ADVANCED MACHINE LEARNING PROJECT

|  |  |  |
| --- | --- | --- |
|  | ID | Full Name |
| 1 | 20211087 | يوسف عماد سيد على |
| 2 | 20210363 | زياد حسام حسن ابراهيم |
| 3 | 20211081 | يوسف عبدالمنعم حمدي عبدالمنعم |
| 4 | 20211054 | يسر حسن محمود ابراهيم |
| 5 | 20210371 | زياد محمد السيد نصر |
| 6 | 20210379 | زياد هاني احمد فرغلي |
| 7 | 20210402 | سالم جمال صلاح عزام |

OCR of Handwritten Arabic Digits using DTS & RTS. (Classification)

House prices. (Regression)

1. **The Dataset employed**

**What is A Dataset?**

**a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.**

--------------------------------------------A picture containing text, receipt, screenshot

Description automatically generated

**The dataset used in classification is Arabic digits:** <https://www.https://www.kaggle.com/datasets/mloey1/ahdd1.com/datasets/mloey1/ahdd1>

**It contains two main files** "Arabic Handwritten Digits Dataset CSV/csvTrainImages 60k x 784.csv" and "Arabic Handwritten Digits Dataset CSV/csvTrainLabel 60k x 1.csv"

**One contains the actual images**

Text

Description automatically generated with medium confidence

**Represented as 28\*28 array each index representing a pixel opacity from (0 to 255)**

A picture containing text

Description automatically generated

**And the other contains the labels for these images.**

**The dataset used in regression is house prices:**

[**House Prices - Advanced Regression Techniques | Kaggle**](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques)

**What is Decision tree (classification)?**

**Splitting**: **The decision tree algorithm starts with the entire dataset at the root node. It then selects the best feature to split the data into smaller subsets. This process is repeated recursively for each subset, creating a tree-like structure.**

**Node Selection**: **At each step, the algorithm selects the best feature to split the data based on a criterion such as Gini impurity or information gain.**

**Stopping Criteria**: **The recursion continues until one of the stopping criteria is met, such as reaching a maximum depth, having a minimum number of samples in a node, or no further improvement can be made.**

**Leaf Node**: **Once the stopping criteria are met, the final nodes of the tree are called leaf nodes, and they contain the predicted class label.**

**Decision tree Architecture**

**A diagram of a tree

Description automatically generated**

**Hyperparameter used:(Best parameters)**

**1-Max\_depth=20. 2- min\_samples =1.**

**3-min\_samples\_split=2**

**Results:**

**1-accuracy = 93 % (test)**

**What is Neural network(classification)?**

**Input Layer: The layer where the input data is fed into the network. Each neuron in this layer represents an input feature.**

**Hidden Layers: Layers between the input and output layers where the computational processing occurs. These layers extract features from the input data through a series of transformations. Deep neural networks have multiple hidden layers, allowing them to learn complex representations.**

**Output Layer: The layer where the network produces its final output. The number of neurons in this layer depends on the type of task the network is designed for. For example, in classification tasks, each neuron may represent a class label, while in regression tasks, there may be a single neuron for continuous output.**

**Connections: Neurons in adjacent layers are connected by weighted connections. Each connection has an associated weight that determines the strength of the connection. During training, these weights are adjusted based on the network's performance on training data.**

**Activation Functions: Each neuron typically applies an activation function to its weighted sum of inputs to introduce non-linearity into the network. Common activation functions include sigmoid, tanh, ReLU (Rectified Linear Unit), and SoftMax.**

**Neural network Architecture**

**A diagram of a network

Description automatically generated**

**Hyperparameter used:**

1. **Batch size =128 2-no. of epochs = 3**
2. **Verbose = 1**

**Results:**

**Accuracy = 98.5 % (test)**

**What is support vector machine(regression)?**

**Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. In classification, SVMs are particularly effective for binary classification problems, but they can also be extended to handle multi-class classification tasks.**

**The main idea behind SVM is to find the optimal hyperplane that best separates the data points belonging to different classes in the feature space. This hyperplane was chosen to maximize the margin, which is the distance between the hyperplane and the nearest data points (support vectors) from each class.**

**SVM can handle both linearly separable and non-linearly separable data by using different kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid. These kernel functions transform the input features into a higher-dimensional space, where the data points are more likely to be linearly separable.**

**In addition to classification, SVMs can also be used for regression tasks, where the goal is to predict a continuous target variable instead of class labels. In regression SVM, the algorithm tries to fit a hyperplane that best approximates the relationship between the input features and the target variable while minimizing the error.**

**Overall, SVM is a versatile and powerful algorithm that is widely used in various fields such as image classification, text classification, bioinformatics, and more.**

**Support vector machine Architecture**

**A diagram of support vector

Description automatically generated**

**Results:**

**Accuracy = 90 % (test)**