DAR Coursework Solution

1. Statistical learning methods

Answer(a): Parametric methods, Non-parametric methods may underperform because they don't make strong assumptions about the mapping function, making them more flexible. However, they require a large amount of data to estimate accurately and can suffer from overfitting in scenarios with a small number of observations and a large number of predictors.

Answer(b): Non-parametric methods, Non-parametric methods have the advantage of utilizing a large sample size and numerous predictors to estimate intricate relationships without being limited by assumptions regarding the functional form. In contrast, parametric methods may face challenges in accurately representing the complexity of the data due to their strict assumptions.

Answer (c) Parametric methods, Parametric techniques rely on a predetermined function form (such as linear), and when this assumption holds true (in cases of strong linearity), they can yield robust findings even with a limited sample size.

Answer (d) Non-parametric methods, Non-parametric techniques refrain from making strong assumptions regarding the distribution of errors and possess the ability to adjust more effectively to the increased variability present in the data.

---

2. Linear regression

Answwe (a)

```
#Answer-
# Install and load the ISLR package
library(ISLR)
data(Auto)

model <- lm(acceleration ~ cylinders, data=Auto)
summary(model)
```

```
##
## Call:
## lm(formula = acceleration ~ cylinders, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4778 -1.7428 -0.2428  1.3897  8.7222
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.0078     0.4052   49.38   <2e-16 ***
## cylinders    -0.8163     0.0707  -11.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.385 on 390 degrees of freedom
## Multiple R-squared:  0.2547, Adjusted R-squared:  0.2528
## F-statistic: 133.3 on 1 and 390 DF,  p-value: < 2.2e-16
```

Answer i. -There exists a correlation between the predictor variable (cylinders) and the response variable (acceleration). This is evident from the noteworthy p-value linked to the coefficient for cylinders ($p < 0.001$).

Answer ii. - The R-squared value of 0.2547 implies that roughly 25.47% of the variation in acceleration can be attributed to the number of cylinders in the model.

Answer iii. -negative.coefficient of 'cylinders'((-0.8163)), The relationship between the number of cylinders and acceleration is .

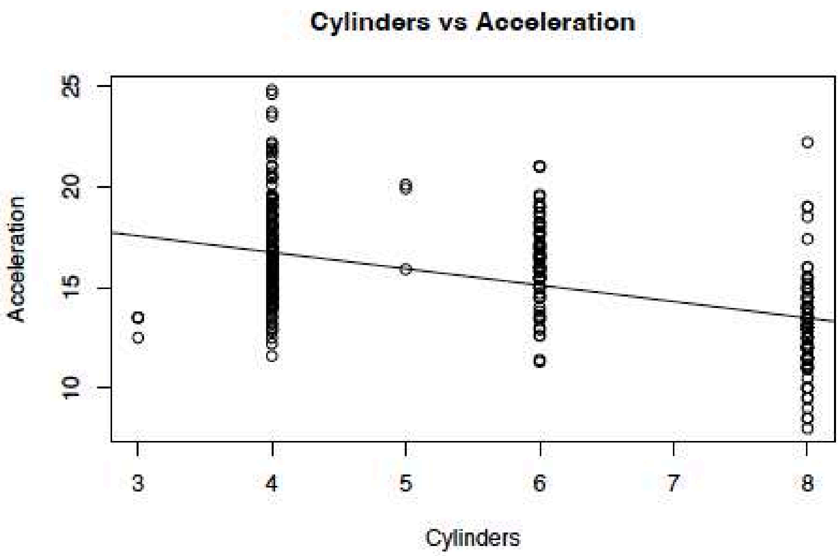
Answer iv.

```
#Predicted acceleration for cylinders = 3.0
new_data <- data.frame(cylinders = 3.0)
predict(model, newdata = new_data, interval = "prediction", level = 0.99)
```

```
##        fit      lwr      upr
## 1 17.55906 11.36164 23.75648
```

Answer (b)

```
plot(Auto$cylinders,Auto$acceleration,main = "Cylinders vs Acceleration",xlab = "Cylinders",ylab = "Acc
abline(model)
```



Answer (b)
```{r}
plot(Auto$cylinders,Auto$acceleration,main = "Cylinders vs Acceleration",xlab = "Cylinders",ylab = "Acceleration")
abline(model)
```

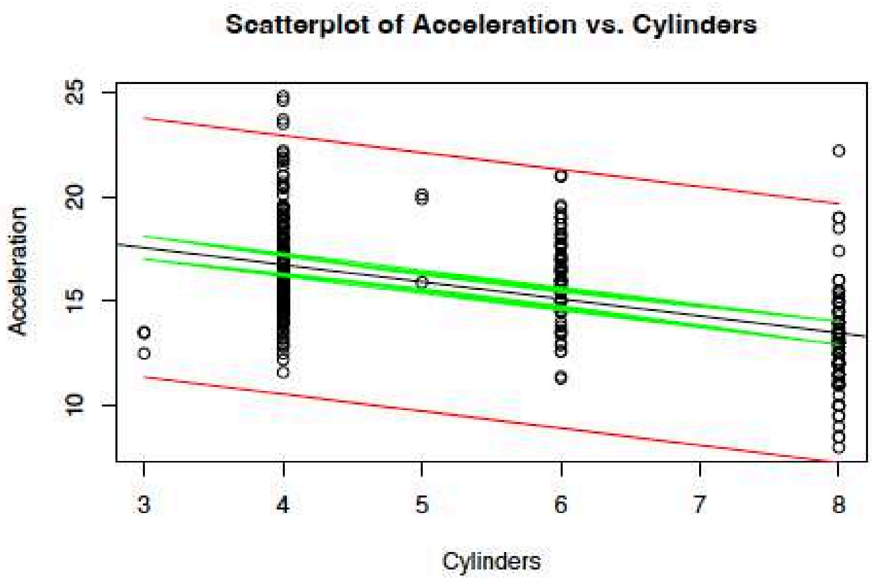
Answer (c)

Answer (c)

```{r warning=FALSE}
```

```
plot(Auto$cylinders, Auto$acceleration, xlab = "Cylinders", ylab = "Acceleration",
    main = "Scatterplot of Acceleration vs. Cylinders")
abline(model)

conf_intr <- predict(model, interval = "confidence", level = 0.99)
pred_intr<- predict(model, interval = "prediction", level = 0.99)

lines(Auto$cylinders, conf_intr[, "lwr"], col = "green")
lines(Auto$cylinders, conf_intr[, "upr"], col = "green")

lines(Auto$cylinders, pred_intr[, "lwr"], col = "red")
lines(Auto$cylinders, pred_intr[, "upr"], col = "red")
```

```r
plot(Auto$cylinders, Auto$acceleration, xlab = "Cylinders", ylab = "Acceleration", main = "Scatterplot (
abline(model)

conf_intr <- predict(model, interval = "confidence", level = 0.99)
pred_intr<- predict(model, interval = "prediction", level = 0.99)

lines(Auto$cylinders, conf_intr[, "lwr"], col = "green")
lines(Auto$cylinders, conf_intr[, "upr"], col = "green")

lines(Auto$cylinders, pred_intr[, "lwr"], col = "red")
lines(Auto$cylinders, pred_intr[, "upr"], col = "red")
```



3. Bayesian networks and naïve Bayes classifiers. Answer a) -

Q.3 (a.)

→ CPT for student

$P(\text{student} = \text{True} \mid \text{Buy compute} = \text{yes}) = \cancel{\frac{6}{12}}\ \frac{1}{2} = 0.5$

$P(\text{student} = \text{False} \mid \text{Buy comput} = \text{yes}) = \frac{1}{2} = 0.5$

$P(\text{student} = \text{True} \mid \text{Buy compute} = \text{NO}) = \cancel{\frac{12}{18}}\ \frac{2}{3} = 0.667$

$P(\text{studen} = \text{False} \mid \text{Buy comput} = \text{No})\ \cancel{\frac{6}{18}}\ \frac{1}{3} = 0.333$

	Student	
Buy compute	True	false
yes	0.5	0.5
NO	0.667	0.33

→ CPT for income.

$P(\text{income} = \text{high} \mid \text{Buy compute} = \text{yes}) = \frac{5}{12} = 0.416$

$P(\text{income} = \text{Low} \mid \text{Buy compute} = \text{yes}) = \frac{7}{12} = 0.583$

$P(\text{income} = \text{high} \mid \text{Buy compute} = \text{NO}) = \frac{7}{15} = 0.466$

$P(\text{income} = \text{low} \mid \text{Buy com} = \text{No}) = \frac{11}{18} = 0.611$

	Income	
Buy compute	Low	high
yes	0.583	0.416
NO	0.611	0.466

→ ~~P(Buy computer~~

P(Buy compute)

yes	no
~~$\frac{1}{3}$ = 0.33~~	~~0.667~~

$\frac{12}{30}$ $\frac{4}{3}$ ~~12~~ ~~10~~ = 0.666 = 0.4

$\cancel{\frac{18}{30}}\ \frac{9}{\cancel{30}}$ = $\frac{15}{30}\ \frac{1}{2}$ $\frac{18}{30}\ \frac{3}{5}$ = 0.6

yes	No
0.4	0.6

* CPT for credit raiting — Credit raiting = CR, Buy computer = BC (3)

① P(CR = fair | income = H, stud = T, BC = y) = 1/2 = 0.5

② P(CR = Excel | income = H, stud = T, BC = y) = 1/2 = 0.5

③ P(CR = fair | income = H, stud = T, BC = N) = 3/4 = 0.75

④ P(CR = Excel | ") = 1/4 = 0.25

⑤ P(CR = fair | income = H, student = F, BC = y) = 1/3 = 0.333

⑥ P(CR = Excellent | ") = 2/3 = 0.667

⑦ P(CR = fair | income = H, Student = F, BC = N) = 1/3 = 0.333

⑧ P(CR = Excellent | ") = 2/3 = 0.667

⑨ P(CR = fair | income = L, Student = T, BC = y) = 1/2 = 0.5

⑩ P(CR = Excellent | ") = 1/2 = 0.5

⑪ P(CR = fair | income = L, Student = T, BC = N) = 3/4 = 0.75

⑫ P(CR = Excellent | ") = 1/4 = 0.25

⑬ P(CR = fair | income = L, Student = F, BC = y) = 1/3 = 0.33

⑭ P(CR = Excellent | ") = 2/3 = 0.667

⑮ P(CR = fair | income = L, Student = F, BC = N) = 2/3 = 0.667

⑯ P(CR = Excellent | ") = 1/3 = 0.33

			credit raiting	
Income	Student	Buy computer	fair	Excellent
High	True	yes	0.5	0.5
High	True	No	0.75	0.25
High	False	yes	0.33	0.66
High	False	No	0.33	0.66
Low	True	yes	0.5	0.5
Low	True	No	0.75	0.25
Low	False	yes	0.33	0.667
Low	False	No	0.667	0.33

* CPT table for Buying Computer

yes	NO
$\frac{12}{20}$ $\frac{2}{5}$ = 0.4	$\frac{18}{30} = \frac{3}{5} = 0.6$

Q.3 (b) Objservation 31, Income = Low, Student = True, Credit Rati = Excellent (3)

⇒ (1) P(Buy computer = yes | Income = low, Student = True, credit rate = Excellent)

= P(Credit rating = Excellent | income = Low, Student = True, Buy comp = yes) *

P(student = True | Buy computer = yes) * P(income = Low | Buy com = yes)

* P(Buy computer = yes)

= 0.5 * 0.5 * 0.583 * 0.4 = 0.0583 ~~0.06~~

⇒ P(Buy computer = NO | Income = low, student = True, CR = Excellent) =

P(CR = Excellent | income = low, Student = True, Buy comp = NO) *

P(student = True | Buy comput = NO) * P(income = Low | Buy com = NO)

P(Buy compu = NO)

⇒ 0.25 * 0.667 * 0.611 * 0.6 = 0.0611.

So, for objservation 31, ~~To~~ predict $\boxed{\text{Buy computer = yes.}}$

P(Buy computer = NO, income = Low, studen = True, CR = Excellent) >

P(Buy computer = yes, income = Low, student = True, CR = Excellent)

Ans: $\boxed{\text{So, for observation 31, to predict Buy computer = NO}}$

②

Q.3 (b) (ii) Observation 32 → Income = High, Student = True, CR = fair, BC = ?

BC = ?

BC → Buy computer
CR → credit rating

If BC = yes

P(BC = yes, Income = high, Student = true, CR = fair) =

P(Student = True | BC = yes) * P(income = high | BC = yes) *

P(CR = fair | income = high, student = True, BC = yes) * P(Buy comp = yes)

⇒ 0.5 * 0.416 * 0.5 * 0.4

⇒ 0.0416

If BC = NO

P(BC = NO, Income = high, Student = true, CR = fair) = P(Student = true | BC = yes)

* P(income = high | BC = NO) * P(CR = fair | income = high, Student = T, BC = N) * P(BC = N)

⇒ 0.667 * 0.466 * 0.75 * 0.6

⇒ 0.13986 ≃ 0.14

P(BC = yes; income = high, Stud = True, CR = fair) < P(BC = NO, Income = high, Stud = True, CR = fair) ⇒ 0.0416 < 0.14

Ans Hence Prediction is Buy computer = NO

Testing obs	Inc	Inc	Cred R	Buy com
ob 31				NO
ob 32				NO

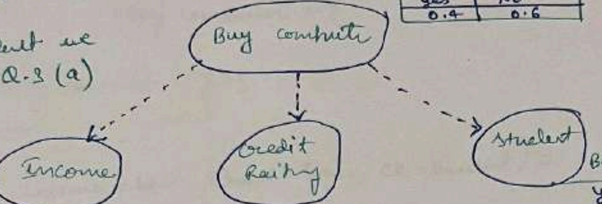
Q.3(c)

CR → Credit Raiting
BC → Buy computer

→ Estimating the joint probability by assuming that all predictors are independent with each other.

Buy com	
yes	NO
0.4	0.6

→ Income, Student we can use from Q.3 (a)



Income

Buy computer	Low	High
yes	0.583	0.416
NO	0.611	0.466

Student

BC	True	False
yes	0.5	0.5
NO	0.667	0.33

Joint probability for Income, credit Raiting, Student, Buy computer.

$$P(\text{Buy computer} \mid \text{student}, \text{income}, \text{credit raiting}) \propto P(\text{income} \mid \text{Buy computer})$$

$$P(\text{student} \mid \text{Buy computer})\, P(\text{credit raeting} \mid \text{Buy computer})\, P(\text{Buy computer})$$

$\Rightarrow P(CR \mid \text{Buy computer}) = ?$

$P(Cr = \text{fair} \mid \text{Buy comput} = \text{yes}) = \frac{5}{14} \Rightarrow \frac{7}{12}\ \frac{5}{12}$

$P(Cr = \text{fair} \mid \text{Buy comput} = NO) = \frac{12}{17} \Rightarrow \frac{6}{18} = \frac{1}{3}$

$P(Cr = \text{Excellent} \mid BC = \text{yes}) = \frac{7}{13} \Rightarrow \frac{7}{12}$

$P(Cr = \text{Excellent} \mid BC = NO) = \frac{6}{13} \Rightarrow \frac{12}{18}\ \frac{2}{3}$

$P(cr = \text{fair} \mid BC = \text{yes}) = \frac{5}{12}$

$P(cr = \text{Excellent} \mid BC = \text{yes}) = \frac{7}{12}$

$P(cr = \text{fair} \mid BC = NO) = \frac{12}{18}\ \frac{2}{3}$

$P(cr = \text{Excellent} \mid BC = NO) = \frac{6}{18}\ \frac{1}{3}$

Credit Raiting

Buy computer	fair	Excellent
yes	5/12 = 0.4166	7/12 = 0.5833
No	2/3 = 0.666	1/3 = 0.333

(8)

Question 3(d)

(i) Observation 31 → Income = low, Student = True, Credit Rating = Excellent

Buy computer = ?

For Buy compute = yes

$\Rightarrow P(BC = yes, income = low, Student = True, CR = Excellent) =$

$P(income = low | BC = yes) \times P(Student = True | BC = yes) \times P(CR = Excellent | BC = yes)$

$\times P(Buy\ compute = yes)$

$\Rightarrow 0.583 * 0.5 * 0.5833 * 0.4 = 0.0680$

for Buy compute = NO

$P(BC = NO, income = low, student = True, CR = Excellent) =$

$P(income = low | BC = NO) * P(student = True | BC = NO) \times P(CR = Excellent | BC = NO)$

$* P(Buy\ compute = NO)$

$\Rightarrow 0.611 * 0.667 * 0.333 * 0.6 \Rightarrow 0.08142$

$P(BC = NO, income = low, student = True, CR = E) > P(BC = NO, income = low,$

$student = True, CR = Excellent)$

$0.08142 > 0.0680$

$\Rightarrow$ Hence Prediction of Buy computer = NO for obser 31

Q.3 (d) ⑦

(ii) Observation 33, income = High, Student = True, CR = Fair, BC = ?

for Buy computer = yes

P(BC=yes, income = High, Student = True, CR = fair) = P(income = High | BC = yes)

* P(Student = True | BC=yes) * P(CR=fair | BC = yes) * P(BC = yes)

$\Rightarrow$ 0.416 * 0.5 * 0.5833 * 0.4

$\Rightarrow$ 0.0485

for Buy compute = NO

P(BC=NO, income = High, Student = True, CR = fair) = P(income = High | BC=NO)

* P(Student = True | BC = NO) * P(CR = fair | BC = NO) * P(BC = NO)

$\Rightarrow$ 0.466 * 0.667 * 0.666 * 0.4

$\Rightarrow$ 0.083

Hence, P(BC=NO, income = High, Student = True, CR = fair) >

P(BC = yes, income = High, Student = True, CR = fair)

$\Rightarrow$ 0.083 > 0.0485

Here Predition of $\boxed{\text{Buy compute} = \text{NO}}$

Result

Testing obs	Income	Student	Credit Rating	Buy computer
Observation 31	Low	True	Excellent	NO
observation 32	High	True	fair	NO

4. Predicting wine quality by using support vector machine classification algorithm.

```
#prep data
library(e1071)
library(ROCR)
# Reading the data
training_data <- read.csv('WineQuality_Training.txt', header = TRUE, sep = ",")
```

```
test_data <- read.csv('WineQuality_Testing.txt', header = TRUE, sep = ",")

training_data$quality <- as.factor(training_data$quality)
test_data$quality <- as.factor(test_data$quality)
```

Answer a.-

```
set.seed(1)
tune.grid <- expand.grid(C = c(0.01, 1, 100))
svm_tune <- tune.svm(x = training_data[, -ncol(training_data)],
                     y = training_data$quality,
                     kernel = "linear",
                     cost = tune.grid$C)

svm_model <-  svm_tune$best.model
print(svm_model)
```

```
##
## Call:
## best.svm(x = training_data[, -ncol(training_data)], y = training_data$quality,
##     cost = tune.grid$C, kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##
## Number of Support Vectors:  1710
```

Answer b:

```
svm_linear_model <- svm(quality ~ ., data = training_data, kernel = "linear", cost = 1)

predictions <- predict(svm_linear_model, test_data)
accuracy <- mean(predictions == test_data$quality)
confusion_matrix <- table(predicted = predictions, actual = test_data$quality)
print(confusion_matrix)
```

```
##          actual
## predicted Bad Good
##      Bad  104   89
##      Good  38  169
```

```
cat("Prediction accuracy:", accuracy,"\n")
```

```
## Prediction accuracy: 0.6825
```

Answer c:

```r
set.seed(1)
ranges <- list(cost = c(0.01, 1, 100), gamma = c(0.01, 1, 100))
svm_cv_r <- tune(svm, quality ~ ., data = training_data, kernel = "radial", ranges = ranges)
summary(svm_cv_r)
```

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost gamma
##   100     1
##
## - best performance: 0.1556667
##
## - Detailed performance results:
##    cost gamma     error dispersion
## 1 1e-02 1e-02 0.2826667 0.02647151
## 2 1e+00 1e-02 0.2340000 0.01755415
## 3 1e+02 1e-02 0.2003333 0.03233505
## 4 1e-02 1e+00 0.5060000 0.04870673
## 5 1e+00 1e+00 0.1623333 0.03067351
## 6 1e+02 1e+00 0.1556667 0.02923088
## 7 1e-02 1e+02 0.5120000 0.03103960
## 8 1e+00 1e+02 0.3253333 0.03006988
## 9 1e+02 1e+02 0.3253333 0.03006988
```

Answer d:

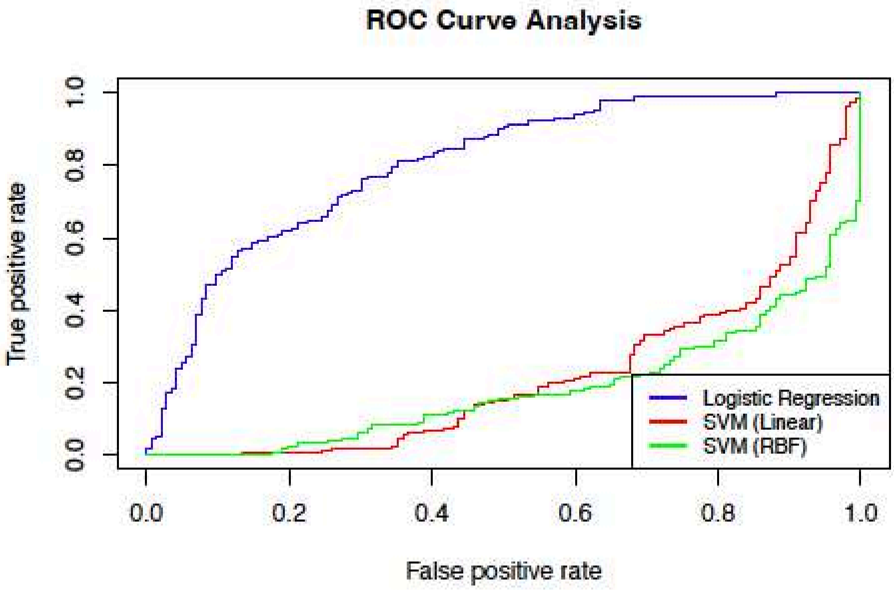
```r
svm_model <- svm(quality ~ ., data = training_data, kernel = "radial",cost=100,gamma=1)
predictions <- predict(svm_model, newdata = test_data)
accuracy <- mean(predictions == test_data$quality)
#conf_matrix <- table(Actual = test_data$quality, Predicted = predictions)
#conf_matrix
print(paste("Classification Accuracy:", accuracy))
```

```
## [1] "Classification Accuracy: 0.64"
```

Answer e:

```r
logit_model <- glm(quality ~ ., data = training_data, family = "binomial")
logit_predictions <- predict(logit_model, newdata = test_data, type = "response")

svm_linear_model <- svm(quality ~ ., data = training_data, kernel = "linear", probability = TRUE)
svm_linear_predictions <- attr(predict(svm_linear_model, newdata = test_data, probability = TRUE),
                               "probabilities")

svm_r_model <- svm(quality ~ ., data = training_data, kernel = "radial", probability = TRUE)
svm_r_predictions <- attr(predict(svm_r_model, newdata = test_data, probability = TRUE),
                          "probabilities")
```

```r
roc_logit <- prediction(logit_predictions, test_data$quality)
roc_svm_linear <- prediction(svm_linear_predictions[,2], test_data$quality)
roc_svm_rbf <- prediction(svm_r_predictions[,2], test_data$quality)

roc_logit_perf <- performance(roc_logit, "tpr", "fpr")
roc_svm_linear_perf <- performance(roc_svm_linear, "tpr", "fpr")
roc_svm_rbf_perf <- performance(roc_svm_rbf, "tpr", "fpr")

plot(roc_logit_perf, col = "blue", main = "ROC Curve Analysis")
plot(roc_svm_linear_perf, col = "red", add = TRUE)
plot(roc_svm_rbf_perf, col = "green", add = TRUE)


legend("bottomright", legend = c("Logistic Regression", "SVM (Linear)", "SVM (RBF)"),
       col = c("blue", "red", "green"), lwd = 2, cex = 0.8)
```



»»Based on ROC curves, Logistic regression model has the highest overall cumulative error rate (AUC), followed by RBF regression model and then linear regression model. This indicates that the Logistic regression model is better at the classification of positive and negative instances than the other 2 models in this dataset.