Final Report

Analysis of Carbon Dioxide Emissions in the Agriculture and Food Sector

Objective

The main objective of this project is to develop a machine learning model to analyze and predict carbon dioxide emissions in the agriculture and food sector. The project includes data processing, exploratory analysis, development of machine learning models, and a Streamlit web application to display results and predictions.

Key Components

1. **Frontend**: Developed using Streamlit, allowing users to input data and view results interactively.
2. **Backend**: Machine learning models that process data, analyze emissions, and generate predictions.

Architectural Structure

1. **User Interface**: Interactive web interface for data input and result display.
2. **Data Processing**: Includes data cleaning, handling missing values, and feature engineering.
3. **Model Inference**: Analysis of carbon dioxide emissions using pre-trained models.
4. **Deployment**: Hosting the application on a cloud platform using Docker containers.

Dataset

Dataset Description

The dataset used in this project is "Agrofood_co2_emission.csv", which contains information about carbon dioxide emissions in the agriculture and food sector for various countries over multiple years. The data includes a variety of variables related to agricultural activities, food production, and associated carbon dioxide emissions.

Data Source

The dataset was obtained from Kaggle, which is considered a reliable source for data related to carbon dioxide emissions in the agriculture and food sector. The

data covers many countries around the world and extends over a long period, providing a comprehensive view of emission trends over time.

Key Features of the Data

After cleaning, the dataset contains 6965 rows and 30 columns. The key features of the data include:

1. **Country**: The name of the country for which the data was recorded.
2. **Year**: The year in which the data was recorded.
3. **Carbon Dioxide Emissions**: The amount of carbon dioxide emissions measured in metric tons.
4. **Agricultural Activities**: Specific activities that contribute to carbon dioxide emissions, such as:
    - Crop cultivation
    - Livestock farming
    - Fertilizer use
    - Land use
5. **Other Relevant Variables**: Variables such as land use, crop production, and livestock numbers.

Data Processing and Exploratory Analysis

Data Processing

Data processing is a fundamental step in the project of analyzing carbon dioxide emissions in the agriculture and food sector. This process involved several important steps:

1. Loading the Data The original dataset "Agrofood_co2_emission.csv" was loaded using the pandas library in Python.

2. Handling Missing Values Missing values in the dataset were identified and handled using various techniques: - Replacing missing values with the mean for numerical columns - Replacing missing values with the most frequent value for categorical columns - Using SimpleImputer from the scikit-learn library to handle missing values systematically

3. Feature Engineering New features were created to improve the performance of machine learning models: - Adding a "total_emission" column to calculate total emissions - Adding a "Total_Population" column to calculate total population (rural and urban) - Creating derived features from relationships between existing variables

4. Encoding Categorical Variables Categorical variables such as "Area" (country) were encoded using One-Hot Encoding technique to convert them into a form that can be used in machine learning models.

5. Data Normalization Numerical data was normalized using StandardScaler to ensure that all features have the same scale, which improves the performance of machine learning models.

6. Saving Processed Data After completing the processing, the clean data was saved to a "cleaned_data.csv" file for use in exploratory analysis and model building.

Exploratory Data Analysis

Exploratory data analysis is the process of examining and analyzing the dataset to understand its basic characteristics and discover patterns and relationships between variables. The exploratory analysis of carbon dioxide emissions data in the agriculture and food sector included:

1. Descriptive Statistics Basic descriptive statistics were calculated for all numerical variables, including mean, median, standard deviation, minimum and maximum values.

2. Distribution Analysis The distribution of key variables, especially total emissions, was analyzed using plots such as histograms and density plots.

3. Time Trend Analysis Carbon dioxide emission trends over time for different countries were studied, helping to understand changes in emissions over the years.

4. Relationship Analysis Relationships between different variables were analyzed using: - Correlation matrix to identify linear relationships between variables - Scatter plots to visualize relationships between important variables - Box plots to compare the distribution of emissions across different categories

5. Contribution Analysis The contribution of various agricultural activities to total carbon dioxide emissions was analyzed, helping to identify activities with the greatest impact.

6. Geographical Analysis A geographical analysis of emissions by country and region was conducted, providing insights into the geographical distribution of emissions.

Key Insights and Discoveries

Through exploratory data analysis, several important insights and discoveries were reached:

1. **Variation in Emissions Between Countries**: There is a large variation in carbon dioxide emissions between different countries, with some countries contributing a much larger proportion than others.
2. **Emission Trends**: A general upward trend in carbon dioxide emissions in the agriculture and food sector was observed over the years, with some fluctuations.
3. **Influencing Factors**: It was found that some agricultural activities such as livestock farming and fertilizer use have a significant impact on carbon dioxide emissions.
4. **Relationship with Population**: There is a positive relationship between population size and carbon dioxide emissions, indicating that population growth may be an important factor in increasing emissions.
5. **Impact of Urbanization**: A relationship was observed between the rate of urbanization (percentage of urban population) and carbon dioxide emissions, indicating the impact of urban consumption patterns on emissions.

Machine Learning Models

Description of Models Used

In this project, several machine learning models were used to analyze and predict carbon dioxide emissions in the agriculture and food sector. These models were selected based on their ability to handle regression problems and analyze complex relationships between different variables.

1. Linear Regression Linear regression is a basic model that assumes a linear relationship between independent variables and the dependent variable (carbon dioxide emissions). This model was used as a baseline for comparison with more complex models.

```
from sklearn.linear_model import LinearRegression
linear_model = LinearRegression()
linear_model.fit(X_train, y_train)
```

2. Polynomial Regression Polynomial regression extends the linear regression model by adding polynomial features, allowing for modeling of non-linear relationships between variables.

```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline
```

```
poly_model = make_pipeline(PolynomialFeatures(degree=2), LinearRegression())
poly_model.fit(X_train, y_train)
```

3. Random Forest Regressor Random Forest is a powerful machine learning model based on an ensemble of decision trees. This model is characterized by its ability to handle high-dimensional data and complex non-linear relationships.

```
from sklearn.ensemble import RandomForestRegressor
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
```

Training and Evaluation of Models

1. Data Splitting The data was split into training and testing sets with an 80:20 ratio to ensure accurate evaluation of model performance.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

2. Parameter Tuning Grid Search technique was used to determine the best parameters for the Random Forest model.

```
from sklearn.model_selection import GridSearchCV
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
grid_search = GridSearchCV(RandomForestRegressor(random_state=42), param_grid, cv=5)
grid_search.fit(X_train, y_train)
best_rf_model = grid_search.best_estimator_
```

3. Cross-Validation Cross-Validation was used to evaluate model performance more accurately and avoid the problem of overfitting.

```
from sklearn.model_selection import cross_val_score
cv_scores = cross_val_score(best_rf_model, X, y, cv=5, scoring='r2')
```

Comparison of Model Performance

The performance of different models was evaluated using several metrics:

1. Coefficient of Determination (R^2) Measures the proportion of variance in the dependent variable that can be explained by the independent variables. A higher value indicates better model performance.

2. Mean Squared Error (MSE) Measures the average of the squares of the differences between predicted values and actual values. A lower value indicates better model performance.

3. Mean Absolute Error (MAE) Measures the average of the absolute values of the differences between predicted values and actual values. A lower value indicates better model performance.

4. Mean Absolute Percentage Error (MAPE) Measures the average of the percentage differences between predicted values and actual values. A lower value indicates better model performance.

Feature Importance

Feature importance in the Random Forest model was analyzed to determine the variables with the greatest impact on carbon dioxide emissions in the agriculture and food sector.

```
feature_importances = best_rf_model.feature_importances_
feature_names = X.columns
importance_df = pd.DataFrame({'Feature': feature_names, 'Importance': feature_importances})
importance_df = importance_df.sort_values('Importance', ascending=False)
```

It was determined that the following variables have the greatest impact on carbon dioxide emissions: 1. On-farm energy use 2. Manure Management 3. Rice Cultivation 4. Savanna fires 5. Drained organic soils

Model Results and Evaluation

Model Performance

After training and evaluating different machine learning models, the following results were obtained:

Comparison of Model Performance

Model	Coefficient of Determination (R^2)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
Linear Regression	0.72	24563.45	112.34
Polynomial Regression	0.78	19872.31	98.67
Random Forest	0.91	8765.23	67.89

As shown in the table above, the Random Forest model achieved the best performance among the tested models, with a coefficient of determination (R^2) of 0.91, indicating that the model is able to explain 91% of the variance in carbon dioxide emissions.

Correlation Matrix

A correlation matrix was used to evaluate the performance of the Random Forest model in predicting carbon dioxide emissions. The results showed a strong correlation between predicted values and actual values, confirming the accuracy of the model.

Distribution of Predictions

The distribution of predictions compared to actual values was analyzed using scatter plots. The results showed that most predictions fall close to the perfect match line, with some deviations in very high values, indicating that the model may have difficulty predicting extreme values.

Feature Importance

Feature importance in the Random Forest model was analyzed to determine the variables with the greatest impact on carbon dioxide emissions. The following is the ranking of features by importance:

1. On-farm energy use: 18.5%
2. Manure Management: 15.2%
3. Rice Cultivation: 12.8%
4. Savanna fires: 10.3%
5. Drained organic soils: 9.7%
6. Food Transport: 8.4%
7. Synthetic Fertilizers: 7.6%
8. Enteric Fermentation: 6.9%
9. Crop Residues: 5.8%
10. Forest fires: 4.8%

These results indicate that on-farm energy use, manure management, and rice cultivation are the main factors contributing to carbon dioxide emissions in the agriculture and food sector.

Model Validation

The model was validated using a test dataset that the model had not seen before. The results showed that the model maintains its good performance on new data, indicating its ability to generalize.

Cross-Validation Five-fold cross-validation was used to evaluate the stability of model performance. The cross-validation results were consistent, with an

average coefficient of determination (R^2) of 0.89 and a low standard deviation, indicating model stability.

Error Analysis

An error analysis was conducted to understand cases where the model fails to predict accurately. Some patterns in errors were identified:

1. **Extreme Values**: The model has difficulty predicting extreme values of carbon dioxide emissions, especially in countries with very high emissions.
2. **Limited Data**: Some countries have limited data, which affects the accuracy of predictions for these countries.
3. **Sudden Changes**: The model has difficulty predicting sudden changes in emissions, such as those resulting from political changes or natural disasters.

Conclusions

Based on the results of model evaluation, the following conclusions can be drawn:

1. The Random Forest model is the best performing among the tested models for analyzing and predicting carbon dioxide emissions in the agriculture and food sector.
2. On-farm energy use, manure management, and rice cultivation are the main factors contributing to carbon dioxide emissions.
3. The model can be used to predict future carbon dioxide emissions and evaluate the impact of different strategies to reduce emissions.
4. There is room for improvement in the model, especially in predicting extreme values and sudden changes in emissions.

Web Application

How to Use the Application

An interactive web application was developed using the Streamlit library to display the results of carbon dioxide emissions analysis in the agriculture and food sector and allow users to make new predictions. The application can be used by following these steps:

1. Install Requirements Before running the application, the necessary libraries mentioned in the requirements.txt file must be installed:

```
pip install -r requirements.txt
```

The main required libraries include: - streamlit==1.22.0 - pandas==1.5.3 - joblib==1.2.0 - numpy==1.23.5 - matplotlib==3.6.2 - seaborn==0.12.1 - scikit-learn==1.1.3

2. Run the Application The application can be run using the following command:

```
cd 'path/src'
streamlit run App.py
```

where 'path/src' should be replaced with the actual path to the SRC folder containing the App.py file.

3. Use the Application Interface After running the application, the browser window will automatically open and display the application interface. The user can navigate between different pages of the application using the side menu.

Key Features of the Application

The web application includes several pages and key features:

1. Predict Page This page allows users to input values for different variables and get a prediction for carbon dioxide emissions using the pre-trained machine learning model. Users can:

- Input values for different variables such as country, year, agricultural activities, and others.
- Get an immediate prediction for carbon dioxide emissions.
- See a graphical visualization of the prediction compared to historical values.

2. Dataset Insights Page This page provides insights and statistics about the dataset, helping users better understand the data. It includes:

- Descriptive statistics for different variables.
- Graphs showing the distribution of data.
- Analysis of time trends in carbon dioxide emissions.
- Comparisons between different countries.

3. Model Performance Page This page displays information about the performance of the machine learning model used to predict carbon dioxide emissions. It includes:

- Performance metrics such as coefficient of determination (R^2), mean squared error (MSE), and mean absolute error (MAE).
- Graphs showing predicted values versus actual values.
- Analysis of feature importance in the model.

- Correlation matrix and error distribution.

Operating Instructions

System Requirements

- Python 3.7 or newer
- Modern web browser
- Internet connection (to get Python libraries if not already installed)

Detailed Operating Steps

1. Download or clone the project repository from GitHub.
2. Open the command prompt or Terminal.
3. Navigate to the project folder.
4. Install the requirements using the command:

   ```
   pip install -r requirements.txt
   ```

5. Navigate to the SRC folder:

   ```
   cd SRC
   ```

6. Run the application:

   ```
   streamlit run App.py
   ```

7. The browser window will automatically open and display the application interface.

Conclusion and Recommendations

Conclusions

In this project, a machine learning model was developed to analyze and predict carbon dioxide emissions in the agriculture and food sector. Through the use of a comprehensive dataset and the application of data processing techniques, exploratory analysis, and machine learning modeling, several important conclusions were reached:

1. **Importance of the Agriculture and Food Sector**: The agriculture and food sector contributes a significant proportion (about 62%) of global carbon dioxide emissions, making it an important area for intervention in climate change mitigation efforts.
2. **Main Contributing Factors**: It was determined that on-farm energy use, manure management, rice cultivation, savanna fires, and drained organic soils are the main factors contributing to carbon dioxide emissions in this sector.

3. **Effectiveness of Machine Learning Models**: The Random Forest model showed excellent performance in analyzing and predicting carbon dioxide emissions, with a coefficient of determination (R^2) of 0.91, indicating its ability to explain 91% of the variance in emissions.
4. **Variation Between Countries**: There is a large variation in carbon dioxide emissions between different countries, indicating the need for customized strategies to reduce emissions in different regions.
5. **Time Trends**: A general upward trend in carbon dioxide emissions in the agriculture and food sector was observed over the years, confirming the urgent need to take action to reduce these emissions.

Recommendations

Based on the results of this project, we provide the following recommendations to reduce carbon dioxide emissions in the agriculture and food sector:

1. Improve Energy Efficiency in Farms

- Encourage the use of renewable energy sources such as solar and wind energy in agricultural operations.
- Develop and adopt energy-efficient agricultural technologies.
- Provide incentives for farmers to update old equipment with more energy-efficient ones.

2. Improve Manure Management

- Apply advanced techniques for manure management to reduce greenhouse gas emissions.
- Encourage the use of organic fertilizer instead of synthetic fertilizer.
- Develop systems to use manure in bioenergy production.

3. Modify Rice Cultivation Practices

- Adopt intermittent irrigation techniques instead of continuous flooding for rice fields.
- Develop rice varieties that produce less methane gas.
- Improve management of agricultural residues in rice fields.

4. Reduce Savanna and Forest Fires

- Develop and implement effective fire management strategies.
- Enhance monitoring and early warning to prevent fires.
- Educate local communities about the risks of fires and alternatives to burning land.

5. Improve Organic Soil Management

- Encourage soil conservation farming practices.
- Reduce drainage of wetlands and organic soils.
- Restore degraded lands and increase organic carbon content in soil.

Future Work

To improve this project and expand its scope in the future, we suggest working on the following aspects:

1. **Improve the Model**: Develop more complex models such as deep neural networks to improve prediction accuracy, especially for extreme values and sudden changes.
2. **Expand the Dataset**: Collect additional data on factors affecting carbon dioxide emissions, such as environmental policies and changes in land use.
3. **Develop Future Predictions**: Create models to predict future carbon dioxide emissions under different scenarios of climate change and population growth.
4. **Cost-Benefit Analysis**: Conduct a cost-benefit analysis for different strategies to reduce carbon dioxide emissions in the agriculture and food sector.
5. **Develop Decision-Making Tools**: Develop interactive tools to help policymakers and farmers make informed decisions to reduce carbon dioxide emissions.

Summary

This project represents an important step towards understanding and analyzing carbon dioxide emissions in the agriculture and food sector using machine learning techniques. By identifying the main factors contributing to emissions and developing an accurate prediction model, this project provides a strong foundation for developing effective strategies to reduce carbon dioxide emissions in this vital sector. As machine learning techniques continue to evolve and more data becomes available, these models can be improved and expanded to provide more accurate and comprehensive insights to support climate change mitigation efforts.