

**MachineLearning - Project Report Document**

|  |  |
| --- | --- |
| **Student Name** | Tella Krishna Prasad |
| **Batch** | AI Elite 18 |
| **Project Name** | Recognition of Handwritten Alphabets |
| **Project Domain** | Computer Vision |
| **Type of Machine Learning** | Supervised |
| **Type of Problem** | Classification |
| **Project Methodology** | CRISP-DM |
| **Stages Involved** | * Business Understanding * Data Collection and Understanding * Data Preparation * Model |

**Stage 1: Business Understanding:**

MNIST ("Modified National Institute of Standards and Technology") is the de facto “hello world” dataset of computer vision. Since its release in 1999, this classic dataset of handwritten images has served as the basis for benchmarking classification algorithms. As new machine learning techniques emerge, MNIST remains a reliable resource for researchers and learners alike.

In this SPRINT, your goal is to preprocess the image files given to you and create a pandas data frame.

**Stage 2: Data Collection and Understanding**

1. **Data Collection:**

The data was provided to us by the client.

1. **Data Understanding:**

|  |  |  |
| --- | --- | --- |
| **S No** | **Feature Name** | **Data Type** |
| 1 | Pixels | int |
| 2 | Labels | char |

**Stage 3: Data Preparation  
  
a) Exploratory Data Analysis:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Type** | **Feature Names** | **Observation** |
| 1 | Missing Values | NA | NA |
| 2 | Duplicates | NA | NA |
| 3 | Outliers | NA | NA |

**b) Data Cleaning/wrangling:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S no** | **Type of Cleaning** | **Technique** | **Feature Name** | **Reason** |
| 1 | Image Editing | Flattening layer | Image to array | Image data can be analysed by converting raw images into flatten values |

**Stage 4: Model Building:**

|  |  |  |
| --- | --- | --- |
| **S No** | **Type of Problem** | **Algorithm Name** |
| 1 | Classification | Logistic Regression |
| 2 | Classification | Decision Tree Classifier |
| 3 | Classification | Random Forest Classifier |

**Logistic Regression:**

Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables.

**Decision Tree Classifier:**

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

**Random Forest Classifier:**

Random forest is a commonly-used machine learning algorithm, trademarked by Leo Bierman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

**Stage 5: Model Training:**

|  |  |  |
| --- | --- | --- |
| **S No** | **Algorithm Name** | **Metric used for Evaluation** |
| 1 | Logistic Regression | Accuracy |
| 3 | Decision Tree Classifier | Accuracy |
| 4 | Random Forest Classifier | Accuracy |

**Stage 5: Model Evaluation:**

|  |  |  |
| --- | --- | --- |
| **S No** | **Algorithm Name** | **Metric Score** |
| 1 | Logistic Regression | 0.877761 |
| 3 | Decision Tree Classifier | 0.947279 |
| 4 | Random Forest Classifier | 0.985490 |

**Challenges Faced:**

1. It took more time to load the data with CPU RAM (12GB) later by converting TPU v2 the processing is spread up.

**Conclusion:**

Based on the Accuracy metrics, the best model for the classification problem appears to be Logistic Regression model. Although all the other models had similar accuracy scores, Logistic Regression model had the highest Accuracy.

While Logistic Regression model had a good accuracy score of 0.985490. Therefore, based on the metrics evaluated Logistic Regression model appears to be the best model for this classification problem.