A Global Analysis and Prediction of Food Insecurity and Hunger Crisis Outbreaks using Machine Learning

Bhavan K. Basavaraju, Namratha S. Kumar, Shreya Chikatmarla, Sowjanya Pamulapati

Project Background

- In today's world, hunger and food scarcity are on the rise as more people in the world face the challenge of securing a meal on a regular basis, leading to malnutrition and food deprivation.
- According to Tarasuk (2001), food insecurity is a fundamental human need that is unaffected by a person's gender, colour, or ethnicity.
- Hernandez et al. (2017) report that 27% of adults, aged 36 on average, suffer food insecurity. This emphasizes the need for immediate monitoring and eradication.
- Previous studies have restricted its scope and narrowed the focus to particular regions.
- This area of study intends to improve global food intake estimates by addressing shortcomings, including a broader scope that includes post-pandemic scenarios, building upon the work of Martini et al. (2022).

Executive Summary

- Understanding and addressing the global food issues, with a focus on malnutrition and food insecurity.
- Diverse data for effective algorithm designing and model development.
- Time-series biassed algorithms (LSTM, ARIMA, Random Forest, Prophet) are used for precise predictions.
- To evaluate algorithm performance and tweak as needed, metrics like as MAE, MSE, R2 Score, and Explained Variance Score are used.
- Integration of initial analysis, data visualisation, and a predictive model to provide an overview of the global hunger situation and its progression over time.

Motivation

- The project is motivated by the global prevalence of those battling for a daily meal.
- Tarasuk (2001) and Hernandez et al. (2017) underline the severity of food insecurity. Emphasizing the urgency and need for study and deliberate elimination of a food crisis.
- The shortcomings of current prediction models, particularly missing out on post-pandemic scenarios such as COVID-19, further motivate our project.
- To address the global nature of food insecurity, extending outside existing geographical boundaries.

Problem Statement

- Addressing global food hunger is a significant concern with consequences for both individuals and the nation.
- Current studies utilising machine learning approaches exhibit restrictions in terms of using diverse algorithms, efficiency in data procurement, and region-specific analysis.
- Our proposed approach intends to overcome these gaps by incorporating unique algorithms and using diversified data sources, with the objective of enhancing the precision of forecasts while offering actionable insights for successful mitigation strategies.

Deliverables

- Project Proposal
- Project Management
- Project Effort Estimates
- Data collection/apprehension
- Modelling/Prototype
- Evaluations
- Deployment
- Project Final Report

Project Requirements

- Addressing the hunger crisis and food insecurity.
- Identifying factors and inclusion of post-pandemic data
- Diverse dataset collection: census (2010-2021), RASFF (2000-2023), OECD survey(1970-2022), IMF trade data (1980-present)
- Time-series biased algorithm selection: LSTM, ARIMA, Prophet, and Random Forest.
- Algorithmic workflow: preprocessing, splitting, configuration, training, and evaluation.
- Evaluation metrics: MSE, MAE, R2 Score, Explained Variance Score.
- Comparative study to determine the best algorithm performance.
- Organizing project workflow and task division for optimal collaboration.

Technology & Literature Survey

PAPER TITLE	AUTHORS	DATASET	MODELS	RESULTS
Food security and agricultural challenges in West-African rural communities: a machine learning analysis.	Ahn et al. (2022)	Food Security and Production Assessment of Ghana, Senegal, and Liberia	Random Forest Algorithm, Decision tree and Chi-Square Automatic Interaction Detection (CHAID)	The decision tree achieved an accuracy of 73.7% Classification for Senegal using decision trees was more precise than for Liberia and the misclassification rate was 20%. Random forest employed for Ghana had a 9% misclassification rate and 0.01 standard error indicating efficient model performance and precision.
Forecasting transitions in the state of food security with machine learning using transferable features	Westerveld et al. (2021)	food security transitions on a monthly basis in Ethiopia	“Extreme Gradient Boosting”, “Random Forest”, and “CatBoost”	Welch based T-test between the three algorithms is performed using normal distribution. It was found that with a t (198) =19.86 and t (198) =86.33 with p < 0.001 across comparison of F1 scores between XGboost and Catboost, XGBoost and Random Forest respectively.
Food security prediction from heterogeneous data combining machine and deep learning methods	Deléglise et al. (2020)	The datasets used in the study included time series data, meteorological data, population density data, World Bank economic data, and the Normalized Difference Vegetation	Random Forest (RF), Convolutional Neural Networks (CNNs), and Long-term and Short-term Memory (LSTM) models	The framework outperforms all competing methods, with model (b) surpassing model (a). The $\hat{R}2$ values obtained for (0.469) and (0.434) are statistically significant, demonstrating the benefits of integrating various data science techniques. The WFP framework's results in Burkina Faso are relatively modest (0.34 for and 0.30 for), while Lentz et al.'s study shows even lower $\hat{R}2$ values below 0.2
Modeling and Forecasting of Food Security for Wheat in Egypt Through Year 2025 by Using time Series ARIMA Models	Negm et al. (2018)	Data on wheat production, consumption, and imports in Egypt obtained from sources	ARIMA	wheat consumption in Egypt is projected to increase by 18.54 million tons based on the ARIMA (0, 0, 1) model

Project Resource Requirements

Hardware Requirements

Software Requirements

Tools and Licenses

Hardware	Memory	Configuration	Purpose
Apple M2, macOS Sonoma Version 14.0	16 GB	494.38 GB	ML Model
Apple M1 Pro, macOS Ventura Version 13.5.1	16 GB	494.38 GB	ML Model
Apple M2, macOS Ventura Version 13.0	16 GB	994.66 GB	Computation engine for running ML Models in Jupyter Notebook

Libraries/Packages	Purpose	Version
Python	Data Cleaning Data Preparation Data Analysis	3.8
Pandas	Data Manipulation and analysis Data Structures	1.3.2
Numpy	Fundamental Computing for Numerical data	3.2
Matplotlib	Data Visualization using plots and charts	3.2
Seaborn	Informative Statistical Graphics for Data Exploration	0.12.0
sci-kit-learn	Machine learning Model building	1.1.2

Tools	Purpose	License
Jupyter NoteBook	Code Design and Development	Open Source
IntelliJ	Software Development	Open Source
GitHub	Create and storing our Project	Open Source
Draw.io	Aiding Design	Free
Zoom	Team Collaboration meeting	Free
Google Meet	Team Collaboration meetings	Free
Google Docs	Documentation	Free
Google Drive	Data Storage	Free

Project Management Tools & WBS

WBS

WBS is a Project management tool for hierarchical structure based on scope assessment

It breaks down complex deliverables into simpler, manageable tasks and defines project objectives effectively.

Establishes milestones and divides deliverables into work packages and identifies dependencies for successful project delivery.

Integration of CRISP-DM methodology enhances work breakdown structure.

Gantt Chart

They are essential project management tools that visually represent tasks, schedules, and dependencies of a project.

They provide a clear overview of the project schedule, help identify milestones and due dates, facilitate communication and collaboration among team members , and allow for tracking task progress

Creating a Gantt chart involves listing tasks, determining task durations and dependencies, and using software tools such as WBS - Gantt chart plugin in JIRA.

In our project, we follow the structure of Epics, User stories, Tasks, and, Sub Tasks which are all represented on the Gantt chart to track progress and ensure timely completion.

PERT CHART

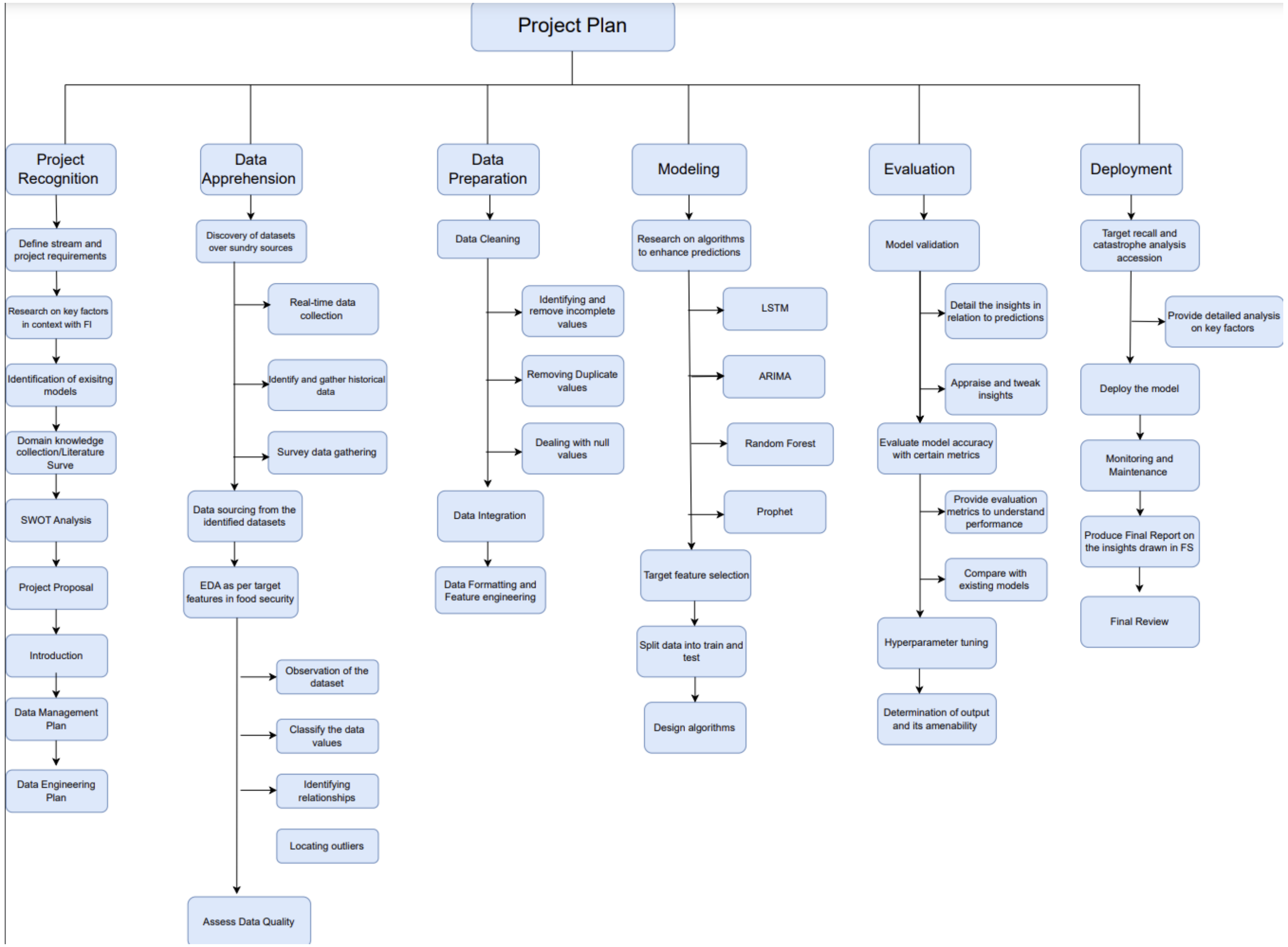
PERT is a project management tool for handling complex projects effectively and provides detailed explanations and addresses project complexities.

They help to determine task order, estimate task durations, identify critical paths, and manage project risks.

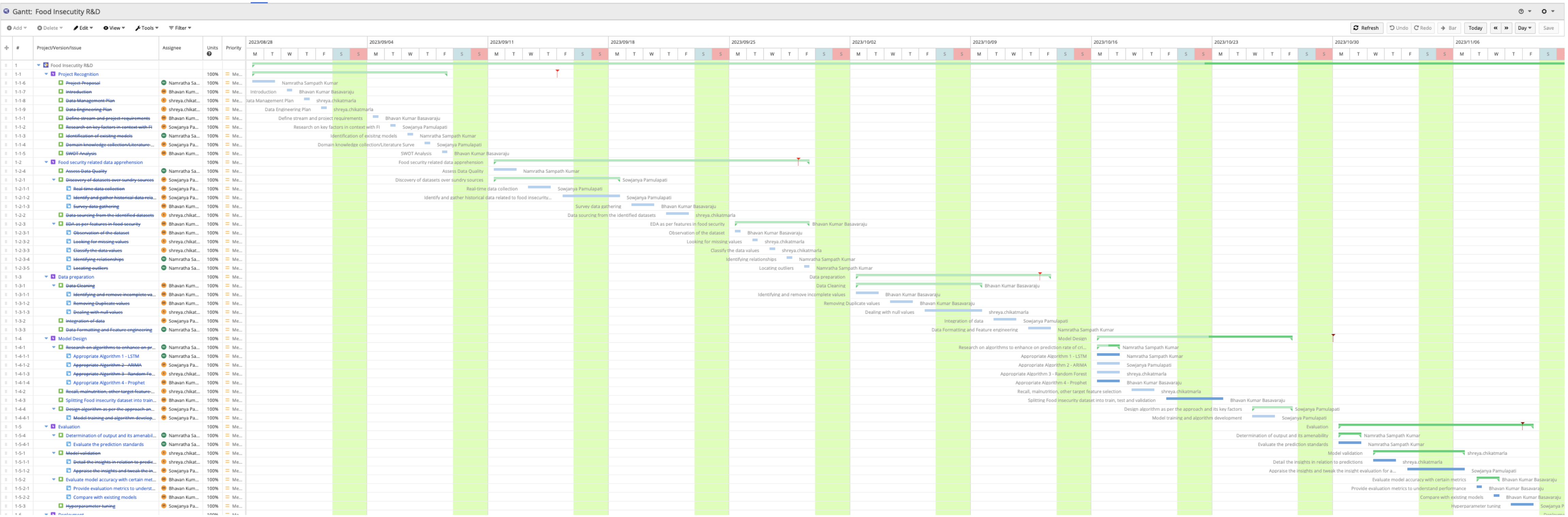
Critical path analysis is crucial in identifying tasks that can impact the overall project timeline.

They have limitations and can be challenging for large projects, but they remain a valuable tool for handling complex and risky projects.

Work Breakdown Structure

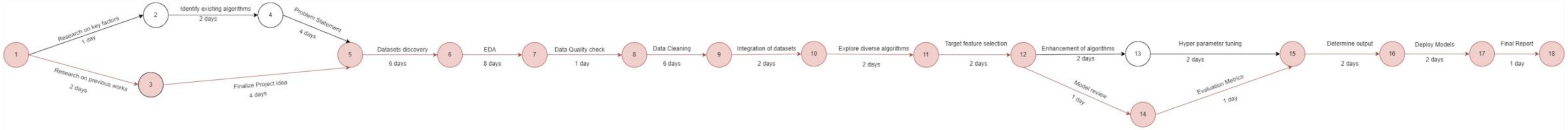


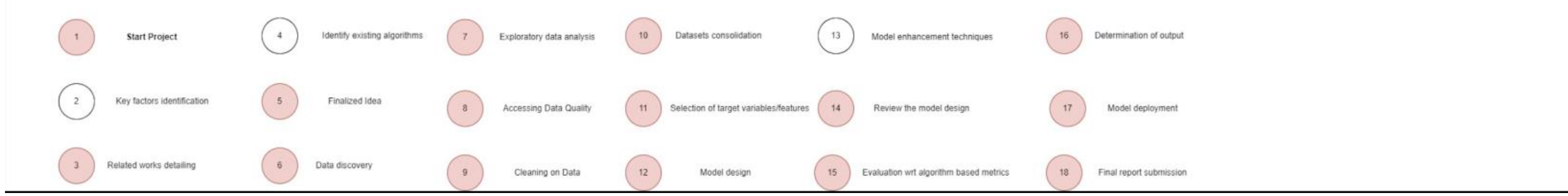
Gantt Chart



Gantt: Food Insecutity R&D
Add
Delete
Edit
View
Tools
Filter
Refresh
Undo
Redo
Bar
Today
Day
Save
Project/Version/Issue Assignee Units Priority
2023/08/28
2023/09/04
2023/09/11
2023/09/18
2023/09/25
2023/10/02
2023/10/09
2023/10/16
2023/10/23
2023/10/30
2023/11/06
1 Food Insecutity R&D
1-1 Project Recognition 100% Me...
1-1-6 ~~Project Proposal~~ Namratha Sa... 100% Me...
1-1-7 ~~Introduction~~ Bhavan Kum... 100% Me...
1-1-8 ~~Data Management Plan~~ shreya.chikat... 100% Me...
1-1-9 ~~Data Engineering Plan~~ shreya.chikat... 100% Me...
1-1-1 ~~Define stream and project requirements~~ Bhavan Kum... 100% Me...
1-1-2 ~~Research on key factors in context with FI~~ Sowjanya Pa... 100% Me...
1-1-3 ~~Identification of exisitng models~~ Namratha Sa... 100% Me...
1-1-4 ~~Domain knowledge collection/Literature...~~ Sowjanya Pa... 100% Me...
1-1-5 ~~SWOT Analysis~~ Bhavan Kum... 100% Me...
1-2 Food security related data apprehension 100% Me...
1-2-4 ~~Assess Data Quality~~ Namratha Sa... 100% Me...
1-2-1 ~~Discovery of datasets over sundry sources~~ Sowjanya Pa... 100% Me...
1-2-1-1 ~~Real-time data collection~~ Sowjanya Pa... 100% Me...
1-2-1-2 ~~Identify and gather historical data rela...~~ Sowjanya Pa... 100% Me...
1-2-1-3 ~~Survey data gathering~~ Bhavan Kum... 100% Me...
1-2-2 ~~Data sourcing from the identified datasets~~ shreya.chikat... 100% Me...
1-2-3 ~~EDA as per features in food security~~ Bhavan Kum... 100% Me...
1-2-3-1 ~~Observation of the dataset~~ Bhavan Kum... 100% Me...
1-2-3-2 ~~Looking for missing values~~ shreya.chikat... 100% Me...
1-2-3-3 ~~Classify the data values~~ shreya.chikat... 100% Me...
1-2-3-4 ~~Identifying relationships~~ Namratha Sa... 100% Me...
1-2-3-5 ~~Locating outliers~~ Namratha Sa... 100% Me...
1-3 Data preparation 100% Me...
1-3-1 ~~Data Cleaning~~ Bhavan Kum... 100% Me...
1-3-1-1 ~~Identifying and remove incomplete va...~~ Bhavan Kum... 100% Me...
1-3-1-2 ~~Removing Duplicate values~~ Bhavan Kum... 100% Me...
1-3-1-3 ~~Dealing with null values~~ shreya.chikat... 100% Me...
1-3-2 ~~Integration of data~~ Sowjanya Pa... 100% Me...
1-3-3 ~~Data Formatting and Feature engineering~~ Namratha Sa... 100% Me...
1-4 Model Design 100% Me...
1-4-1 ~~Research on algorithms to enhance on pr...~~ Namratha Sa... 100% Me...
1-4-1-1 Appropriate Algorithm 1 - LSTM Namratha Sa... 100% Me...
1-4-1-2 ~~Appropriate Algorithm 2 - ARIMA~~ Sowjanya Pa... 100% Me...
1-4-1-3 ~~Appropriate Algorithm 3 - Random Fo...~~ shreya.chikat... 100% Me...
1-4-1-4 Appropriate Algorithm 4 - Prophet Bhavan Kum... 100% Me...
1-4-2 ~~Recall, malnutrition, other target feature...~~ shreya.chikat... 100% Me...
1-4-3 Splitting Food insecurity dataset into train... Bhavan Kum... 100% Me...
1-4-4 ~~Design algorithm as per the approach an...~~ Sowjanya Pa... 100% Me...
1-4-4-1 ~~Model training and algorithm develop...~~ Sowjanya Pa... 100% Me...
1-5 Evaluation 100% Me...
1-5-4 Determination of output and its amenabil... Namratha Sa... 100% Me...
1-5-4-1 Evaluate the prediction standards Namratha Sa... 100% Me...
1-5-1 ~~Model validation~~ shreya.chikat... 100% Me...
1-5-1-1 ~~Detail the insights in relation to predic...~~ shreya.chikat... 100% Me...
1-5-1-2 ~~Appraise the insights and tweak the in...~~ Sowjanya Pa... 100% Me...
1-5-2 Evaluate model accuracy with certain met... Bhavan Kum... 100% Me...
1-5-2-1 Provide evaluation metrics to underst... Bhavan Kum... 100% Me...
1-5-2-2 Compare with existing models Bhavan Kum... 100% Me...
1-5-3 ~~Hyperparameter tuning~~ Sowjanya Pa... 100% Me...

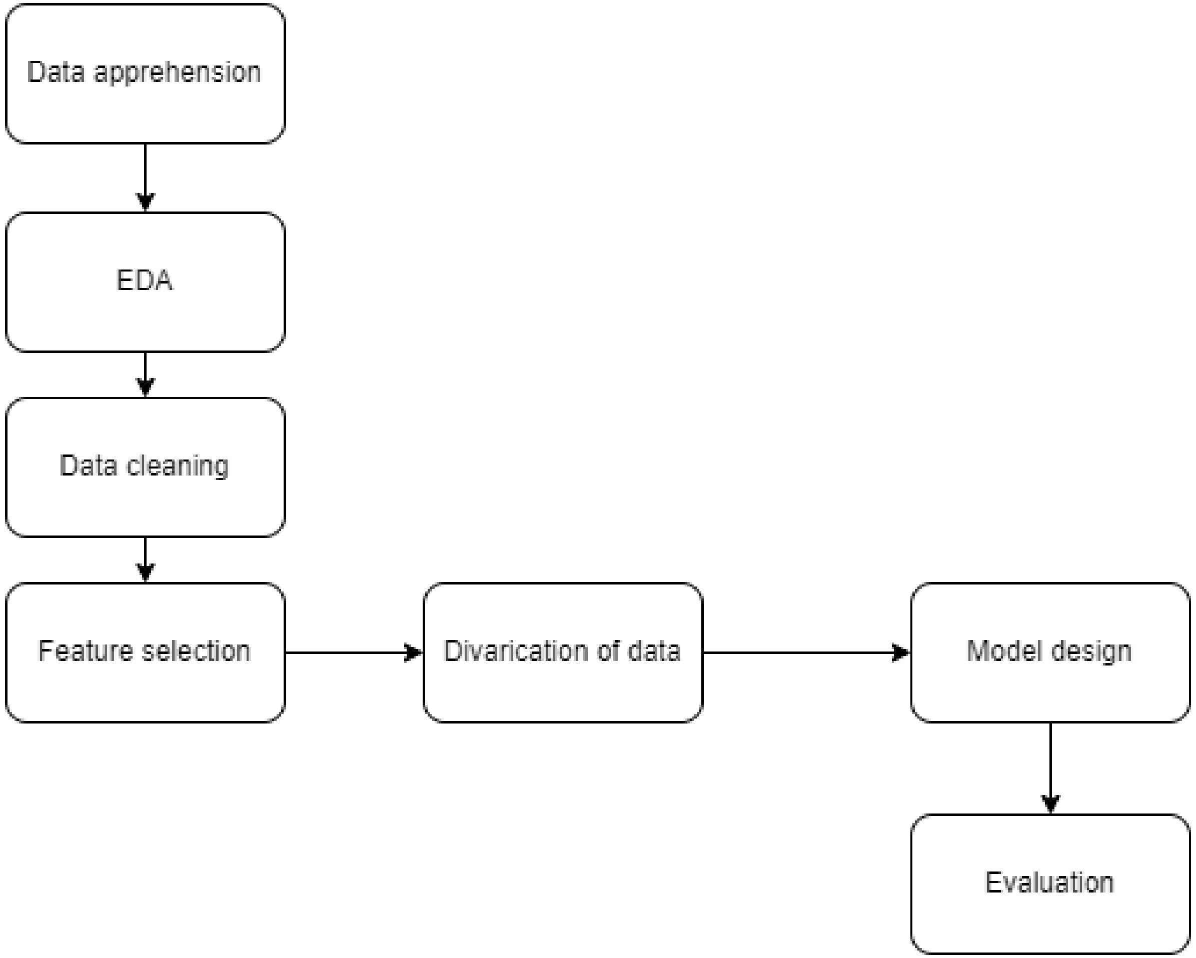
PERT CHART





Activity	Description	Predecessor/Dependency
1	Start Project	_
2	Key factors identification	_
3	Related works detailing	_
4	Identify existing algorithms	_
5	Finalized Idea	1,3
6	Data discovery	5
7	Exploratory data analysis	6
8	Accessing Data Quality	7
9	Cleaning on Data	8
10	Datasets consolidation	9
11	Selection of target variables/features	10
12	Model Design	11
13	Model Enhancement Techniques	12
14	Review the Model Design	12
15	Evaluation wrt algorithm based metrics	12,14
16	Determination of output	15
17	Model deployment	16
18	Final report submission	17

Project Workflow



Data Collection

- Census information on food security(2010 to 2021)

  Datatypes - int64, float64

  Number of columns - 507

  Data Size - 485.4 MB

- RASSF(Ras association domain family) dataset EC (Europe Commision) data from 2017to present)

  Datatypes - object, float64

  Number of columns - 14

  Data Size - 546.6+ KB

- Gross domestic product (GDP) survey data - (OECD sourced survey data from 1970 to 2022)

  Datatypes - int64, float64, object

  Number of columns - 23

  Data Size - 12.0+ MB

- Commodity Terms of Trade dataset (IMF sourced circadian data from 1980 to present)

  Datatypes- int64, float64, object

  Number of columns - 592

  Data Size - 19.5+ MB

Raw Dataset Samples

Census information on *Food security dataset*

HRHHID	HRMONTH	HRYEAR4	HURESPLI	HUFINAL	FILLER	HETENURE	HEHOUSUT	HETELHHD	HETELAVL	HEPHONEO	HEFAMINC	HUTYPEA	HUTYPB	HUTYPC	HWHHWGT
1.1E+14	12	2020	1	201		2	1	1	-1	1	12	-1	-1	-1	22105070
1.1002E+11	12	2020	1	201		2	1	1	-1	1	13	-1	-1	-1	20140310
6.1091E+14	12	2020	1	201		1	1	1	-1	1	9	-1	-1	-1	24068502
6.1091E+14	12	2020	1	201		1	1	1	-1	1	9	-1	-1	-1	24068502
6.1091E+14	12	2020	1	201		1	1	1	-1	1	9	-1	-1	-1	24068502
6.1091E+14	12	2020	1	201		1	1	1	-1	1	9	-1	-1	-1	24068502
6.1091E+14	12	2020	1	201		1	1	1	-1	1	9	-1	-1	-1	24068502
1.0701E+14	12	2020	1	201		1	1	1	-1	1	14	-1	-1	-1	16644183
1.0701E+14	12	2020	1	201		1	1	1	-1	1	14	-1	-1	-1	16644183
9.3141E+14	12	2020	1	201		1	1	1	-1	0	9	-1	-1	-1	22105105
9.3141E+14	12	2020	1	201		1	1	1	-1	0	9	-1	-1	-1	22105105
9.3141E+14	12	2020	1	201		1	1	1	-1	0	9	-1	-1	-1	22105105
7.6108E+14	12	2020	1	201		1	1	1	-1	0	10	-1	-1	-1	17110393
6.7092E+13	12	2020	-1	226		-1	5	-1	-1	0	-1	-1	1	-1	0
6.9172E+14	12	2020	-1	225		1	5	-1	-1	0	-1	-1	-1	-1	0
5.0501E+14	12	2020	-1	226		-1	1	-1	-1	0	-1	-1	1	-1	0
8.1096E+14	12	2020	-1	226		-1	1	-1	-1	0	-1	-1	1	-1	0
5.1001E+14	12	2020	-1	226		-1	1	-1	-1	0	-1	-1	1	-1	0
1.4724E+14	12	2020	-1	226		-1	1	-1	-1	0	-1	-1	1	-1	0
2.0135E+14	12	2020	-1	226		-1	1	-1	-1	0	-1	-1	1	-1	0
5.1E+14	12	2020	1	201		1	1	1	-1	1	9	-1	-1	-1	17891595
5.1E+14	12	2020	1	201		1	1	1	-1	1	9	-1	-1	-1	17891595
1.1018E+11	12	2020	1	226		-1	1	-1	-1	0	-1	-1	1	-1	0
1.5213E+14	12	2020	1	201		1	1	1	-1	1	2	-1	-1	-1	20653384
1.5505E+14	12	2020	1	201		1	1	1	-1	1	16	-1	-1	-1	21551244
1.5505E+14	12	2020	1	201		1	1	1	-1	1	16	-1	-1	-1	21551244

RASSF dataset

3928	2023	prepared dishes and snacks	food	Sensory deviation in cheese compone	17-11-2022 17:52:27	Germany	information notification for follow-up	undecided	Germany	Germany	Italy	Germany	Italy
3929	2023	fruits and vegetables	food	Unauthorized substance omethoate (M	17-11-2022 17:49:48	Germany	information notification for attention	undecided	Germany			Egypt,Ge	Egypt
3930	2023	fruits and vegetables	food	Cadmium in Cepes	17-11-2022 17:14:36	Netherlar	alert notification	serious	Belgium,Malta,Netherla	Belgium,INFOSAN	Malta,Netherlands,Poland	Belgium,I	Poland
3931	2023	prepared dishes and snacks	food	Undeclared allergens in wheat snacks	17-11-2022 17:11:05	Sweden	information notification for follow-up	undecided	Sweden	Sweden	Poland	Poland,S	Poland
3932	2023	prepared dishes and snacks	food	Listeria monocytogenes in ready meal (	17-11-2022 16:57:59	Italy	alert notification	serious	Italy	Italy,Spain	Spain	Italy,Spai	Spain
3933	2023	feed materials	feed	Unauthorized substance chlorprophan	17-11-2022 16:37:12	Germany	information notification for follow-up	potentially serious	Austria,Germany	Austria	Austria,Hungary	Austria,G	Austria
3934	2023	cereals and bakery products	food	ergot alkaloids in rye flour	17-11-2022 16:36:30	Belgium	alert notification	serious	Belgium,Netherlands	Belgium,Netherlands	Belgium,Germany,Netherlar	Belgium,I	Belgium,Germar
3935	2023	pet food	feed	Lead in complete feed for dogs from th	17-11-2022 16:00:02	Germany	alert notification	serious	Austria,Belgium,Czech	Germany,INFOSAN	Austria,Belgium,Czech Repu	Germany	United Kingdom
3936	2023	bivalve molluscs and products	food	Norovirus genogroup II in clams (Chame	17-11-2022 15:55:20	Spain	information notification for attention	undecided	Spain	Italy		Italy,Spai	Italy
3937	2023	meat and meat products (othe	food	Salmonella in dry sausages from Franc	17-11-2022 14:45:30	France	alert notification	serious	Denmark,French Polyn	INFOSAN	Denmark,Germany	France	France
3938	2023	fruits and vegetables	food	Malathion in Split Broad Beans from Eg	17-11-2022 12:29:57	Cyprus	border rejection notification	undecided	Cyprus			Cyprus,E	Egypt
3939	2023	herbs and spices	food	Pyrrolizidine alkaloids in dried oregano	17-11-2022 12:07:21	France	border rejection notification	serious	United States			France,T	TÃ¼rkiye
3940	2023	cereals and bakery products	food	Foreign bodies (metal parts) in basmati	16-11-2022 18:00:08	Germany	alert notification	serious	Belgium,Germany,Swit:	Belgium,Germany,Swit:	Belgium,Switzerland	Belgium,I	Belgium,Germar
3941	2023	nuts, nut products and seeds	food	Salmonella Oslo und Salmonella Korov	16-11-2022 17:01:22	Germany	alert notification	serious	Austria,Germany,Irelan	INFOSAN	Austria,Ireland,Luxembourg	Germany	Germany,Ugand
3942	2023	dietetic foods, food suppleme	food	Too high content of vitamin D3, zinc ar	16-11-2022 16:49:52	Slovenia	alert notification	serious	Serbia	INFOSAN	Bulgaria,Slovakia	Slovakia	Bulgaria
3943	2023	poultry meat and poultry meat	food	detection of salmonella enteritidis on cl	16-11-2022 16:46:40	France	alert notification	serious	Belgium,France,Germa	Belgium,France,Germa	Belgium,France,Germany,Ir	France,S	Spain
3944	2023	fish and products thereof	food	Mercurio por encima del lÃmite permitic	16-11-2022 15:52:17	Spain	information notification for attention	serious		INFOSAN		Mexico,S	Mexico
3945	2023	food contact materials	food c	Poor stability of the material in melamin	16-11-2022 15:51:46	Finland	border rejection notification	not serious	Finland	Belgium,Finland,France,Italy,Netherlands,Sweden		Finland,L	China
3946	2023	dietetic foods, food suppleme	food	Sildenafil and Tadalafil in food supplem	16-11-2022 15:42:41	Poland	alert notification	serious	Hungary,Poland,Slovak	INFOSAN,Poland	Germany,Hungary,Slovakia	Germany	United Kingdom
3947	2023	poultry meat and poultry meat	food	Salmonella spp. (in 9 out of 10 samples)	16-11-2022 15:36:29	Lithuania	information notification for attention	undecided	Lithuania	Poland		Lithuania	Poland
3948	2023	other food product / mixed	food	unauthorised operator (for dairy produc	16-11-2022 12:30:20	France	information notification for follow-up	not serious	Belgium,Finland,France,Germany,Netherlands		Belgium,Finland,Germany,N	France,G	Malaysia
3949	2023	milk and milk products	food	Used by date error on milk chocolate m	16-11-2022 12:25:13	France	information notification for follow-up	not serious	Belgium,Luxembourg	Belgium,Luxembourg		France	France
3950	2023	feed materials	feed	Dioxins in fatty acid distillate from Germ	16-11-2022 12:08:31	Netherlar	information notification for follow-up	not serious		Netherlands	Germany	Germany	Germany
3951	2023	herbs and spices	food	Ethylene oxide in dried rosemary from M	16-11-2022 11:58:38	Germany	alert notification	serious	Belgium,Germany,Neth	INFOSAN,Netherlands	Belgium,Netherlands	Germany	Morocco
3952	2023	fish and products thereof	food	Listeria monocytogenes in fried fish fille	16-11-2022 11:41:12	Denmark	alert notification	serious	Denmark,Germany		Denmark,Germany,Poland	Denmark	Poland
3953	2023	feed materials	feed	Salmonella in post-extraction rapesee	16-11-2022 11:31:42	Poland	information notification for attention	serious	Poland			Lithuania	Ukraine
3954	2023	cereals and bakery products	food	Unauthorized substance Thiamethoxa	16-11-2022 08:59:28	Cyprus	border rejection notification	not serious	Cyprus	INFOSAN		Cyprus,L	India
3955	2023	fish and products thereof	food	Listeria monocytogenes in slices smok	15-11-2022 17:21:44	Italy	alert notification	serious	Italy,Spain	Italy	Spain	Italy,Spai	Spain
3956	2023	meat and meat products (othe	food	Salmonella Typhimurium (in 1 out of 5 u	15-11-2022 17:19:07	Latvia	information notification for attention	not serious	Latvia	Latvia,Poland		Latvia,Pc	Poland
3957	2023	cocoa and cocoa preparation	food	Increased Tetrahydrocannibinol (THC)	15-11-2022 17:16:37	Germany	alert notification	serious	Germany,Italy,Netherla	Germany	Italy,Netherlands,Slovenia,S	Germany	Germany
3958	2023	herbs and spices	food	Salmonella in rosemary from Egypt	15-11-2022 16:12:21	Poland	information notification for attention	serious	Poland	INFOSAN		Egypt,Po	Egypt
3959	2023	meat and meat products (othe	food	Duck meat with Listeria monocytogene	15-11-2022 15:44:26	Netherlar	information notification for attention	not serious	Netherlands	France		Belgium,I	France
3960	2023	dietetic foods, food suppleme	food	Unauthorized substance magnesium ir	15-11-2022 14:50:24	Czech Re	information notification for follow-up	undecided	Czech Republic		Netherlands	Czech Re	United States
3961	2023	meat and meat products (othe	food	Salmonella typhimurium in raw beef pat	15-11-2022 14:41:32	Denmark	alert notification	serious	Denmark,Finland	Denmark,Finland,Italy	Italy	Finland,It	Italy

Gross domestic product (GDP) survey data

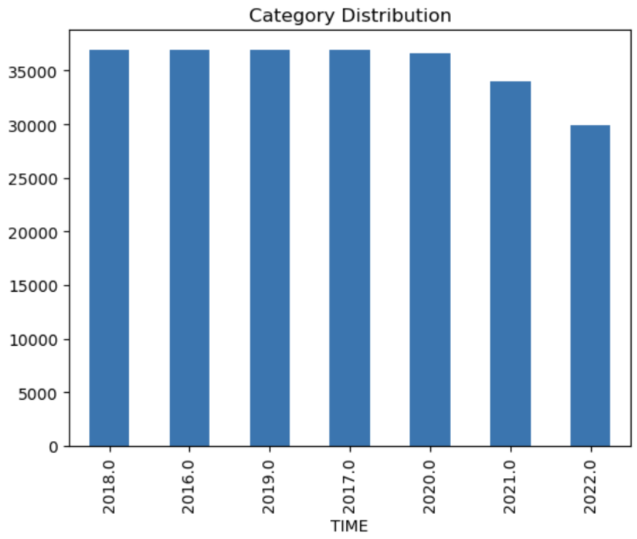
9675	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2010	2010	PC	Percentag	0	Units	USA_100	USA=100	81.89861
9676	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2011	2011	PC	Percentag	0	Units	USA_100	USA=100	81.95616
9677	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2012	2012	PC	Percentag	0	Units	USA_100	USA=100	81.33711
9678	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2013	2013	PC	Percentag	0	Units	USA_100	USA=100	83.59872
9679	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2014	2014	PC	Percentag	0	Units	USA_100	USA=100	83.75768
9680	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2015	2015	PC	Percentag	0	Units	USA_100	USA=100	87.05372
9681	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2016	2016	PC	Percentag	0	Units	USA_100	USA=100	92.87049
9682	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2017	2017	PC	Percentag	0	Units	USA_100	USA=100	93.37004
9683	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2018	2018	PC	Percentag	0	Units	USA_100	USA=100	91.59572
9684	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2019	2019	PC	Percentag	0	Units	USA_100	USA=100	92.13431
9685	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2020	2020	PC	Percentag	0	Units	USA_100	USA=100	85.53511
9686	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2021	2021	PC	Percentag	0	Units	USA_100	USA=100	82.92613
9687	ISL	Iceland	T_GDPPOI	GDP per head of population	PCTUS	As % of the USA	2022	2022	PC	Percentag	0	Units	USA_100	USA=100	87.26818
9688	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1970	1970	PC	Percentag	0	Units	USA_100	USA=100	64.19368
9689	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1971	1971	PC	Percentag	0	Units	USA_100	USA=100	68.11027
9690	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1972	1972	PC	Percentag	0	Units	USA_100	USA=100	69.79901
9691	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1973	1973	PC	Percentag	0	Units	USA_100	USA=100	71.54312
9692	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1974	1974	PC	Percentag	0	Units	USA_100	USA=100	74.82006
9693	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1975	1975	PC	Percentag	0	Units	USA_100	USA=100	72.92791
9694	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1976	1976	PC	Percentag	0	Units	USA_100	USA=100	74.12664
9695	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1977	1977	PC	Percentag	0	Units	USA_100	USA=100	81.41664
9696	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1978	1978	PC	Percentag	0	Units	USA_100	USA=100	84.96242
9697	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1979	1979	PC	Percentag	0	Units	USA_100	USA=100	88.75992
9698	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1980	1980	PC	Percentag	0	Units	USA_100	USA=100	91.5433
9699	ISL	Iceland	T_GDPHRS	GDP per hour worked	PCTUS	As % of the USA	1981	1981	PC	Percentag	0	Units	USA_100	USA=100	88.99241

Commodity Terms of Trade dataset

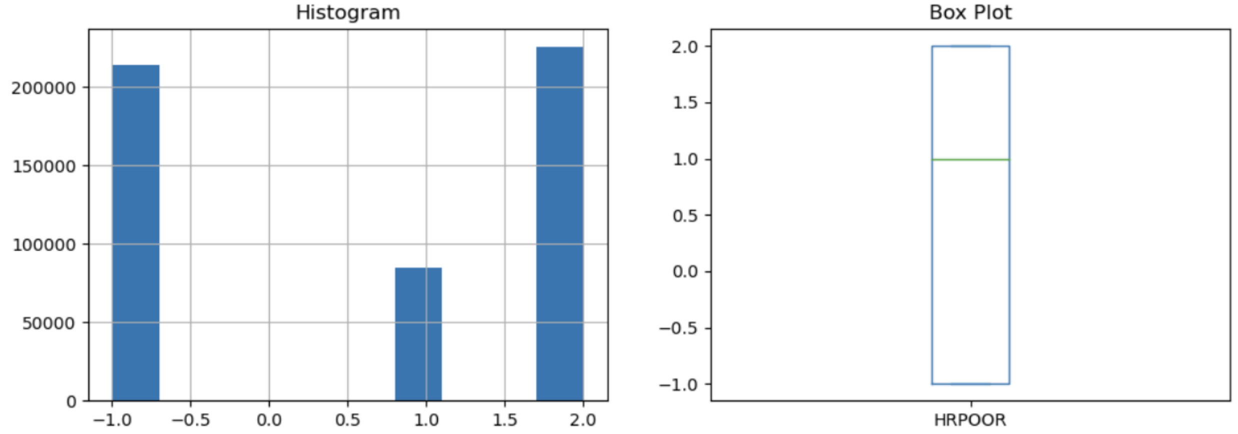
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Country N	Country C	Indicator I	Indicator (	Type Nam	Type Code	Attribute	1980M1	1980M2	1980M3	1980M4	1980M5	1980M6	1980M7	1980M8	1980M9	1980M10	1980M11	1980M12	1981M1	1981M
2	Peru	293	Commodi	m	Recent, Fi	R_FW_IX	Value	97.66689	96.9931	92.79942	90.37221	91.7209	91.00529	92.903	92.54916	90.89473	93.15162	96.62324	93.29862	93.42048	90.37
3	Peru	293	Commodi	xm	Recent, Fi	R_FW_IX	Value	93.0569	95.4979	87.31962	83.30352	82.12994	81.5274	82.50726	81.29194	81.47253	80.78108	79.45136	76.80608	76.53876	74.29
4	Peru	293	Commodi	x_gdp	Recent, Fi	R_FW_IX	Value	98.55697	98.90469	97.35847	96.56189	96.4059	96.33775	96.62002	96.36697	96.35487	96.30241	96.16021	95.4985	95.47326	94.91
5	Peru	293	Commodi	m_gdp	Recent, Fi	R_FW_IX	Value	99.81856	99.78288	99.64299	99.55042	99.59032	99.56464	99.62064	99.5953	99.51903	99.60579	99.74296	99.64232	99.64445	99.53
6	Peru	293	Commodi	xm_gdp	Recent, Fi	R_FW_IX	Value	98.73598	99.11977	97.70718	96.99786	96.80235	96.75888	96.98782	96.75844	96.82044	96.68343	96.40791	95.84119	95.81381	95.35
7	Namibia	728	Commodi	m_gdp	Recent, Fi	R_FW_IX	Value														
8	Namibia	728	Commodi	xm_gdp	Recent, Fi	R_FW_IX	Value														
9	Niger	692	Commodi	x	Recent, Fi	R_FW_IX	Value	148.7939	140.9966	127.9018	118.4221	117.2909	115.3884	115.547	110.6276	104.7057	103.072	102.322	97.69827	92.39689	91.29
10	Niger	692	Commodi	m	Recent, Fi	R_FW_IX	Value	105.9437	106.6407	103.4103	102.1088	105.8307	104.9294	102.0442	101.043	98.53271	101.3127	105.5385	103.3796	103.11	100.7
11	Niger	692	Commodi	xm	Recent, Fi	R_FW_IX	Value	124.7621	120.3203	114.6647	109.8141	107.6695	106.9212	108.1299	105.6465	103.0577	100.9805	98.98293	96.97677	93.72466	93.88
12	Niger	692	Commodi	x_gdp	Recent, Fi	R_FW_IX	Value	103.2429	102.7881	101.9819	101.3425	101.2608	101.1222	101.1303	100.7746	100.3209	100.1944	100.1411	99.77152	99.31374	99.22
13	Niger	692	Commodi	m_gdp	Recent, Fi	R_FW_IX	Value	100.2726	100.3021	100.1551	100.093	100.2594	100.2171	100.0846	100.035	99.9132	100.0443	100.2438	100.1498	100.1384	100.0
14	Niger	692	Commodi	xm_gdp	Recent, Fi	R_FW_IX	Value	102.9622	102.4784	101.824	101.2483	100.9988	100.903	101.0447	100.7393	100.408	100.15	99.89745	99.62218	99.1764	99.1
15	Nigeria	694	Commodi	x	Recent, Fi	R_FW_IX	Value	82.45204	79.43845	78.48428	76.73038	75.97845	75.50039	72.81089	70.05552	65.88019	68.20306	72.21207	73.68949	73.62174	72.0
16	Nigeria	694	Commodi	m	Recent, Fi	R_FW_IX	Value	148.3413	152.772	143.3414	141.019	148.878	147.1679	146.9981	149.1385	149.5042	153.8272	153.4333	143.0994	142.1916	135.
17	Nigeria	694	Commodi	xm	Recent, Fi	R_FW_IX	Value	81.39101	78.49332	77.99667	76.49612	75.49872	75.13033	72.6791	70.08239	66.229	68.21674	71.88369	73.61191	73.57594	72.41
18	India	534	Commodi	xm	Recent, Fi	R_FW_IX	Value	114.9633	115.8441	118.5594	120.3107	120.8097	120.1015	121.0552	123.7394	125.2454	124.9649	122.9173	122.8978	122.3668	124.2
19	India	534	Commodi	x_gdp	Recent, Fi	R_FW_IX	Value	99.81889	99.76785	99.67818	99.61529	99.60187	99.58588	99.57418	99.55293	99.46034	99.51281	99.58566	99.54256	99.52235	99.47
20	India	534	Commodi	m_gdp	Recent, Fi	R_FW_IX	Value	98.74718	98.61668	98.30952	98.11698	98.06017	98.12353	98.06583	97.87745	97.6964	97.76398	97.94579	97.86555	97.88686	97.70
21	India	534	Commodi	xm_gdp	Recent, Fi	R_FW_IX	Value	101.0852	101.1673	101.3921	101.527	101.5721	101.4902	101.538	101.7117	101.8054	101.7887	101.6742	101.7135	101.6707	101.8
22	Ireland	178	Commodi	x	Recent, Fi	R_FW_IX	Value	133.9984	130.531	117.9958	113.1532	110.7431	112.0223	119.9825	124.3317	119.4683	120.84	122.3588	115.5677	112.8761	10
23	Ireland	178	Commodi	m	Recent, Fi	R_FW_IX	Value	96.58919	94.58177	91.06554	88.33367	87.45601	86.54319	87.38233	86.09525	83.15221	84.07378	86.4295	84.9812	84.78668	83.35
24	Ireland	178	Commodi	xm	Recent, Fi	R_FW_IX	Value	116.4312	116.5461	113.9731	113.8971	113.486	114.7341	117.5989	120.4846	120.8464	120.6417	119.3279	117.4896	116.44	115.8
25	Ireland	178	Commodi	x_gdp	Recent, Fi	R_FW_IX	Value	101.2738	101.1652	100.7215	100.5445	100.4495	100.5001	100.7998	100.9612	100.796	100.8452	100.8995	100.648	100.5464	100.3

PCTOT_10-04-2023 16-42-10-15_ti

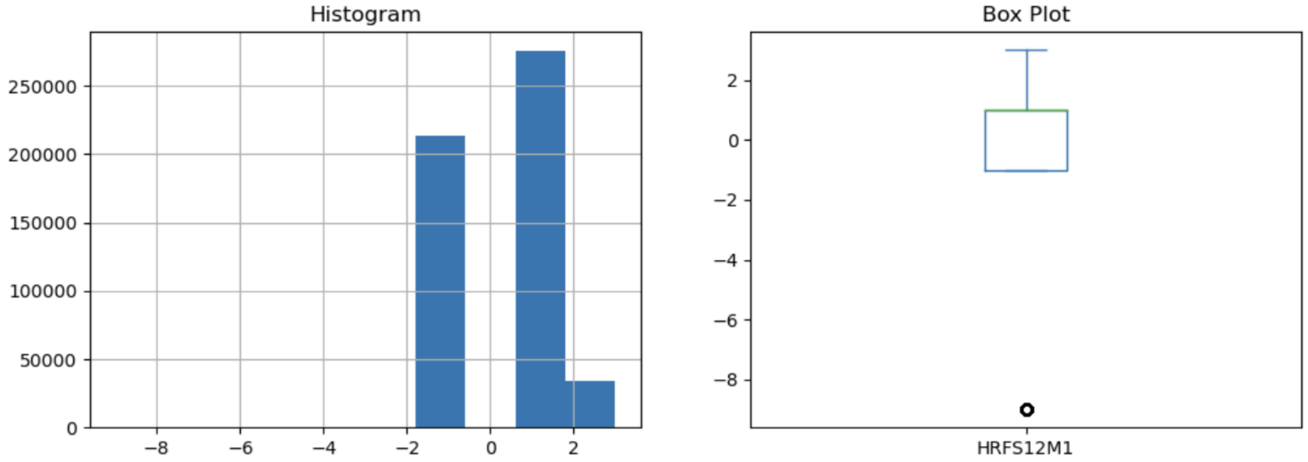
Data Preprocessing: Exploratory Data Analysis



Distribution of time series data

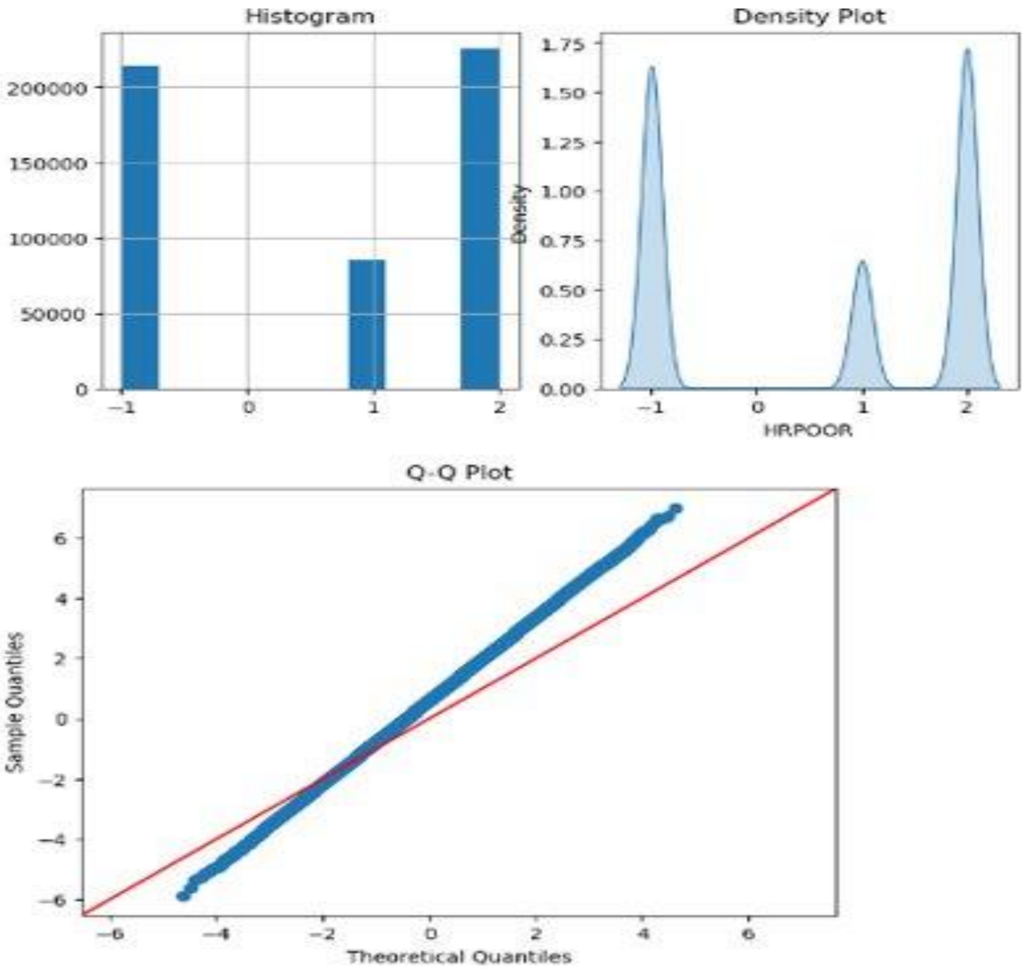


Sample HRPOOR feature analysis

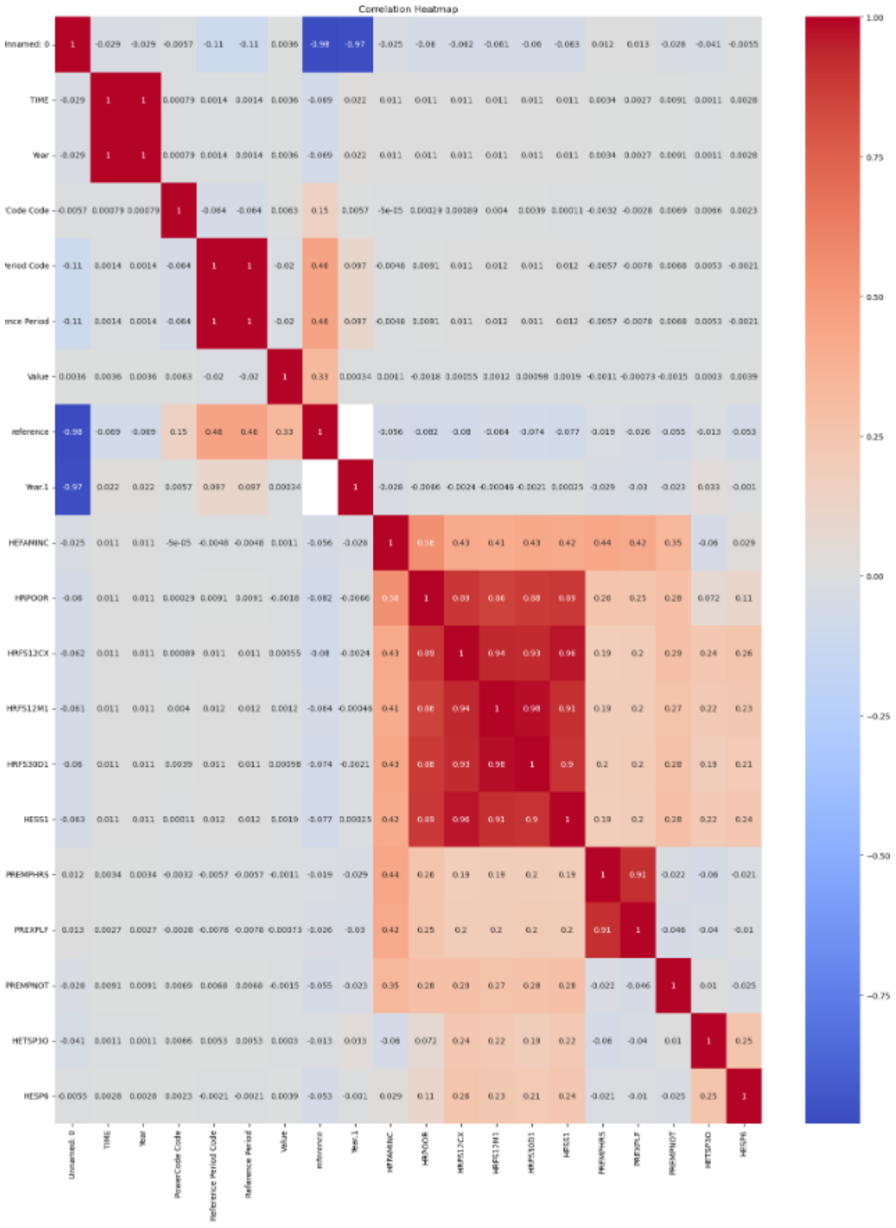


Sample HRFS12M1 feature analysis

Data Preprocessing: Exploratory Data Analysis



HRPOOR features multivariate analysis



Correlation matrix

Data Cleaning: Handling of Inconsistent Data

	reference	category	type	subject	date
0	2023.7463	fruits and vegetables	food	Buprofezin in lemons from Türkiye	2023-01-11 09:48:00
1	2023.7457	dietetic foods, food supplements and fortified...	food	Novel Food in Food Supplement	2023-10-31 18:27:17
2	2023.7456	fruits and vegetables	food	Perchlorate in radish	2023-10-31 17:55:46
3	2023.7455	fruits and vegetables	food	Chlormequat in Nashi Pears	2023-10-31 17:46:25
4	2023.7454	soups, broths, sauces and condiments	food	Consignment Mayonaises not presented for veter...	2023-10-31 17:32:54
...	...	...	...	...	...
4995	2022.5568	crustaceans and products thereof	food	Ruptura de la cadena de frío en Litopenaeus va...	2022-09-26 15:11:33
4996	2022.5567	confectionery	food	Too high content of fumaric acid in gummy bear...	2022-09-26 15:07:40
4997	2022.5566	nuts, nut products and seeds	food	SALMONELLA IN SESAME SEEDS FROM TURKEY	2022-09-26 14:25:06
4998	2022.5565	nuts, nut products and seeds	food	SALMONELLA IN SESAME SEEDS FROM TURKEY	2022-09-26 14:18:42
4999	2022.5564	nuts, nut products and seeds	food	SALMONELLA IN SESAME SEEDS FROM TURKEY	2022-09-26 14:12:57

Time in dd-m-y : hr format

```
def parse_date(date_str):
    try:
        # Try one format
        return pd.to_datetime(date_str, format="%m/%d/%Y %H:%M")
    except ValueError:
        try:
            # Try another format
            return pd.to_datetime(date_str, format="%d-%m-%Y %H:%M:%S")
        except ValueError:
            # Handle any other format or errors
            return None
df2['date'] = df2['date'].apply(parse_date)
df2
```

Sample code snippet

Year
2023.0
2023.0
2023.0
2023.0
2023.0
...
2022.0
2022.0
2022.0
2022.0
2022.0

Year extracted from time format

Data Cleaning: Handling of Incomplete & Missing Data

```
Unnamed: 0                    0
LOCATION                 275831
Country                  275831
TRANSACT                 275831
Transaction              275831
MEASURE                  275831
Measure                  275831
TIME                     275831
Year                     275831
Unit Code                275831
Unit                     275831
PowerCode Code           275831
PowerCode                275831
Reference Period Code    397267
Reference Period         397267
Value                    275831
Flag Codes               497908
Flags                    497908
reference                518979
category                 518979
type                     518979
subject                  518979
date                     518979
notifying_country        518979
classification           518979
risk_decision            518979
distribution             520547
forAttention             521408
forFollowUp              521667
operator                 518986
origin                   519105
hazards                  520356
Year.1                        0
HEFAMINC                      0
HRPOOR                        0
HRFS12CX                      0
HRFS12M1                      0
HRFS30D1                      0
HESS1                         0
PREMPHRS                      0
PREXPLF                       0
PREMPNOT                      0
HETSP3O                       0
HESP6                         0
dtype: int64
12650054
```

Null values count

```
clean_df=new_df.dropna()
```

```
clean_df
```

	Unnamed: 0	LOCATION	Country	TRANSACT	Transaction	MEASURE	Measure	TIME	Year	Unit Code	...	HRPOOR	HRFS12CX	HRFS12M1	HRFS30D1
0	0	AUS	Australia	B1_GA	Gross domestic product (output approach)	C	Current prices	2016.0	2016.0	AUD	...	-1.0	-1.0	-1.0	-1.0
1	1	AUS	Australia	B1_GA	Gross domestic product (output approach)	C	Current prices	2017.0	2017.0	AUD	...	-1.0	-1.0	-1.0	-1.0
2	2	AUS	Australia	B1_GA	Gross domestic product (output approach)	C	Current prices	2018.0	2018.0	AUD	...	-1.0	-1.0	-1.0	-1.0
3	3	AUS	Australia	B1_GA	Gross domestic product (output approach)	C	Current prices	2019.0	2019.0	AUD	...	-1.0	-1.0	-1.0	-1.0
4	4	AUS	Australia	B1_GA	Gross domestic product (output approach)	C	Current prices	2020.0	2020.0	AUD	...	-1.0	-1.0	-1.0	-1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
248142	248142	TUR	Türkiye	P32S13	Collective consumption expenditure of general ...	DOB	Deflator	2018.0	2018.0	IDX	...	-1.0	-1.0	-1.0	-1.0
248143	248143	TUR	Türkiye	P32S13	Collective consumption expenditure of general ...	DOB	Deflator	2019.0	2019.0	IDX	...	-1.0	-1.0	-1.0	-1.0
248144	248144	TUR	Türkiye	P32S13	Collective consumption expenditure of general ...	DOB	Deflator	2020.0	2020.0	IDX	...	-1.0	-1.0	-1.0	-1.0
248145	248145	TUR	Türkiye	P32S13	Collective consumption expenditure of general ...	DOB	Deflator	2021.0	2021.0	IDX	...	-1.0	-1.0	-1.0	-1.0
248146	248146	TUR	Türkiye	P32S13	Collective consumption expenditure of general ...	DOB	Deflator	2022.0	2022.0	IDX	...	-1.0	-1.0	-1.0	-1.0

248147 rows × 26 columns

Sample data after clearing Null values

```
<class 'pandas.core.frame.DataFrame'>
Index: 248147 entries, 0 to 248146
Data columns (total 26 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Unnamed: 0      248147 non-null  int64
 1   LOCATION        248147 non-null  object
 2   Country         248147 non-null  object
 3   TRANSACT        248147 non-null  object
 4   Transaction     248147 non-null  object
 5   MEASURE         248147 non-null  object
 6   Measure         248147 non-null  object
 7   TIME            248147 non-null  float64
 8   Year            248147 non-null  float64
 9   Unit Code       248147 non-null  object
 10  Unit            248147 non-null  object
 11  PowerCode Code  248147 non-null  float64
 12  PowerCode       248147 non-null  object
 13  Value           248147 non-null  float64
 14  Year.1          248147 non-null  float64
 15  HEFAMINC        248147 non-null  float64
 16  HRPOOR          248147 non-null  float64
 17  HRFS12CX        248147 non-null  float64
 18  HRFS12M1        248147 non-null  float64
 19  HRFS30D1        248147 non-null  float64
 20  HESS1           248147 non-null  float64
 21  PREMPHRS        248147 non-null  float64
 22  PREXPLF         248147 non-null  float64
 23  PREMPNOT        248147 non-null  float64
 24  HETSP3O         248147 non-null  float64
 25  HESP6           248147 non-null  float64
dtypes: float64(16), int64(1), object(9)
memory usage: 51.1+ MB
```

Dataset after clearing null values

Data Cleaning: Handling of Noisy Data

```
[7530 rows x 17 columns]
```

Row count for outliers

```python
z_scores = (numerical_df - numerical_df.mean()) / numerical_df.std()
print(z_scores)
outliers_df = numerical_df[~(z_scores < 3).all(axis=1)]

# Rows containing outliers
print("Rows with outliers:")
print(outliers_df)

# Rows after eliminating outliers
res_df = numerical_df[(z_scores < 3).all(axis=1)]
print("\nRows after eliminating outliers:")
print(res_df)
```

Sample code snippet

```
[240617 rows x 17 columns]
```

Total row count after removing outliers

To summarize:

- Exploratory data analysis is performed to understand how the data stands after merging the datasets which helps in understanding correlation between columns, dispersion of datatypes, etc.
- Data cleaning is performed in three phases,
- As we have many yearly insights to be drawn, the data is manipulated to a yearly format
- Next, we clean the data of all the null values
- Finally, noisy data is cleared using the z-scores from all the numerical columns

Data Transformation: Label Encoding

```
unique_values = sorted(final_clean_df['Country'].unique())

# Print the unique values
print(unique_values)
```

```
['Argentina', 'Australia', 'Austria', 'Belgium', 'Brazil', 'Bulgaria', 'Cameroon', 'Canada', 'Chile', "China (Peopl
e's Republic of)", 'Colombia', 'Costa Rica', 'Croatia', 'Cyprus', 'Czech Republic', 'Denmark', 'Estonia', 'Euro area
(19 countries)', 'European Union – 27 countries (from 01/02/2020)', 'Finland', 'France', 'Germany', 'Greece', 'Hungar
y', 'Iceland', 'India', 'Indonesia', 'Ireland', 'Israel', 'Italy', 'Japan', 'Korea', 'Latvia', 'Lithuania', 'Luxembou
rg', 'Malta', 'Mexico', 'Netherlands', 'New Zealand', 'Norway', 'OECD - Total', 'Poland', 'Portugal', 'Romania', 'Rus
sia', 'Saudi Arabia', 'Senegal', 'Slovak Republic', 'Slovenia', 'South Africa', 'Spain', 'Sweden', 'Switzerland', 'Tü
rkiye', 'United Kingdom', 'United States']
```

Sample code snippet to identify unique values in a column to be encoded

```
from sklearn.preprocessing import LabelEncoder
country_encoder = LabelEncoder()
# Fit and transform the column
final_clean_df['Country_Column_encoded'] = country_encoder.fit_transform(final_clean_df['Country'])

# Print or use the encoded column
print(final_clean_df['Country_Column_encoded'])
```

```
0          1
1          1
2          1
3          1
4          1
          ..
240612    53
240613    53
240614    53
240615    53
240616    53
Name: Country_Column_encoded, Length: 240617, dtype: int32
```

Sample code snippet of label encoding a column

Data Normalization

```
In [5]: import pandas as pd
        from sklearn.preprocessing import MinMaxScaler, StandardScaler
        num_df = final_clean_df.select_dtypes(include=['number'])

        # Z-score Normalization
        z_score_scaler = StandardScaler()
        df_z_score_normalized = pd.DataFrame(z_score_scaler.fit_transform(num_df), columns=num_df.columns)
```

Sample code snippet to perform Z-score normalization

	TIME	Year	PowerCode Code	Value	Year.1	HEFAMINC	HRPOOR	HRFS12CX	HRFS12M1	HRFS30D1	...	PREXPLF	PREMPNOT	HETSP3O	
0	-1.472603	2016.0	0.487835	-0.007763	0.988656	-1.695627	-1.134884	-1.12064	-1.071966	-1.087396	...	-0.802531	-1.108711	-0.149102	-0.
1	-0.962859	2017.0	0.487835	-0.007583	0.988656	-1.695627	-1.134884	-1.12064	-1.071966	-1.087396	...	-0.802531	-1.108711	-0.149102	-0.
2	-0.453115	2018.0	0.487835	-0.007359	0.988656	-1.695627	-1.134884	-1.12064	-1.071966	-1.087396	...	-0.802531	-1.108711	-0.149102	-0.
3	0.056629	2019.0	0.487835	-0.007289	0.988656	-1.695627	-1.134884	-1.12064	-1.071966	-1.087396	...	-0.802531	-1.108711	-0.149102	-0.
4	0.566373	2020.0	0.487835	-0.007072	0.988656	-1.695627	-1.134884	-1.12064	-1.071966	-1.087396	...	-0.802531	-1.108711	-0.149102	-0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
240612	-0.453115	2018.0	-2.251247	-0.011545	-1.011474	-1.695627	-1.134884	-1.12064	-1.071966	-1.087396	...	-0.802531	-1.108711	-0.149102	-0.
240613	0.056629	2019.0	-2.251247	-0.011545	-1.011474	-1.695627	-1.134884	-1.12064	-1.071966	-1.087396	...	-0.802531	-1.108711	-0.149102	-0.
240614	0.566373	2020.0	-2.251247	-0.011545	-1.011474	-1.695627	-1.134884	-1.12064	-1.071966	-1.087396	...	-0.802531	-1.108711	-0.149102	-0.
240615	1.076117	2021.0	-2.251247	-0.011545	-1.011474	-1.695627	-1.134884	-1.12064	-1.071966	-1.087396	...	-0.802531	-1.108711	-0.149102	-0.
240616	1.585861	2022.0	-2.251247	-0.011545	-1.011474	-1.695627	-1.134884	-1.12064	-1.071966	-1.087396	...	-0.802531	-1.108711	-0.149102	-0.

240617 rows × 22 columns

Sample look at dataset after normalization

Data Regularization: L1 regularization

```
[14]: import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler

numeric_columns = final_clean_df.select_dtypes(include=[np.number]).columns

df_numeric = final_clean_df[numeric_columns]
X = df_numeric.drop('HRFS30D1', axis=1)
y = df_numeric['HRFS30D1']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Standardize the data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Apply L2 regularization
model = LogisticRegression(penalty='l2', solver='liblinear')
model.fit(X_train_scaled, y_train)

print(X_train_scaled)

# Evaluate the model
accuracy = model.score(X_test_scaled, y_test)
print(f'Accuracy: {accuracy}')
```

Sample code snippet to perform L1 regularization

```
[[ 1.58656589  1.58656589  0.48719587 ... -0.52537944 -1.86778544
   0.12749048]
 [-0.96196616 -0.96196616  0.48719587 ... -0.52537944  1.24083575
   0.12749048]
 [ 0.56715307  0.56715307  0.48719587 ... -0.52537944 -1.97497928
   0.12749048]
 ...
 [ 0.56715307  0.56715307  0.48719587 ... -0.52537944 -1.33181627
   0.12749048]
 [ 0.05744666  0.05744666  0.48719587 ... -0.17354301  0.81206041
   0.12749048]
 [-0.45225975 -0.45225975  0.48719587 ... -1.40497053 -1.33181627
   0.12749048]]
```

Sample data after L1 regularization

```
Accuracy: 0.9742747901255091
```

Accuracy for L1 regularization

Data Regularization: L2 regularization

```python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler

numeric_columns = final_clean_df.select_dtypes(include=[np.number]).columns

df_numeric = final_clean_df[numeric_columns]
X = df_numeric.drop('HRFS30D1', axis=1)
y = df_numeric['HRFS30D1']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Standardize the data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Apply L2 regularization
model = LogisticRegression(penalty='l2', solver='liblinear')
model.fit(X_train_scaled, y_train)

print(X_train_scaled)

# Evaluate the model
accuracy = model.score(X_test_scaled, y_test)
print(f'Accuracy: {accuracy}')
```

Sample code snippet to perform L2 regularization

```
[[ 1.58656589  1.58656589  0.48719587 ... -0.52537944 -1.86778544
   0.12749048]
 [-0.96196616 -0.96196616  0.48719587 ... -0.52537944  1.24083575
   0.12749048]
 [ 0.56715307  0.56715307  0.48719587 ... -0.52537944 -1.97497928
   0.12749048]
 ...
 [ 0.56715307  0.56715307  0.48719587 ... -0.52537944 -1.33181627
   0.12749048]
 [ 0.05744666  0.05744666  0.48719587 ... -0.17354301  0.81206041
   0.12749048]
 [-0.45225975 -0.45225975  0.48719587 ... -1.40497053 -1.33181627
   0.12749048]]
```

Sample data after L2 regularization

```
Accuracy: 0.9741916715152522
```

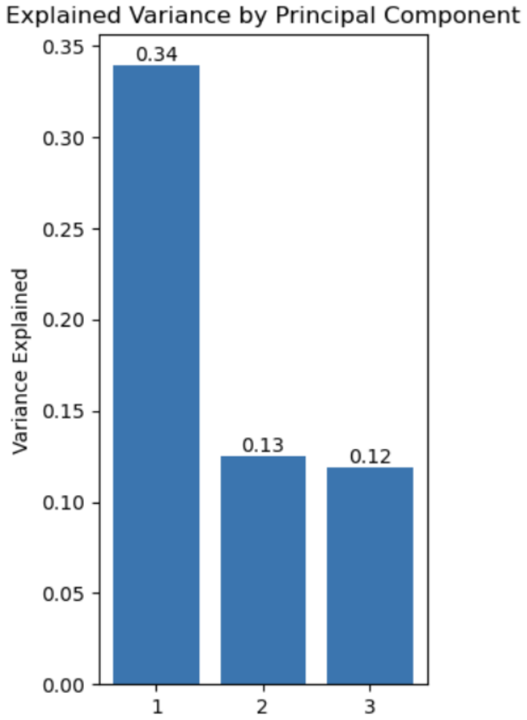
Accuracy for L2 regularization

Principal Component Analysis

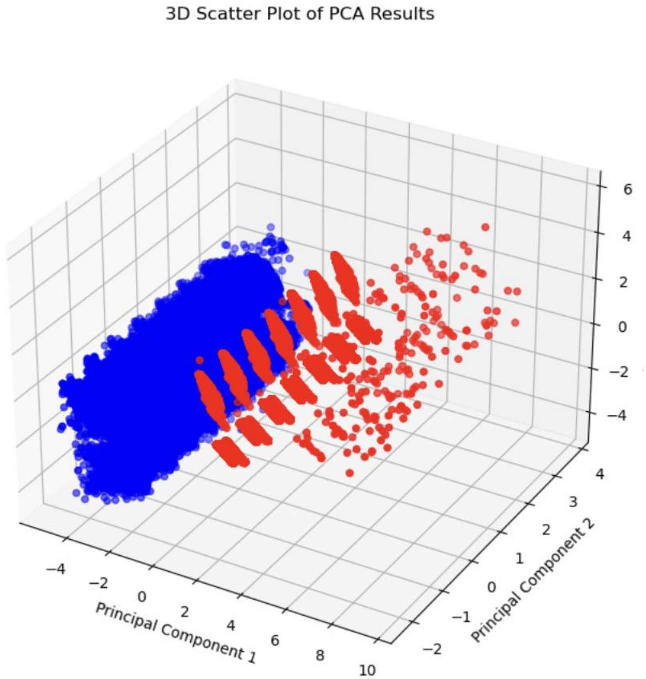
```
             PC1       PC2       PC3
0       3.212763 -1.986734 -0.747473
1       3.204051 -1.267029 -0.714400
2       3.195338 -0.547324 -0.681326
3       3.186625  0.172381 -0.648253
4       3.177912  0.892086 -0.615180
...          ...       ...       ...
240612  3.213045 -0.615550 -0.568933
240613  3.204332  0.104156 -0.535859
240614  3.195619  0.823862 -0.502786
240615  3.186906  1.543567 -0.469712
240616  3.178193  2.263273 -0.436639

[240617 rows x 3 columns]
```

Sample PC distribution

2D scatter plot



Explained variance with respect to PC



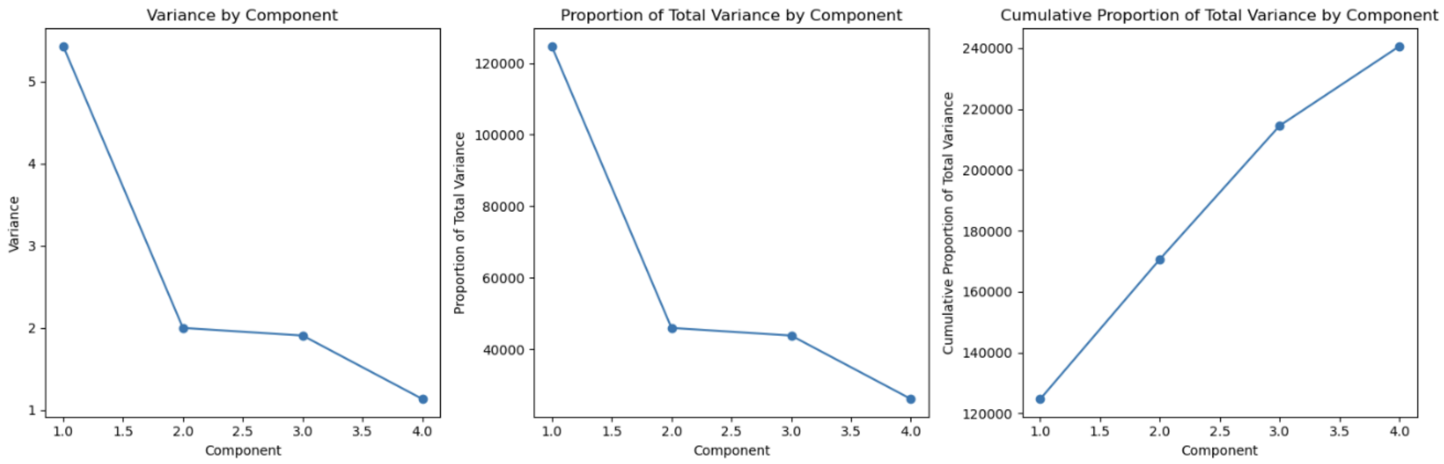
3D scatter plot

Singular Value Decomposition

	Component 1	Component 2	Component 3
0	3.212763	-1.986734	-0.747473
1	3.204051	-1.267029	-0.714400
2	3.195338	-0.547324	-0.681326
3	3.186625	0.172381	-0.648253
4	3.177912	0.892086	-0.615180
...	...	...	...
240612	3.213045	-0.615550	-0.568933
240613	3.204332	0.104156	-0.535859
240614	3.195619	0.823862	-0.502786
240615	3.186906	1.543567	-0.469712
240616	3.178193	2.263273	-0.436639

[240617 rows x 3 columns]

Sample SVD distribution



Explained variance with respect to PC

To summarize:

- PCA is usually used to work on linear relationships, thus there seems to be lower variance extraction from the 2 principal components
- SVD has shown efficiency in dealing with non-linear data, it better explains about data for non-linear relationships, works for broader range of applications, and hold numerical stability over PCA
- SVD aids working with time series data through extraction of relevant features, isolation of anomalous patterns or trends, etc. making it an efficient tool to work with time series-based algorithms and data.

Train, Test, and Validation Datasets

```
from sklearn.model_selection import train_test_split

features = transformed_df.drop('HRFS30D1', axis=1)
target = transformed_df['HRFS30D1']

# Split the data into training (80%) and temporary set (20%)
temp_features, test_features, temp_target, test_target = train_test_split(features, target, test_size=0.2, random_state=42)

# Further split the temporary set into training (60%) and validation (20%)
train_features, val_features, train_target, val_target = train_test_split(temp_features, temp_target, test_size=0.25, random_sta

print(f"Training set size: {len(train_features)}")
print(f"Validation set size: {len(val_features)}")
print(f"Test set size: {len(test_features)}")

print("\nSample from Training Set:")
print(train_features.head())

print("\nSample from Validation Set:")
print(val_features.head())

print("\nSample from Test Set:")
print(test_features.head())
```

Code snippet for splitting the dataset

```
Training set size: 144369
Validation set size: 48124
Test set size: 48124
```

Dataset size after splitting

Model Development

Long Short-Term Memory (LSTM)

- It is a recurrent neural network (RNN) architecture that is well-suited for processing and making predictions based on sequential data, such as time series data and natural language.
- It also has popular choices for various applications, including time series forecasting, language modeling, speech recognition, and more.

ARIMA

- It is a well-known time series model for forecasting. It's divided into three parts: autoregression (AR), differencing (I), and moving average (MA).
- It is responsible for capturing a variety of standard temporal structures in time series data. It is a well-known time series model for forecasting. It's divided into three parts: autoregression (AR), differencing (I), and moving average (MA).
- It is responsible for capturing a variety of standard temporal structures in time series data.

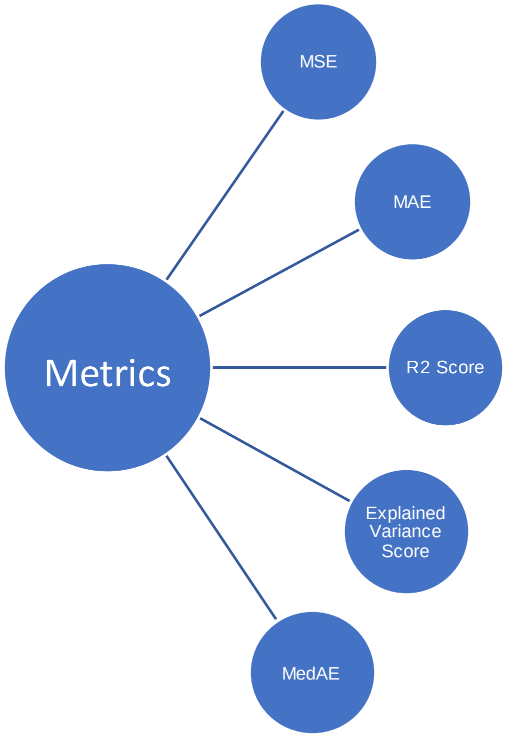
Prophet

- It is a process for forecasting time series data that is based on an additive model where non-linear trends are fit with yearly, weekly, and daily.
- It is designed to work best and strong with time series and historical data.
- It is available in both R and Python and used in many applications for providing forecasts.

Random Forest

- It is an ensemble machine learning model because during training, it gives both features and targets to predict the data.
- It is commonly used for both regression and classification and is also trained on random subset of the data points and features at each node.

Model Evaluation Metrics



LSTM
- MSE : 1.27
- MAE : 1.05
- R2 Score : -0.05
- Variance Score : -0.01
- MedAE : 0.96

ARIMA
- MSE : 1.17
- MAE : 1.02
- R2 Score : 0.00
- Variance Score : 0.00
- MedAE : 0.81

Prophet
- MSE : 1.17
- MAE : 1.02
- R2 Score : 0.00
- Variance Score : 0.00
- MedAE : 0.83

Random Forest
- MSE : 0.012
- MAE : 0.021
- R2 Score : 0.989
- Variance Score : 0.989
- MedAE : 0.00

Mean Squared Error (MSE)
Mean Absolute Error (MAE)
Median Absolute Error (MedAE)

Model Justification

- The use of diverse machine learning algorithms, including Prophet, Long short-term memory (LSTM), Autoregressive Integrated Moving Average (ARIMA) demonstrates a comprehensive approach to model selection.
- The Random Forest model was chosen as the optimal model for this project due to its high performance, as evidenced by a 0.98 R2 Score and 0.98 Explained Variance Score.
- This indicates that the model can explain 98% of the variance in the data, suggesting a strong ability to make accurate predictions.

Limitations

- The availability of comprehensive and up-to-date data, especially for post-pandemic scenarios, may pose challenges to the accuracy and reliability of the predictions.
- The model primarily focuses on key influencing factors and may not fully capture the complexity of socio-political, cultural, and environmental factors contributing to food insecurity.
- It provides a global analysis, but regional variations and unique challenges faced by different regions may not be fully captured.
- The interpretation of results and the implementation of policies to address food insecurity require additional considerations.

Summary

The project focuses on global analysis and prediction of food insecurity and hunger crisis outbreaks using machine learning techniques.

The goal is to address the constant rise in malnutrition and food deprivation by widening the scope of analysis from a regional to a global scale.

The proposed solution aims to incorporate a prediction-based approach that considers various factors, including economic shocks, extreme weather events, conflicts, and the impact of the COVID-19 pandemic.

References

- Hernandez, D. C., Reesor, L., & Murillo, R. (2017). Food insecurity and adult overweight/obesity: Gender and race/ethnic disparities. Appetite, 117, 373–378. https://doi.org/10.1016/j.appet.2017.07.010
- Martini, G., Bracci, A., Riches, L., Jaiswal, S., Corea, M., Rivers, J., Husain, A., & Omodei, E. (2022). Machine learning can guide food security efforts when primary data are not available. Nature Food, 3(9), 716–728. https://doi.org/10.1038/s43016-022-00587-8
- Tarasuk, Valerie. (2001). Health Canada. (2002, December 4). ARCHIVED - Discussion Paper on Household and Individual Food Insecurity - Executive Summary. Canada.ca. https://www.canada.ca/en/health-canada/services/food-nutrition/healthy-eating/nutrition-policy-reports/executive-summary-discussion-paper-household-individual-food-insecurity.html
- Ahn, J., Briers, G. E., Baker, M., Price, E., Djebou, D. C. S., Strong, R. L., Piña, M., & Kibriya, S. (2022). Food security and agricultural challenges in West-African rural communities: a machine learning analysis. *International Journal of Food Properties*, *25*(1), 827–844. https://doi.org/10.1080/10942912.2022.2066124
- Westerveld, J. J., Van Den Homberg, M., Nobre, G. G., Van Den Berg, D. L., Teklesadik, A., & Stuit, S. (2021). Forecasting transitions in the state of food security with machine learning using transferable features. *Science of the Total Environment*, *786*, 147366. https://doi.org/10.1016/j.scitotenv.2021.147366
- Deléglise, H., Interdonato, R., Bégué, A., D'Hôtel, É. M., Teisseire, M., & Roche, M. (2022). Food security prediction from heterogeneous data combining machine and deep learning methods. Expert Systems With Applications, 190, 116189. https://doi.org/10.1016/j.eswa.2021.116189
- Negm, M. M. (2018). Modeling and Forecasting of Food Security for Wheat in Egypt Through Year 2025 by Using time Series ARIMA Models. *The Journal of Social Sciences Research*, *5*, 510–518. https://doi.org/10.32861/jssr.spi5.510.518
- *Commodity Terms of Trade dataset.* (2023). International Monetary Fund. https://data.imf.org/?sk=2cddccb8-0b59-43e9-b6a0-59210d5605d2
- *Food security dataset.* (2021). United States Census Bureau. https://www.census.gov/data/datasets/time-series/demo/cps/cps-supp_cps-repwgt/cps-food-security.html
- *Gross domestic product (GDP) survey data.* (2022). Organization for Economic Co-operation and Development. https://stats.oecd.org/
- *RASSF dataset.* (2022). European Commission. RASFF Database (Version 2.4.1) https://webgate.ec.europa.eu/rasff-window/screen/search

THANK YOU