Assignment 2 – Part A: PRML Project

Md. Zunaed Mazumdar

2025-09-24

Data Selection

I selected the Australian weather dataset from Kaggle. The original training set contained 99,516 observations with 23 variables, while the testing set had 42,677 observations with 22 variables. After performing data cleaning and preprocessing, the training set was reduced to 2,225 observations with 9 features, and the testing set to 960 observations.

Data Preparation

```
#loading necessary libraries
library(dplyr)
library(tidyr)
library(ggplot2)
library(caret)
```

Now, i will perform the data cleaning and data prepration. In our original data set we have data for many region in Australia, but i will select only the data for the Canberra region. I will drop some variables which are not necessary for our analysis

After performing the data cleaning, we found our final dataset.

```
dim(train)
```

```
## [1] 2393    9
```

```
dim(test)
```

```
## [1] 1025    8
```

Removing the missing value

Now, i will check if i have any missing value in my final data set.

```
colSums(is.na(train))
```

```
##     Rainfall  Evaporation     Sunshine    RainToday RainTomorrow      TempAvg
##           13         1120         1346           13            0            0
## WindSpeedAvg  HumidityAvg  PressureAvg
##          153            2          152
```

I found that, Evaporation and Sunshine has more than 1000 missing value, so i will drop this variable. Rainfall, Raintoday, windspeedAvg, pressureAvg has 13, 13, 153 and 152 missing values respectiably. Now i will remove the missing value from the data set.

```
train <- train %>%
  select(-Evaporation, -Sunshine)
```

```
test <- test %>%
   select(-Evaporation, -Sunshine)
```

```
train <- train %>%
  drop_na(WindSpeedAvg, PressureAvg, Rainfall, RainToday, HumidityAvg)
test  <- test %>%
  drop_na(WindSpeedAvg, PressureAvg, Rainfall, RainToday)

dim(train)
```

```
## [1] 2225    7
```

```
dim(test)
```

```
## [1] 960   6
```

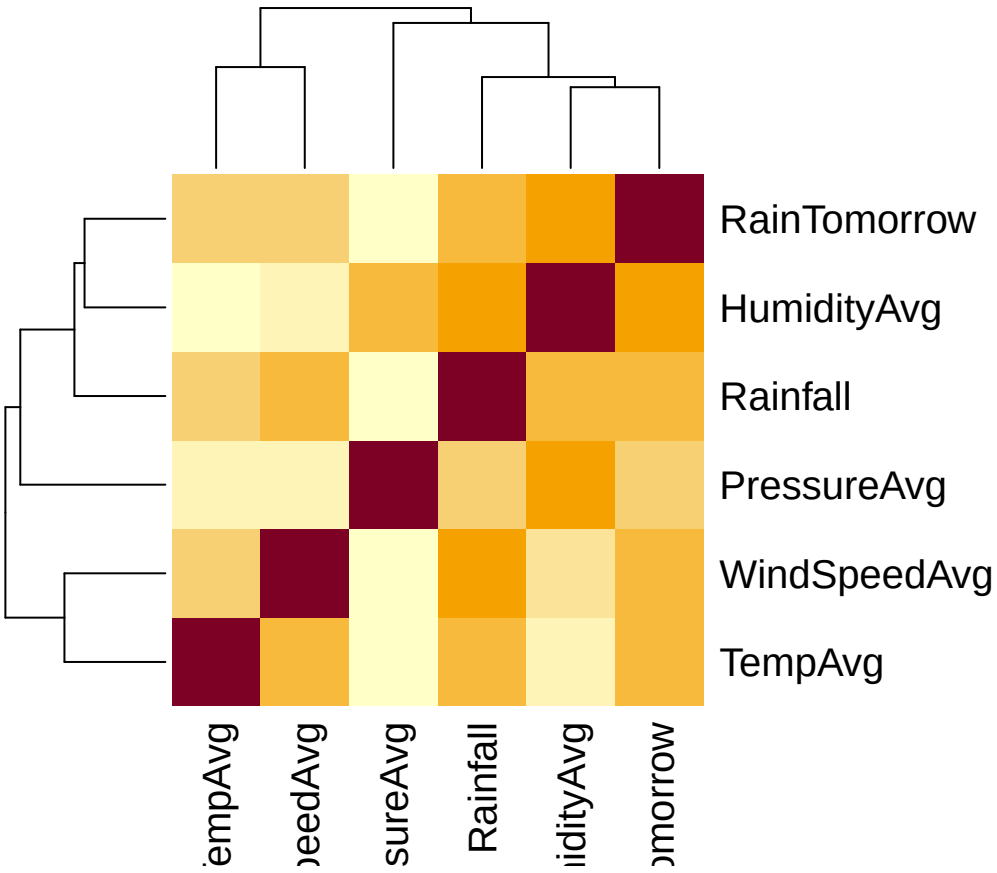
Explanatory data analysis

Now i will check the correlation between the variables in our final dataset.

```
num_train <- train %>%
  select(where(is.numeric))

cor_mat <- cor(num_train, use = "pairwise.complete.obs")

heatmap(cor_mat)
```



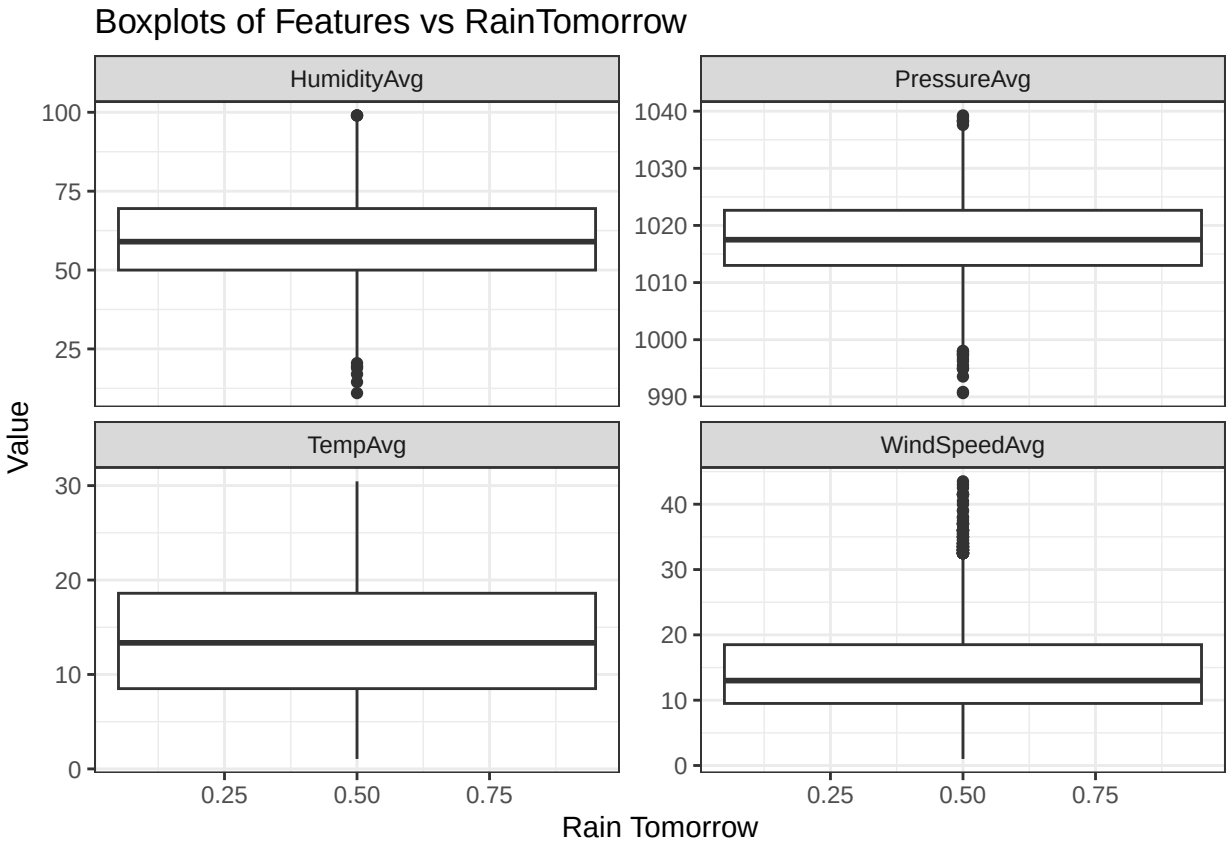
We can see from the correlation heatmap, humidity, windpressure has strong correlation with rainfall and

raintomorrow.

Now i will check the boxplot of the features and rain tomorrow.

```
df_long <- train %>%
  select(RainTomorrow, TempAvg, WindSpeedAvg, HumidityAvg, PressureAvg) %>%
  pivot_longer(cols = c(TempAvg, WindSpeedAvg, HumidityAvg, PressureAvg),
               names_to = "Feature", values_to = "Value")

# plot with facets
ggplot(df_long, aes(x = RainTomorrow, y = Value, fill = RainTomorrow)) +
  geom_boxplot() +
  facet_wrap(~ Feature, scales = "free_y") +
  labs(title = "Boxplots of Features vs RainTomorrow",
       x = "Rain Tomorrow", y = "Value") +
  theme_bw()
```



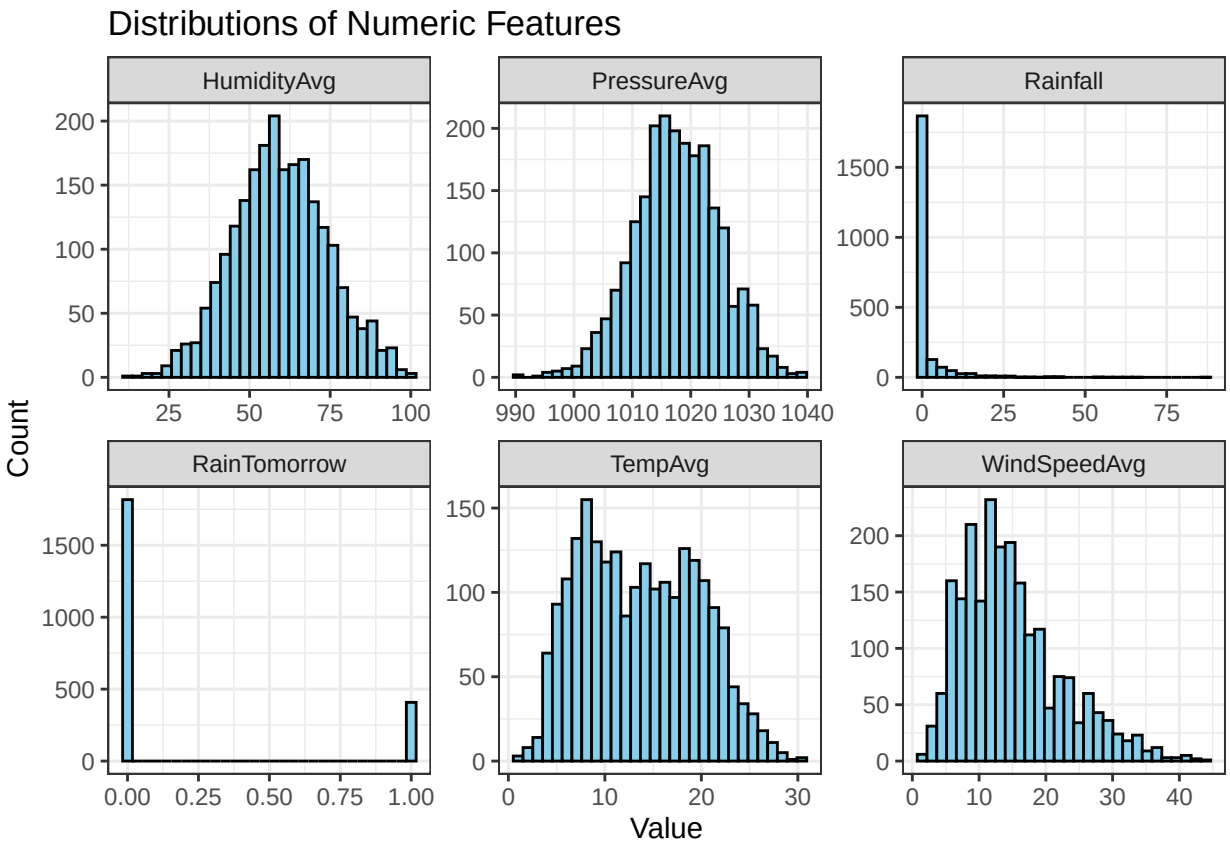
The boxplots compare the distributions of average temperature, wind speed,humidity, and pressure for days with and without rain. Humidity is generally higher and pressureslightly lower on rainy days, while temperature tends to be lower when it rains. Wind speed shows more variability on rainy days, suggesting these features may carry predictive power for rainfall.

Data distribution of the variables

```
# Select numeric variables only
num_vars <- train %>% select(where(is.numeric))
```

```r
# Reshape to long format
df_long <- num_vars %>%
  pivot_longer(cols = everything(), names_to = "Feature", values_to = "Value")

# Plot histograms for all numeric features
ggplot(df_long, aes(x = Value)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  facet_wrap(~ Feature, scales = "free") +
  labs(title = "Distributions of Numeric Features", x = "Value", y = "Count") +
  theme_bw()
```



The histograms show that most features are approximately symmetric and bell-shaped,such as HumidityAvg and PressureAvg, while TempAvg is fairly spread but still close to normal. WindSpeedAvg is slightly right-skewed, with more lower values and fewer high speeds. Rainfall is highly right-skewed, with the majority of days´having no or very little rain

Algorithm Selection and Justification.

We have select **Linear regression** to model the temp and **Logistic regression** to model the raintoday.

We selected linear regression to model the temparature because it is a widely used and interpretable method for predicting continuous outcomes. It allows us to examine how weather variables such as rainfall, humidity, pressure, and wind speed contribute to temparature levels. Linear regression provides straightforward coefficient estimates, making the relationships between predictors and temparature easy to understand and communicate.

For predicting RainToday, we used logistic regression since the target variable is binary (Yes/No). Logistic

regression is the standard choice for classification problems of this nature, as it estimates the probability of rainfall occurring on a given day. The method ensures probability outputs between 0 and 1 and enables evaluation using metrics such as accuracy, sensitivity, and ROC analysis.

Model implementation

Linear Regression

First we will implement the linear regression to model the tempAvg.

```
model_temp <- lm(TempAvg ~ HumidityAvg + WindSpeedAvg + PressureAvg + Rainfall,
                 data = train)

summary(model_temp)
```

```
##
## Call:
## lm(formula = TempAvg ~ HumidityAvg + WindSpeedAvg + PressureAvg +
##     Rainfall, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.948  -3.505  -0.036   3.696  13.919
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  467.069770  16.482291  28.338  < 2e-16 ***
## HumidityAvg   -0.148573   0.007371 -20.157  < 2e-16 ***
## WindSpeedAvg  -0.219954   0.015295 -14.381  < 2e-16 ***
## PressureAvg   -0.433700   0.016150 -26.854  < 2e-16 ***
## Rainfall       0.057537   0.018699   3.077  0.00212 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.781 on 2220 degrees of freedom
## Multiple R-squared:  0.3719, Adjusted R-squared:  0.3708
## F-statistic: 328.6 on 4 and 2220 DF,  p-value: < 2.2e-16
```

The linear regression model for TempAvg is statistically significant ($p < 2.2e\text{-}16$) with an R^2 of about 0.37, meaning it explains roughly 37% of the variability in temperature. All predictors are highly significant, with humidity, wind speed, and pressure showing negative associations with temperature, while rainfall has a small positive effect. The residual standard error of 4.8 indicates an average deviation of around 5°C, suggesting moderate predictive accuracy but also room for improvement with more complex models.

Now we will use the test dataset to see how the model perform to test data.

```
pred_temp <- predict(model_temp, newdata = test)


####### evalutaing the model

actual_temp <- test$TempAvg

# RMSE
rmse <- sqrt(mean((pred_temp - actual_temp)^2, na.rm = TRUE))

# MAE
```

```r
mae <- mean(abs(pred_temp - actual_temp), na.rm = TRUE)

# R-squared
SSE <- sum((pred_temp - actual_temp)^2, na.rm = TRUE)
SST <- sum((actual_temp - mean(actual_temp, na.rm = TRUE))^2, na.rm = TRUE)
rsq <- 1 - SSE/SST

list(RMSE = rmse, MAE = mae, R2 = rsq)
```

```
## $RMSE
## [1] 4.706325
##
## $MAE
## [1] 3.867072
##
## $R2
## [1] 0.4152184
```

The model's performance on the test data shows an RMSE of about 4.7 and an MAE of about 3.9, meaning predictions deviate from actual temperatures by roughly 4–5°C on average. The R^2 of 0.41 indicates the model explains around 41% of the variance in unseen data. This suggests the model generalizes reasonably well, though its predictive power is moderate and could be improved with additional features or more flexible models.

Logistic regreesion

Now, i will use the logistic regression to model the probability of raining today.

```r
log_model_today <- glm(RainToday ~ TempAvg + WindSpeedAvg + HumidityAvg + PressureAvg,
                       data = train,
                       family = binomial)

summary(log_model_today)
```

```
##
## Call:
## glm(formula = RainToday ~ TempAvg + WindSpeedAvg + HumidityAvg +
##     PressureAvg, family = binomial, data = train)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  96.448729  12.232545   7.885 3.16e-15 ***
## TempAvg       0.053459   0.013544   3.947 7.91e-05 ***
## WindSpeedAvg  0.064628   0.009709   6.657 2.80e-11 ***
## HumidityAvg   0.080419   0.005385  14.934  < 2e-16 ***
## PressureAvg  -0.103204   0.011874  -8.691  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2068.6  on 2224  degrees of freedom
## Residual deviance: 1534.7  on 2220  degrees of freedom
## AIC: 1544.7
##
```

```
## Number of Fisher Scoring iterations: 5
```

The logistic regression model for RainToday is highly significant, with all predictors contributing meaningfully ($p < 0.001$). Higher temperature, wind speed, and humidity increase the likelihood of rainfall today, while higher pressure reduces it, which aligns with meteorological expectations. The drop in deviance from 2068.6 to 1534.7 and the AIC of 1544.7 indicate a substantial improvement over the null model, suggesting the predictors have good explanatory power.

Now i will use the model to predict the temp in our testing dataset and evaluate the model performance using confushion matrix.

```
pred_prob <- predict(log_model_today, newdata = test, type = "response")

# Convert to Yes/No with threshold 0.5
pred_class <- ifelse(pred_prob > 0.5, "Yes", "No")


confusionMatrix(as.factor(pred_class), test$RainToday, positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  735 124
##        Yes  37  64
##
##                Accuracy : 0.8323
##                  95% CI : (0.8071, 0.8554)
##     No Information Rate : 0.8042
##     P-Value [Acc > NIR] : 0.01435
##
##                   Kappa : 0.3546
##
##  Mcnemar's Test P-Value : 1.221e-11
##
##             Sensitivity : 0.34043
##             Specificity : 0.95207
##          Pos Pred Value : 0.63366
##          Neg Pred Value : 0.85565
##              Prevalence : 0.19583
##          Detection Rate : 0.06667
##    Detection Prevalence : 0.10521
##       Balanced Accuracy : 0.64625
##
##        'Positive' Class : Yes
##
```

The logistic regression model achieves an overall accuracy of about 83%, which is better than the no-information rate. However, the sensitivity is low (34%), meaning it misses many rainy days, while the specificity is high (95%), showing it correctly identifies non-rainy days. This indicates the model is biased toward predicting "No Rain," which is expected given the class imbalance, and improvements may require resampling techniques or alternative models.