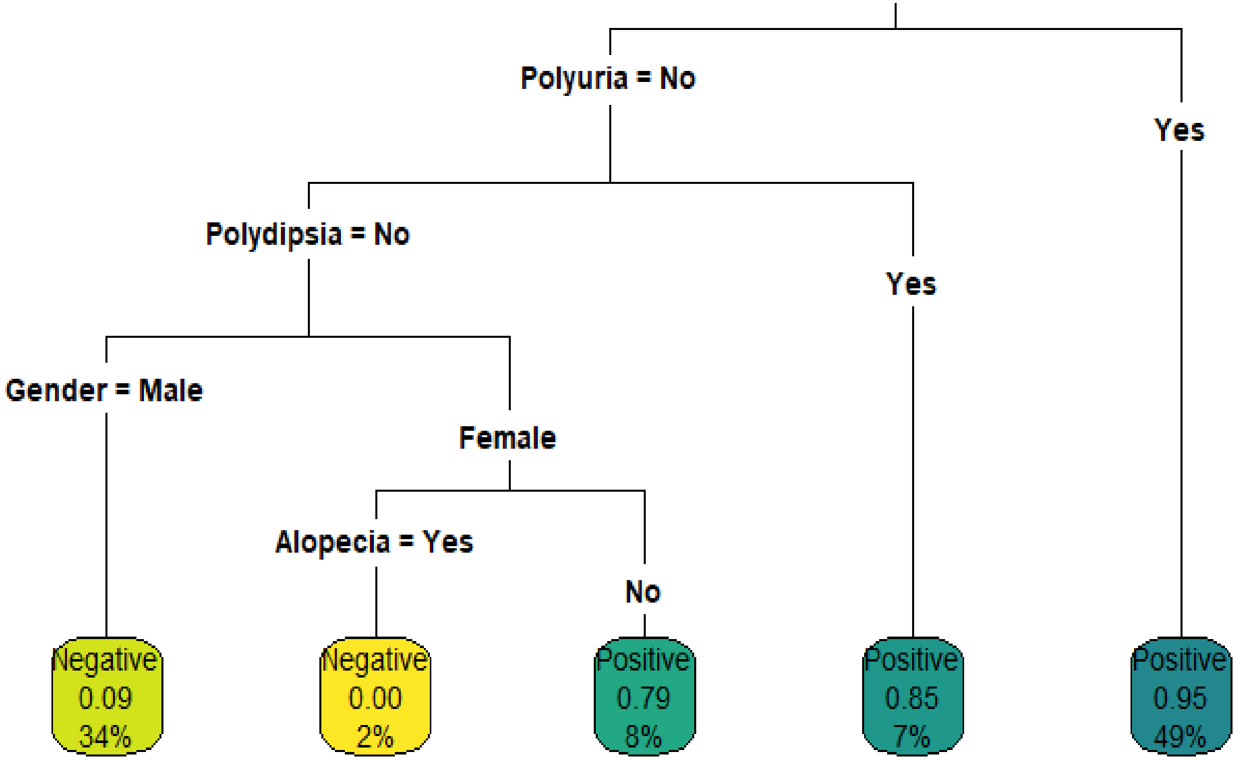


3 Pruning the Tree

- The earlier decision tree achieves a training accuracy of 93.68%, but a testing accuracy of 89.74%, **hinting at a potential overfitting problem**
- The idea of pruning is to achieve a smaller tree with fewer splits, possibly **lowering the variance**, improving the interpretability, although **at the cost of a little more bias**
- An optimal complexity parameter (CP) is chosen from the below plot



3 Pruning the Tree



- The pruned tree **increases the test accuracy** (91.67%) and sensitivity (96.88%)

Training Dataset			Test Dataset		
	Reference			Reference	
Prediction	Negative	Positive	Prediction	Negative	Positive
Negative	121	11	Negative	50	3
Positive	19	213	Positive	10	93
Accuracy :	**0.9176**		Accuracy :	**0.9167**	
Kappa :	0.8240		Kappa :	0.8200	
Sensitivity :	**0.9509**		Sensitivity :	**0.9688**	
Specificity :	0.8643		Specificity :	0.8333	
Balanced Accuracy :	0.9076		Balanced Accuracy :	0.9010	

4 Random Forest (RF)

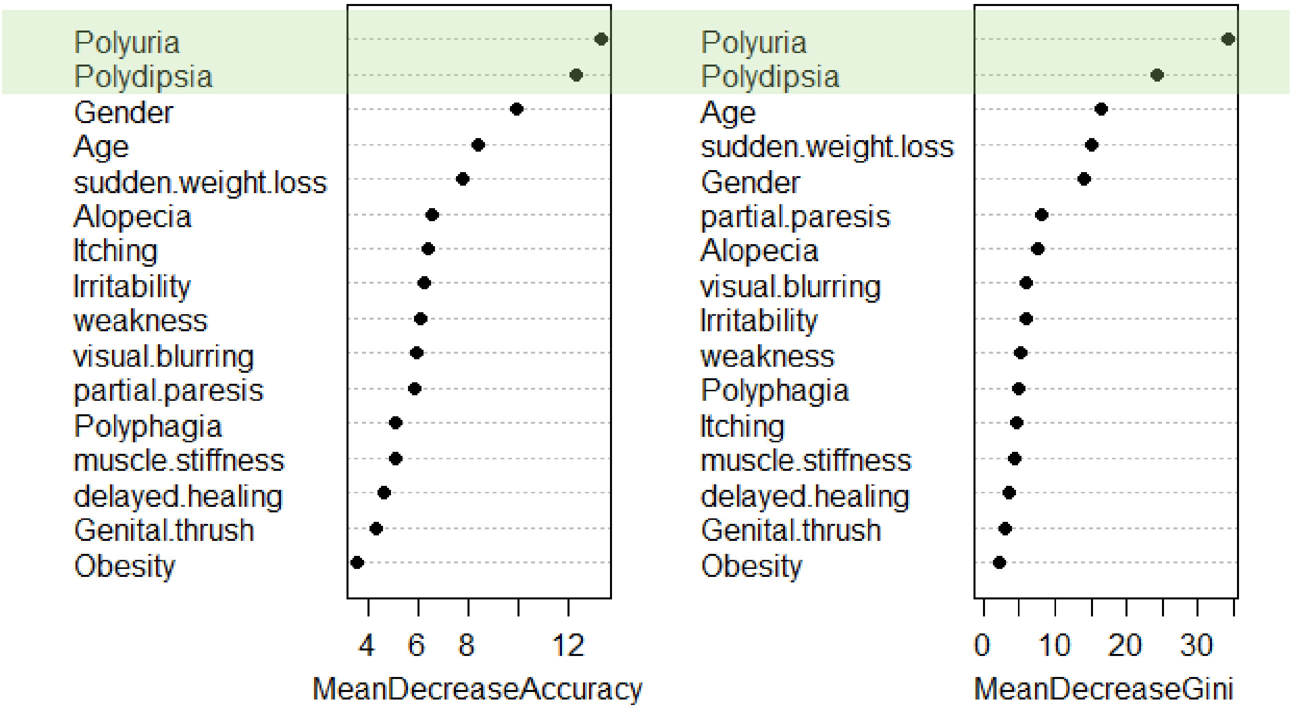
- Bagging is a general procedure to **reduce the variance** of a statistical learning method (and thus improve the model's performance)
- This is achieved by taking repeated samples from the same training dataset, **building multiple trees, and taking the average of all predictions**
- RF improves on bagged tree by **decorrelating** the trees thus reducing the variance further
- This is achieved by **randomizing the selection of features available** to the model at each split

- The RF achieves the **highest test accuracy** (93.59%) and sensitivity (97.92%)

Training Dataset			Test Dataset		
	Reference			Reference	
Prediction	Negative	Positive	Prediction	Negative	Positive
Negative	**139**	**0**	Negative	**52**	**2**
Positive	**1**	**224**	Positive	**8**	**94**
Accuracy :	**0.9973**		Accuracy :	**0.9359**	
Kappa :	0.9942		Kappa :	0.8620	
Sensitivity :	**1.0000**		Sensitivity :	**0.9792**	
Specificity :	0.9929		Specificity :	0.8667	
Balanced Accuracy :	0.9964		Balanced Accuracy :	0.9229	

4 Random Forest (RF)

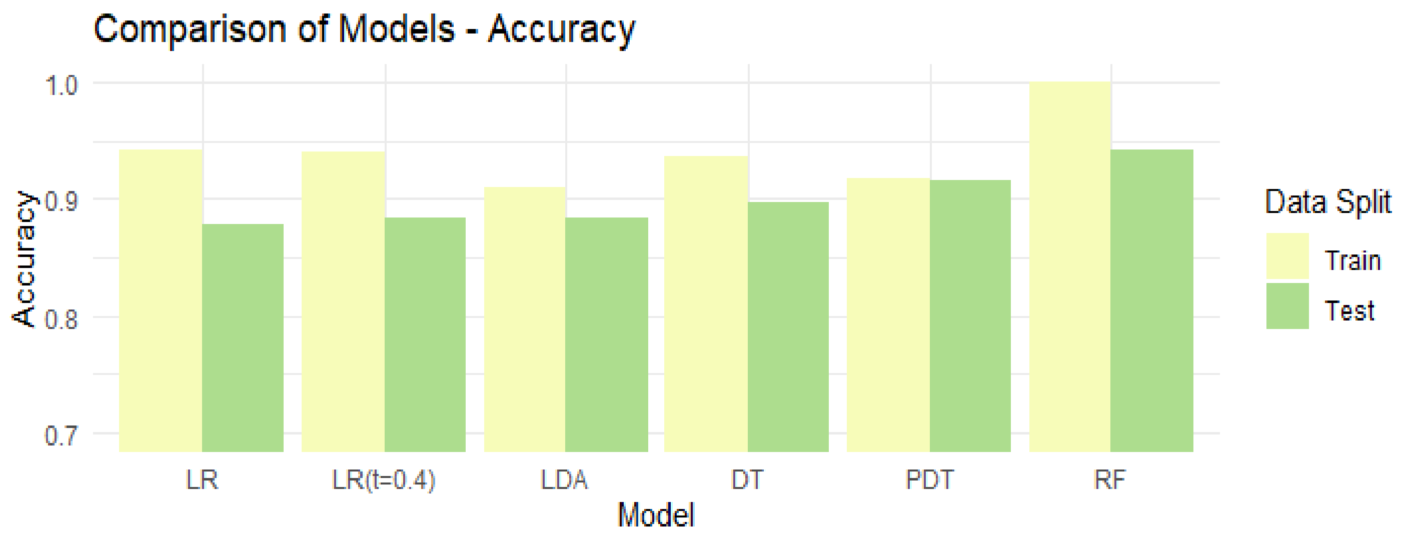
- RF produces a **Variable Importance Plot**, indicating which variables have the most influence on the performance of the model

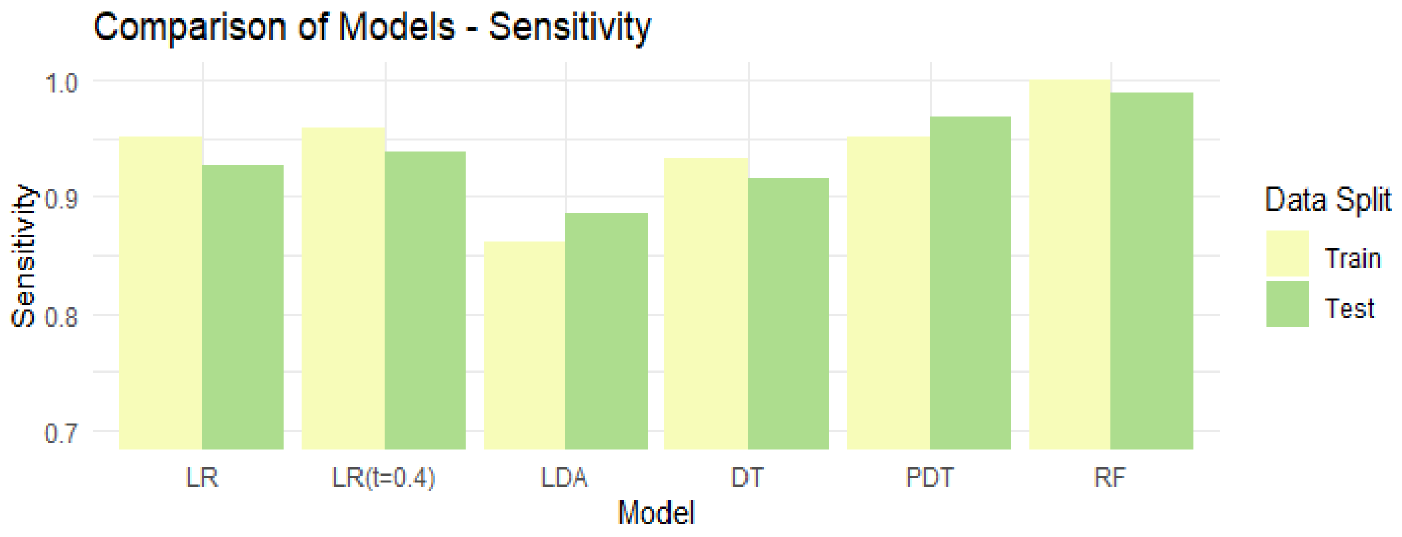


Key Findings

Model Comparison

- All models achieve high accuracy and sensitivity results, confirm the **validity of the concept of early detection of diabetes from the presence of certain symptoms**
- RF **achieves highest** accuracy and sensitivity
- PT is in close second place, while being the **most interpretable** model (easiest for individuals to apply, compared to the math of LR and LDA, or the digitally stored RF)





Key Results

- **Polyuria** (excessive urination) and **polydipsia** (excessive thirst) are by far the **strongest indicators of diabetes** to watch out for

- **Sudden unintentional weight loss** is also a significant factor in most models at predicting diabetes

- The higher **the number of physiological symptoms** present from the set studied the more likely a person is diabetic

- Minimal training of community volunteers, the community, and even individuals to simply **keep an eye out for the presence of a few symptoms provides a strong and free predictor** to identify diabetes at an early stage