빅데이터 시각화 기말고사 최종 보고서

정재원

2025-05-27

1. 서론

1.1 주제 : 당뇨병 진단지표

본 주제 : 피마족의 당뇨병 데이터 분석

본 과제 목표 : 인슈린 농도(Insulin)가 당뇨 여부와 관계를 보이는지 확인하고 이를 기반으로 당뇨병 진단에 유용한 지표를 시각화하고 해석하는 것을 목표로 한다.

핵심 질문 : 인슈린 농도(Insulin)가 다른 지표(Glucose, BMI, Age 등)와 함께 당뇨병 진단에 얼마나 유의미한 정보를 제공하는가?

1.2. 데이터 설명

데이터 셋 출처 : UCI Machine Learning Repository CDC Diabetes Health Indicators

2. 자료 설명

데이터 셋의 제시, 시각화

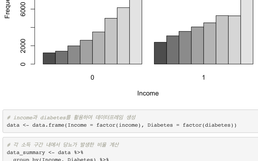
- Diabetes_binary : 당뇨병 유무 (0 : 정상, 1 : 당뇨병)
- HighChol : 콜레스테롤 이상증 유무 (0 : 정상, 1 : 이상증)
- CholCheck : 콜레스테롤 검사 여부 (0 : 미검사, 1 : 검사)
- BMI : 체질량 지수
- Smoker : 흡연 여부
- Stroke : 뇌졸중 여부(0 : 없음, 1 : 있음)
- HeartDiseaseorAttack : 심장질환 및 심근경색 유무 (0 : 없음, 1 : 있음)
- PhysActivity : 신체활동 유무
- Fruits : 과일 섭취 유무
- Veggies : 채소 섭취 유무
- HvyAlcoholConsump : 과음 여부 (0 : 아니오, 1 : 예)
- AnyHealthcare : 의료보험 가입 유무 (0 : 없음, 1 : 있음)
- NoDocbcost : 의료비 때문에 의사 방문 여부 (0 : 아니오, 1 : 예)
- GenHlth : 일반 건강 상태 (1 : 매우 좋음, 2 : 좋음, 3 : 보통, 4 : 나쁨, 5 : 매우 나쁨)
- MentHlth : 정신 건강 상태 (0 : 없음, 1 : 있음)
- PhysHlth : 신체 건강 상태 (0 : 없음, 1 : 있음)
- DiffWalk : 보행 장애 유무 (0 : 없음, 1 : 있음)
- Sex : 성별 (0 : 여성, 1 : 남성)
- Age : 연령 (1 : 18-24, 2 : 25-29, 3 : 30-34, 4 : 35-39, 5 : 40-44, 6 : 45-49, 7 : 50-54, 8 : 55-59, 9 : 60-64, 10 : 65-69, 11 : 70-74, 12 : 75-79, 13 : 80 이상)
- Education : 교육 수준 (1 : 미취학 또는 유치원, 2 : 초등학교, 3 : 중학교, 4 : 고등학교, 5 : 대학 1-3년, 6 : 대학 4년 이상)
- Income : 소득 수준 (1 : 10,000 미만, 2 : 10,000-15,000, 3 : 15,000-20,000, 4 : 20,000-25,000, 5 : 25,000-35,000, 6 : 35,000-50,000, 7 : 50,000-75,000, 8 : 75,000 이상)

원 데이터 셋 불러오기

```
# 데이터 셋이 있는 폴더 지정
data_path <- "C:/Users/jwjeong/Desktop/data_science/"

# 전체 파일 목록 확인
list.files(path = data_path, pattern = "\.csv$")
```

```
## [1] "Cardiac_Arrest.csv"
## [2] "Diabetes.csv"
## [3] "Diabetes_binary_5050split.csv"
## [4] "Diabetes_binary_BRFSS2015.csv"
## [5] "Diabetes_binary_BRFSS2021.csv"
## [6] "Diabetes_binary_BRFSS2023.csv"
## [7] "Diabetes_binary_BRFSS2024.csv"
## [8] "Diabetes_binary_BRFSS2025.csv"
## [9] "Diabetes_binary_Kaggle.csv"
## [10] "Diabetes_binary_Kaggle2.csv"
## [11] "Diabetes_binary_Kaggle3.csv"
## [12] "Diabetes_binary_Kaggle4.csv"
## [13] "Diabetes_binary_Kaggle5.csv"
## [14] "Diabetes_binary_Kaggle6.csv"
## [15] "Diabetes_binary_Kaggle7.csv"
## [16] "Diabetes_binary_Kaggle8.csv"
## [17] "Diabetes_binary_Kaggle9.csv"
## [18] "Diabetes_binary_Kaggle10.csv"
## [19] "Diabetes_binary_Kaggle11.csv"
## [20] "Diabetes_binary_Kaggle12.csv"
## [21] "Diabetes_binary_Kaggle13.csv"
## [22] "Diabetes_binary_Kaggle14.csv"
##
```

```
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  
```

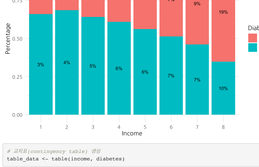



```
# income과 diabetes를 활용하여 데이터프레임 생성
data <- data.frame(Income = factor(income), Diabetes = factor(diabetes))

# 각 소득 구간 내에서 당뇨가 발생한 비율 계산
data_summary <- data %>%
  group_by(Income, Diabetes) %>%
  summarise(count = n(), .groups = 'drop') %>%
  mutate(percent = count / sum(count) * 100)

data_summary
```

```
## # A tibble: 16 × 4
##    Income Diabetes count percent
##    <fct>  <fct>   <int>   <dbl>
## 1  1      0        1228    1.74
## 2  1      1        2383    3.37
## 3  2      0        1412    2.00
## 4  2      1        3086    4.37
## 5  3      0        1989    2.81
## 6  3      1        3568    5.05
## 7  4      0        2604    3.68
## 8  4      1        4054    5.73
## 9  5      0        3506    4.96
## 10 5      1        4504    6.37
## 11 6      0        4996    7.07
## 12 6      1        5291    7.48
## 13 7      0        6160    8.71
## 14 7      1        5265    7.45
## 15 8      0       13451   19.0
## 16 8      1        7195   10.2
```

3. 소득구간과 당뇨여부 시각화 - 2

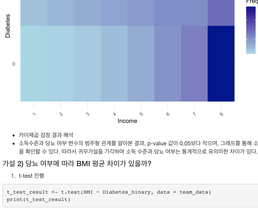
```
ggplot(data_summary, aes(x = Income, y = percent, fill = Diabetes)) +
  geom_bar(stat = "identity", position = "fill") +
  geom_text(aes(label = paste0(round(percent), "%")), position = position_fill(vjust = 0.5), size = 3) +
  labs(title = "Income Level vs. Diabetes Status", x = "Income", y = "Percentage") +
  scale_fill_discrete(name = "Diabetes", labels = c("No", "Yes"))
```




```
# 교차표(contingency table) 생성
table_data <- table(income, diabetes)

# 교차표 데이터 프레임으로 변환
table_df <- as.data.frame(table_data)
names(table_df) <- c("Income_Group", "Diabetes", "Freq")
```

4. 소득구간과 당뇨여부 시각화 - 3

```
ggplot(table_df, aes(x = Income_Group, y = Diabetes, fill = Freq)) +
  geom_tile() +
  labs(title = "Heatmap of Income Level vs Diabetes Status", x = "Income", y = "Diabetes") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



- 카이제곱 검정 결과 해석
- 소득수준이 당뇨 여부 한수의 방추한 관계를 알아본 결과, p-value 값이 0.05보다 작으며, 그래프를 통해 소득 수준이 높을 수록 당뇨가 아닐 확률이 높아지는 것을 확인할 수 있다. 따라서 귀무가설을 기각하며 소득 수준과 당뇨 여부는 통계적으로 유의미한 차이가 있다.

가설 2) 당뇨 여부에 따라 BMI 평균 차이가 있을까?

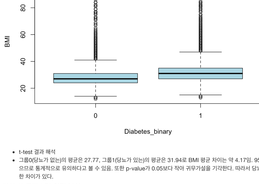
1. t-test 진행

```
t_test_result <- t.test(BMI ~ Diabetes_binary, data = team_data)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  BMI by Diabetes_binary
## t = -81.591, df = 68653, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -4.274321 -4.073781
## sample estimates:
## mean in group 0 mean in group 1
##        27.76996        31.94401
```

2. t-test 시각화

```
boxplot(BMI ~ Diabetes_binary, data = team_data, col = "lightblue", main = "Boxplot of BMI by Diabetes_binary")
```



- t-test 결과 해석
- 그룹0(당뇨가 없는)의 평균은 27.77, 그룹1(당뇨가 있는)의 평균은 31.94로 BMI 평균 차이는 약 4.17임. 95% 신뢰구간은 -4.274 ~ -4.074로 0이 포함되지 않으므로 통계적으로 유의미다고 볼 수 있음. 또한 p-value가 0.05보다 작아 귀무가설을 기각한다. 따라서 당뇨 여부에 따라서 BMI의 평균에 통계적으로 유의미한 차이가 있다.

가설 3) 로지스틱 회귀 분석

1. 로지스틱 회귀 모델 훈련

```
logistic_model <- glm(Diabetes_binary ~ Income * BMI, data = team_data, family = "binomial")
```

2. 모델 요약 정보

```
summary_info <- summary(logistic_model)
summary_info
```

```
##
## Call:
## glm(formula = Diabetes_binary ~ Income * BMI, family = "binomial",
##     data = team_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7455  -1.0423  -0.1837   1.0600   2.2644
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.9624017  0.1628859 -11.609  < 2e-16 ***
## Income2     0.4114052  0.2279937   1.812  0.07005 .
## Income3     -0.1185649  0.2139898  -0.540  0.58890
## Income4     -0.1567946  0.2113427  -0.742  0.45815
## Income5     -0.4746860  0.2058565  -2.316  0.02058 *
## Income6     -0.8253945  0.1994898  -4.137 3.52e-05 ***
## Income7     -1.3000746  0.1978425  -6.571 4.99e-11 ***
## Income8     -1.9417743  0.1875681 -10.352  < 2e-16 ***
## BMI          0.0962337  0.0055536  15.527  < 2e-16 ***
## Income2:BMI -0.0094833  0.0075125  -1.262  0.20683
## Income3:BMI  0.0021442  0.0072272   0.295  0.76827
## Income4:BMI -0.0052578  0.0070281  -0.037  0.97074
## Income5:BMI  0.0044891  0.0068449   0.656  0.51193
## Income6:BMI  0.0098491  0.0066176   1.488  0.13667
## Income7:BMI  0.0186232  0.0065606   2.839  0.00453 **
## Income8:BMI  0.0268144  0.0062285   4.305  1.67e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98000  on 70691  degrees of freedom
## Residual deviance: 88093  on 70676  degrees of freedom
## AIC: 88125
##
## Number of Fisher Scoring iterations: 4
```

3. 각 계수에 대한 p-value 확인

```
p_values <- summary_info$coefficients[, "Pr(>|z|)"]
p_values
```

```
##  (Intercept)      Income2      Income3      Income4      Income5      Income6
## 1.4525616e-31 7.0064416e-02 5.8930076e-01 6.5814944e-01 2.0578896e-02 3.8209550e-05
##      Income7      Income8          BMI Income2:BMI Income3:BMI Income4:BMI
## 4.9939726e-11 4.6926646e-25 2.2424946e-54 2.0683076e-01 7.6826646e-01 9.7073646e-01
##  Income5:BMI  Income6:BMI  Income7:BMI  Income8:BMI
## 5.1193196e-01 1.3666659e-01 4.5304546e-03 1.6491358e-05
```

4. 테스트 데이터로 예측

```
predicted_probs <- predict(logistic_model, newdata = team_data, type = "response")
```

5. ROC 커브 계산

```
roc_curve <- roc(team_data$Diabetes_binary, predicted_probs)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

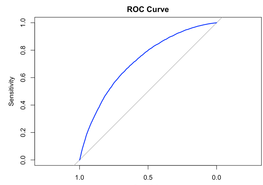
6. AUC 출력

```
auc_value <- auc(roc_curve)
cat("AUC:", auc_value, "\n")
```

```
## AUC: 0.7172288
```

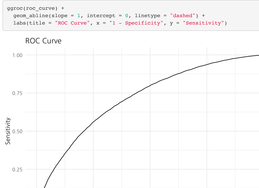
7. ROC 커브 그리기

```
plot(roc_curve, main = "ROC Curve", col = "blue", lwd = 2)
```



8. ggplot을 활용한 ROC 커브 그래프

```
ggroc(roc_curve) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  labs(title = "ROC Curve", x = "1 - Specificity", y = "Sensitivity")
```



- 로지스틱 회귀 해석
- 소득은 -2, BMI는 +2의 계수를 보이며 두 변수 모두 p-value값이 0.05보다 작아 유의한 관계성을 보인다고 할 수 있다. 또한 AUC값도 0.7로 높은 값을 보이고 있다.

2) 결론

- 카이제곱 검정, t-test, 로지스틱 회귀분석 모두 p-value가 매우 작았으며, 그래프로 상에서도 유의미한 결과를 도출할 수 있었다. 따라서 소득이 낮은 사람과 BMI가 높은 사람이 당뇨에 더 취약할 것이라는 가설을 다양한 분석 기법 및 시각화를 통해 유의미함을 입증하였다.