# train test split

2024-12-07

## Can we successfully predict the outcome of a match given the match statistics?

```
match_data_team <- read.csv("match_data_team.csv")
#Make tie the reference variable
match_data_team$result <- as.factor(match_data_team$result)
match_data_team$result <- relevel(match_data_team$result, ref = "0")
#Create df to use in model
#Remove variables that would make such predictions too easy (vars related to
#number of goals)
#Also remove variables time, data, team
match_data_team_model <- match_data_team[,-which(names(match_data_team) %in%
                                                   c("team",
                                        "number.of.goals.team", "date",
                                        "hour", "category", "team_num", "outcome",
                                        "conceded.team",
                                        "goal.inside.the.penalty.area.team",
                                        "goal.outside.the.penalty.area.team",
                                        "own.goals.team", "assists.team",
                                        "penalties.scored.team"))]


set.seed(42)
train_indices <- sample(1:nrow(match_data_team_model),
                        0.8 * nrow(match_data_team_model))
train_data <- match_data_team_model[train_indices, ]
test_data <- match_data_team_model[-train_indices, ]

write.csv(train_data, "train_data.csv", row.names = FALSE)
write.csv(test_data, "test_data.csv", row.names = FALSE)
```