INFO H-501 proposal

Bofu Dong[1], Krish, Shah[1], Shruti Sujit Sonawane[1], Varun, Rapolu[1], Mayank Gupta[1]
[1]Indiana University Indianapolis
bofudong@iu.edu

1 Can machine learn about constellation

1.1 Problem Statement

For centuries, people have looked up at the night sky and connected stars into constellations. These shapes are not dictated by physics but by human imagination: a small shift in brightness or spacing can change the figure entirely. Today we have precise measurements from projects like Gaia DR3 (Gaia Collaboration, 2022), which record where stars are in the sky, how bright they are, and even their color. This lets us ask: can a computer take this information and draw its own constellation-like patterns?

In our problem setting, each star comes with a few key pieces of information. The first is its position in the sky, given by two coordinates called right ascension and declination–like longitude and latitude. The second is brightness, measured in what astronomers call "G magnitude," which tells us how bright the star looks from Earth. Finally, we can use color, expressed through a "BP–RP index," which is simply the difference between how bright the star looks in blue light and in red light. Together these values let us picture not just where a star is, but also how it appears.

The challenge we set ourselves is to take this information and connect the stars with lines based on simple rules–such as only linking stars that are close enough, or limiting how many lines can come out of each star. By adjusting these parameters, we can see how the resulting patterns change. Sometimes the shapes may look sparse and chain-like, other times dense and web-like. Comparing these results against traditional constellation maps, and also looking at the statistical properties of the graphs we generate, will help us understand how different choices influence what kinds of "constellations" emerge.

Put simply, the problem is to explore whether a machine can recreate the human idea of constellations using nothing more than star positions, brightness, and color, and to see how sensitive those patterns are to the rules we give it.

1.2 Data Preference

For this project, we will use a curated subset of the Gaia Data Release 3 (DR3) catalogue, available through Kaggle[1]. The official documentation for this release is provided in the Gaia DR3 summary paper (Gaia Collaboration, 2022).

Detailed description. Gaia DR3 is one of the most comprehensive stellar catalogues ever compiled. It contains precise astrometric, photometric, and astrophysical data for more than 1.8 billion celestial objects, including stars, galaxies, quasars, and binaries. The Kaggle subset we will use provides over 600,000 rows of individual star entries with 50+ columns describing their properties.

Purpose of collection. The Gaia mission, launched by the European Space Agency, aims to create the most accurate 3D map of the Milky Way by measuring the positions, motions, distances, and brightness of stars with unprecedented precision. This information supports research in stellar evolution, galactic dynamics, and cosmology.

Column explanations. A selection of key columns from the dataset includes:

- RA_ICRS, DE_ICRS: The coordinates of each star on the sky (right ascension and declination), analogous to longitude and latitude on Earth.
- Gmag: The brightness of a star as seen from Earth in a broad visible-light filter (G band).
- BPmag, RPmag: Brightness measured through blue (BP) and red (RP) filters, used to estimate stellar color.

[1]Gaia DR3 Stars Dataset on Kaggle

- BP–RP: The color index, calculated as the difference between BP and RP magnitudes; bluer stars have smaller values, redder stars larger.
- Plx (parallax): A measurement used to infer distance to the star in parsecs.
- pmRA, pmDE: Proper motion components, describing how the star appears to move across the sky in right ascension and declination directions.
- SpType–ELS: Estimated spectral type, such as "G" or "M," which categorizes stars by temperature and color.
- Pstar, PGal, PWD, Pbin: Probabilities that the object is a star, galaxy, white dwarf, or binary system.

There are more columns for astrological analysis, but those will not be used for this project, for example: Evol stands for Evolution stage of the star.

Source. All measurements originate from the European Space Agency's Gaia mission, with the DR3 catalogue published and described by Gaia Collaboration, 2022. The Kaggle dataset is a repackaged subset of this release for easier access and analysis.

1.3 Explanation

Use and value. This project addresses the problem by turning the Gaia DR3 star data into constellation-like patterns that can be explored interactively. By plotting stars using their positions and visual properties, and then connecting them with edges according to simple geometric rules, we can see how different choices lead to different shapes. This allows us to study how constellations emerge from basic data and to communicate those patterns visually in a way that is easy for others to understand. The application has value for education, because it demonstrates how star maps are built, and for exploration, because it highlights how subjective rules can generate different yet plausible constellations.

Goals. The main goals are:

- Build a pipeline that takes Gaia DR3 star positions and generates graphs connecting stars.
- Provide visual outputs, including all-sky views and small patches, with edges overlaid on the star positions.
- Allow parameter variation, such as changing the maximum edge length, the number of allowed connections per star, or brightness filters, to show how the resulting constellations change.
- Compare the generated graphs against known constellation line sets or analyze their structure with basic graph statistics.
- Build website that allow user to modify parameters and observe changes in constellations.

Constraints. There are several practical limits to consider:

- The full Gaia DR3 catalogue is extremely large, so we will intentionally use the Kaggle subset and may still need to sample or filter stars for manageable computation and clear visuals.
- Parallax-derived distances can be noisy, so the primary focus will be on angular positions and brightness rather than precise 3D structure.
- The project will rely on classical geometric and graph methods, which means the results will be interpretable but may not capture the full richness of human constellation-making traditions.

2 Can machine taste wine?

2.1 Problem Statement

Wine tasting has traditionally been considered a human skill that blends chemistry, sensory experience, and cultural knowledge. A glass of wine can be judged by its aroma, flavor, balance, and finish, and these judgments are often summarized into a single "quality score." But what if a machine could do the same? Using chemical measurements instead of taste buds, can we predict how a panel of human tasters would rate the quality of a wine?

The Wine Quality dataset[2] (Cortez et al., 2009b) provides exactly this opportunity. For each sample of red or white Portuguese wine, the dataset records chemical properties such as acidity, sugar, pH, sulfur content, alcohol, and more, along with a quality score assigned by human experts. The challenge is to see whether patterns in these measurable variables can be linked to the subjective quality ratings.

The problem we address is to build and evaluate a model that predicts wine quality from its chemical features. This will allow us to explore whether a machine can "taste" wine in a numerical sense, by learning the connection between laboratory measurements and human judgment. The project also opens questions about which chemical factors matter most in shaping perceived quality, and whether such predictions can be accurate enough to be useful in wine production or quality control.

2.2 Data Preference

For this project, we will use the Wine Quality dataset (Cortez et al., 2009b)[2]. The dataset was originally described and analyzed in Cortez et al., 2009a.

Detailed description. The dataset contains physicochemical measurements of both red and white variants of Portuguese "Vinho Verde" wine. It includes over 6,000 rows, with each row corresponding to a unique wine sample and 11 measured input variables. Alongside these variables is a target column representing the quality score assigned by a panel of human tasters, rated on a scale from 0 (very poor) to 10 (excellent).

Purpose of collection. The dataset was compiled to study the relationship between measurable chemical properties of wine and subjective human quality assessments. It serves as a benchmark for regression and classification tasks in data mining, and has been widely used to explore predictive modeling of sensory evaluation.

Column explanations. Key variables in the dataset include:

- Fixed acidity, volatile acidity, citric acid: measures of the wine's acid composition.
- Residual sugar: amount of sugar left after fermentation.
- Chlorides: concentration of salt in the wine.
- Free sulfur dioxide, total sulfur dioxide: measures of sulfur dioxide levels, which affect preservation.
- Density: relative density of the wine compared to water.
- pH: measure of acidity or alkalinity.
- Sulphates: concentration of sulphates, which can affect flavor.
- Alcohol: percentage of alcohol by volume.
- Quality: the target variable, an integer score given by human tasters.

Source. All measurements originate from Portuguese wine samples collected for research on modeling wine preferences, with the dataset curated and distributed through the UCI Machine Learning Repository (Cortez et al., 2009b)[2]. The accompanying study and documentation are described by Cortez et al., 2009a.

2.3 Explanation

Use and value. This project uses the Wine Quality dataset (Cortez et al., 2009b)[2] to test whether chemical measurements of wine can be used to reproduce the quality scores normally given by human tasters. By building predictive models on the dataset, we can evaluate how well a machine can "taste" wine in a numerical sense. The application has practical value for winemaking and quality control, where quick chemical tests may serve as a substitute for costly or inconsistent human evaluation. It also has educational value, since it demonstrates the relationship between measurable chemical properties and subjective human judgment.

[2] UCI Machine Learning Repository: Wine Quality

Goals. The main goals of the project are:

- Train and evaluate predictive models that map wine chemistry to human-rated quality scores.
- Identify which chemical properties are most strongly associated with quality, providing insights into how composition influences taste.
- Compare different modeling approaches (e.g., regression versus classification) to see which captures the relationship most effectively.
- Produce clear visualizations that show how variables such as acidity, alcohol, and residual sugar correlate with perceived quality.
- Create a web application to demonstrate above examples interactively.

Constraints. There are some limitations that must be acknowledged:

- The dataset is limited to Portuguese "Vinho Verde" wines and may not generalize to other wine regions or styles.
- Quality scores are based on a small panel of tasters, so they may contain subjective bias and noise.
- The quality ratings are integers from 0 to 10, but in practice most scores cluster between 3 and 8, creating class imbalance.
- Models will be based only on chemical features and cannot capture sensory aspects such as aroma or texture, which may also influence quality.

References

P. Cortez, Antonio Luíz Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009a. Modeling wine preferences by data mining from physicochemical properties. *Decis. Support Syst.*, 47:547–553.

P. Cortez, Antonio Luíz Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009b. Wine Quality. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56S3T.

Gaia Collaboration. 2022. Gaia data release 3: Summary of the content and survey properties. *Astronomy & Astrophysics*, 674:A1.