Bayesian Network for Crop Yield Prediction

Alessandro Dario, Andrea Scanu, Riccardo Zerbinati

# Project Idea

Adapting to global warming is one of the greatest challenges for agriculture today. In this project, we developed a Bayesian Network model to predict crop yields under varying environmental conditions. By combining historical data on crop production, temperature, precipitation, soil composition, and extreme weather events, we built an interpretable probabilistic framework that estimates the likelihood of achieving a good harvest for each specific crop, based on user-defined constraints.

# Potential Impact

• Estimate the probability of low yields under forecasted weather conditions, supporting insurance and risk mitigation strategies.

• Identify the most suitable crops to cultivate in a given region and climate scenario.

• Promote adaptation strategies aligned with the UN Sustainable Development Goals (SDG 2: Zero Hunger, SDG 9: Industry, Innovation and Infrastructure, SDG 13: Climate Action).

# Data and Preprocessing

We collected open-source data from ISTAT and other public archives, covering urban indicators, climate variables (temperature and precipitation), soil properties, and agricultural statistics for Italian provinces over multiple years. The raw data were cleaned, standardized, and linearly interpolated to handle missing yearly records. All sources were then merged into a single, consistent database that directly links crop yields with local climate and soil conditions.

# Bayesian Modelling and Causal Discovery

Starting from this integrated dataset, we first built a **discrete Bayesian Network**. Continuous variables such as temperature, rainfall, or soil nutrients were discretized into five quantile-based categories (from q0, the lowest 20%, to q4, the highest 20%), allowing us to capture how different value ranges influence yield. For this, we focused on the **yield ratio**, which is the crop production divided by the productive land area. Using the pgmpy library, we applied **constraint-based structure learning** to automatically discover plausible causal relationships while enforcing logical constraints (e.g., soil variables cannot influence climate variables).

A diagram of a diagram of a plant

AI-generated content may be incorrect.

Figure 1: Discrete Bayesian network diagram.

This discrete model revealed, for example, that wheat does not respond well to dry days and high temperatures (consistent with known impacts of climate change) while heavy rain tends to have a smaller negative effect but increases yield variability. Moreover, all soil fertility factors show strong positive correlations with yield ratio, with nitrogen content being a key indicator. This network enables probabilistic inference and generates visual outputs such as probability heatmaps to illustrate how the chance of high or low yield ratios changes with specific factors. For each crop, we first trained the model on 80% of data and tested on the remaining 20%, obtaining an average precision of 54% on crops with at least 100 data samples; around 70% on wheat or other crops with large amount of data. We then trained on the whole dataset.

A chart of a number of colors

AI-generated content may be incorrect.A chart of a temperature

AI-generated content may be incorrect.

Figure 2: Heatmap P(yield\_ratio|dry\_days) Figure 3: Heatmap P(yield\_ratio|mean\_temp)

A chart of a number of colors

AI-generated content may be incorrect.A chart of a number of colors

AI-generated content may be incorrect.Figure 4: Heatmap P(yield\_ratio|heavy\_rain\_days) Figure 5: Heatmap P(yield\_ratio|soil\_nitrogen)

After validating the causal relationships with the discrete model, we extended our approach with a **continuous Bayesian Network** to capture more detailed numeric effects. Here, using Linear Gaussian CPDs and a score-based learning algorithm (Hill Climb Search with BIC scoring), we modeled the yield-to-productive-area ratio as a continuous target variable.

A diagram of a diagram of a plant

AI-generated content may be incorrect.

Figure 6: Continuous Bayesian network diagram.

This variant enabled us to analyze direct quantitative impacts of each climate or soil variable through regression plots and updated causal graphs. In summary, we first discovered the causal structure with the discrete model, then refined it using continuous variables to gain deeper insights creating a robust, interpretable tool for informed and resilient agricultural planning.

A graph showing the effect of a temperature

AI-generated content may be incorrect.

Figure 7: Scatter plot mean temperature on yield ratio.

The data reveals a concerning trend: as temperatures rise, crop yields tend to decrease. This negative correlation underscores the impact of **climate change**, which is causing more frequent heat waves and unpredictable weather patterns. These changes threaten agricultural productivity, posing serious challenges for food security and sustainable farming.