**Machine Learning Mini Project**

**Bachelor of Technology**

**in**

**Computer Science & Engineering**

Submitted by

Anand Kumar Kandu

Roll No:1821510004

**(Affiliated to Dr. A.P.J. Abdul Kalam Technical University, Lucknow)**

**ACADMIC SESSION**

**(2020-2021)**

****

**Sanskar College of Engineering and Technology**

**Opp. Jindal Pipes Ltd. NH-24, Ghaziabad, Ph: 0120-2809200**

**ACKNOWLEDGEMENT**

It is my proud privilege and duty to acknowledge the kind of help and guidance received from several people in preparation of this project. It would not have been possible to prepare this report in this form without their valuable help, co-operation and guidance.

First and foremost, I would like to record my sincere gratitude to “**Nextwing Infotech**” coordinators for their constant support and encouragement in preparation of this report and for making available online lectures and interface facilities needed to prepare this report.

The online lectures and doubt sessions was very helpful to us in giving the necessary background information and inspiration in choosing this project. Their contributions and technical support in preparing this report are greatly acknowledged.

Last but not the least, I wish to thank my parents for financing my studies in this college as well as for constantly encouraging me to learn engineering. Their personal sacrifice in providing this opportunity to learn engineering is greatly acknowledged.

**Anand Kumar Kandu**

**TABLE OF CONTENTS**

Abstract ---------------------------------------------------------------------------- 4

Introduction ---------------------------------------------------------------------------- 5

Purpose ---------------------------------------------------------------------------- 6

Configuration ---------------------------------------------------------------------------- 6

Library Required ---------------------------------------------------------------------------- 7

* Numpy
* Pandas
* Seaborn
* Matplotlib

Algorithm Used ----------------------------------------------------------------- 8

* Linear Regression (sklearn)

Regression Evaluation Metrics --------------------------------------------------------------- 9

* Mean Absolute Error (MAE)
* Mean Squared Error (MSE)
* Root Mean Squared Error (RMSE)

Coefficient of determination ------------------------------------------------------------------- 10

Bibliography ------------------------------------------------------------------- 11

**ABSTRACT**

In this project, we were asked to experiment with a real-world dataset, and to explore how machine learning algorithms can be used to find the patterns in data. We were expected to gain experience using a common data-mining and machine learning library, Weka, and were expected to submit a report about the dataset and the algorithms used. After performing the required tasks on a dataset of my choice, herein lies my final report.

Keywords: Machine Learning, Pattern Recognition, Classification, Supervised learning, Artificial Intelligence

**Machine Learning Mini Project**

**Introduction:**

Machine learning is a sub-domain of computer science which evolved from the study of pattern recognition in data, and also from the computational learning theory in artificial intelligence. It is the first-class ticket to most interesting careers in data analytics today. As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions.

Machine Learning can be thought of as the study of a list of sub-problems, viz: decision making, clustering, classification, forecasting, deep-learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic algorithms, sparse dictionary learning, etc. Supervised learning, or classification is the machine learning task of inferring a function from a labeled data. In Supervised learning, we have a training set, and a test set. The training and test set consists of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a function that maps the input vector to the output vector with minimal error. In an optimal scenario, a model trained on a set of examples will classify an unseen example in a correct fashion, which requires the model to generalize from the training set in a reasonable way. In layman’s terms, supervised learning can be termed as the process of concept learning, where a brain is exposed to a set of inputs and result vectors and the brain learns the concept that relates said inputs to outputs. A wide array of supervised machine learning algorithms is available to the machine learning enthusiast, for example Neural Networks, Decision Trees, Support Vector Machines, Random Forest, Naïve Bayes Classifier, Bayes Net, Majority Classifier etc.

### Purpose:

We have multiple featured data set,

Variables in order:

* CRIM per capita crime rate by town
* ZN proportion of residential land zoned for lots over 25,000 sq.ft.
* INDUS proportion of non-retail business acres per town
* CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
* NOX nitric oxides concentration (parts per 10 million)
* RM average number of rooms per dwelling
* AGE proportion of owner-occupied units built prior to 1940
* DIS weighted distances to five Boston employment Centre’s
* RAD index of accessibility to radial highways
* TAX full-value property-tax rate per $10,000
* PTRATIO pupil-teacher ratio by town B 1000(Bk - 0.63) ^2 where Bk is the proportion of blacks by town
* LSTAT % lower status of the population
* MEDV Median value of owner-occupied homes in $1000's

Using these features, we have to predict the prices of the houses.

### Configuration:

<https://jupyter.org/install>

##### **Library Required:**

* **NumPy:** Numpy is most suitable for performing basic numerical computations such as mean, median, range, etc. Alongside, it also supports the creation of multi-dimensional arrays.
* **Pandas:** Pandas is best at handling tabular data sets comprising different variable types (integer, float, double, etc.). In addition, the pandas library can also be used to perform even the most naive of tasks such as loading data or doing feature engineering on time series data**.**
* **Seaborn:** Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
* **Matplotlib:** Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.
* **Sklerrn:** The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction. Please note that sklearn is used to build machine learning models.

**Algorithm Used:**

* **Linear Regression (sklearn):**

what is regression? and how its work?

Regression is a parametric technique used to predict continuous (dependent) variable given a set of independent variables. It is parametric in nature because it makes certain assumptions (discussed next) based on the data set. If the data set follows those assumptions, regression gives incredible results.

**Y = βo + β1X + ∈**

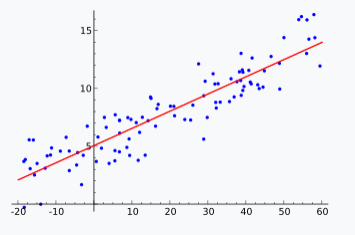
where, Y - Dependent variable

X - Independent variable

βo - Intercept

β1 - Slope

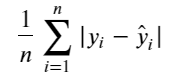
∈ - Error



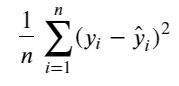
**Regression Evaluation Metrics:**

Here are three common evaluation metrics for regression problems:

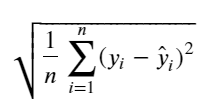
* **Mean Absolute Error (MAE**)**:** is the mean of the absolute value of the errors:



* **Mean Squared Error (MSE):** is the mean of the squared errors:



* **Root Mean Squared Error (RMSE):** is the square root of the mean of the squared errors:



**MAE** is the easiest to understand, because it's the average error.

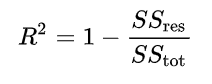
**MSE** is more popular than MAE, because MSE "punishes" larger errors, which tends to be useful in the real world.

**RMSE** is even more popular than MSE, because RMSE is interpretable in the "y" units.

**Coefficient of determination:**

In statistics, the coefficient of determination, denoted R2 or r2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

this metric explains the percentage of variance explained by covariates in the model. It ranges between 0 and 1. Usually, higher values are desirable but it rests on the data quality and domain. For example, if the data is noisy, you'd be happy to accept a model at low R² values. But it's a good practice to consider adjusted R² than R² to determine model fit.



Basically R-square is a score to measure the accuracy of the model. Generally, the model accuracy is measure on the scale to 0 to 1. if R-square produce result as 0.79 then your model is about 79% accurate.

**Bibliography**

Following are the various references used in the preparation of this project

* <https://www.hackerearth.com/practice/machine-learning/prerequisites-of-machine-learning/basic-probability-models-and-rules/tutorial/>
* <https://en.wikipedia.org/wiki/Linear_regression>
* <https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>
* <https://matplotlib.org/>
* <https://www.geeksforgeeks.org/image-classifier-using-cnn/>
* <http://cs231n.stanford.edu/>
* <https://www.udemy.com/>
* <https://en.wikipedia.org/wiki/Coefficient_of_determination>