장비 에러와 정비 데이터의 상관관계 분석을 통한 예측 정비 시스템 구축 가능성 평가

이두원*, 서태원*
*고려대학교 컴퓨터정보통신대학원 소프트웨어공학과
e-mail : ldoowon@korea.ac.kr

Evaluation of possibility of constructing predictive maintenance system through correlation analysis between equipment error and maintenance data

Doo-Won Lee*, Tae-Weon Suh*
*Dept. of Software Engineering, Korea University Graduate School of Computer & Information Technology
**Dept. of Computer Science and Engineering, Korea University

요 약

제조 산업에서 4 차 산업혁명의 바람이 불면서 다양한 시도들이 진행되고 있다. 이러한 노력에도 불구하고 실제 제조 환경에서 효과적으로 시스템을 구축하는데 어려움을 겪고 있는 장비 정비 예측 시스템에 대한 새로운 시도를 통해 그 가능성을 평가해 보았다. 이 논문에서는 최근 여러 분야에서 성능 적으로 상당한 성과를 올리고 있는 Machine Learning 기반으로 예측 성능 평가를 진행 했다.

1. 서론

최근 제조 산업에선 Smart Factory 를 혁신 목표로 제시하여 Industry 4.0 Platform 선점을 위해 노력 중이다. Smart Factory 를 구성하는 주요 기능 중 PMS(Predictive Maintenance System)라는 장비 예측 정비 시스템에 대한 관심이 높아지고 있다. 이는 미리 장비의 데이터를 실시간으로 수집하여 Machine Learning 을 통해 장비의 정비 시점을 미리 예측하여 장비를 효율적으로 관리하고 생산 스케줄과 연동하여 생사 계획을 최적화 할 수 있는 기반 정보로 활용될 수 있다. 이번 데이터 분석을 통해 그 동안 수집하여 조회하는 용도로만 사용된 데이터의 연관성을 실행을 통해 조사할 수 있는 기회가 되었으며 예측 가능성을 분석하여 새로운 방향을 제시할 수 있는 기회가 될 것이다.

2. 평가를 위한 데이터 선정

데이터는 실제 제조 환경에서 장비 로부터 수집됐으며, 반도체 공정 중 Wire Bonding 공정의 A 사 장비업체의 I 모델을 선정하였다. 종속변수는 장비 별 Maintenance Count 값이며 독립변수로 장비 가동 정보 중 에러 관련 정보를 데이터로 선정하였다. 일주일간의 데이터를 수집하여 에러 종류 별 Maintenance 와의 연관성을 분석하고 특정 에러의 Input 에 따른 Maintenance 의 예측이 가능한지를 조사하는데 의미가 있다. 데이터는 OE(Overall Equipment Efficiency) 시스템에서 추출하였으며, 장비 가동률과 Maintenance 의 연관성을 확인할 수 있는 기회가 될 것이다. <표 1> 에 독립 변수 별 의미를 설명하였다.

<표 1> 예측 정비 상관관계 변수와 설명

변수	설명
MFG_TIME	– 생산 시간(Run + Idle Time)
DEVICE_QTY	– 생산 수량(Output)
GROSS_UPH	– Unit Per Hour(MFT Time 동안의 생산 수량)
NET_UPH	– Unit Per Hour(Run Time 동안의 생산 수량)
BONDING_ALARM	– Bonding 에러
EFO_ALARM	– EFO 에러
INSPECTION_ALARM	– Inspection 에러
MHS_ALARM	– MHS 에러
MATERIAL_ALARM	– Material 에러
NSOL_ALARM	– NSOL 에러
NSOP_ALARM	– NSOP 에러
PR_ERROR_ALARM	– PR 에러
SHOR_TAIL_ALARM	– Shor Tail 에러
SYSTEM_ALARM	– System 에러
IDLE	– 장비 대기 시간
RUN	– 장비 가동 시간
DOWN	– 장비 다운 시간
POWER_OFF	– 장비 Shutdown 시간

3. 데이터 상관관계 분석

3.1 카이제곱 분석을 통한 연관성

R 프로그램을 통해 카이제곱 분석을 통해 연관성을 확인한다. (p-value 0.05 이하)
Chisq.test 명령을 통해 MAINT_COUNT 에 대한 개별 독립 변수들의 상관관계 결과를 <표 2> 에 작성하였다. 변수들 만으로 연관성의 깊이를 예측 가능한 부분도 있으므로 예측과 실제 결과에 대한 비교도 포함하였다.

<표 2> 카이제곱의 연관성 분석 결과 및 해석

변수	연관성 분석		해석
	예상	결과	
MAINT_COUNT			
MAINT_WAIT_TIME	높음	높음	정비 건수 만큼 정비 대기 시간은 비례
MAINT_TIME	높음	높음	정비 건수 만큼 정비 시간은 비례
ASSIST_ALARM_COUNT	높음	낮음	정비 건수와 경알람은 연관성이 낮음
FAILURE_ALARM_COUNT	높음	높음	정비 건수와 중알람은 연관성이 높음
UNDEFINED_ALARM_COUNT	낮음	높음	아직 정의되지 않음 알람 중 정비를 필요로 하는 알람이 많음을 의미, 알람의 세분화 등록 필요
MTBA	낮음	낮음	경 알람 발생 건수가 낮을 수록 MTBA 는 증가하므로 정비와는 반비례 관계
MTBF	낮음	낮음	중 알람 발생 건수가 낮을 수록 MTBA 는 증가하므로 정비와는 반비례 관계
TOTAL	무관	낮음	무의미
MFG_TIME	낮음	낮음	제조 시간과 정비시간은 반비례 증명
DEVICE_QTY	낮음	낮음	생산량과 정비 건수는 반비례 증명
GROSS_UPH	낮음	낮음	생산량 수치도 정비 건수와 반비례 증명
NET_UPH	낮음	낮음	생산량 수치도 정비 건수와 반비례 증명
BONDING_ALARM	높음	높음	정비 건수와 관련있는 에러로 판단됨
EFO_ALARM	높음	높음	정비 건수와 관련있는 에러로 판단됨
INSPECTION_ALARM	높음	높음	정비 건수와 관련있는 에러로 판단됨
MHS_ALARM	높음	높음	정비 건수와 관련있는 에러로 판단됨
MATERIAL_ALARM	높음	매우높음	정비 건수와 특별히 관련이 높음. 주의깊게 모니터 필요한 에러항목으로 판단됨
NSOL_ALARM	높음	낮음	에러 중 정비와 무관한 에러로 판단됨
NSOP_ALARM	높음	높음	정비 건수와 관련있는 에러로 판단됨
PR_ERROR_ALARM	높음	낮음	에러 중 정비와 무관한 에러로 판단됨
SHORT_TAIL_ALARM	높음	높음	정비 건수와 관련있는 에러로 판단됨
SYSTEM_ALARM	높음	높음	정비 건수와 관련있는 에러로 판단됨
IDLE	높음	높음	자재 대기 시간도 정비 건수와 비례, 정비로 인한 장비 대기 시간이 높음을 의미
RUN	낮음	낮음	장비 가동 시간과 정비 건수를 반비례 증명
DOWN	높음	높음	정비 대기 시간과 정비 건수 비례
POWER_OFF	높음	낮음	장비 Power Off 는 정비와는 무관함을 의미

3.2 다중공선성 분석

회귀분석에서 정확한 선형 관계 분석을 위한 과정으로 회귀분석을 수행하기 전 변수들의 다중 공선성 확인을 통해 마이너스 적 변수를 확인한다. VIF 값이 10 미만의 수치를 보인다면 다중 공선성(Multicollinearity) 문제가 없다고 판단할 수 있다. 즉, 선정된 데이터는 분석을 위해 사용해도 무방할 정도의 데이터 분포가 균형을 이룬다고 볼 수 있다. 이번 분석에서 선택된 데이터 중 'GROSS_UPH'는 거의 10에 가까운 값을 보이지만, 이번 연구에서 중요한 변수로 사용되지 않으므로 10 미만의 값으로써 그대로 사용하기로 하였다. 다음은 수행 결과를 나타낸다.

<표 3> 다중공선성 마이너스적 변수

	Variables	VIF
1	MAINT_WAIT_TIME	1.437326e+00
2	MAINT_TIME	1.155012e+00
3	ASSIST_ALARM_COUNT	1.343148e+06
4	FAILURE_ALARM_COUNT	Inf
5	UNDEFINED_ALARM_COUNT	1.205276e+00
6	MTBA	3.312388e+00
7	MTBF	1.110674e+00
8	TOTAL	1.113700e+00
9	MFG_TIME	3.428816e+07
10	DEVICE_QTY	1.185445e+01
11	GROSS_UPH	9.370455e+01
12	NET_UPH	7.169960e+01
13	BONDING_ALARM	3.896982e+03
14	EFO_ALARM	5.597343e+04
15	INSPECTION_ALARM	1.309769e+04
16	MHS_ALARM	1.486302e+05
17	MATERIAL_ALARM	2.992176e+04
18	NSOL_ALARM	1.180347e+05
19	NSOP_ALARM	1.825546e+05
20	PR_ERROR_ALARM	4.444685e+05
21	SHORT_TAIL_ALARM	7.028243e+04
22	SYSTEM_ALARM	Inf
23	IDLE	2.500682e+00
24	RUN	3.294921e+07
25	DOWN	3.418406e+06
26	POWER_OFF	5.956304e+00

3.3 회귀분석 수행

*이 많을수록 깊은 영향력이 있음을 의미한다. T 값의 양수는 클수록 영향력이 크고 비례하여 증가하며, 음수는 반대로 반비례관계를 의미한다. 최적의 모델을 뽑으면 Adjusted R-squared 값이 더 낮아 진다.

<표 4> 회귀 분석을 통한 영향력 평가

```
Residuals:
     Min       1Q   Median       3Q      Max
-7.9188 -1.3683 -0.2525  1.0059 16.8175

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.579e+04  6.813e+03  -2.318   0.0209 *
Data$MAINT_WAIT_TIME     2.994e-02  1.350e-03  22.174   <2e-16 ***
Data$MAINT_TIME          2.010e-03  5.474e-05  36.710   <2e-16 ***
Data$ASSIST_ALARM_COUNT  5.173e-01  8.265e-01   0.626   0.5317
Data$FAILURE_ALARM_COUNT 5.489e-02  2.771e-02   1.981   0.0482 *
Data$UNDEFINED_ALARM_COUNT 1.256e-02 1.105e-02  1.137   0.2562
Data$MTBA               -6.584e-03  1.131e-02  -0.582   0.5608
Data$MTBF                1.424e-05  6.128e-05   0.232   0.8164
Data$TOTAL               1.075e+02  4.639e+01   2.318   0.0209 *
Data$MFG_TIME           -1.452e+01  2.288e+01  -0.635   0.5259
Data$DEVICE_QTY          5.860e-06  1.150e-05   0.510   0.6106
Data$GROSS_UPH          -6.926e-03  3.372e-03  -2.054   0.0405 *
Data$NET_UPH             5.459e-03  2.681e-03   2.036   0.0423 *
Data$BONDING_ALARM      -5.091e-01  8.268e-01  -0.616   0.5384
Data$EFO_ALARM          -5.031e-01  8.269e-01  -0.608   0.5432
Data$INSPECTION_ALARM   -5.110e-01  8.263e-01  -0.618   0.5366
Data$MHS_ALARM          -5.150e-01  8.266e-01  -0.623   0.5335
Data$MATERIAL_ALARM     -5.022e-01  8.268e-01  -0.607   0.5439
Data$NSOL_ALARM         -5.098e-01  8.263e-01  -0.617   0.5375
Data$NSOP_ALARM         -5.094e-01  8.266e-01  -0.616   0.5380
Data$PR_ERROR_ALARM     -5.131e-01  8.264e-01  -0.621   0.5350
Data$SHORT_TAIL_ALARM   -5.174e-01  8.265e-01  -0.626   0.5316
Data$SYSTEM_ALARM               NA         NA      NA       NA
Data$IDLE               -1.948e-02  1.659e-02  -1.174   0.2408
Data$RUN                 1.452e+01  2.288e+01   0.635   0.5260
Data$DOWN                1.451e+01  2.288e+01   0.634   0.5262
Data$POWER_OFF           1.274e-02  1.060e-02   1.201   0.2303
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.518 on 473 degrees of freedom
Multiple R-squared:  0.8638,    Adjusted R-squared:  0.8566
F-statistic:   120 on 25 and 473 DF,  p-value: < 2.2e-16
```

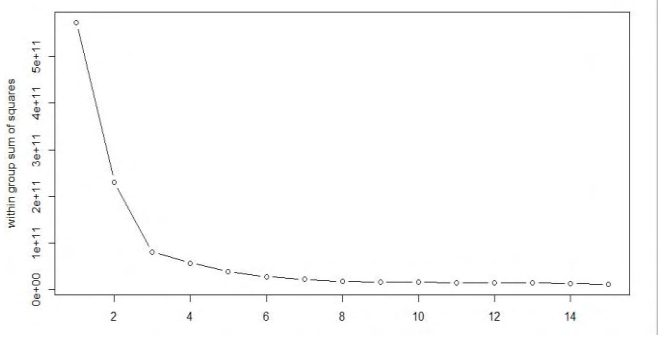
3.4 Backword 후진 제거

후진 제거 방법을 통한 기여도가 낮은 변수를 제거하여 최적의 모델을 찾을 수 있다.

<표 5> 후진 제거를 통한 기여도 평가

```
> summary(reduced)

Call:
lm(formula = Data$MAINT_COUNT ~ MAINT_WAIT_TIME + MAINT_TIME +
    FAILURE_ALARM_COUNT + UNDIFINED_ALARM_COUNT + EFO_ALARM +
    MATERIAL_ALARM + NSOL_ALARM + NSOP_ALARM + PR_ERROR_ALARM +
    MAINT_OVER5, data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7223 -1.2385 -0.1092  1.0375 17.5630

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.105e-01  2.099e-01   2.431  0.01540 *
MAINT_WAIT_TIME          2.253e-02  1.239e-03  18.189  < 2e-16 ***
MAINT_TIME               1.768e-03  4.891e-05  36.150  < 2e-16 ***
FAILURE_ALARM_COUNT      4.586e-02  2.355e-02   1.948  0.05201 .
UNDIFINED_ALARM_COUNT 1.403e-02  9.222e-03   1.521  0.12884
EFO_ALARM                1.022e-02  3.259e-03   3.136  0.00182 **
MATERIAL_ALARM           1.142e-02  4.338e-03   2.632  0.00876 **
NSOL_ALARM               3.763e-03  2.136e-03   1.762  0.07870 .
NSOP_ALARM               5.316e-03  1.825e-03   2.912  0.00375 **
PR_ERROR_ALARM           1.884e-03  1.105e-03   1.706  0.08874 .
MAINT_OVER5Yes           3.550e+00  2.660e-01  13.345  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.192 on 488 degrees of freedom
Multiple R-squared:  0.8935,    Adjusted R-squared:  0.8913
F-statistic: 409.4 on 10 and 488 DF,  p-value: < 2.2e-16
```

3.5 K-Means clustering Algorithm



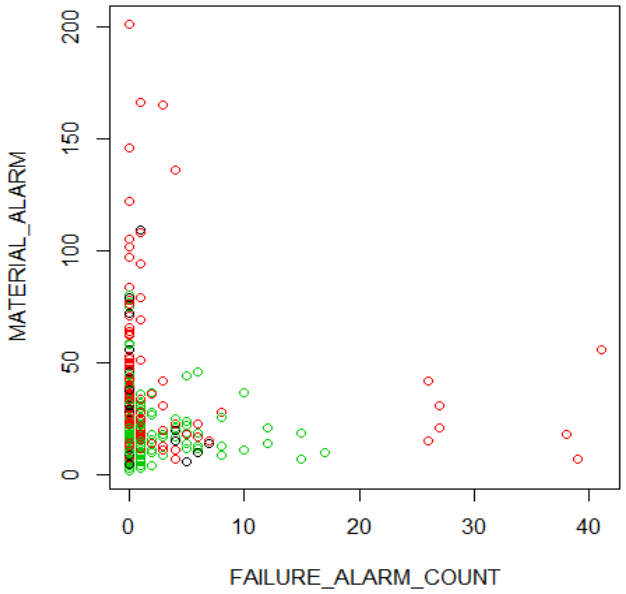
(a) Number of Clusters

Fig.1. Eblow Method 클러스터 설정 선택

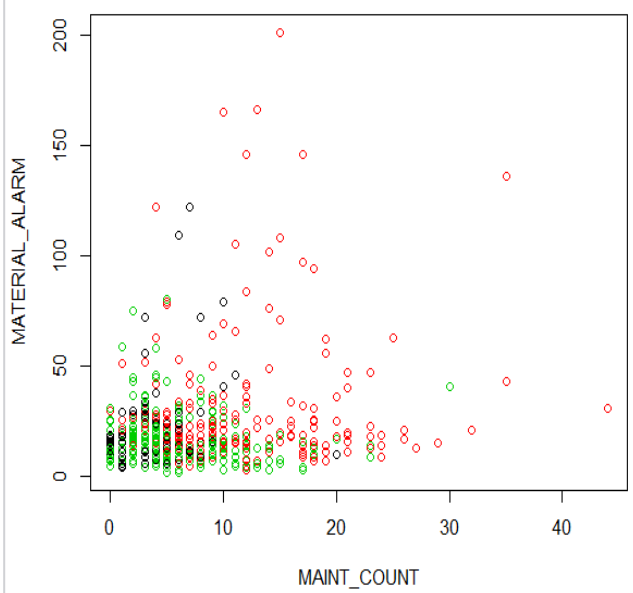
Elbow Method 결과에 따라 K-Means 클러스터링을 수행한다.

```
> Data_cluster$centers
  MAINT_COUNT MAINT_WAIT_TIME MAINT_TIME ASSIST_ALARM_COUNT FAILURE_ALARM_COUNT UNDIFINED_ALARM_COUNT
1  -0.5778458      -0.4587775 -0.3196433         -0.3144855         -0.19069401           -0.2538347
2   0.6307042       0.4307280  0.3762565          0.9645183          0.13198433            0.2854524
3  -0.2609571      -0.1467470 -0.1679813         -0.6783036         -0.02038187           -0.1218820
        MTBA       MTBF      TOTAL   MFG_TIME DEVICE_QTY  GROSS_UPH    NET_UPH BONDING_ALARM    EFO_ALARM
1  0.5946316 -0.06912393 -0.0420354  0.3900603  1.3957671  1.2592858  1.1843031   -0.50571947 -0.08790472
2 -0.6391582  0.02699927 -0.1681186  0.2930950 -0.1389005 -0.1141323 -0.1139646    0.29512199  0.45561061
3  0.2600202  0.01060388  0.1657424 -0.4443521 -0.5648160 -0.5192430 -0.4826038   -0.00666273 -0.35016345
  INSPECTION_ALARM  MHS_ALARM MATERIAL_ALARM NSOL_ALARM NSOP_ALARM PR_ERROR_ALARM SHORT_TAIL_ALARM
1      -0.02356799 -0.1103147     -0.1225129 -0.2978347 -0.2348181     0.08499209      -0.37583610
2       0.08264096  0.3479187      0.3690512  0.3416341  0.6515541     0.47837822       0.26281931
3      -0.05977465 -0.2462094     -0.2584671 -0.1487935 -0.4472327    -0.45463392      -0.04249519
  SYSTEM_ALARM       IDLE         RUN        DOWN  POWER_OFF
1  -0.19069401 -0.37923067  0.518799960 -0.3753122 -0.1417147
2   0.13198433  0.04408811  0.002761283  0.9196995 -0.2092133
3  -0.02038187  0.14798050 -0.256889201 -0.6097762  0.2501149
```

K-Means 클러스터링 결과는 z-score 로 계산되어 0 을 평균으로 평균보다 높은지, 낮은지 확인할 수 있다. 군집에 따라 독립변수들의 만족도 파악 및 군집의 시각화를 표현한다.



(a) Material Alarm



(b) Material Alarm

Fig.2. Visualization

4. KNN 예측 분석 실험 및 고찰

데이터 정리를 통해 연관성을 고려하여 변수 재 정리 진행 하였다.

- MAINT_COUNT
- MAINT_WAIT_TIME
- MAINT_TIME
- FAILURE_ALARM_COUNT
- UNDIFINED_ALARM_COUNT
- BONDING_ALARM
- EFO_ALARM
- INSPECTION_ALARM
- MHS_ALARM
- MATERIAL_ALARM
- NSOL_ALARM
- NSOP_ALARM
- PR_ERROR_ALARM
- SHORT_TAIL_ALARM
- SYSTEM_ALARM
- MAINT_OVER5

[실험 #1]

에러 빈도수에 따른 정비 건수 6 건 이상을 1 로, 5 건 이하를 0 으로 표현한 것이다.

```
normalize <- function(x){return((x - min(x)) / (max(x) - min(x)))}
Data_n <- as.data.frame(lapply(Data[1:15], normalize))
summary(Data$MAINT_COUNT)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	3.00	6.00	7.91	11.00	44.00

```
Data_train <- Data_n[1:450, ]
Data_test <- Data_n[451:499, ]
Data_train_labels <- Data[1:450, 16]
Data_test_labels <- Data[451:499, 16]
library(class)
Data_test_pred <- knn(train = Data_train, test = Data_test, cl = Data_train_labels, k=22)
library(gmodels)
CrossTable(x = Data_test_labels, y = Data_test_pred, prop.chisq = FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  49


                | Data_test_pred
Data_test_labels |         0 |         1 | Row Total |
-----------------|-----------|-----------|-----------|
               0 |        23 |         1 |        24 |
                 |     0.958 |     0.042 |     0.490 |
                 |     0.852 |     0.045 |           |
                 |     0.469 |     0.020 |           |
-----------------|-----------|-----------|-----------|
               1 |         4 |        21 |        25 |
                 |     0.160 |     0.840 |     0.510 |
                 |     0.148 |     0.955 |           |
                 |     0.082 |     0.429 |           |
-----------------|-----------|-----------|-----------|
    Column Total |        27 |        22 |        49 |
                 |     0.551 |     0.449 |           |
-----------------|-----------|-----------|-----------|
```

1 의 예측 정확도 = 23/24(95.84%)

0 의 예측 정확도 = 21/25(84%)

이번 실험에서 중요한 것은 에러 건수 별 정비 필요성에 대한 예측이므로 0 보다는 1 의 정확도가 중요하다. [실험 #2] 를 통해 6 건 이상의 정비 건이 발생하는 경우는 예측 정확도가 높게 나타났으며 동일 조건에서 특정 에러에 대해서도 거의 동일한 정확도를 확인 할 수 있었다.

5. Decision Tree Classification Algorithm

이번에는 의사결정트리 알고리즘을 활용하여 트레이닝 데이터를 통해 예측 정확도를 측정해보도록 하겠다

```
> summary(Data_model)

Call:
C5.0.default(x = Data_train[-9], y = Data_train$MAINT_OVER5)


C5.0 [Release 2.07 GPL Edition]        Sun Jun 18 14:27:53 2017
-----------------------------

Class specified by attribute `outcome'

Read 450 cases (16 attributes) from undefined.data

Decision tree:

MAINT_OVER5 = No: No (211)
MAINT_OVER5 = Yes: Yes (239)


Evaluation on training data (450 cases):

          Decision Tree
        ----------------
        Size      Errors

           2    0( 0.0%)   <<


         (a)   (b)    <-classified as
        ----  ----
         211          (a): class No
               239    (b): class Yes


        Attribute usage:

        100.00% MAINT_OVER5


Time: 0.0 secs
```

트리 성능 확인 및 성능 평가 테이블 생성

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  49


             | predicted left
 actual left |        No |       Yes | Row Total |
-------------|-----------|-----------|-----------|
          No |        26 |         0 |        26 |
             |     0.531 |     0.000 |           |
-------------|-----------|-----------|-----------|
         Yes |         0 |        23 |        23 |
             |     0.000 |     0.469 |           |
-------------|-----------|-----------|-----------|
Column Total |        26 |        23 |        49 |
-------------|-----------|-----------|-----------|
```

6. SVM Classification Algorithm

Training Data = 80%, Test Data = 20%

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  49

                | Data_predictions
Data_test$MAINT_OVER5 |        No |       Yes | Row Total |
----------------------|-----------|-----------|-----------|
                   No |        24 |         0 |        24 |
                      |     1.000 |     0.000 |     0.490 |
                      |     0.960 |     0.000 |           |
                      |     0.490 |     0.000 |           |
----------------------|-----------|-----------|-----------|
                  Yes |         1 |        24 |        25 |
                      |     0.040 |     0.960 |     0.510 |
                      |     0.040 |     1.000 |           |
                      |     0.020 |     0.490 |           |
----------------------|-----------|-----------|-----------|
         Column Total |        25 |        24 |        49 |
                      |     0.510 |     0.490 |           |
----------------------|-----------|-----------|-----------|
```

No의 예측 정확도 = 24/24(100%)

Yes의 예측 정확도 = 24/25(96%)

7. 결론과 연구 방향

이번 연구를 통해 예측 정비 시스템을 구축하기 위한 정비 이력 및 정비 별 에러와의 상관관계를 파악할 수 있었던 좋은 기회가 되었으며, 에러의 실시간 데이터 수집에 따른 예측 정비의 가능성을 평가해볼 수 있었다. 또한, 결과를 도출하면서 Training Data 에 대한 중요성을 깨닫고 신뢰할 수 있는 데이터를 바탕으로 평가 진행해야지만 보다 정확한 예측 정확도를 추출할 수 있을 것이라는 판단을 내릴 수 있었다. 현재 예측 정비 분야는 여러 시도를 통해 그 가능성을 평가해오고 있고 최적의 솔루션을 찾기 위한 노력이 진행되고 있으므로 앞으로도 다양한 이론을 통한 다양한 접근법으로 예측 정비 시스템의 효과적인 구축에 한발 다가설 수 있는 기회가 되기를 바란다.

참고문헌

[1] Gongde Guo. ‘KNN Model-Based Approach in Classification”

[2] Qing-yun Dai. “Research of Decision Tree Classification Algorithm in Data Mining”