

FAERS Team

Daiwik Bommireddipally, Sydney Chau,

Ryan Mersereau, Carina Fu

Introduction

- FAERS (FDA Adverse Event Reporting System) contains information on medication errors and adverse event reports.
- Voluntary self reporting by healthcare professionals, consumers, and manufacturers.
- Designed to support the FDA's post marketing safety surveillance program for drug and therapeutic products.
- Used to help evaluate safety concerns and create regulatory actions to improve product safety and public health.
- There may be duplicate reports due to submissions by a consumer and a sponsor.

Introduction

- Data from **faers_ascii_2018q4** and **faers_ascii_2020q4**.
- Response variables of interest were **dechal** and **rechal**
- **dechal**: Dechallenge code, indicating if reaction abated when drug therapy was stopped.
- **rechal**: Rechallenge code, indicating if reaction recurred when drug therapy was restarted.

CODE MEANING_TEXT

----	-----
Y	Positive rechallenge
N	Negative rechallenge
U	Unknown
D	Does not apply

Predictor Variables of Interest

Variable	Description	Example
Age	Patient Age at event	30
Age code (age_cod)	Unit abbreviation for age	YR (years), MON (months)
Sex	Patient sex	M (Male)
Weight (wt)	Numeric value of patient weight	150
Weight code (wt_code)	Unit abbreviation for weight	LBS (pounds)
To_mfr	Whether the reporter also notified the manufacturer	Y (Yes)

Predictor Variables of Interest Cont.

Variable	Description	Example
Occp_cod	Reporter occupation	MD - Physician PH - Pharmacist OT - Other LW - Lawyer CN - Consumer
Reporter_country	Country of the reporter	US
Occr_country	Country where event occurred	US
Role_cod	Code for drugs role in event	PS - Primary suspect drug
Route	Route of drug administration	Oral
Outc_cod	Code for patient outcome	DE - Death

Cleaning

1. `library(dplyr)`

`library(readr)`

Read in 7 .txt files and remove variables not in the email. Also, OUTC20Q4.txt file has variable `outc_cod` and RPSR20Q4.txt file has variable `rpsr_cod` that match the same `primaryid`.

	primaryid	outc_cod
1	37363932	OT
2	39629172	OT, DS
3	41627247	CA, OT
4	41923193	HO
5	42021114	DE
6	56585445	DE, HO
7	57118383	HO
8	57528163	HO
9	59057815	HO
10	59297543	HO, OT, DE
11	62923272	HO
12	63523522	OT, CA
13	63657203	DE, OT
14	63690602	OT
15	63716933	HO, OT
16	63890722	OT
17	63926532	OT, DS
18	64138616	OT, DE

Showing 1 to 19 of 281,559 entries, 2 total

	primaryid	rpsr_cod
1	183341011	CSM
2	183341351	CSM
3	183341411	HP
4	183341581	HP
5	183341621	HP
6	183341641	CSM
7	183344721	HP
8	183344761	HP
9	183344811	HP
10	183344861	CSM
11	183344891	HP
12	183344981	CSM
13	183345041	HP
14	183345551	HP
15	183345601	HP
16	183345731	CSM
17	183346331	HP
18	183346591	HP

Showing 1 to 19 of 14,287 entries, 2 total columns

Cleaning

2. Find unique primaryid: `distinct(primaryid)`

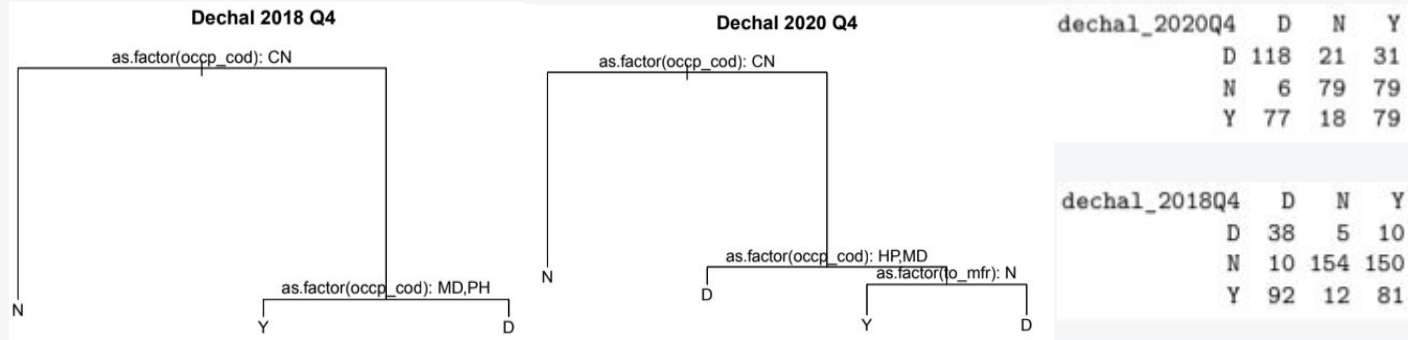
3. Find the intersection of unique primaryid: `common_primaryid <- Reduce(intersect, list(of unique primaryid find above))`

4. Merged data and filter 12 + 2 variables we are interested in by `filter(!is.na(var name))`

But there are many sequence variables e.g. `caseversion`, `drug_seq`, `indi_drug_seq`, `dsg_drug_seq` → cause duplicates

5. Remove rows with duplicate primaryid, only keep the first appeared primaryid.
Most sequences are 1.

Classification Trees (Dechal)

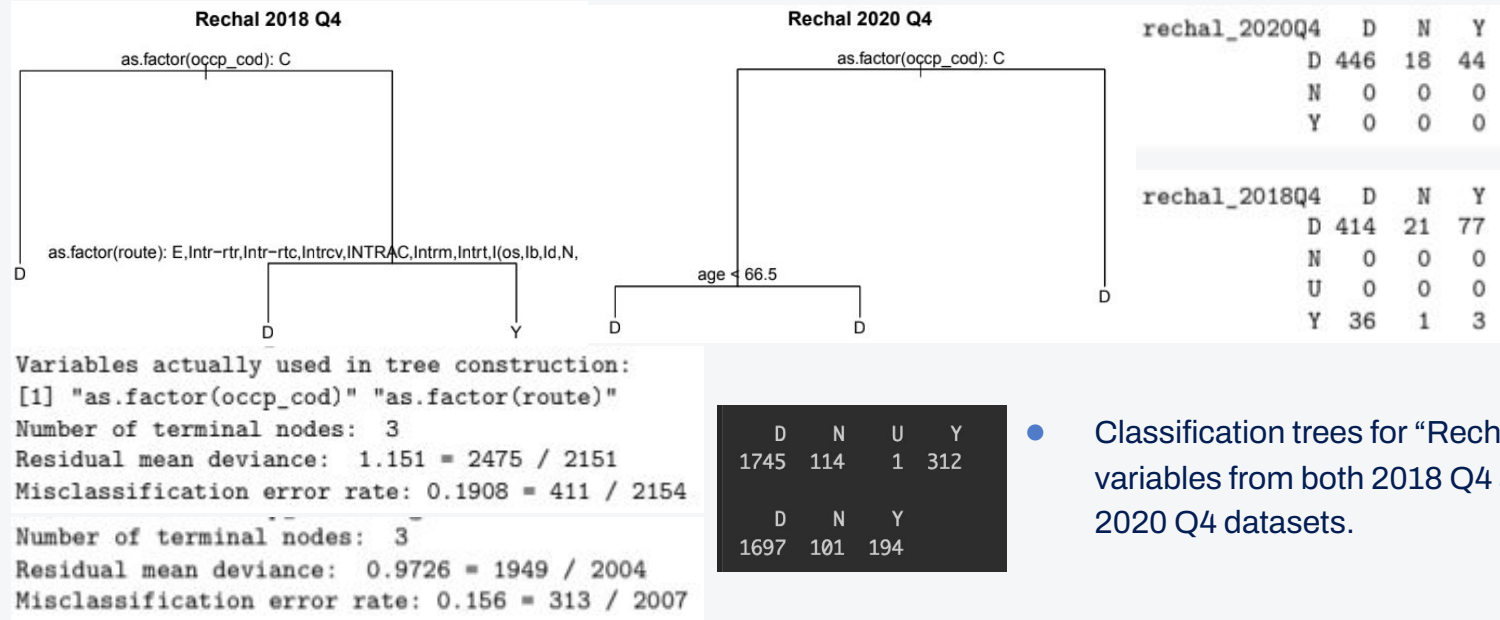


Variables actually used in tree construction:
[1] "as.factor(occp_cod)"
Number of terminal nodes: 3
Residual mean deviance: 1.711 = 3680 / 2151
Misclassification error rate: 0.4884 = 1052 / 2154

Variables actually used in tree construction:
[1] "as.factor(occp_cod)" "as.factor(to_mfr)"
Number of terminal nodes: 4
Residual mean deviance: 1.702 = 3410 / 2003
Misclassification error rate: 0.418 = 839 / 2007

- Classification trees for “Dechal” variables from both 2018 Q4 and 2020 Q4 datasets.

Classification Trees (Rechal)



- Classification trees for “Rechal” variables from both 2018 Q4 and 2020 Q4 datasets.

Classification Models for 2020Q4 data

Logistic Regression (Dechal)

Coefficients: (4 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-6.252e+01	4.404e+03	-0.014	0.988675
age	-4.732e-02	3.217e-03	-14.709	< 2e-16 ***
as.factor(age_cod)YR	1.907e+01	3.261e+03	0.006	0.995335
as.factor(sex)M	3.766e-01	1.262e-01	2.983	0.002852 **
wt	3.200e-02	2.617e-03	12.228	< 2e-16 ***
as.factor(wt_cod)LBS	1.548e+01	1.111e+03	0.014	0.988880
as.factor(to_mfr)Y	-2.720e-01	1.712e-01	-1.589	0.112016
as.factor(occip_cod)HP	1.310e+00	3.285e-01	3.989	6.63e-05 ***
as.factor(occip_cod)M0	4.116e+00	4.253e-01	9.678	< 2e-16 ***
as.factor(occip_cod)PH	4.594e+00	2.305e-01	19.929	< 2e-16 ***
as.factor(reporter_country)DE	1.993e+01	2.306e+03	0.009	0.993104
as.factor(reporter_country)GB	-2.698e+00	9.786e-01	-2.757	0.005839 **
as.factor(reporter_country)IN	-1.567e+01	1.017e+03	-0.015	0.987711
as.factor(reporter_country)US	1.335e+00	3.891e-01	3.431	0.000602 ***
as.factor(occur_country)DE	NA	NA	NA	NA
as.factor(occur_country)GB	NA	NA	NA	NA
as.factor(occur_country)IN	NA	NA	NA	NA
as.factor(occur_country)US	NA	NA	NA	NA
as.factor(role_cod)SS	8.203e-01	1.681e-01	4.880	1.06e-06 ***
as.factor(route)INTRA-AURAL	4.224e+01	4.790e+03	0.009	0.992965
as.factor(route)INTRACAVITY	2.146e-01	1.856e+03	0.000	0.999908
as.factor(route)Intradermal	3.722e+01	6.649e+03	0.006	0.995534
as.factor(route)Intramuscular	2.027e+01	1.292e+03	0.016	0.987484
as.factor(route)Intravenous (not otherwise specified)	1.936e+01	1.292e+03	0.015	0.988048
as.factor(route)Intravenous bolus	3.996e+01	2.400e+03	0.017	0.986718
as.factor(route)Intravenous drip	2.031e+01	1.292e+03	0.016	0.987460
as.factor(route)Occlusive dressing technique	3.740e+01	3.981e+03	0.009	0.992506
as.factor(route)Ophthalmic	2.352e+01	1.292e+03	0.018	0.985479
as.factor(route)Oral	2.281e+01	1.292e+03	0.018	0.985921
as.factor(route)Respiratory (inhalation)	2.378e+01	1.292e+03	0.018	0.985318
as.factor(route)Subcutaneous	2.215e+01	1.292e+03	0.017	0.986328
as.factor(route)Sublingual	4.030e+01	1.666e+03	0.024	0.980703
as.factor(route)Topical	2.261e+01	1.292e+03	0.017	0.986043
as.factor(route)Transdermal	2.706e+01	1.292e+03	0.021	0.983295
as.factor(route)Vaginal	2.640e+01	1.292e+03	0.020	0.983705

as.factor(outc_cod)DE	1.769e+01	2.663e+03	0.007	0.994699
as.factor(outc_cod)DS	1.811e+01	2.663e+03	0.007	0.994573
as.factor(outc_cod)H0	1.888e+01	2.663e+03	0.007	0.994343
as.factor(outc_cod)LT	1.809e+01	2.663e+03	0.007	0.994580
as.factor(outc_cod)OT	1.945e+01	2.663e+03	0.007	0.994171
as.factor(outc_cod)RI	3.359e+01	2.682e+03	0.013	0.990009

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5681.2 on 4274 degrees of freedom

Residual deviance: 2721.1 on 4238 degrees of freedom

(1372 observations deleted due to missingness)

AIC: 2795.1

Number of Fisher Scoring iterations: 17

> |

Logistic Regression (Rechal)

Coefficients: (4 not defined because of singularities)

```
Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.645e+01 4.397e+03 0.004 0.99702
age -7.575e-03 2.878e-03 -2.632 0.00848 **
as.factor(age_cod)YR -1.831e+01 3.261e+03 -0.006 0.99552
as.factor(sex)M -5.194e-01 1.311e-01 -3.961 7.47e-05 ***
wt -1.811e-03 2.189e-03 -0.827 0.40797
as.factor(wt_cod)LBS 1.812e+01 1.463e+03 0.012 0.99012
as.factor(to_mfr)Y 7.129e-02 1.787e-01 0.399 0.68990
as.factor(occip_cod)HP -1.830e+00 2.537e-01 -7.215 5.40e-13 ***
as.factor(occip_cod)MD -2.561e+00 2.247e-01 -11.397 < 2e-16 ***
as.factor(occip_cod)PH -3.331e+00 1.639e-01 -20.323 < 2e-16 ***
as.factor(reporter_country)DE 2.179e+01 2.306e+03 0.009 0.99246
as.factor(reporter_country)GB 4.678e+00 7.862e-01 5.949 2.69e-09 ***
as.factor(reporter_country)IN 2.308e+01 9.949e+02 0.023 0.98149
as.factor(reporter_country)US 4.049e+00 3.108e-01 13.028 < 2e-16 ***
as.factor(occr_country)DE NA NA NA NA
as.factor(occr_country)GB NA NA NA NA
as.factor(occr_country)IN NA NA NA NA
as.factor(occr_country)US NA NA NA NA
as.factor(role_cod)SS -5.723e-02 1.446e-01 -0.396 0.69234
as.factor(route)INTRA-AURAL 3.373e+01 4.784e+03 0.007 0.99437
as.factor(route)INTRACAVITY 3.616e+01 1.839e+03 0.020 0.98431
as.factor(route)Intradermal -1.633e+00 6.645e+03 0.000 0.99980
as.factor(route)Intramuscular 1.678e+01 1.269e+03 0.013 0.98945
as.factor(route)Intravenous (not otherwise specified) 1.658e+01 1.269e+03 0.013 0.98957
as.factor(route)Intravenous bolus -1.646e+00 2.705e+03 -0.001 0.99951
as.factor(route)Intravenous drip 1.670e+01 1.269e+03 0.013 0.98950
as.factor(route)Occlusive dressing technique 4.225e-01 3.974e+03 0.000 0.99992
as.factor(route)Ophthalmic 1.761e+01 1.269e+03 0.014 0.98893
as.factor(route)Oral 1.673e+01 1.269e+03 0.013 0.98948
as.factor(route)Respiratory (inhalation) 3.332e+01 1.439e+03 0.023 0.98152
as.factor(route)Subcutaneous 1.435e+01 1.269e+03 0.011 0.99098
as.factor(route)Sublingual 3.402e+01 1.783e+03 0.019 0.98478
as.factor(route)Topical 1.536e+01 1.269e+03 0.012 0.99034
as.factor(route)Transdermal 3.304e+01 1.328e+03 0.025 0.98014
as.factor(route)Vaginal 1.647e+01 1.269e+03 0.013 0.98964
as.factor(out_c_cod)DE -1.552e+01 2.663e+03 -0.006 0.99535
as.factor(out_c_cod)DS -1.630e+01 2.663e+03 -0.006 0.99512
as.factor(out_c_cod)HO -1.664e+01 2.663e+03 -0.006 0.99501
```

```
as.factor(out_c_cod)LT -1.692e+01 2.663e+03 -0.006 0.99493
as.factor(out_c_cod)OT -1.522e+01 2.663e+03 -0.006 0.99544
as.factor(out_c_cod)RI -1.669e+01 2.663e+03 -0.006 0.99500
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5817.5 on 4274 degrees of freedom
Residual deviance: 2757.7 on 4238 degrees of freedom
(1372 observations deleted due to missingness)
AIC: 2831.7

Number of Fisher Scoring iterations: 17

>

LDA

```
Call:
lda(rechal ~ sex + to_mfr + occp_cod, data = training)

Prior probabilities of groups:
      D      N      Y
0.80919540 0.05057471 0.14022989

Group means:
sexM   to_mfrY occp_codMD occp_codOT occp_codPH
D 0.4267045 0.1176136 0.04829545 0.11193182 0.33693182
N 0.3727273 0.1272727 0.03636364 0.06363636 0.40000000
Y 0.3344262 0.1344262 0.01967213 0.03934426 0.08852459

Coefficients of linear discriminants:
      LD1      LD2
sexM    -0.03455616 1.1845157
to_mfrY  -0.02314959 -0.3781145
occp_codMD -1.85454532 1.0686860
occp_codOT -1.94387384 1.9643566
occp_codPH -2.12804817 -1.0960542

Proportion of trace:
      LD1      LD2
0.9573 0.0427
> p1 <- predict(linear, testing)$class
> tab1 <- table(Predicted = p1, Actual = testing$rechal)
> sum(diag(tab1))/sum(tab1)
[1] 0.809434
>
```

```
Call:
lda(dechal ~ sex + to_mfr + occp_cod, data = training)

Prior probabilities of groups:
      D      N      Y
0.24000000 0.3213793 0.4386207

Group means:
sexM   to_mfrY occp_codMD occp_codOT occp_codPH
D 0.4597701 0.1187739 0.09386973 0.31034483 0.53065134
N 0.3505007 0.1258941 0.02575107 0.01716738 0.08869814
Y 0.4287212 0.1174004 0.02935010 0.04402516 0.34067086

Coefficients of linear discriminants:
      LD1      LD2
sexM    0.2014743 0.27851440
to_mfrY -0.1940921 -0.01798082
occp_codMD -2.4004316 -1.04875934
occp_codOT -3.5050634 -1.79804992
occp_codPH -2.2217364 1.44208714

Proportion of trace:
      LD1      LD2
0.9655 0.0345
>
> p2 <- predict(linear2, testing)$class
> tab2 <- table(Predicted = p2, Actual = testing$dechal)
> sum(diag(tab2))/sum(tab2)
[1] 0.4735849
<
```

```
Call:
lda(rechal ~ sex + to_mfr + occp_cod, data = training2)

Prior probabilities of groups:
      D      N      Y
0.85107446 0.04697651 0.10194983

Group means:
sexM   to_mfrY occp_codHP occp_codMD occp_codPH
D 0.4650617 0.2166765 0.11802791 0.08220787 0.5079272
N 0.4787234 0.1489362 0.08510638 0.04255319 0.5744681
Y 0.3774510 0.1274510 0.05882353 0.04901961 0.1568627

Coefficients of linear discriminants:
      LD1      LD2
sexM    0.04804544 -0.01258204
to_mfrY  -0.11872410 1.78387298
occp_codHP -2.02255697 1.01341299
occp_codMD -1.89950145 1.89623367
occp_codPH -2.26534204 -0.75862868

Proportion of trace:
      LD1      LD2
0.9621 0.0379
>
> p3 <- predict(linear3, testing2)$class
> tab3 <- table(Predicted = p3, Actual = testing2$rechal)
> sum(diag(tab3))/sum(tab3)
[1] 0.8581946
>
```

```
Call:
lda(dechal ~ sex + to_mfr + occp_cod, data = training2)

Prior probabilities of groups:
      D      N      Y
0.3848076 0.2298851 0.3853073

Group means:
sexM   to_mfrY occp_codHP occp_codMD occp_codPH
D 0.5090909 0.3181818 0.19740260 0.13896104 0.6220779
N 0.3739130 0.1760870 0.05652174 0.02608696 0.1913043
Y 0.4539559 0.1076524 0.05577173 0.04539559 0.4980545

Coefficients of linear discriminants:
      LD1      LD2
sexM    -0.01984342 -0.03880481
to_mfrY  -0.56405886 -2.03892688
occp_codHP -2.83850052 -0.61113026
occp_codMD -2.97969830 -0.33766248
occp_codPH -2.22509082 1.36557374

Proportion of trace:
      LD1      LD2
0.9071 0.0929
>
> p4 <- predict(linear4, testing2)$class
> tab4 <- table(Predicted = p4, Actual = testing2$dechal)
> sum(diag(tab4))/sum(tab4)
[1] 0.5836576
>
```

QDA

```
> quadratic
Call:
qda(rechal ~ sex + to_mfr + occp_cod, data = training)

Prior probabilities of groups:
      D      N      Y
0.80919540 0.05057471 0.14022989

Group means:
      sexM    to_mfrY occp_codMD occp_cod0T occp_codPH
D 0.4267045 0.1176136 0.04829545 0.11193182 0.33693182
N 0.3727273 0.1272727 0.03636364 0.06363636 0.40000000
Y 0.3344262 0.1344262 0.01967213 0.03934426 0.08852459
>
> predicted <- predict(quadratic, testing)
> mean(predicted$class==testing$rechal)
[1] 0.5528302
> |
```

```
Call:
qda(rechal ~ sex + to_mfr + occp_cod, data = training2)

Prior probabilities of groups:
      D      N      Y
0.85107446 0.04697651 0.10194903

Group means:
      sexM    to_mfrY occp_codHP occp_codMD occp_codPH
D 0.4650617 0.2166765 0.11802701 0.08220787 0.5079272
N 0.4787234 0.1489362 0.08510638 0.04255319 0.5744681
Y 0.3774510 0.1274510 0.05882353 0.04901961 0.1568627
>
> predicted3 <- predict(quadratic3, testing2)
> mean(predicted3$class==testing2$rechal)
[1] 0.8501946
> |
```

```
Call:
qda(dechal ~ sex + to_mfr + occp_cod, data = training)

Prior probabilities of groups:
      D      N      Y
0.24000000 0.3213793 0.4386207

Group means:
      sexM    to_mfrY occp_codMD occp_cod0T occp_codPH
D 0.4597701 0.1187739 0.09386973 0.31034483 0.53065134
N 0.3505007 0.1258941 0.02575107 0.01716738 0.08869814
Y 0.4287212 0.1174004 0.02935010 0.04402516 0.34067086
>
> predicted2 <- predict(quadratic2, testing)
> mean(predicted2$class==testing$dechal)
[1] 0.4735849
> |
```

```
Call:
qda(dechal ~ sex + to_mfr + occp_cod, data = training2)

Prior probabilities of groups:
      D      N      Y
0.3848076 0.2298851 0.3853073

Group means:
      sexM    to_mfrY occp_codHP occp_codMD occp_codPH
D 0.5090909 0.3181818 0.19740260 0.13896104 0.6220779
N 0.3739130 0.1760870 0.05652174 0.02608696 0.1913043
Y 0.4539559 0.1076524 0.05577173 0.04539559 0.4980545
>
> predicted4 <- predict(quadratic4, testing2)
> mean(predicted4$class==testing2$dechal)
[1] 0.5836576
> |
```

K-Nearest Neighbors (KNN)

CODE MEANING_TEXT

Y	Positive rechallenge
N	Negative rechallenge
U	Unknown <small>Not in the dataset</small>
D	Does not apply

Reference group:

SexF, to_mfrN, occp_codCN, dechalD, rechalD

Example code:

```
# Perform kNN classification
```

```
knn.pred <- knn(train = train_data[, c("sexM", "to_mfrY",
"occp_codHP", "occp_codMD", "occp_codPH")],
test = test_data[, c("sexM", "to_mfrY",
"occp_codHP", "occp_codMD", "occp_codPH")],
cl = train_data[, which(colnames(train_data) ==
"dechalY")],
k = 10)
```

Problem:

occp_codHP in the dataset but not in the introduction file provided on the website.

Abbreviation for the reporter's type of occupation in the latest version of a case.	
CODE	MEANING_TEXT
----	-----
MD	Physician
PH	Pharmacist
OT	Other health-professional
LW	Lawyer
CN	Consumer

Not in the dataset

dechalY Predicted

Actual	0	1
0	359	125
1	164	107

[1] 0.6172185

dechalN Predicted

Actual	0	1
0	567	13
1	161	14

[1] 0.7695364

rechalY Predicted

Actual	0	1
0	684	0
1	71	0

[1] 0.9059603

rechalN Predicted

Actual	0	1
0	719	0
1	36	0

[1] 0.9523179

Naive Bayes → dechal

```
# Split the data and set seed.
set.seed(123)
train_indices <- sample(seq_len(nrow(data1)), 0.7 * nrow(data1))
train_data <- data1[train_indices, ]
test_data <- data1[-train_indices, ]

# NB model
nb_model <- naiveBayes(dechal ~ sex + to_mfr + occp_cod, data=data1)

# Make predictions
predictions <- predict(nb_model, newdata = test_data)

# Print out confusion_matrix and accuracy
confusion_matrix <- table(predictions, test_data$dechal)
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(confusion_matrix)
cat("Accuracy:", accuracy, "\n")
```

predictions	D	N	Y
D	257	37	123
N	8	93	84
Y	44	45	64

Accuracy: 0.5483444

Naive Bayes → rechal

```
# Split the data and set seed.
set.seed(123)
train_indices <- sample(seq_len(nrow(data1)), 0.7 * nrow(data1))
train_data <- data1[train_indices, ]
test_data <- data1[-train_indices, ]

# NB model
nb_model2 <- naiveBayes(rechal~ sex + to_mfr + occp_cod,data=data1)

# Make predictions
predictions2 <- predict(nb_model2, newdata = test_data)

# Print out confusion_matrix and accuracy
confusion_matrix2 <- table(predictions, test_data$rechal)
accuracy2 <- sum(diag(confusion_matrix2)) / sum(confusion_matrix2)
print(confusion_matrix2)
cat("Accuracy:", accuracy2, "\n")
```

predictions	D	N	Y
D	386	19	12
N	136	5	44
Y	126	12	15

Accuracy: 0.5377483

Challenges

1. Large dataset with too many variables and missing values.
2. Variable not in the introduction file exist in the null data.
3. rept_cod, mfr_sndr, and val_vbm have only one unique observed value, while drugname, prod_ai, indi_pt, pt, start_dt, and end_dt have more than 20 unique observed values
4. Logistic regression models run but do not converge
5. The following variables needed to be removed for LDA/QDA because of problems encountered: age, age_cod, wt, wt_cod, reporter_country, occr_country, role_cod, route, outc_cod

Null counts for selected fields from Demographic file.

Column Name	Total Count	Missing Count
AGE	436148	206147
AGE_COD	436148	206119
AGE_GRP	436148	360350
CASE	436148	0
EVENT_DT	436148	217138
FDA_DT	436148	0
MFR_DT	436148	17647
MFR_NUM	436148	17640
MFR_SNDR	436148	0
WT	436148	353036
WT_COD	436148	353036
REPT_DT	436148	123
REPORTER_COUNTRY	436148	0
OCCR_COUNTRY	436148	685