imodel_dataset

create dataset

```
linear_dgp_fun <- function(n_train, n_test, p, beta, noise_sd) {
  n <- n_train + n_test
  X <- matrix(rnorm(n * p), nrow = n, ncol = p)
  y <- X %*% beta + rnorm(n, sd = noise_sd)
  data_list <- list(
    X_train = X[1:n_train, , drop = FALSE],
    y_train = y[1:n_train],
    X_test = X[(n_train + 1):n, , drop = FALSE],
    y_test = y[(n_train + 1):n]
  )
  return(data_list)
}
linear_dgp <- create_dgp(
  .dgp_fun = linear_dgp_fun, .name = "Linear DGP",
  # additional named parameters to pass to .dgp_fun()
  n_train = 350, n_test = 120, p = 4, beta = c(1,2,1.5,3), noise_sd = 1
)

dataset_dgp_fun <- function(datasetname){

  address <- "C:/Users/pyk/Desktop/nus/RA/project/imodels-data-master/data_cleaned/"
  file <- paste0(datasetname,".csv")
  file_path <- paste0(address,file)
  df <- read.csv(file_path)
  x <- df[, -ncol(df)]
  y <- df[, ncol(df)]

  train_indices <- createDataPartition(y, p = 0.8, list = FALSE)
```

```
  data_list <- list(
    X_train <- x[train_indices, ],
    y_train <- y[train_indices],
    X_test <- x[-train_indices, ],
    y_test <- y[-train_indices]
  )
  return(data_list)
}
dataset_dgp <- create_dgp(.dgp_fun = dataset_dgp_fun,.name = 'heart',
                          datasetname = "heart")
```

build BART model

```
BART_fun <- function(X_train, y_train, X_test, y_test, df,k,q) {
  train_X <- data.frame(X_train)
  test_X <- data.frame(X_test)
  t <- bench::mark(fit <- wbart(x.train = train_X,
                                y.train = y_train,
                                x.test = test_X,
                                k = k,
                                sigdf = df,
                                sigquant = q
                                ))
  time <- mean(t$time[[1]])
  predictions <- colMeans(fit$yhat.test)
  mse_score <- mean((y_test - predictions)^2)

  return(list(time = time, mse=mse_score))
}

dbarts_fun <- function(X_train, y_train, X_test, y_test, df,k,q){
  train_X <- data.frame(X_train)
  test_X <- data.frame(X_test)
  t <- bench::mark(bart_model <- bart(x.train = train_X,
                                      y.train = y_train,
                                      x.test = test_X,
                                      k = k,
                                      sigdf = df,
                                      sigquant = q))
  time <- mean(t$time[[1]])
```

```r
  predictions <- colMeans(bart_model$yhat.test)
  mse_score <- mean((y_test - predictions)^2)

  return(list(time = time, mse=mse_score))
}

bartMachine_fun <- function(X_train, y_train, X_test,y_test,df,k,q){
  train_X <- data.frame(X_train)
  test_X <- data.frame(X_test)
  t <- bench::mark(bart_model <- bartMachine(
          X = train_X,
          y = y_train,
          k = k,
          nu = df,
          q=q))
        # The value of calculating the time required for modeling
  time <- mean(t$time[[1]])
  predictions <- predict(bart_model,test_X,type = "prob")
  mse_score <- mean((y_test - predictions)^2)

  return(list(time = time, mse=mse_score))
}

SoftBart_fun<- function(X_train, y_train, X_test,y_test,num_trees,alpha,beta){
  train_X <- data.frame(X_train)
  test_X <- data.frame(X_test)
  t <-  bench::mark({bart_model <- softbart(X = train_X, Y = y_train, X_test = test_X, hyper
        #print(t)
  time <- mean(t$time[[1]])
  predictions <- bart_model$y_hat_test_mean
  mse_score <- mean((y_test - predictions)^2)

  return(list(time = time, mse=mse_score))
}
RF_fun <- function(X_train, y_train, X_test,y_test){
  train_X <- data.frame(X_train)
  test_X <- data.frame(X_test)
  t <- bench::mark({rf_model <- randomForest(x=train_X, y=y_train)})
  time <- mean(t$time[[1]])
  predictions <- predict(rf_model, test_X)
  mse_score <- mean((y_test - predictions)^2)
  return(list(time = time, mse=mse_score))
```

```
}
```

create evaluation

```
posterior_mse <- function(fit_results,truth_col,estimate_col){
  y_test = fit_results$truth_col
  pred = fit_results$estimate_col
  return(mean((y_test - pred)^2))
}

pred_err <- create_evaluator(
  .eval_fun = posterior_mse, .name = 'Posterior MSE',
  # additional named parameters to pass to .eval_fun()
  truth_col = "y_test", estimate_col = "predictions"
)
```

model fitting

```
BART <- create_method(
  .method_fun = BART_fun, .name = "BART",
  # additional named parameters to pass to .method_fun()
  k=2.5,q=0.95,df=4
)
dbarts <- create_method(.method_fun = dbarts_fun,.name = "dbarts",
                        k=2.5,q=0.95,df=4)
bartMachine <- create_method(.method_fun = bartMachine_fun,.name = "bartMachine",
                        k=2.5,q=0.95,df=4)
SoftBart <- create_method(.method_fun = SoftBart_fun,.name = "SoftBart",
                        num_trees=50,alpha=0.95,beta=2)
RF <- create_method(.method_fun = RF_fun,.name = "RandomForest")
# Create experiment
experiment <- create_experiment(name = "Test Experiment") %>%
  add_dgp(linear_dgp) %>%
  add_dgp(dataset_dgp) %>%
  add_method(dbarts) %>%
  add_method(BART) %>%
  add_method(bartMachine) %>%
  add_method(SoftBart) %>%
```

```r
  add_method(RF)%>%
  add_evaluator(pred_err)



results <- run_experiment(experiment, n_reps = 4, save = TRUE)
```

```
Fitting Test Experiment...

Warning: Some expressions had a GC in every iteration; so filtering is
disabled.
Warning: Some expressions had a GC in every iteration; so filtering is
disabled.
Warning: Some expressions had a GC in every iteration; so filtering is
disabled.

Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
```

```
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
Warning in randomForest.default(x = train_X, y = y_train): The response has
five or fewer unique values.  Are you sure you want to do regression?
```

```
Saving fit results...
Fit results saved | time taken: 0.053212 seconds
4 reps completed (totals: 4/4) | time taken: 4.716040 minutes
==============================
Evaluating Test Experiment...
```

```
Warning: Unknown or uninitialised column: `truth_col`.
```

```
Warning: Unknown or uninitialised column: `estimate_col`.
```

```
Evaluation completed | time taken: 0.000149 minutes
Saving eval results...
Eval results saved | time taken: 0.042850 seconds
==============================
No visualizers to visualize. Skipping visualization.
==============================
```

```
# Render automated documentation and view results
#render_docs(experiment)
```

```
result <- results$fit_results
result
```

```
# A tibble: 40 x 5
   .rep  .dgp_name  .method_name time              mse
   <chr> <chr>      <chr>        <list>          <dbl>
 1 1     Linear DGP BART         <bench_tm [1]> 1.40
 2 1     Linear DGP RandomForest <bench_tm [1]> 3.53
 3 1     Linear DGP SoftBart     <bench_tm [1]> 1.41
 4 1     Linear DGP bartMachine  <bench_tm [1]> 1.46
 5 1     Linear DGP dbarts       <bench_tm [1]> 1.33
 6 1     heart      BART         <bench_tm [1]> 0.130
 7 1     heart      RandomForest <bench_tm [1]> 0.131
 8 1     heart      SoftBart     <bench_tm [1]> 0.133
 9 1     heart      bartMachine  <bench_tm [1]> 0.130
10 1     heart      dbarts       <bench_tm [1]> 1.26
# i 30 more rows
```

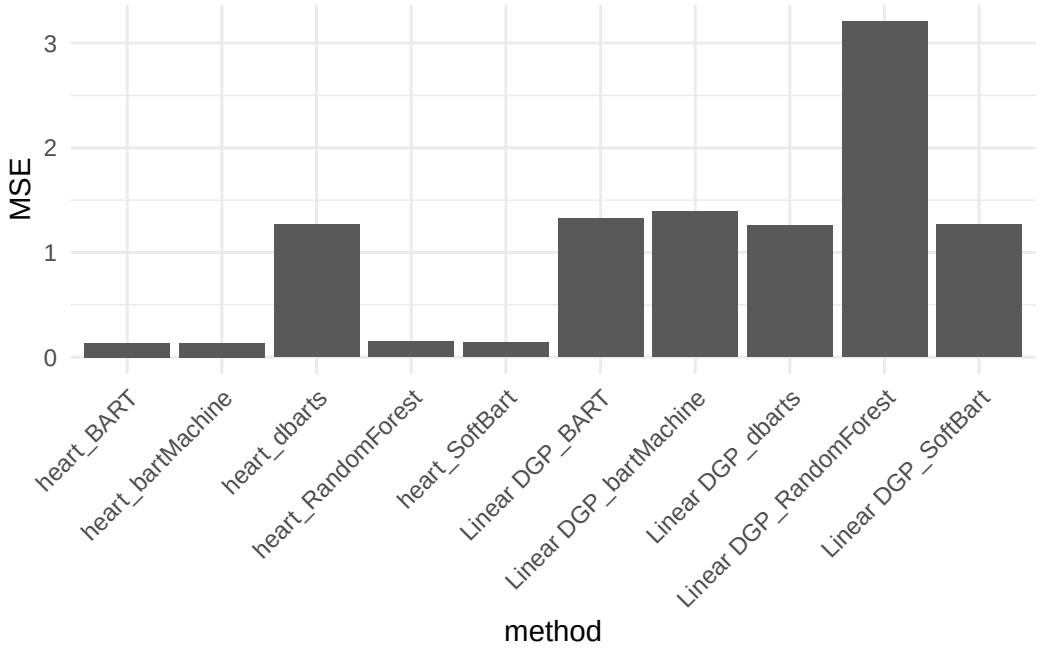
```r
result$time_numeric <- as.numeric(result$time)
```

```r
result$Resource <- paste(result$.dgp_name, result$.method_name, sep="_")

# Calculate MSE for each group
summary <- result %>%
  group_by(Resource) %>%
  summarise(
    Mean_MSE = mean(mse),
    Var_MSE = sd(mse),
    Mean_time = mean(time_numeric),
    Var_time = sd(time_numeric))

print(summary)
```

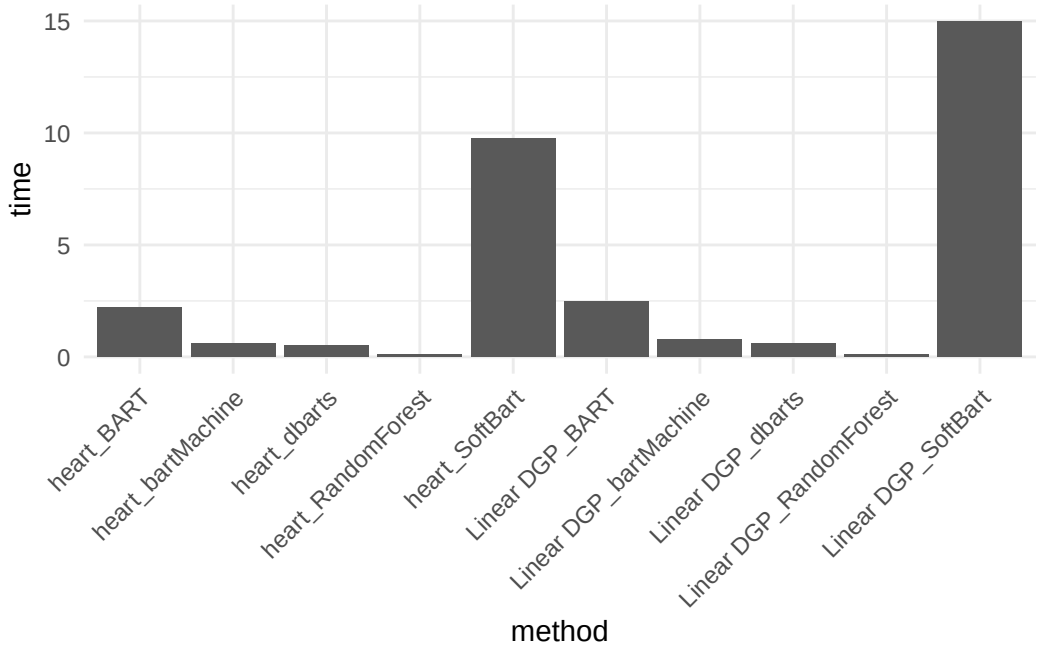
```
# A tibble: 10 x 5
   Resource                Mean_MSE Var_MSE Mean_time Var_time
   <chr>                      <dbl>   <dbl>     <dbl>    <dbl>
 1 Linear DGP_BART            1.33  0.242       2.46   0.0338
 2 Linear DGP_RandomForest    3.21  0.589       0.132  0.0326
 3 Linear DGP_SoftBart        1.27  0.302      15.0    0.0884
 4 Linear DGP_bartMachine     1.39  0.207       0.792  0.0290
 5 Linear DGP_dbarts          1.25  0.152       0.596  0.0160
 6 heart_BART                 0.135 0.00371     2.22   0.0129
 7 heart_RandomForest         0.149 0.0151      0.132  0.00273
 8 heart_SoftBart             0.137 0.00366     9.75   0.0517
 9 heart_bartMachine          0.135 0.00355     0.617  0.0293
```

```
10 heart_dbarts                1.27   0.0709       0.522   0.0144
```

```
ggplot(summary, aes(x = Resource, y = Mean_MSE)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  labs(y = "MSE", x = "method")
```




```
ggplot(summary, aes(x = Resource, y = Mean_time)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  labs(y = "time", x = "method")
```


15
10
5
0
time
heart_BART
heart_bartMachine
heart_dbarts
heart_RandomForest
heart_SoftBart
Linear DGP _BART
Linear DGP _bartMachine
Linear DGP _dbarts
Linear DGP _RandomForest
Linear DGP _SoftBart
method