Mini Project: Supervised Machine Learning

the DEVELOPER Academy

Brief

Select a dataset of your choice and create a supervised machine learning algorithm to provide further insight into the data.

Requirements

Description in general with detailed list of requirements.

1. Start with some exploration of the data – what are the fields, how are they distributed, will they require any transformations?
2. Prepare the dataset for machine learning
3. Create an supervised machine learning model
4. Measure the accuracy of your model

Deliverables

You should submit your code along with some analysis. Your analysis can be presented as presentation slides, or using something like Jupyter Notebooks.

Data

There are three datasets for you to choose from. Pick the one you find most interesting or have an idea for how machine learning could be applied. I have given you some examples of the kind of analysis you could do – but feel free to come up with your own!

Dataset 1 – Fraud Detection

This is a dataset of bank transactions which have been classified as fraudulent or not.

Example insight: Can you create a machine learning model to accurately predict whether a transaction is fraudulent or not?

File name: fraud_detection.csv
Find out more about this dataset here: https://www.kaggle.com/ntnu-testimon/paysim1

Dataset 2 – Poisonous Mushrooms

This dataset contains information on a number of mushrooms, as well as a label indicating if they are poisonous or not.

Example insight: Can you use the mushrooms characteristics to determine if the mushroom is poisonous or not?

File name: mushrooms.csv
Find out more about this dataset here: https://www.kaggle.com/uciml/mushroom-classification

Dataset 3 – Fake News

This dataset contains information on a number of social media posts and classifies them as real or fake news.

Example insight: Can you build a machine learning model to detect which posts are likely to be fake news?

File name: news_articles.csv
Find out more about this dataset here:
https://www.kaggle.com/ruchi798/source-based-news-classification

Extension

To extend the project, try using two different supervised learning algorithms and use methods to measure their accuracy. Which model would you choose? Why?