A blue and green logo

Description automatically generated

Machine Learning Report

Candidate Number: 276270

**1. Features and Labels**

**1.1 FDI MIllıons of USD:**

The reason why the feature is used foreign investment amounts can indicate a country's economic attractiveness and the soundness of its investment environment. That’s why, total fdi values are taken

**1.2 Crop\_Yield\_Per\_100h:**

The whole item was grouped. That’s why, the data is grouped yearly for each country so that it is obtained annual data for each country involved in the dataset. The reason why I picked the group is we care about crop export value. It means that when an increase in the amount of crop export can trigger a number of crop yields. More crop product export means more crop products.

**1.3 Cluster:**

The reason why the cluster feature is used to classify countries for certain criteria is because the model must distinguish the difference between the United States and America due to different economic, and socioeconomic circumstances. First, FDI (Foreign Direct Investment) and Product Productivity data packages are loaded and examined. FDI data are converted into annual total investment amounts. Average values for product efficiency are calculated by using “FDI” and “crop data” combined. The data is prepared for cleaning by K-Means. These operations include organizing data (using Standard Scaler) and selecting features (FDI and product costs). K-Means applications are applied for the number of clusters recorded, which is picked as “3” in the dataset. Then, each data point is assigned to a cluster.

**1.4 Cluster\_0, Cluster\_1, Cluster\_2:**

Cluster classification cannot be directly used for the MLP model. That’s why, It should be applied one hot encoding transformation to make the data to be able to used in the MLP model, which is the only way to tell the machine existence of different classes on the dataset

**1.5 USD Exchange rate:**

exchange\_rate\_data['USD Exchange Rate'] = 1 / exchange\_rate\_data['Value']

Exchange data holds exchange values in their currency for each month and country. In the beginning, each currency is divided to United States of America’s currency so that the value obtained is expressed in one currency. Then, the currency expressed for each month and each country is converted to annual value by calculating their mean value for all months so that each country has its own their currencies by expressing them in terms of USD.Inflation\_High Inflation, Inflation\_Low Inflation, Inflation\_Medium Inflation:

Annual CPI change is grouped yearly for each month and country by calculating them their mean. Then, it is converted to annual change by looking previous year in percent. It is also a special function labeling inflation circumstances of countries. When the function applies to these changes. Labeling outcomes are going to be obtained yearly. Then, one hot encoding transformation is going to be applied to convert them to Boolean type variables because the labeling outcome has nonnumerical outputs. These features is used because of the same reason for cluster\_0, cluster\_1, and cluster\_2.

**1.6 Total Exported Value (Crops):**

It is the variable that will be guessed. In the first, it is determined which items are crop products. Then, items that are crop products are grouped and summed.

**1.7 Fertilizer Use(t):**

Fertilizers are grouped according to the “Fertilizer Use” category. Then, the items which are in the category are summed so that their yearly values are obtained for “Agriculture Use”. The reason why the variable is picked is it there might be a correlation between total exported crops and fertilizer use since when the amount of used fertilizer rises, the total export can have a potential to increase, as well.

**1.8 Agriculture Land (1000 ha):**

“Agriculture Land” is chosen because it has a more general scope than other lands used in the dataset. It is directly used in its yearly value on the main dataset. Agriculture Land can provide characteristics, which means it might be helpful for classifying countries.

**1.9 Agricultural Index Intensity (LAND):**

merged\_agricultural\_land\_fertilizers['Agricultural IndexIntensity']=merged\_agricultural\_land\_fertilizers[Fertilizer Use (t)] / merged\_agricultural\_land\_fertilizers[‘Agricultural Land (1000 ha)’] \* 1000

The calculation mentioned above will provide intensity which will give each country a characteristic so that MLP’s awareness in different types of countries will increase.

**1.10Crop Export Quantity:**

In the correlation analysis, there is a considerable connection between total exported value(crops) and crop export quantity. That’s why, it is a category that needs to be involved in the dataset. First of all, it is picked both export quantity and crop products in order to obtain these features after it is grouped yearly for each country.

**1.11 All Export Quantity:**

The same is done in the crop export quantity. The only difference is there is no filter for foods that are exported.

**1.12 Agriculture Intensity Index(TRADE):**

It is also used to classify countries. The feature is obtained by dividing crop export quantity and all export quantity by each other. Then, it will provide the intensity value for crop production trade.

**1.13 Employment Intensity in Export**

Country-by-country analysis comparing the production rates in the agricultural sector of different countries can provide a better understanding of production capacity and export potential.

**1.4 Relative Year**

main\_ml\_dataset['Relative\_Year'] = main\_ml\_dataset['Year'] – 2000

2000 is accepted as a starting point. That’s why, it is mapped as “0” according to the above calculation. Then, it will be used to tell year differences to the machine learning model.

**1.15 Present State Flags**

FDI\_Million\_USD\_Present

Crop\_Yield\_Per\_100ha\_Present

Cluster\_Present

Exchange Value not USD\_Present

Inflation\_High Inflation\_Present

Inflation\_Low Inflation\_Present

Inflation\_Medium Inflation\_Present

Total Exported Value(Crops)\_Present

Fertilizer Use (t)\_Present

Agricultural Land (1000 ha)\_Present

Agricultural Index Intensity(LAND)\_Present

Crop Export Quantity\_Present

All Products Export Quantity\_Present

Agriculture Intensity Index(TRADE)\_Present

Employment Intensity for Export\_Present

Relative\_Year\_Present

Present state flags are used to tell model absence or existence in the dataset even if these data are filled somehow. The model will know if it is an “actual” or “not actual” value. Then, the machine will have more awareness between “filled values” and “non-filled values”. The features with “0” states will have different weights than features with “1” states. The flag is used for countries with no data for features. Even if a country with partial feature, it will not be the scope of the present state flag. Before the flag process, all data was already filled with “backfill” and “forward fill” operations.

**2. Data Preprocessing**

**2.1 Filling Dataset and Determine Test and Train Datasets**

Data preprocessing started before using flags on the main dataset. When all features are brought together except for Flags. There a plenty of NaN values. The main reason is while some countries have full data in certain features, they might be also lacking in certain years and features. Especially, you can come across certain countries that do not have any single data for some features. That’s why, data processing is a must to fill them all. Data are filled in two stages. In the beginning, Countries with partial missing data at the feature are filled. Filling is done by using the backfill and forward features of pandas. Forward fill was used to fill future values by getting the latest value. Backfill was used to fill past values by getting the earliest value. The approach helps to fill countries’ features with partial missing values. The filling was done according to the first letter of countries. After the backfill and forward-fill process is completed for countries with partial missing features. In the second step, countries with full missing data for the features were detected. Then, features with categorical variables were filled general mod of the dataset. Features with numerical variables were filled with general mean for the interested feature. After features were filled in two steps, was obtained the dataset with full data. The main reason why all missing features were filled was that “MLP” could not work with missing data. Before creating the “MLP” model, it needs to be created “train” and “test” data. Normally, train and test data are created according to the 80:20 ratio. When it comes to time series, it is different. You need to decide forecasted year because the past data is going to be trained and future data is going to be tested depending on the year you are going to forecast. When it comes to my dataset, I made up my mind to pick 2000-2018 for the training dataset. In addition to that, my test dataset was between 2019-2022. The main reason why I picked these years was as train and test was to catch an 80:20 ratio.1.2

Standardize Featurescolumns\_to\_standardize = [

'USD Exchange Rate', 'Crop\_Yield\_Per\_100ha', 'Fertilizer Use (t)', 'Relative\_Year',

'Employment Intensity for Export', 'Agricultural Index Intensity(LAND)',

'Agriculture Intensity Index(TRADE)', 'FDI\_Million\_USD', 'Agricultural Land (1000 ha)',

'Total Exported Value(Crops)', 'All Products Export Quantity', 'Crop Export Quantity'

]

The columns shown above is standartized because all of those columns had numerical values. That’s why, I was selective about what I implement standart scale operation. Flag and dummy variable transformation features was not scope for standart scale operation since “MLP” can use these features directly in the model training.

**3. MLP Model**

**3.1 Model Definition**

The model starts with an embedding layer in order to process the features of different countries. After the layer, there are three layers which is fully connected to each other. Each layer processes the information it receives from the previous layer with a higher level of abstraction. This allows the model to learn more complex patterns and relationships in theory. In addition to that, “RELU” is also used as an activation function in the model. The main reason why ReLU breaks this linearity, allow the model to learn more complex patterns. The model has also 3 output features forecasting “2019”, “2020”, and “2021”. Also, In regression problems, the activation function is not used in the output layer because the predicted values must be in a continuous numerical range. In my case, the value I forecasted was numerical. That’s why, I did not use the activation function at the output layer. Otherwise, the forecasted value would be limited to a certain range. Then, predictions could be affected in an adverse way because all predicted values will stuck into a certain area, which means the model’s generalization ability will be affected.

**“ f(x)=x”**

The formula expresses linear activation because I did not use any activation function at the output. MSE(mean squared error) is used as a loss function, which is expressed.A number and mathematical symbols

Description automatically generated

While yi represents actual value, yi^ shows predicted outputs of the MLP model, which will provide a feedback about accuracy of the model. Units represents number of neurons on model. High number of neurons brings overfitting risk despite the model can potential to learn complex structures on the dataset, which allows the model to learn the most important features in earlier layers and to direct those features to more specific tasks in later layers. That’s why, tapering approach is used for my MLP model.

FC1 128 Units

FC2 64 Units

Output Layer 3 Units

Embedding Layer

It is mentioned above, that the Embedding Layer holds countries' information. These units are related to how the model is going to learn the dataset. Output Layers is given three because it will obtain information of “2019”, “2020”, and “2021” years. 128 Units is used to support the MLP model to learn more complex patterns on the dataset. When it comes to 64 units, it does not focus on complex patterns and applies the filter basically.

**3.2 Steps preventing overfitting**

**Embedding Layer:**

An embedding layer was used in the model to capture cross-country variation. This allows the model to better learn different country characteristics, which increases its generalization and classification ability.

**RELU:**

ReLU allows the model to learn non-linear features. It also helps to minimize gradient loss, as well. These things are important factors when it comes to deep learning.

**Early Stopping:**

In the course of, the loss value is printed at the end of every 10 epochs. This is used to monitor the performance of training and prevent overfitting by stopping it early if necessary.

**Regularization:**

Adam optimizer is used to improve the learning process and abstain from overfitting.

**4.Performance Evaluation**

**4.1 Model Performance Metrics**

The primary in order to evaluate performance is mean squared error which is basically the average squared difference between the estimated values and the actual value.

* A number and mathematical symbols

  Description automatically generated*n* is the number of samples,
* 𝑌𝑖*Yi*​ are the actual values,
* 𝑌^𝑖*Y*^*i*​ are the forecasted values by the model.

Total mean squared error is calculated by getting the difference “2019”, “2020” and “2021”.

**Scaled Value Results:**

Epoch 1/5, Train Loss: 0.867865392539118

Epoch 2/5, Train Loss: 0.21240895647289498

Epoch 3/5, Train Loss: 0.10723949719353446

Epoch 4/5, Train Loss: 0.08721397283287453

Epoch 5/5,

Train Loss: 0.061148758112852065

A graph with a line

Description automatically generated

Mean Squared Error on Train Set: 0.06607449387333222

When the number of epochs increases, MSE shows a decreased tendency, which means the model adapts better and finds better weight and parameters. The graph and obtained mean squared error scores prove it.

**A graph with a line

Description automatically generated**

My test loses also shows the model struggled while learn the test date. Despite

number of epochs is going to raise, the loss increase, as well. Then, it shows a

reduce tendency after the model understand the complex pattern, which is

completely expected. The reason why test data is higher error is train model with

less data and quality of data might can be. All in all, after the number of epochs raise

for both and train, it is a indicator of model can learn patterns on the dataset, which

is a good sign for the model.

**4.2 Data split and the total number of instances and their ratio**

The dataset is originally split into training and testing sets based on the 'Year' attribute. The training dataset involves past data between 2000 and 2018. This ensures that the model will learn to predict future values based on historical data. On the other hand, test data holds information of future data, between 2019 and 2022. It is used so as to evaluate how well the model has learned and can generalize to new, which is not seen data

.Train instance count:3562

Test instace count: :755

Total instance count: 4317

Amount of instances is also given. If you check the number of instances, as you can see that I almost caught an 80:20 ratio according to obtained instance outcomes, which is the most optimum data chosen in terms of dividing years.