CSP 571 Project Proposal

Sai Ram ODURI - A20522183

Ramesh Chandra Reddy MUSILANNA - A20521309

Mouhammad BAZZI - A20522180

Department of Computer Science

Illinois Institute of Technology

I. PROBLEM PROPOSAL

I.1. Description and Goal

The project primarily focuses on the study of the impact of geo-economic factors, such as climate change, economic conditions, social and health conditions, etc. on the average agriculture production of a country.

According to us this is could be a very interesting topic to work on, especially in a context where we predict that it will be even harder to feed the global population due to our way of consuming and also the increase of the population.

Our final goal is to understand how could vary the agriculture production.

I.2. Deliverables

1. **Analyze the impact of climate change on agriculture production**: we plan to investigate how climate change affects agricultural production in different regions. This analysis will include exploring how temperature changes, CO2 emissions, forest and savanna fires, etc. are linked with agriculture production.
2. **Study the relationship between economic conditions and agriculture production**: we plan to analyze the relationship between economic factors such as GDP or credit to agriculture with agriculture production. This analysis could identify potential ways to promote sustainable agriculture practices that balance economic growth and food security.
3. **Examine the impact of social and health conditions on agriculture production**: we plan to investigate how social and health factors such as access to education, healthcare, and sanitation impact agriculture production. This analysis could identify potential solutions to improve food security in vulnerable populations.
4. **Analyze the trend of food production over years**: we plan to explore the trend of food production over years on a global scale, and how it varies across different regions and demographic groups. This analysis could help to identify potential challenges and opportunities for sustainable food production and distribution in the future.

I.3. Methodology

We plan to gather data on agriculture production, climate change, economic social, health factors, population, etc. from the **Food and Agriculture Organization** and **World Health Organization**. We will then merge all the data sets that we will get and perform data cleaning, preprocessing, and exploratory data analysis to understand the patterns and relationships between all the variables that we gather.

After that, we plan to use **regression** (and probably **time series**) analysis to model the relationships between the different variables in order to understand their impact on agriculture production.

Finally, we will gather all our results in order to present a **comprehensive analysis** of the impact of geo-economic factors on agriculture production, including the effects of climate change, economic conditions, social and health factors, population growth, and other variables.

I.4. Metrics

We plan to use statistical measures to analyze our results, such as the correlation matrix, the coefficients of our regressions, the F1 score, R-squared values, etc.

II. PROJECT OUTLINE

II.1. Data Source and Reference Data

- The data about all our social, demographic, and health variables will be found here: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?resource=download
  The data is from the **World Health Organization** and it contains all the data for all the countries from 2000 to 2015. There are some missing values in the dataset.
- All the other data of the other variables can be found on the **Food and Agriculture Organization** here:
  https://www.fao.org/faostat/en/#data/
  And we selected the data from 2000 to 2015 of all the countries to match with the other dataset. There are some missing values in the dataset.

II.2. Data Overview

- Climate Change Parameters:
  - Unique_Key - Unique identifier
  - Forests_Area - Area Covered Under Forest
  - Forests_Area _Units - Hectares
  - Year - Year
  - Country_Name - Name of the Country
  - Country_Area - Area of the Country
  - Temperature_Change - Change in Temperature
  - Temperature_Units - Celsius/Fahrenheit
  - CO2_Emission - Amount of Co2 Emission.

    - CO2_Emission_Unit - Metric tones
- Social / Health Condition Parameters:
    - Life_Expectancy - Life Expectancy in a country
    - Adult_Mortality- Adult Mortality of people used to findout employment.
    - Population - Total Population of the country in millions
    - Education_Background - Education background of people in millions
- Economic Conditions Parameters:
    - GDP - Gross Domestic Product of Country
    - GDP_Units - Millions/ Billions
    - Country_Expendiure_on_Agriculture - Annual Expenditure on Agriculture by country
    - ECA_Units - Millions/ Billions
- Crop Production Parameters:
    - Land_Used_for_Agriculture - Land Used for Agri Purposes by the country.
    - LUA_Units - Hectares
    - Fertilizers_Used_for_Area_of_Agri_land - Fertilizers used in cultivation of crops in terms of land.
    - FUAA_Units - Hectares
    - Employment_in_Agriculture - Number of employees used for agriculture.
    - EAF_Units - Per 1000 people

II.3. Data Pre-processing

- The data collected from the WHO and FAO sources may contain missing or inconsistent values, duplicates, or irrelevant variables. To prepare the data for analysis, we will clean it by identifying and handling these issues using techniques such as dropping rows or columns with missing values, filling missing values with mean or median imputation, and removing duplicates.
- We will then normalize the data in order to get help our models to perform better.
- Finally, to ensure the robustness of our analysis, we will probably identify and remove any outliers that could potentially skew our results.

II.4. Model Selection

- We will use various feature selection techniques, such as correlation analysis and recursive feature elimination (based on AIC or BIC for example), to identify the most significant features that affect agriculture production.
- We plan to use regression analysis to model the relationships between different variables and agriculture production. We will explore different regression approaches, such as classic linear regression, Ridge regression, LASSO regression, etc. And then we will choose the best one.

II.5. Tools

We plan to use the **R Studio** software and many libraries from the **CRAN** package such as ggplot2, glmnet, lmridge, etc.

We will also use **GitHub** to share the code and other files.

II.6. References

- https://www.fao.org/giews/food-prices/research/detail/en/c/235862/
- https://www.fao.org/giews/food-prices/research/detail/en/c/235864/
- https://www.fao.org/giews/food-prices/research/detail/en/c/277327/