HANDWRITTEN DOCUMENTATION

PROJECT IDEA

Optical character recognition or optical character reader (OCR) is the automated conversion of images of typed, handwritten, or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (e.g., the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (e.g., from a television broadcast).

Our project takes the dataset, processing it and recognizing the character through 2 algorithms which are

1. Decision Tree
2. Random Forest

Optical character recognition or optical character reader (OCR) can be done through it many applications like

- Data entry for business documents, Cheque, passport, invoice, bank statement and receipt
- Automatic number plate recognition
- In airports, for passport recognition and information extraction
- Automatic insurance documents key information extraction
- Traffic Sign Recognition
- Extracting business card information into a contact list
- More quickly make textual versions of printed documents, book scanning for (Project Gutenberg)
- Make electronic images of printed documents searchable (Google Books)
- Converting handwriting in real-time to control a computer (Pen Computing)

- Defeating Captcha anti-bot systems, though these are specifically designed to prevent OCR. The purpose can also be to test the robustness of CAPTCHA anti-bot systems.
- Assistive technology for blind and visually impaired users
- Writing the instructions for vehicles by identifying CAD images in a database that are appropriate to the vehicle design as it changes in real time.

MAIN FUNCTIONALITIES

We have used some libraries and it has mentioned at section 7 you can refer it

Decision Tree

- Used two variables each one read a CSV file, first one train file and second one test file.
- Checked the train file and it printed 5 rows x 785 columns.
- Renamed first column as label for both test and train set, checked train data.
- Dropped label column of test dataset because it will generate error, viewed first 5 records of test file.
- List of all digits that are going to be predict viewed first 5 records of train file.
- Defined the number of samples for training set and for validation set.
- Generated training data from train file.
- Generated validation data from train file.
- Used 3 plots, first plot gives the wrong answer, second through the labels in dataset it predicts the letter through index while third u give it an index and he got the right answer with the number in dataset.
- Initialized decision tree classifier with (criterion = 'entropy', max_depth = 14, random_state = 33) then fitting the training data.
- Predict label's value using classifier.
- Printed validation accuracy.
- Printed validation confusion matrix.

Random Forest

- Used two variables each one read a CSV file, first one train file and second one test file.
- Checked the train file and it printed 5 rows x 785 columns.
- Renamed first column as label for both test and train set, checked train data.
- Dropped label column of test dataset because it will generate error, viewed first 5 records of test file.
- List of all digits that are going to be predict viewed first 5 records of train file.
- Defined the number of samples for training set and for validation set.
- Generated training data from train file.
- Generated validation data from train file.
- Used 3 plots, first plot gives the wrong answer, second through the labels in dataset it predicts the letter through index while third u give it an index and he got the right answer with the number in dataset.
- Initialized random forest classifier with then fitting the training data.
- Predict label's value using classifier.
- Printed validation accuracy.
- Printed validation confusion matrix.

SIMILAR APPLICATIONS

1. **Facial recognition system** is a technology capable of matching a human face from a digital image or a video frame against a database of faces, typically employed to authenticate users through ID verification services, works by pinpointing and measuring facial features from a given image.

2. **Object recognition** is a computer vision technique for identifying objects in images or videos. Object recognition is a key output of deep learning and machine learning algorithms. When humans look at a photograph or watch a video, we can readily spot people, objects, scenes, and visual details. The goal is to teach a computer to do what comes naturally to humans: to gain a level of understanding of what an image contains

3. **Defeating Captcha anti-bot systems** though these are specifically designed to prevent OCR. The purpose can also be to test the robustness of CAPTCHA anti-bot systems.
4. Writing the instructions for vehicles by identifying CAD images in a database that are appropriate to the vehicle design as it changes in real time.
5. In airports, for passport recognition and information extraction
6. Automatic number plate recognition
7. Data entry for business documents, Cheque, passport, invoice, bank statement and receipt
8. Making scanned documents searchable by converting them to searchable PDFs

HANDWRITTEN LITERATURE REVIEW

- https://drive.google.com/file/d/1jR9Da2PiI_PC4d24U8GYifUrbYK9UNl-/view?usp=sharing

DATASET EMPLOYED

We have used the dataset which have mentioned in the reference of the projects which is

- https://www.kaggle.com/crawford/emnist

ALGORITHMS & RESULT

We have used two types of algorithms in this project:

- Decision tree

- A decision tree is **a flowchart-like structure in which each internal node represents a "test" on an attribute** (whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes)

- Random Forest

- Random forests or random decision forests are **an ensemble learning method for classification, regression and other tasks** that operates by constructing a multitude of decision trees at training time. ... Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees.

We give the dataset x_train and y_train, both algorithms made to fit this trains set and made a validation then predicted the results through the matplotlib

DEVELOPMENT PLATFORM

This project was devolved using Anaconda platform & Jupyter notebook editor.

It was written in python language (3.9)

The Main libraries used in this project were the following:

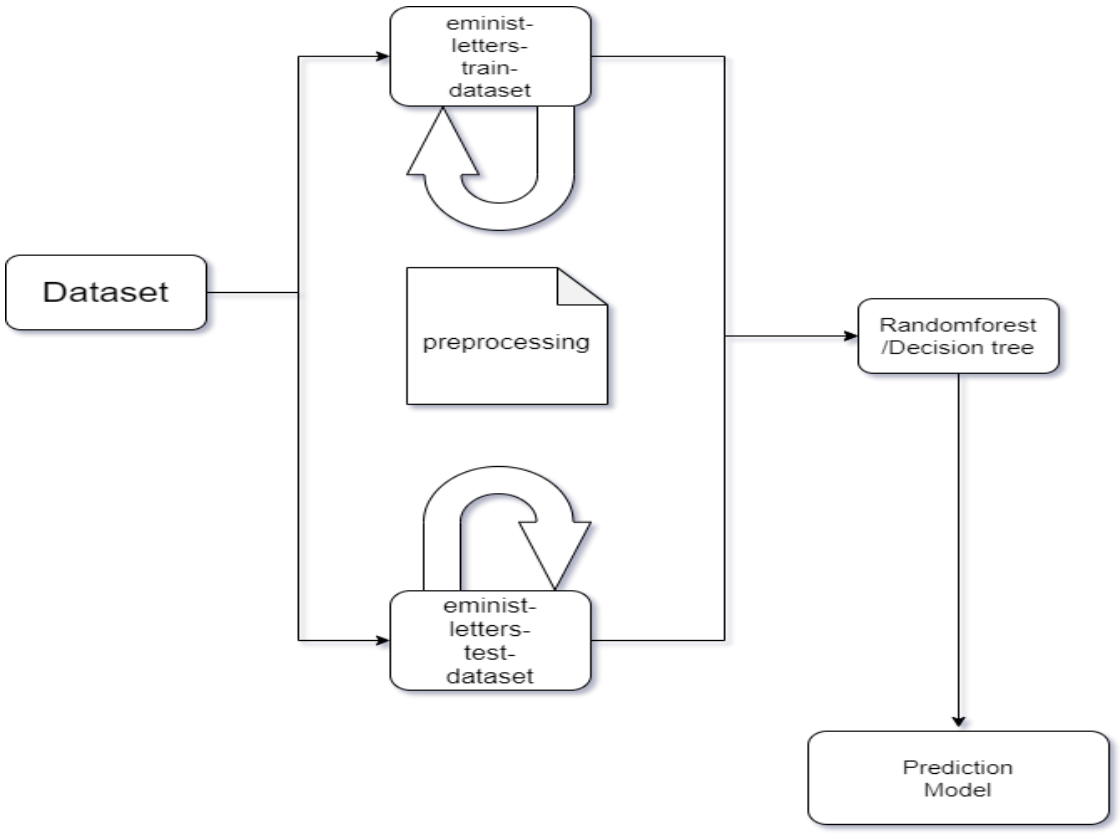
Scikit-learn --- Pandas ---- Numpy

We have used the dataset which have mentioned in the reference of the projects which is

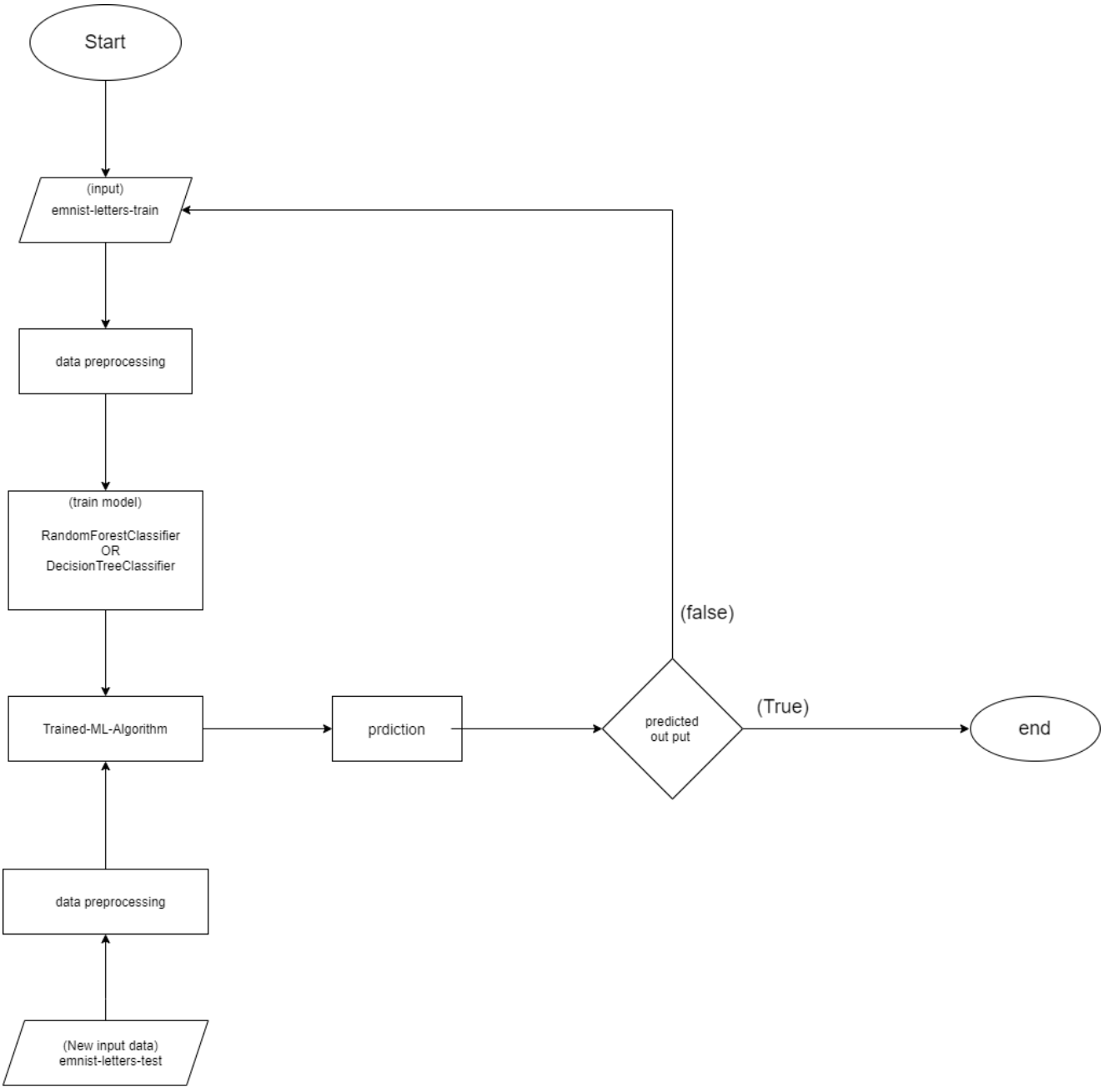
- https://www.kaggle.com/crawford/emnist

DIAGRAMS USED TO ILLUSTRATE THE PROJECT

1) **Block Diagram**



2) **Flow Chart**



Shared Link for Project Material:

https://bit.ly/3qpx9aE

Name (in Arabic)	ID	Level	Major \ Minor
عبدالرحمن خالد طاهر السيد	201900416	Three	Cs \ Is
اسلام رضا محمد عبد القادر	201900140	Three	Cs \ Is
عبدالكريم أنور أحمد محمد	201900451	Three	Cs \ Is
كريم وليد سيد زكي	201900585	Three	Is / Cs
محمد طارق محمد محمد	201900696	Three	Cs \ Is
عبدالرحمن نبيل محمد محمود	201900442	Three	Cs \ Is
محمود عماد عبدالموجود إسماعيل	201900778	Three	Cs \ Is