

# Shahjalal University of Science and Technology

## Department of Computer Science and Engineering



### Machine Learning Lab Project

Md Motaher Hossain

Reg. No.: 2014331013

4<sup>th</sup> year, 1<sup>st</sup> Semester

Md Hasan

Reg. No.: 2014331069

4<sup>th</sup> year, 1<sup>st</sup> Semester

Submitted to

**Ms Ayesha Tasnim**

Assistant Professor

Department of Computer Science and Engineering

# Title

## House price prediction

### Data set

We use “The Boston Housing Dataset” for this problem. This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. The dataset is small in size with only 506 cases. The data was originally published by Harrison, D. and Rubinfeld, D.L. *‘Hedonic prices and the demand for clean air’*, J. Environ. Economics & Management, vol.5, 81-102, 1978.

For each model, we split dataset into two part,

- Training data - 90%
- Testing data - 10%

### Used Model and Result

#### 1. linear regression

linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. Here, scalar response is price and other features are explanatory.

The accuracy of this model is 73.20% .

#### 2. Ensemble

The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator.

Two families of ensemble methods are usually distinguished:

- In **averaging methods**, the driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimator is usually better than any of the single base estimator because its variance is reduced.  
Examples: Bagging methods, Forests of randomized trees, ...
- By contrast, in **boosting methods**, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble. Examples: AdaBoost, Gradient Tree Boosting, ...

We use Gradient tree boosting and accuracy is 92.02% .

### 3. SVM

Support Vector Machine(SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences.

The accuracy of this model is -7.19% .

### 4. Decision Tree

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

The accuracy of this model is 70.08% .

### 5. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate

by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

The accuracy is 85.88% .

## **6. MLPRegression**

A multilayer perceptron is a logistic regressor where instead of feeding the input to the logistic regression you insert a intermediate layer, called the hidden layer, that has a nonlinear activation function (usually tanh or sigmoid) .

The accuracy is 58.79% .

# Title

## Stock price prediction

### Dataset

Here, we use stock price dataset of a company from 2004-08-19 to 2018-03-27. This dataset has 3425 cases/rows .

For each model, we split dataset into two part,

- Training data - 80%
- Testing data - 20%

### Used Model and Result

Model	Accuracy
Linear regression	97.89%
Ensemble	98.45%
SVM	21.96%
Decision Tree	98.28%
Random Forest	97.63%
MLPRegression	92.95%