# **Application of Machine Learning to Age adjusted death rate modeling and forecasting**.

The purpose of this study is to investigate a prediction model that identifies the potential age adjusted mortality rate in United states using machine learning based on leading causes of deaths data collected by the CDC United States from 1999 to 2017. The data details 11 unique causes of deaths and includes states, year, and number of deaths caused by unique causes.

Multiple linear Regression

Random Forest Regression

Xgboost Regression

Transform the Input Data

Pre-Process the Data

Obtain Dataset for Leading Causes of Deaths

Select the best model after hyperparameter tuning

Metric: Accuracy

Metric: RMSE

Figure 1 Diagramatic Approach for the workflow of methodology

**Test and Training Data** The training dataset represents 67% of the original data which was used to build the model; the test dataset represents the remaining 33% of the original data which was used to assess model performance.

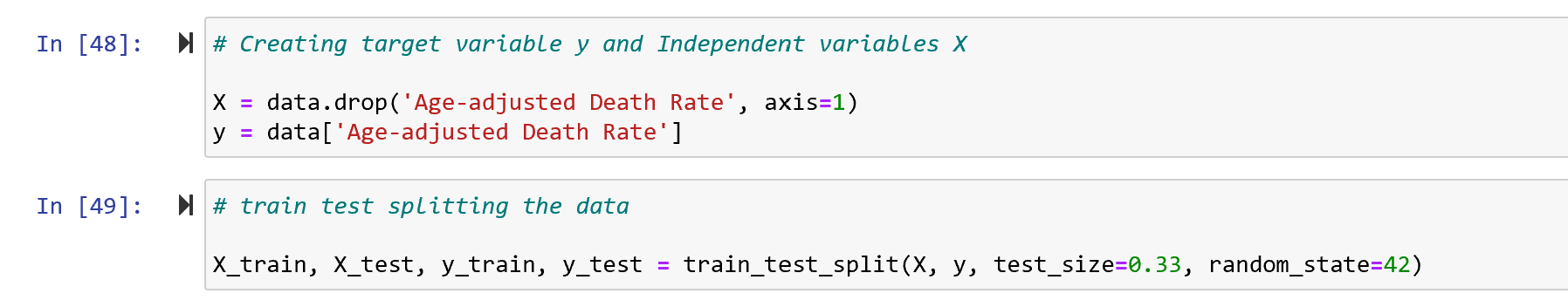


Figure Train Test Splitting data

**Multiple Linear Regression**: Multiple regression analysis was used to determine the relationships between the dependent variables Y to discover which relationships were linear. The following equation shows the relationships between Y and X1, X2, . . ., Xn

Υ = β0 + β1X1 + β2X2 + . . .. +βnXn + e

where Y is the dependent variable, X1–Xn are the independent variables, β0 is a constant (or y-intercept), and β1– βn are the coefficients of the respective variables (loading or partial slopes), which are used to describe the change in the Y value when the X value changes. RMSE calculated for this model was 161.

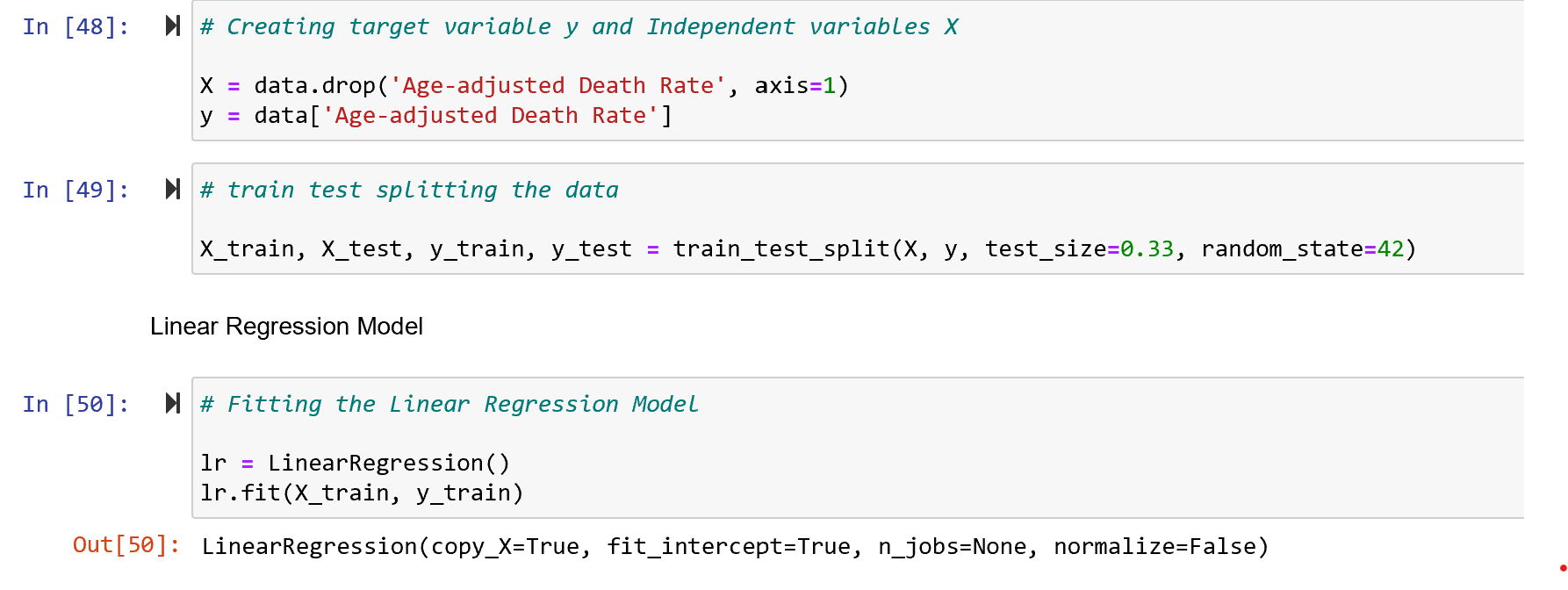
****

Figure Linear Regression

**Random Forest Regression**: - Random forest (RF) is a supervised decision tree algorithm that bags un-pruned trees by using randomly selected covariates at each split. The accuracy for test dataset from this model was 99% and RMSE was found to be 12.

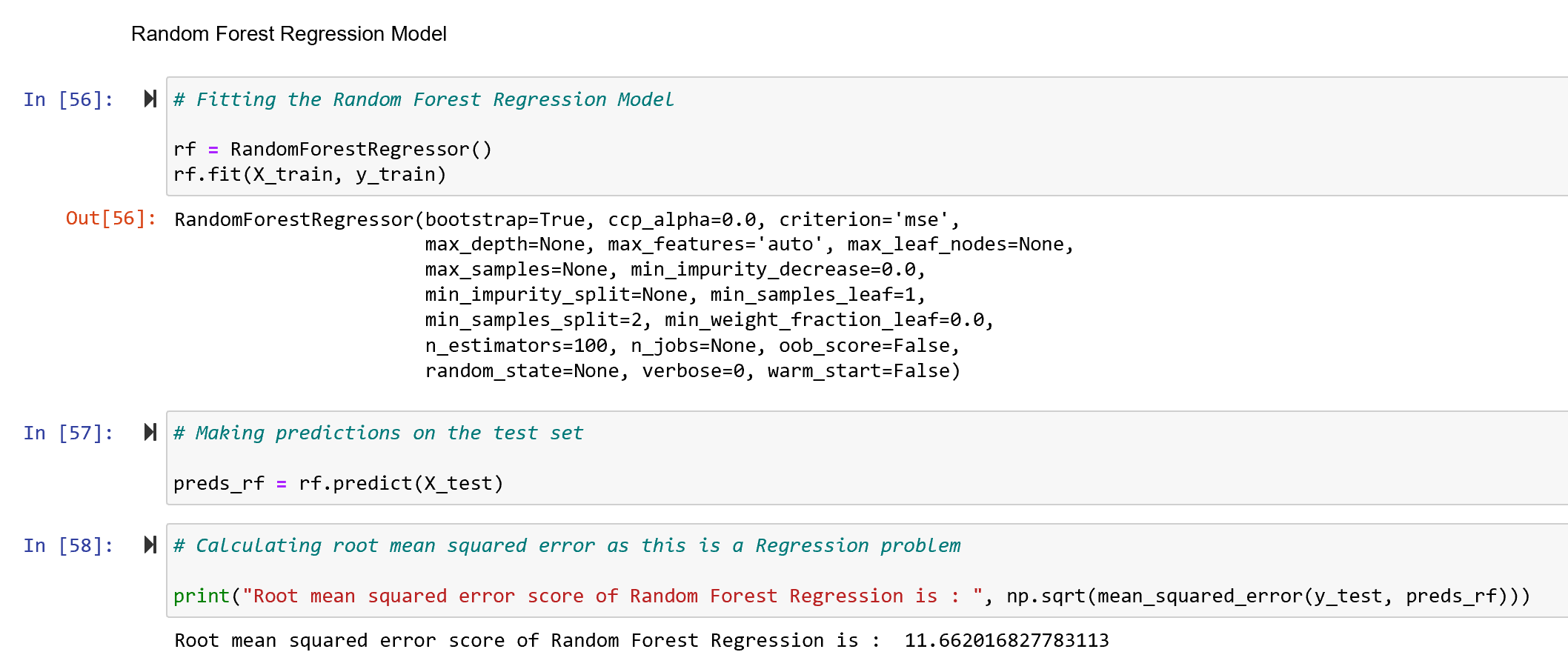


Figure Random Forest Regression

**Xgboost Regression: -** Gradient Boosting (Xgboost) is a successful machine learning library based on a gradient boosting algorithm proposed by Tianqi Chen. It has better control against overfitting by using more regularized model formalization, in comparison to prior algorithms. The Xgboost model was fitted using GridSearch and RMSE was calculated which shows decline in RMSE after hypertuning the parameters for the model.

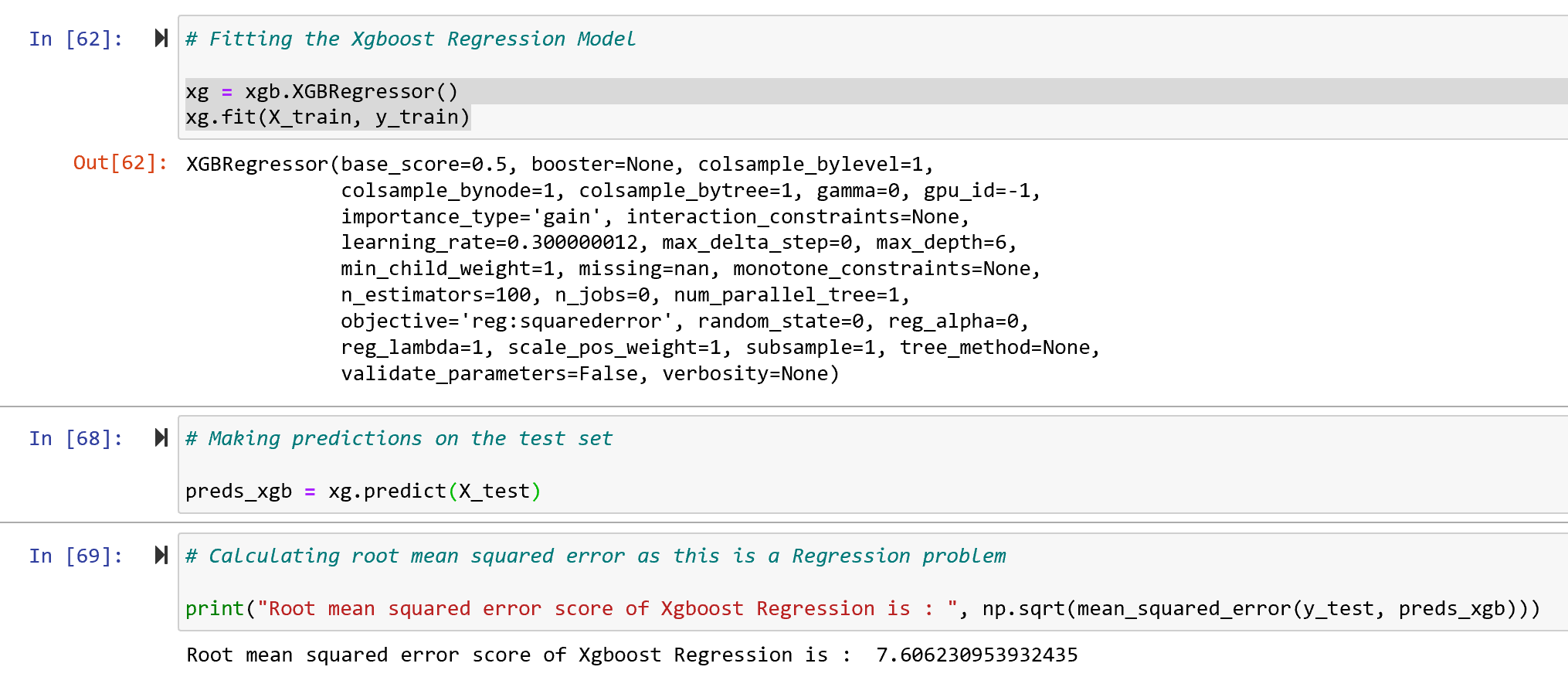


Figure Xgboost Regression

**Root Mean Square Error (RMSE): -** RMSE is the most common statistical metric of a regression model to measure the performance of model.

**Fine tune your model:** -After calculating RMSE for all the models, model with low RMSE was fine-tuned using Scikit learns GridSearch.

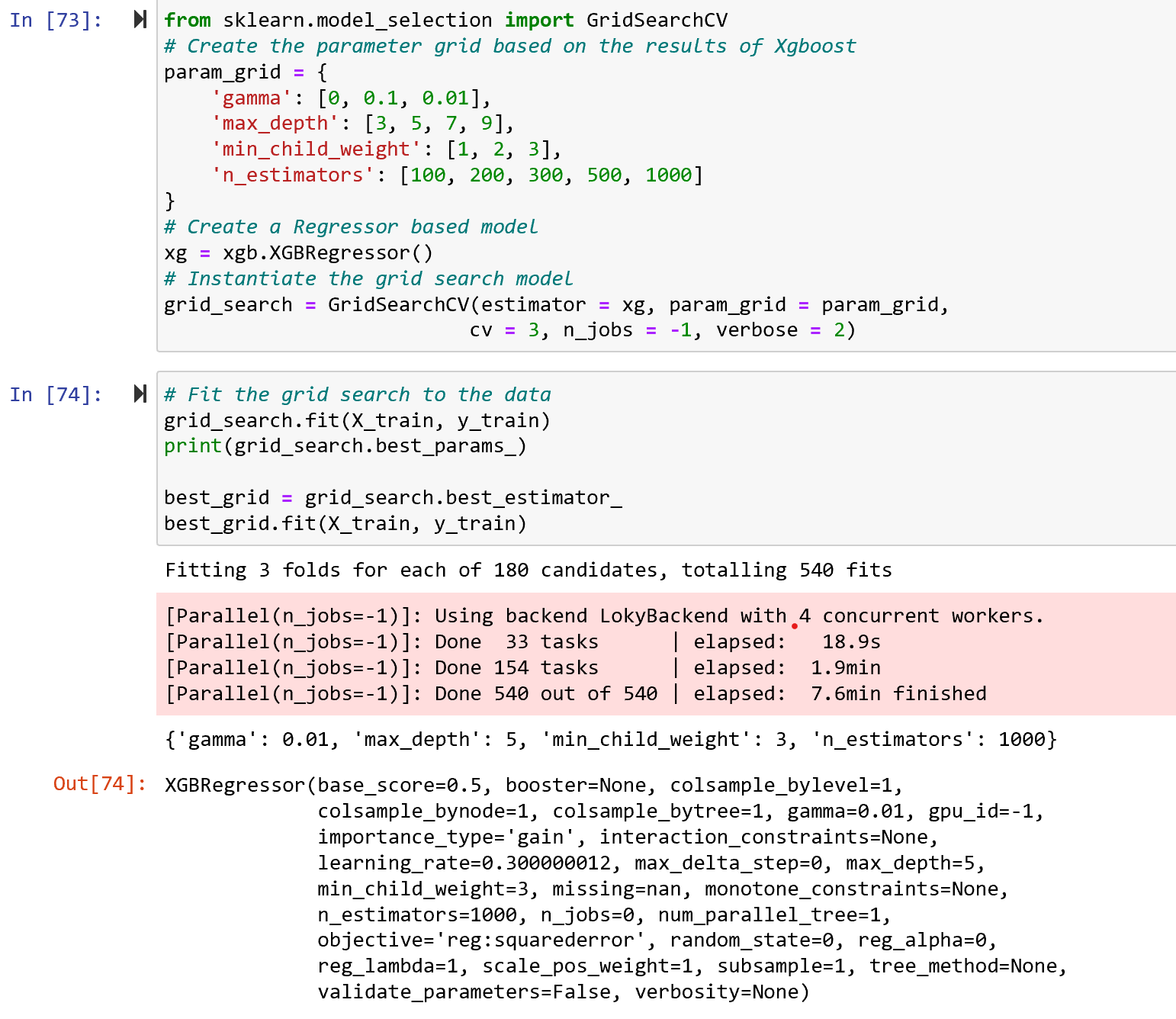


Figure Hypertuning of Parameters

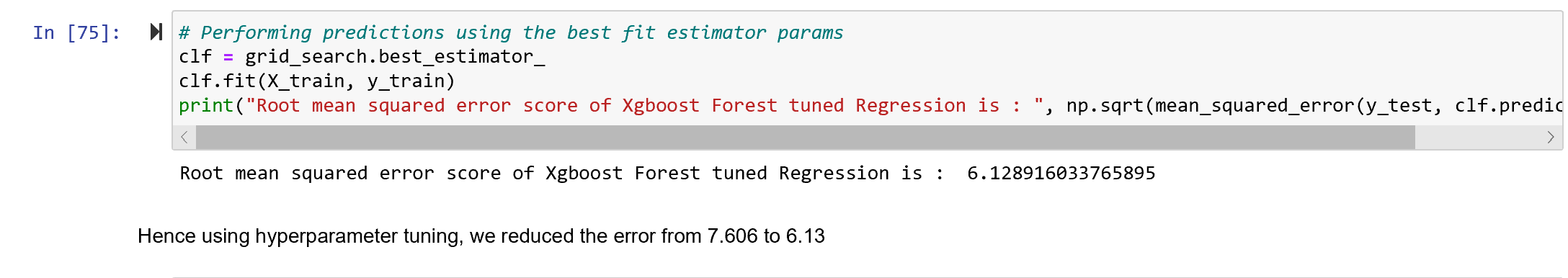


Figure RMSE of Xgboost Regression after Hypetuning

**Findings: -** Upon describing the distribution of the dataset, machine learning methods, such as linear regression, random forest regression, and XgBoost analyses were applied with the derivation of major variables influencing classification in each algorithm. A comparison of the performance of each model showed the root mean square error to be lowest for the Xgboost method, at 6.17, which translates to a 99.9% successful predictive rate in terms of classifying age adjusted death rate. To quantify the usefulness of all the variables in the entire random forest, we can look at the relative importance’s of the variables. The random forest analysis of this study indicates that the cause of deaths are the most influential factors, followed by deaths and state based on feature prediction importance graph shown in figure 2.



Figure 8 Feature Importance

Hence, this study demonstrates the feasibility of machine learning in the predicting age adjusted mortality rate on different states based on leading causes of disease. The results obtained can contribute to the prevention of leading cause of deaths by raising awareness of potential risks, by quantitatively predicting age adjusted fatality and incorporating the findings with healthcare industry.

**References:**

Muller A. Education, income inequality, and mortality: a multiple regression analysis. *BMJ*. 2002;324(7328):23‐25. doi:10.1136/bmj.324.7328.23

Yu W: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making. 2010, 10 (1): 16-10.1186/1472-6947-10-16.

Breiman L: Random forests. Machine learning. 2001, 45 (1): 5-32. 10.1023/A:1010933404324.

Chen, T.; Guestrin, C. XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD ’16, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794