UNIVERSITÀ DEGLI
STUDI DI FIRENZE

# Previsione: Random Forest vs BART

Statistical Learning - Contest

---

**Arianna Russo e Simone Capucci**

Laurea Magistrale in Data Science, Calcolo Scientifico e Intelligenza Artificiale

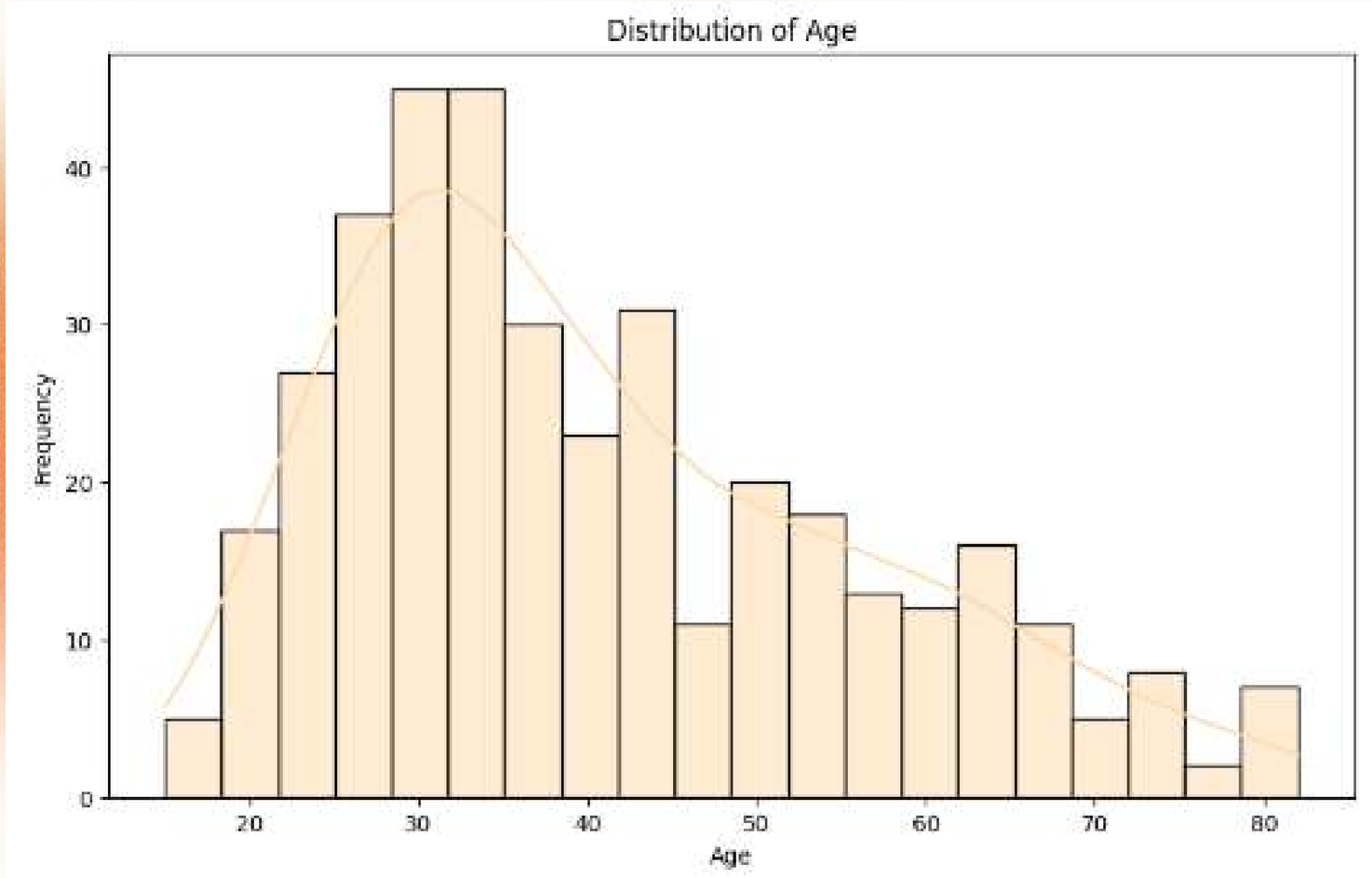
UNIVERSITÀ DEGLI STUDI DI FIRENZE

# Recidiva del cancro alla tiroide

**Obiettivo:** Prevedere la recidiva del cancro alla tiroide

## Il dataset:

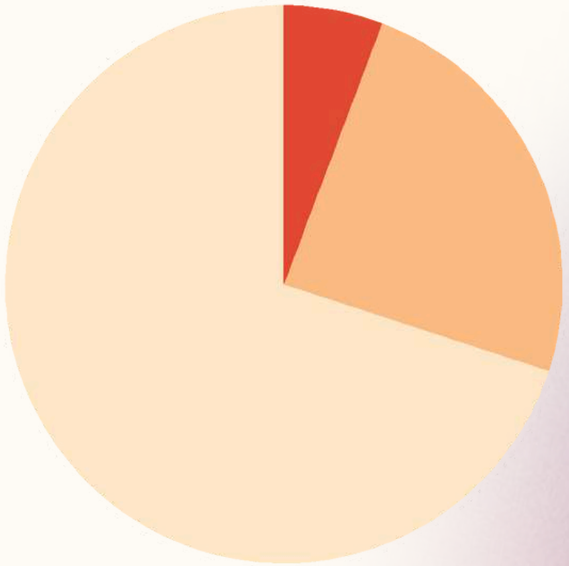
Il dataset si compone di **383 osservazioni su 18 variabili** (collegate o possibilmente collegate al tema trattato) , raccolte da soggetti differenti. Le variabili sono per lo più **categoriche** (a parte la variabile Age).

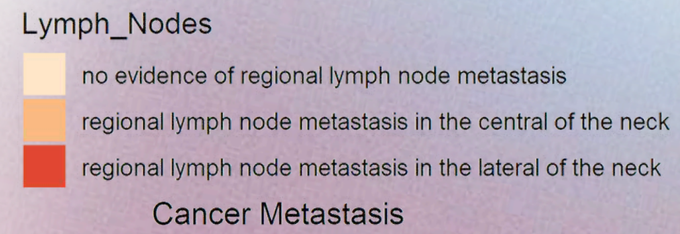


UNIVERSITÀ DEGLI STUDI DI FIRENZE

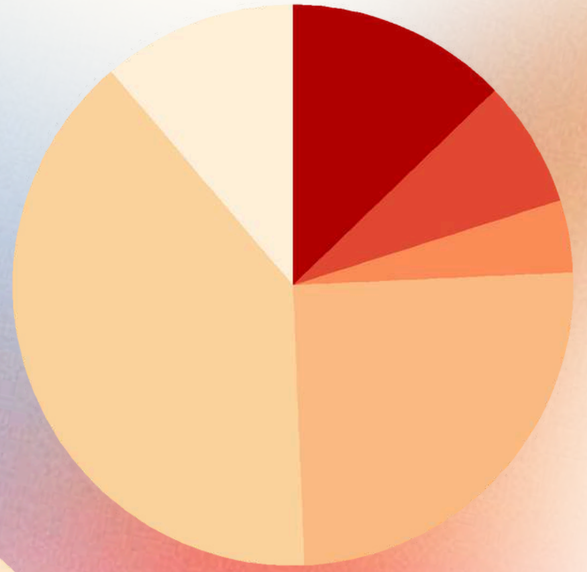
# Data Visualization

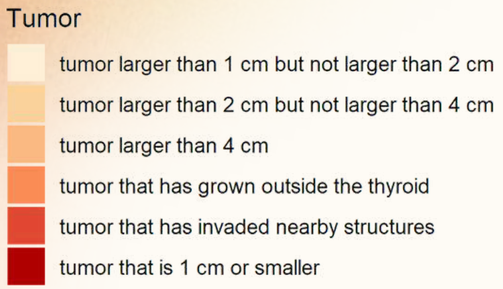
Lymph Nodes

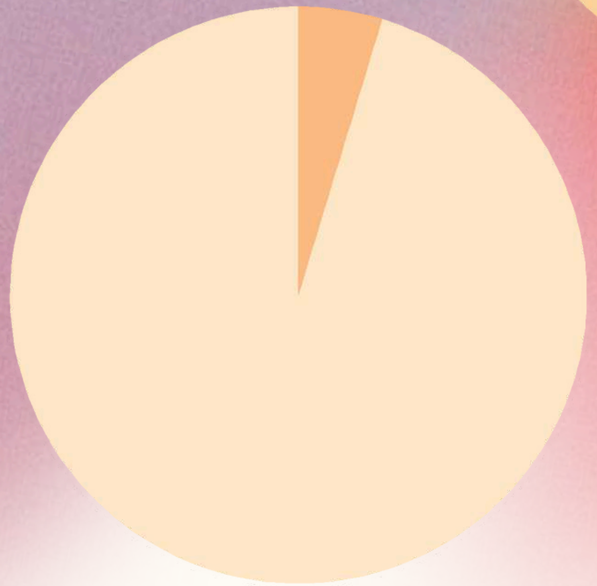


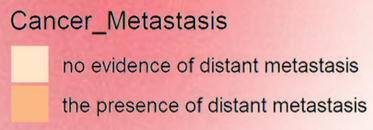



Tumor





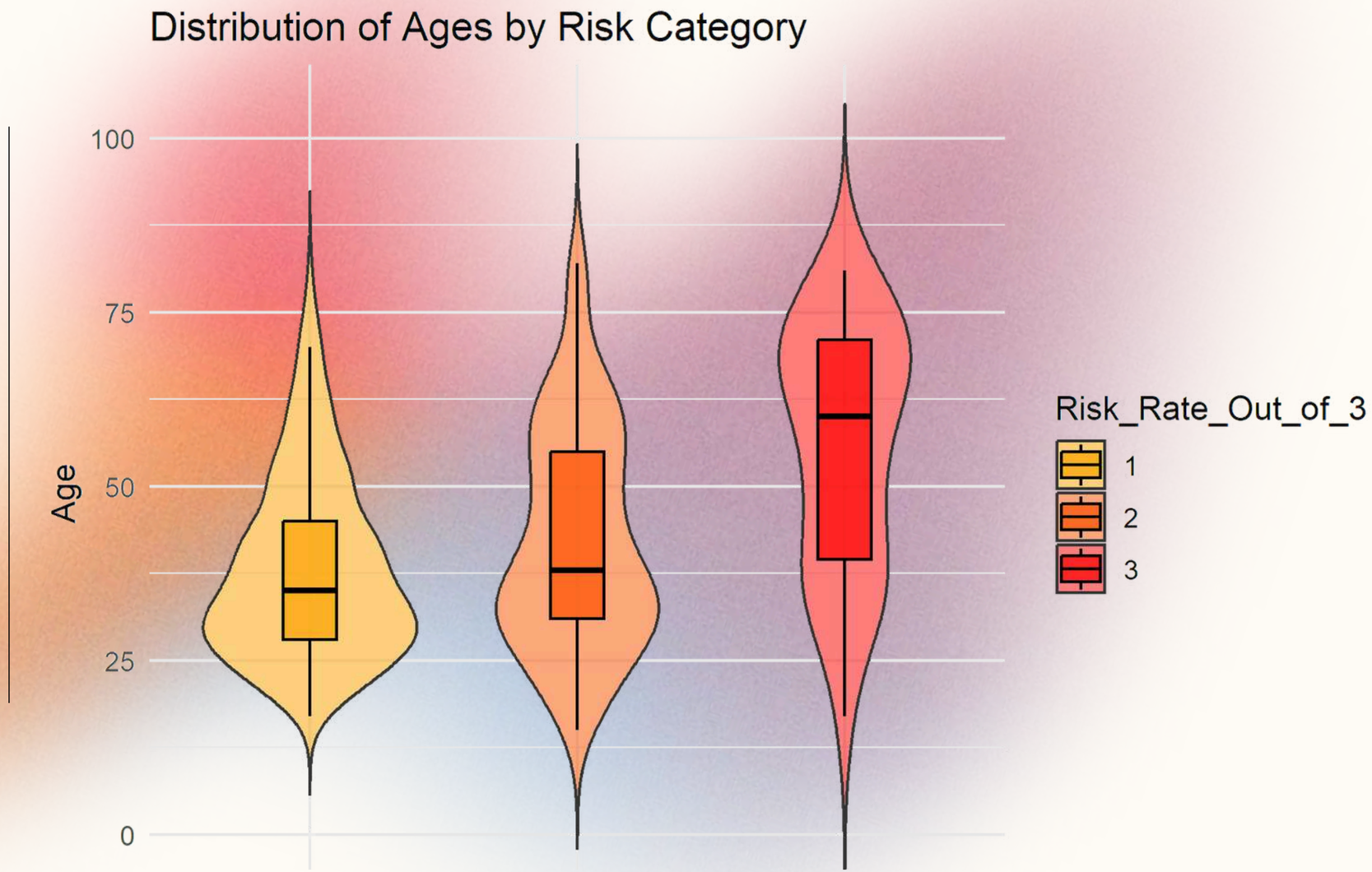





UNIVERSITÀ DEGLI
STUDI DI FIRENZE

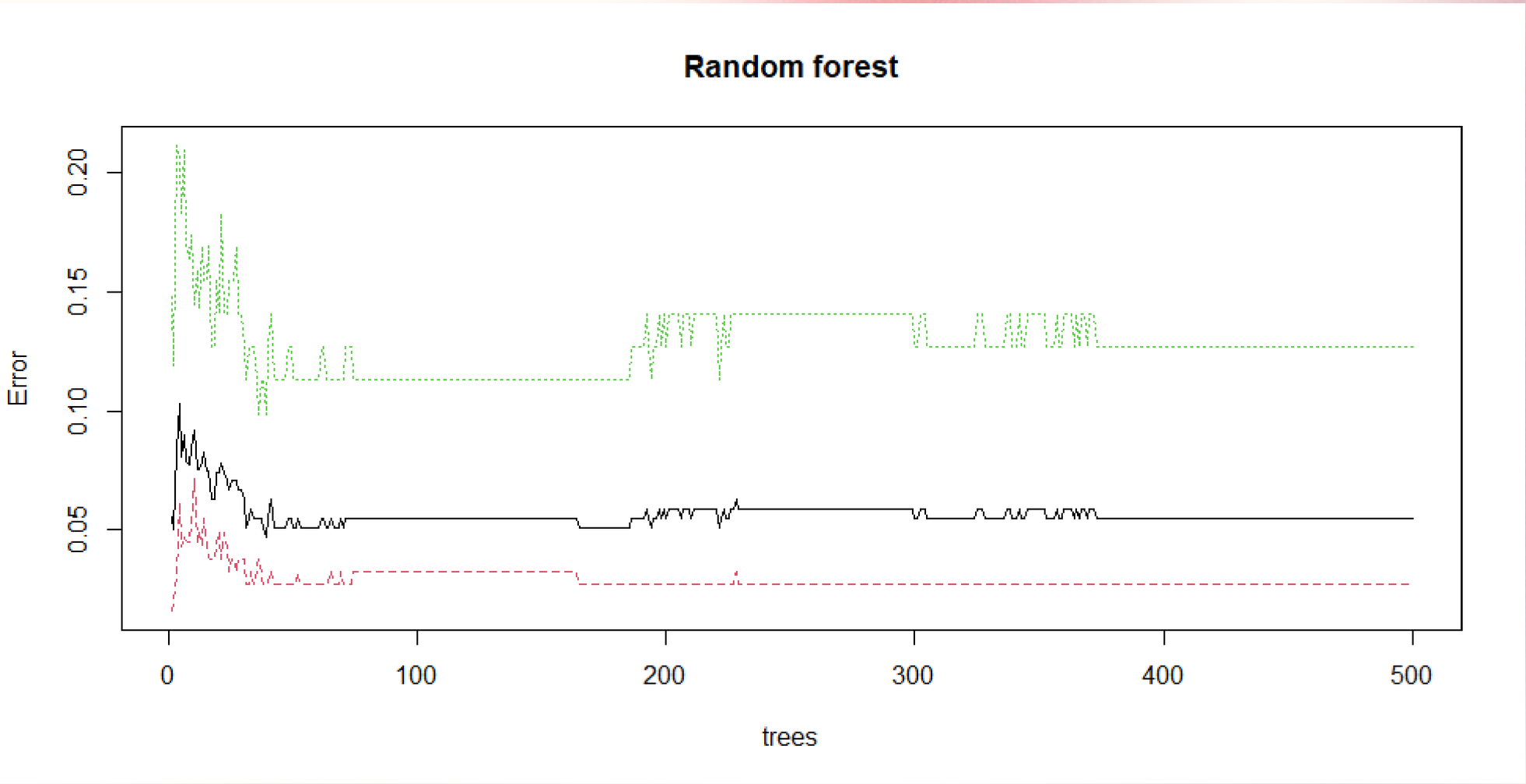
# Data Visualization

All'aumentare dell'età **aumenta anche il rischio** per il paziente associato al tumore



UNIVERSITÀ DEGLI
STUDI DI FIRENZE

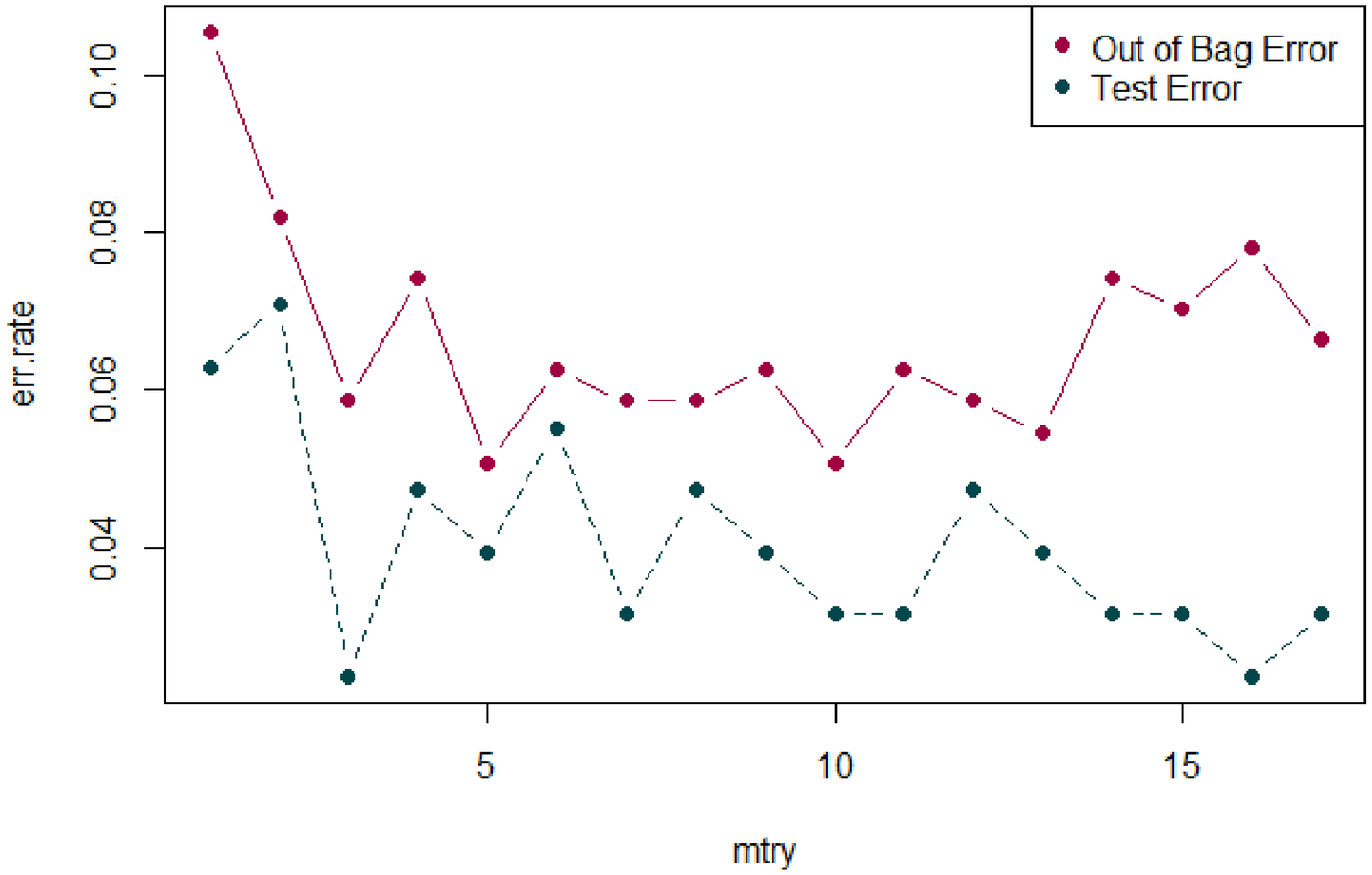
# Random Forest




L’error rate dell’oob si stabilizza con un numero di alberi pari circa a **50**

UNIVERSITÀ DEGLI
STUDI DI FIRENZE

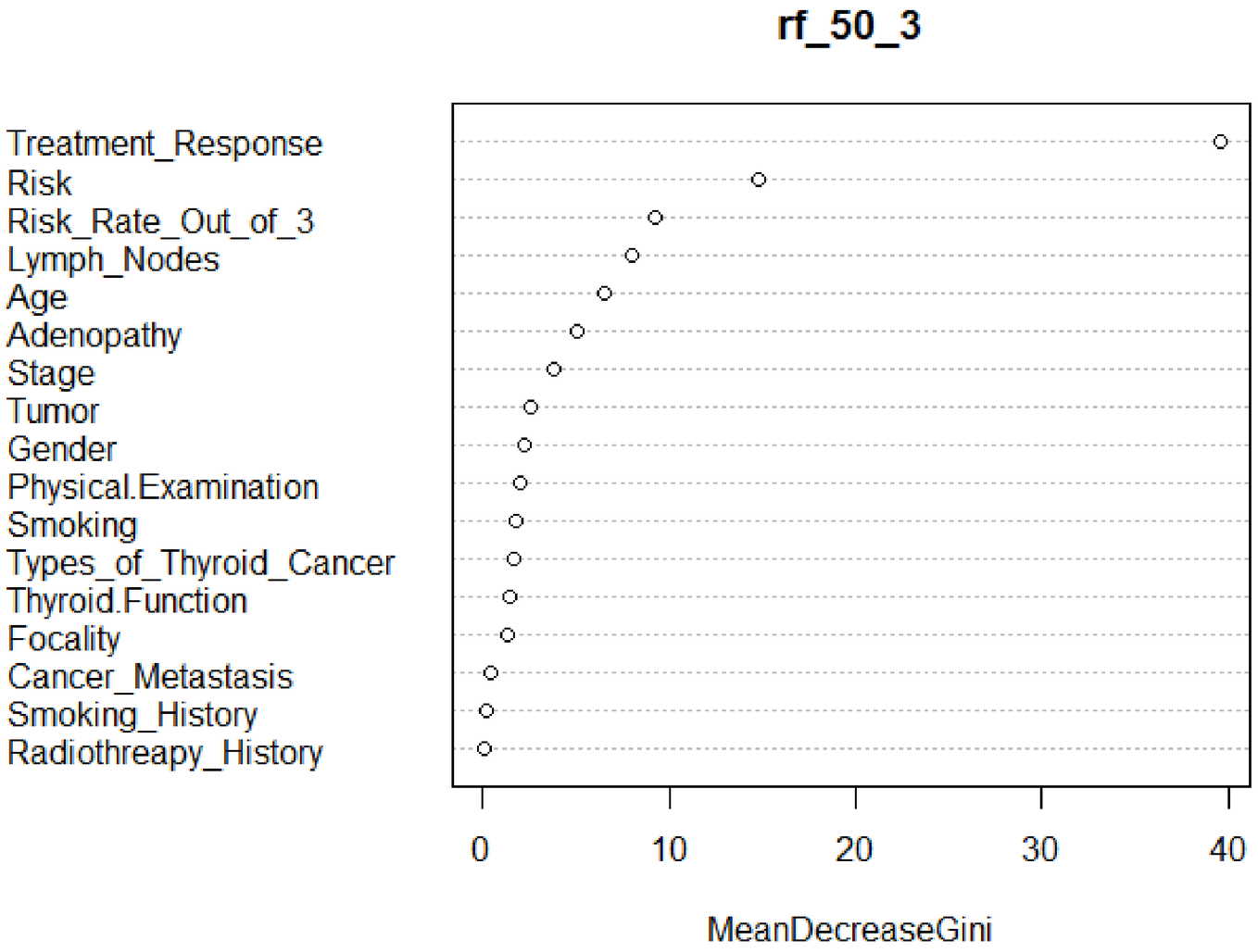
# Random Forest




Il grafico mostra che il valore migliore di mtry (con ntree impostato a 50) è circa **5**

UNIVERSITÀ DEGLI
STUDI DI FIRENZE

# Random Forest




Dal grafico si può notare che la variabile **Treatment response** è la più importante del modello

UNIVERSITÀ DEGLI
STUDI DI FIRENZE

# Bayesian Additive Regression Trees

**Passi inferenziali nell'analisi bayesiana**

- Sulla base delle conoscenze a priori si specifica un modello di probabilità iniziale (prior);
- Si aggiorna il modello sulla base dei dati osservati (posterior);
- Infine, si valuta la bontà del modello e la sensibilità delle conclusioni.

UNIVERSITÀ DEGLI STUDI DI FIRENZE

# Bayesian Additive Regression Trees

**Passi inferenziali nell'analisi bayesiana**

- Sulla base delle conoscenze a priori si specifica un modello di probabilità iniziale (prior);
- Si aggiorna il modello sulla base dei dati osservati (posterior);
- Infine, si valuta la bontà del modello e la sensibilità delle conclusioni.

**Modello assunto:**

$$Y = f(\boldsymbol{x}) + \varepsilon_i \qquad \varepsilon \sim N(0, \sigma^2)$$

$$f(\boldsymbol{x}) = \sum_{j=1}^{m} g\left(\boldsymbol{x}; \mathcal{T}_j, \mathcal{M}_j\right)$$

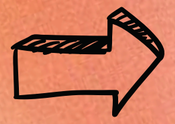
with $\mathcal{T}_j$ = structure of the tree $j$, $m$ = number of trees
$\mathcal{M}_j = (\mu_1, \ldots \mu_{M_j})$: vector of parameters for the leaves

UNIVERSITÀ DEGLI STUDI DI FIRENZE

# Bayesian Additive Regression Trees

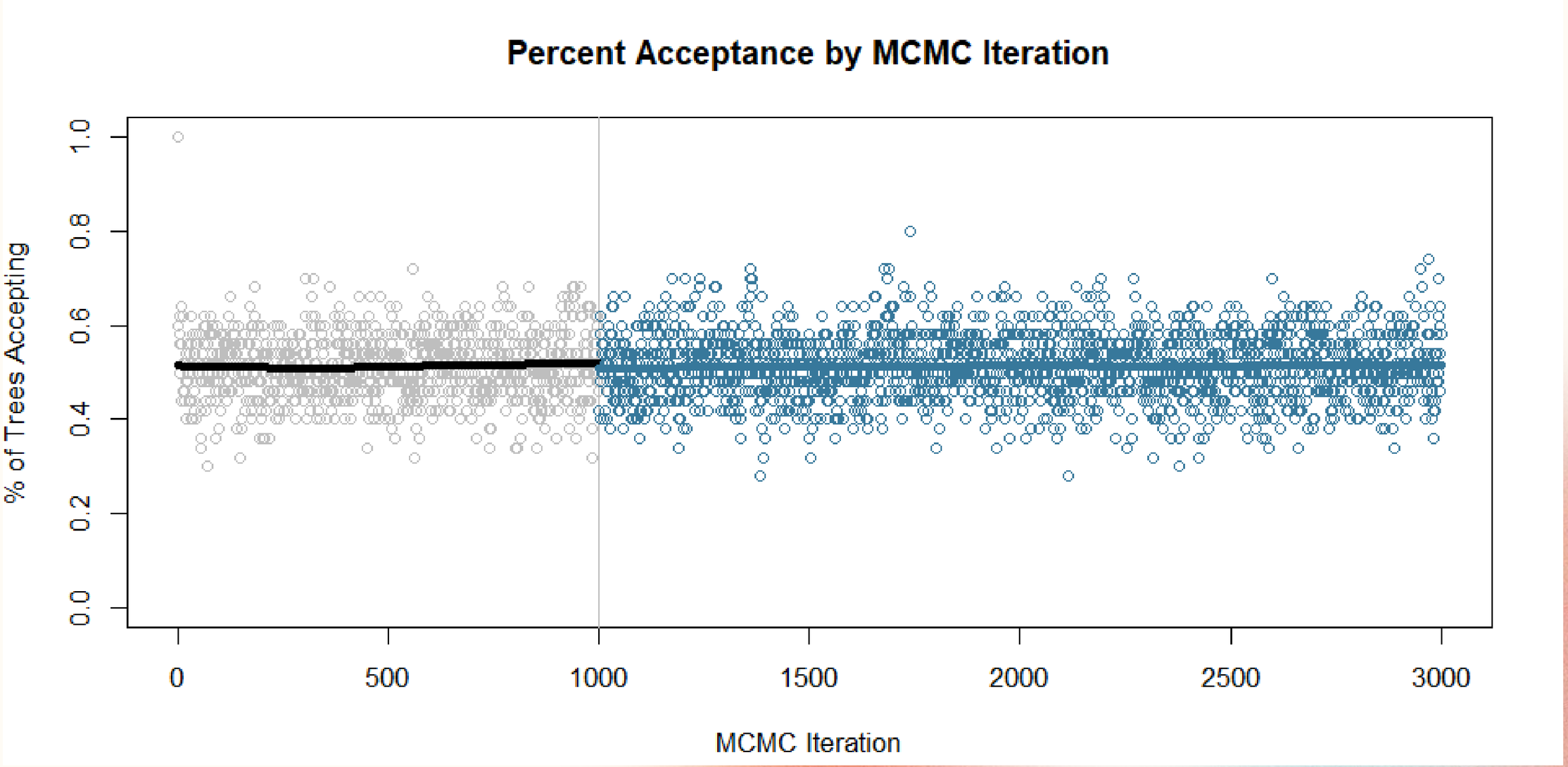
```
> summary(modello_BART)
bartMachine v1.3.4.1 for classification

training data size: n = 256 and p = 57
built in 1.8 secs on 1 core, 50 trees, 1000 burn-in and 2000 post. samples

confusion matrix:

           predicted 1 predicted 0 model errors
actual 1        69.000       7.000        0.092
actual 0         1.000     179.000        0.006
use errors       0.014       0.038        0.031
```

```
> table(predictions, Y_test)
           Y_test
predictions  0  1
          0 94  4
          1  1 28
```

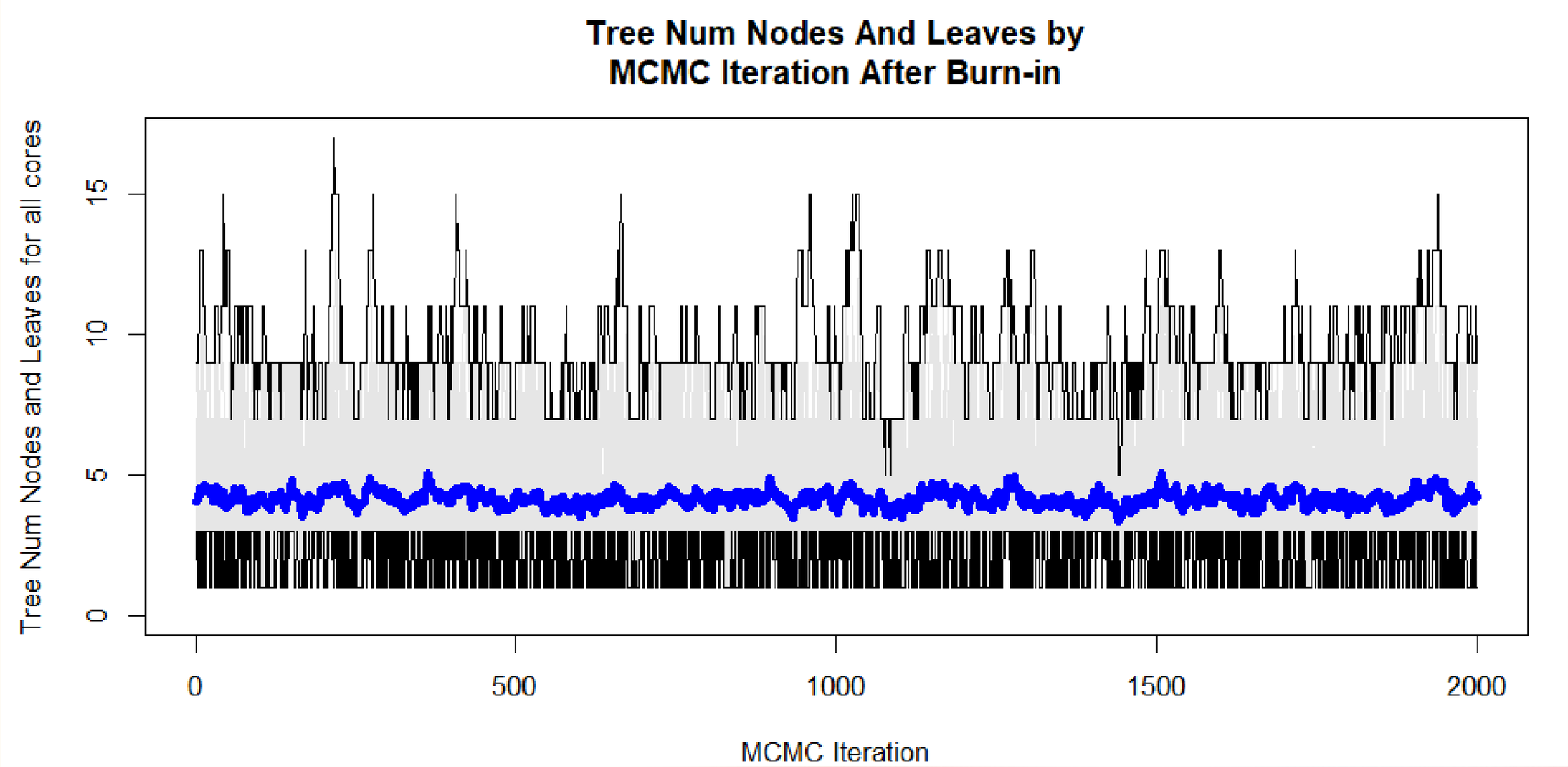


**T_EC:** 0.03937

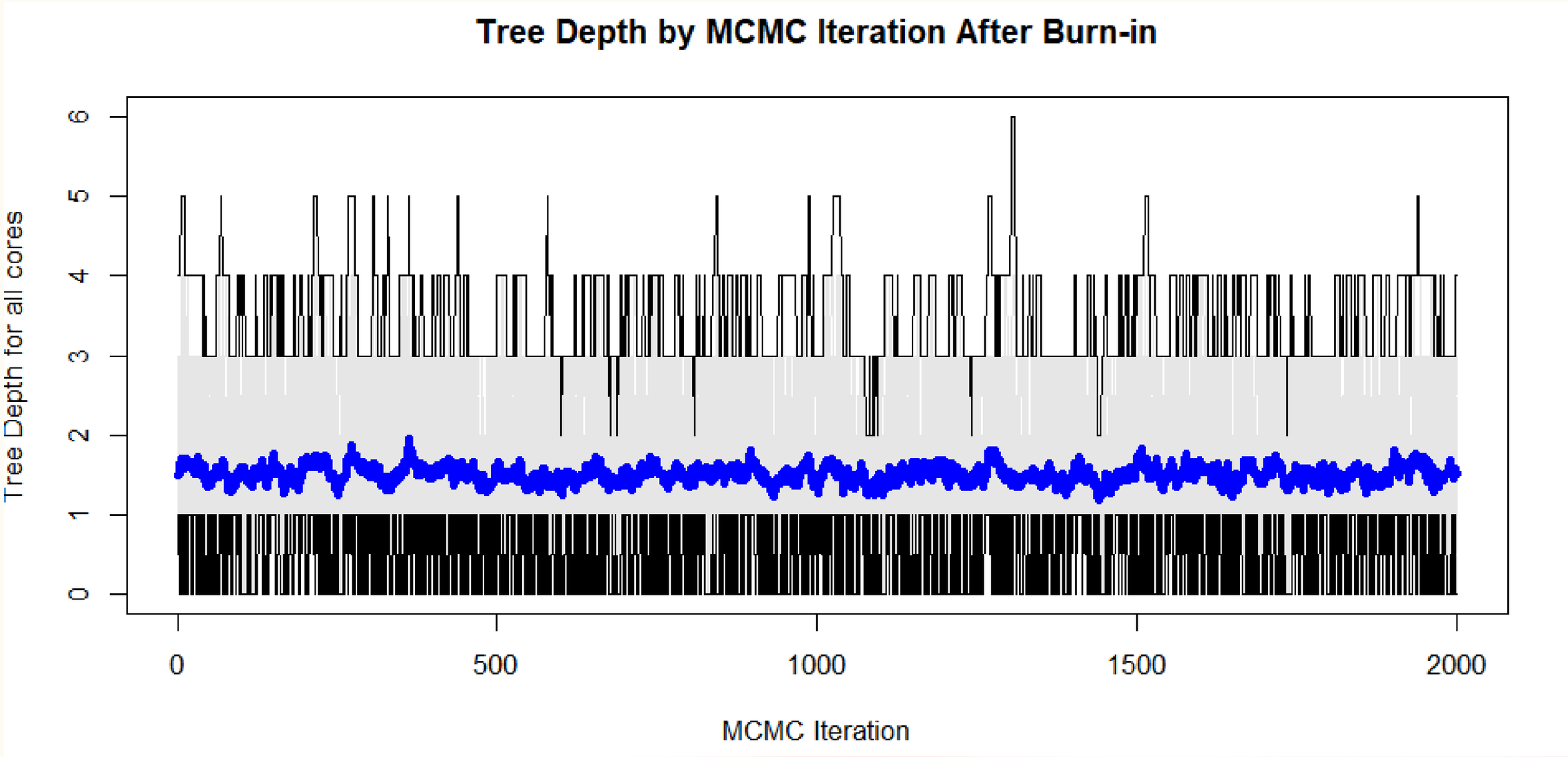
# BART


Percent Acceptance by MCMC Iteration
% of Trees Accepting
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
3000
MCMC Iteration

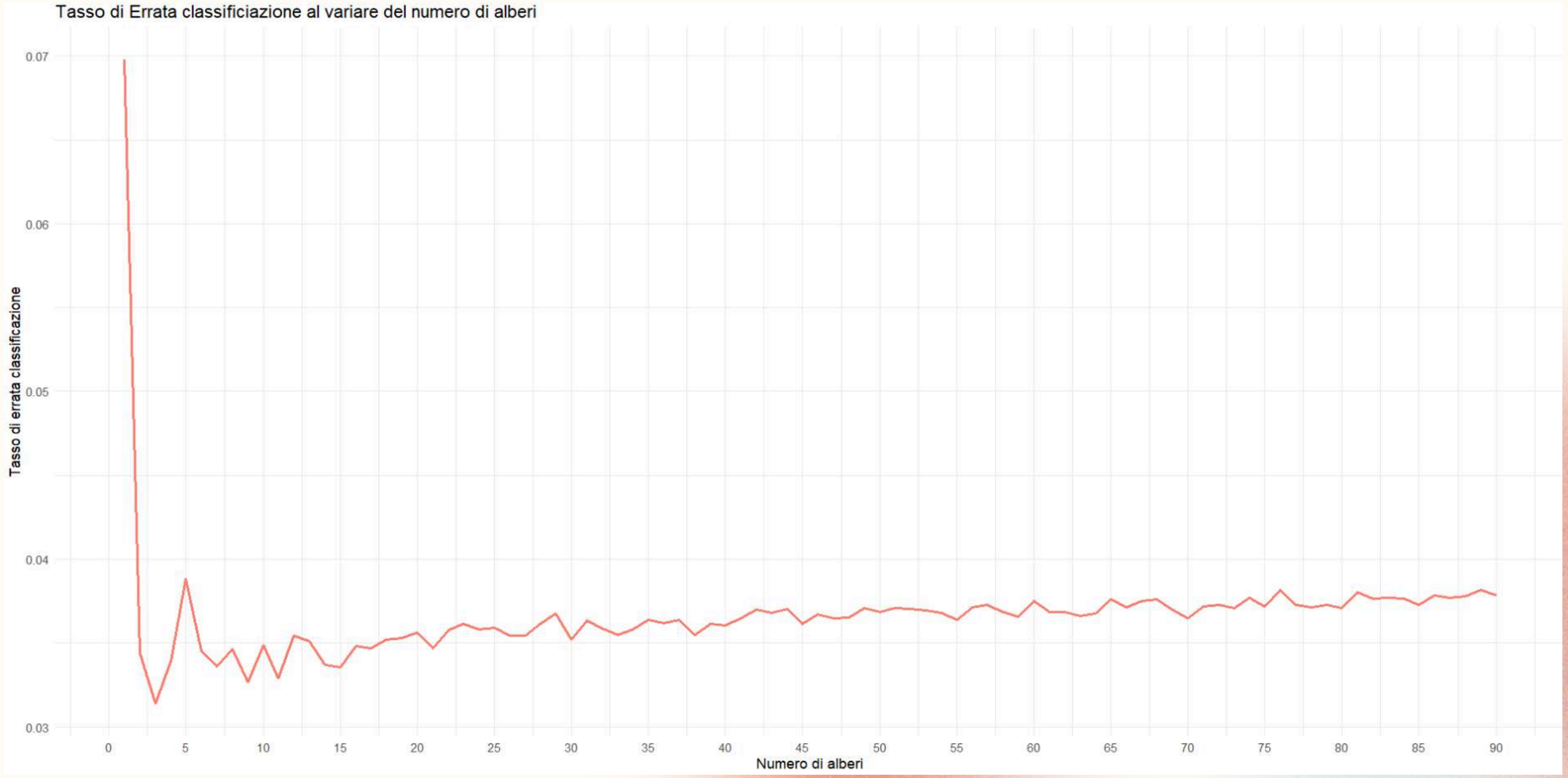
# BART


Tree Num Nodes And Leaves by
MCMC Iteration After Burn-in
Tree Num Nodes and Leaves for all cores
0
5
10
15
0
500
1000
1500
2000
MCMC Iteration

# BART


Tree Depth by MCMC Iteration After Burn-in
Tree Depth for all cores
0
1
2
3
4
5
6
0
500
1000
1500
2000
MCMC Iteration

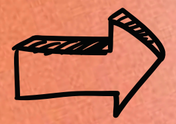
# Cosa succede al variare del numero di alberi ?



# Modello con il numero di alberi ottimale

```
> summary(modello_BART)
bartMachine v1.3.4.1 for classification

training data size: n = 256 and p = 57
built in 0.4 secs on 1 core, 3 trees, 1000 burn-in and 2000 post. samples

confusion matrix:

           predicted 1 predicted 0 model errors
actual 1        69.000       7.000        0.092
actual 0         2.000     178.000        0.011
use errors       0.028       0.038        0.035
```

```
> table(predictions, Y_test)
           Y_test
predictions  0  1
          0 94  3
          1  1 29
```



**T_EC:** 0.03149