HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 3 (2024)

Using Unsupervised Machine Learning to Find the Milky Way's Components

Eshan Guha

High Technology High School, Manalapan, New Jersey, United States

ABSTRACT

Understanding the distributions of stars and their metallicities tells us about the formation of the Milky Way and how it has interacted with satellite galaxies in the past. Three distinct sectors lie in the Milky Way—the thin disk, thick disk, and halo. Because there exists significant overlap in various parameter spaces such as velocity and metallicity, it has been difficult to disentangle those components in the past aside from using empirical methods. In this study, unsupervised machine learning techniques were applied to a Gaia + APOGEE dataset and used to identify the components of the Milky Way. The resulting model was compared to prior methods, highlighting the possible difficulties resulting from applying rigid cuts. Regression was used to analyze the trend in metallicity ratios, which suggests that the rate of supernovae in the Milky Way has changed in its history. The Initial Mass Function and the percentage of halo stars generated from the model were used to approximate the number of neutron stars in the Milky Way. Overall, this study shows that unsupervised machine learning techniques enables the discovery of new trends in the Milky Way's components.

Introduction

There has been a growing interest in understanding the distribution of stars throughout the Milky Way. Previous research and examination has shown that there are three distinct sectors within the Milky Way—the thin disk, the thick disk, and the halo (Kilic et al., 2017). Understanding the distribution of stars across these sectors can have important applications in understanding the age of the Milky Way, its potential merger history with other galaxies, the likely exit velocities gained by neutron stars and black holes during their formation. Through these findings, it is also possible to investigate the metallicity of various stars. In astronomy, metals are defined as anything heavier than helium.

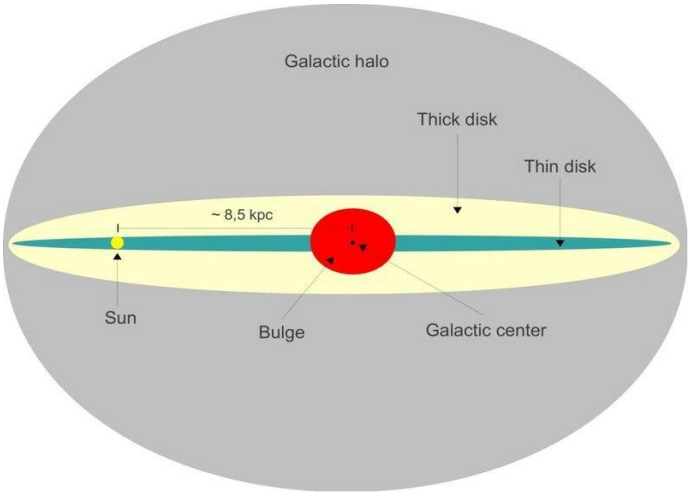

Figure 1. Diagram of the thin disk, thick disk, and halo of the Milky Way (open source figure).

When viewing the Milky Way from a side-on perspective, the thin disk is the portion of stars closest to the horizontal, the thick disk stars are slightly farther away, and the halo stars are few in number, but occupy a vast amount of space surrounding the disks. Stars in the thin and thick disks tend to contain more metals because they are younger stars. Stars in the halo tend to be older, developing at a time in which there are less metals, hence having a lower metallicity (Kilic et al., 2017). The majority of the Milky Way's stars are located in the thin and thick disk, but the galactic halo still makes up an estimated five percent of the total number of stars in the MIlky Way (Kilic et al., 2017). The Toomre diagram is a way to visually differentiate between thin disk stars, thick disk stars, and halo stars. The Toomre diagram has V, essentially the substitute for Y in the Cartesian coordinate system, on the x axis, and the non-azimuthal velocity on the y axis. Non-azimuthal velocity is represented by the equation $(U^2 + W^2)^{1/2}$ with U and W representing the X and Z velocity components of a Cartesian coordinate system. When plotted on the Toomre diagram, the stars closest to the point (0,0) are thin disk stars, The stars slightly farther away are thick disk stars, and finally, the furthest away are halo stars (Kilic et al., 2017).

Not all stars in the Milky Way have the same mass. Salpeter discovered the Initial Mass Function, also referred to as the Salpeter function, which is represented by the following equation: $\xi(M) = \xi 0 * M^{-2.35}$ (Chabrier, 2005). In this equation, ξ_0 is a constant, setting the local star density. M represents the mass of the star. From the IMF function, an expression can be set up to represent the total number of stars formed between masses M_1 and M_2: $N = \xi \int_{M1}^{M2} M^{-2.35} dM$.

The Milky Way is known to contain over 50 satellite galaxies. Satellite galaxies are miniature galaxies that are located within a larger galaxy. Much like how the Earth orbits the Sun, satellite galaxies orbit the center of the Milky Way. Merger galaxies are small galaxies that collide and merge with a larger galaxy. Prior studies have shown that five mergers occurred in the Milky Way's past, but a recent study done using the Gaia dataset has shown evidence for a sixth ("History of Milky Way Mergers revealed in Gaia data"). Satellite galaxies can have important applications in understanding dark matter, as dark matter has been proposed as a possible explanation for the existence of satellite galaxies with gravity that in theory would be too weak to maintain them (Cooper, 2022). Additionally, if more mergers exist in the Milky Way than previously estimated, further constraints can be placed on the total amount of dark matter in the Milky Way.

Astronomical spectroscopy is the way the metallicities of stars and the distance of stars is calculated. Stars reflect certain wavelengths of light that help determine the distance and metallicities. Different wavelengths indicate certain metallicities, and then the amount of red shift or blue shift helps determine their proximity to Earth ("Spectroscopy").

Methods

Data from the APOGEE telescope and Gaia telescope were used in this study. The Gaia telescope is space-based, while the APOGEE survey is part of the ground-based Sloan Digital Sky Survey (SDSS) telescope. APOGEE was used to obtain the metallicities of the stars, while Gaia was used to measure the distances of the stars("APOGEE", "Gaia overview"). Data from these two datasets was combined to form a merged Gaia and APOGEE dataset.

Before data preprocessing, there were 657,214 stars, with 46 fields, or features. Two additional fields were added to this: non-azimuthal velocity and total velocity. Non-azimuthal velocity was calculated by using this equation: $(U^2 + W^2)^{1/2}$. The total velocity was calculated by using this equation: $(U^2 + W^2 + V^2)^{1/2}$. Next, the data underwent preprocessing. After removing stars with null values and larger parallax errors, there were 101,092 stars. Parallax is the inverse distance of the star from the Sun.

The Gaia and APOGEE dataset contained 46 fields. 2 additional fields were added to better visualize the data: non-azimuthal velocity and total velocity. The other fields included positions on the sky, parallax, average velocity with respect to the Sun, proper motion, true velocity, in space in 3D coordinates, data flags, average velocity standard deviation, average velocity uncertainty, optical brightness, the distance of the star from Earth in parsecs, temperature and error, surface gravity and error, and various metallicities. All the data was numerical, with the data flag field using only zeros and ones instead of booleans.

Non-azimuthal velocity and total velocity are different ways of representing the star's velocity in the galaxy, as mentioned previously. The most relevant fields for this study are the velocity of the star, and Fe/H and Mg/Fe metallicity ratios. Metallicity ratios are usually measured in logarithmic scale and that convention is adopted in this paper. Fe/H is the iron to hydrogen ratio and Mg/Fe is the magnesium to iron ratio. The metallicity ratios help to determine the relative age of some of the stars. For instance, stars with a lower Fe/H ratio would contain a lot more hydrogen per iron. As a result, it is likely that these stars formed earlier in the Milky Way's history, and possibly belong to the halo. The velocities of the stars are important in determining if a star got shot out of the disks and into the halo (Lai, 2001). Additionally, one can often be able to predict orbits from the velocity components.

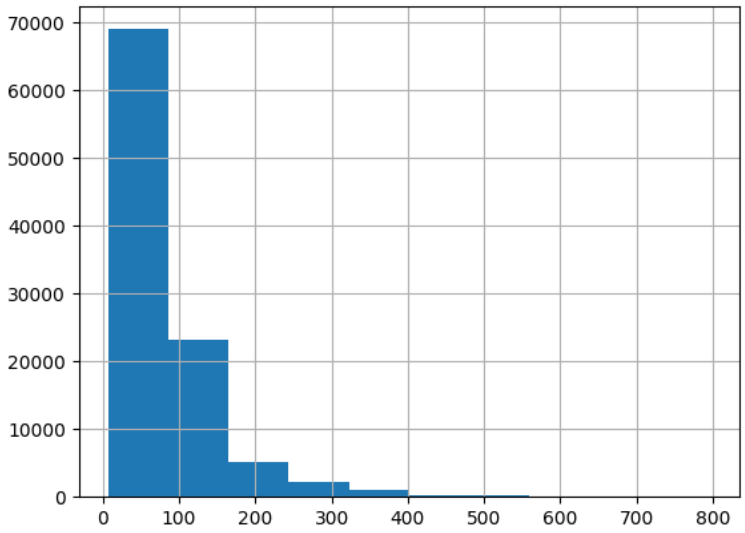

Figure 2. Total velocity histogram. Frequency is on the y-axis and velocity in km/s is on the x-axis.

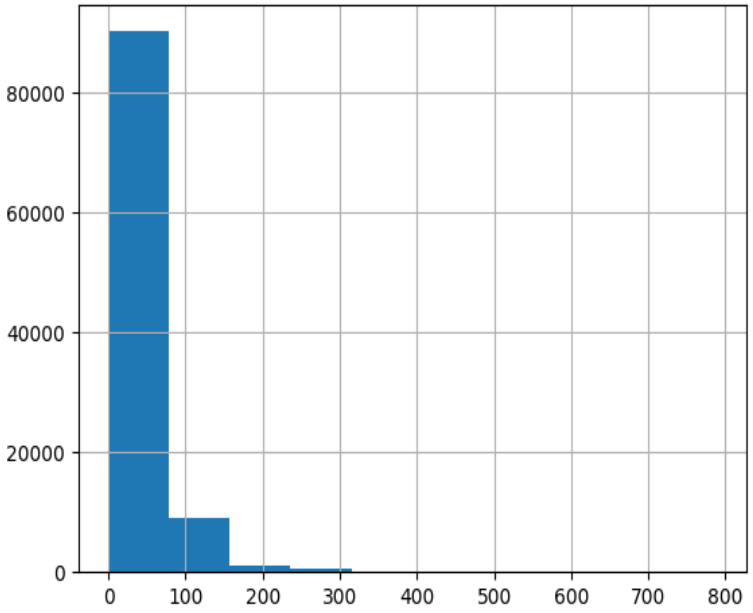

Figure 3. Histogram of the average velocity with respect to the Sun. Frequency is on the y-axis and average velocity in km/s is on the x-axis.

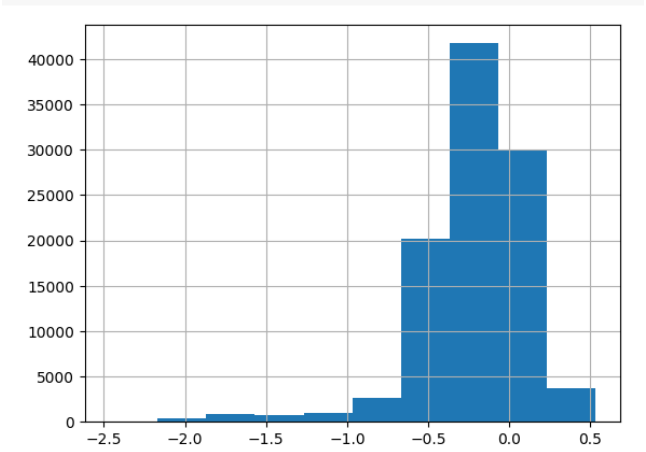

Figure 4. Histogram of the Fe/H metallicity ratio. Frequency is on the y-axis and the metallicity is on the x-axis.

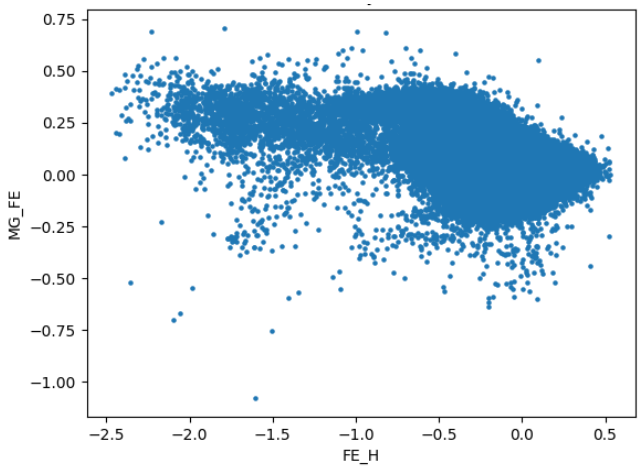

Figure 5. Scatterplot of Fe/H metallicity ratio on x-axis and Mg/Fe metallicity ratio on the y-axis.

Google Collab was used to code and analyze the data. The data was downloaded from the Gaia and APOGEE datasets. Once the data was downloaded, two additional features were added: non-azimuthal velocity and total velocity. In total, there were 657,214 stars. Various libraries were imported from scikit-learn including GaussianMixture, silhouette_score, calinski_harabasz_score, davies_bouldin_score, and LinearRegression. Numpy and Matplotlib were also used.

Next, the Toomre diagram was added to better visualize the data and compare the shape of the data to the Galaxy Halo paper. The result was fairly similar, but there were many outliers, likely caused by faulty data with a large error. In particular, because the APOGEE data comes from a ground-based telescope, it is subject to more potential errors due to weather and atmosphere interference with observations ("APOGEE"). So, the next step was to clean the data to keep only accurate data points.

To remain in the dataset, stars had to pass several requirements. Stars had to have a parallax error of less than one-fifth of the parallax. In other words, the following condition would have to be true: $\frac{Parallax}{Parallax\ Error} > 5$. Additionally, stars had to have a data flag of 0, meaning that the data was of good quality. Furthermore, the average velocity standard deviation of the star would have to be greater than the average velocity uncertainty. The error of the average velocity of the star with respect to the Sun would have to be less than one-third of the average velocity of the star with respect to the Sun. This means that the following relation would have to be true: $\frac{V_{helio}}{V_{error}} > 3$ where V_{helio} represents the average velocity of the star with respect to the Sun and V_{error} represents the error of V_{helio}. Finally, the data was required to have a non-null value for the iron to hydrogen metallicity and magnesium to iron metallicity, since these were the most important metallicities used in the study.

Data that did not fulfill the requirements of the aforementioned criterion was removed from the dataset. This left a total of 101,092 stars. Once the data cleaning was complete, the Toomre diagram was reconstructed. It had the appropriate shape, indicating that the data was accurate.

Next, histograms of the total velocity, distance, average velocity, average velocity standard deviation, and Fe/H were created to better visualize the data. A scatterplot was also created with Fe/H on the x axis and Mg/Fe on the y axis.

The next step was to apply machine learning to find the clusters in the data. To do this, unsupervised machine learning techniques would have to be used because there was no data field explicitly stating whether each star was part of the thin disk, thick disk, or halo. Three different machine learning models were tried: Gaussian Mixture Model, DBSCAN, and Decision Tree.

The Gaussian Mixture Model works using probabilities. The user inputs the specified number of clusters, covariance, and weight parameter. Then, GMM calculates the probability of each point belonging to a certain cluster. It uses these probabilities to create an accurate clustering model. In addition, a major assumption of the Gaussian Mixture model is that different clusters in the data can be described by multivariate Gaussian distributions, with the mean of each distribution corresponding to the center of each cluster ("Gaussian Mixture Model Explained").

DBSCAN works by clustering points that are densely packed together. As a result, it can sometimes prove ineffective with data that is densely packed, since the algorithm may be unable to identify the differences in density ("DBSCAN Clustering Algorithm Demystified").

These three techniques were first applied on the Fe/H and Mg/Fe plots. Then, it was expanded to include non-azimuthal velocity, total velocity, distance, and effective temperature. After this was completed, they were also used on the Toomre diagram. Silhouette scores for each of these models were recorded and kept.

Next, the parameters were revised to contain Fe/H, Mg/Fe, V, and non-azimuthal velocity. Then, the DBSCAN model was rerun and the Gaussian Mixture Model was also rerun. The GMM was rerun with specified clusters n for n = 2,3,4,5,6,7,8,9. The silhouette score was reported for each model in addition to the Calinski-Harabasz and Davies-Bouldin scores. The C-H score and D-B score are density-based metrics that evaluate the model's performance.

Thereafter, the previously theorized line separating the disk stars from the halo stars was graphed on top of the clustered GMM model where n = 3. Next, calculations were made to determine how many disk stars crossed over the line. This was done by looping through each of the disk stars and recording every time a star crossed the line. The disk stars were separated using the clusters created by the GMM model where n = 3. After finding the number of disk stars that crossed the line, they were separated into thin disk stars and thick disk stars. Then, a percentage was calculated using the total number of thin disk stars and thick disk stars, representing the percent of stars that adhere to the theorized line.

Linear regression was then used on the Fe/H, Mg/Fe graph. First, the linear regression model was done on the whole dataset. Then, it was done for two separate sections: Fe/H < -0.4 and Fe/H < -0.4. Finally, it was done across thirteen individuals intervals from -2.1 to 0.5 in intervals of 0.2. The mean squared error and mean absolute error was calculated for each of the regression models. The average mean absolute error of the thirteen intervals was calculated. Then, each model was ranked from highest mean absolute error to lowest mean absolute error.

Next, limits of the number of neutron stars in the Milky Way were calculated. To do this, it was necessary to find the percentage of halo stars in the Milky Way. The individual number of halo stars was divided by the total number of stars and multiplied by 100 to yield a percent. It is estimated that there are 10^{11} stars in the Milky Way, so the percentage was converted to a decimal and multiplied by 10^{11}. This gives the estimated number of halo stars in the Milky Way, thereby setting the maximum threshold for the number of neutron stars in the MIlky Way.

The lower limit of the number of neutron stars in the Milky Way can be calculated using the Initial Mass Function. From it, the following equation can be derived: $N = \xi \int_{M_1}^{M_2} M^{-2.35} dM$ where N is the number of stars, M_1 and M_2 are the masses of a star, and is a constant. To use this function effectively, the value for the constant must be found. Since N is the number of the stars, it can be set to 10^{11}, the approximate number of stars in the Milky Way, and the lower and upper bounds of the mass can be set to 0.07 and 300 solar masses respectively. Once finding the value of the constant, it can then be plugged back in the equation. The goal is to find the number of neutron stars, so N is unknown and M_1 and M_2 can be set to 8 and 15 solar masses respectively, since that's the approximate number of solar masses for neutron star formation. Then, the equation can be solved for N, resulting in the lower limit for the number of neutron stars in the Milky Way.

Results

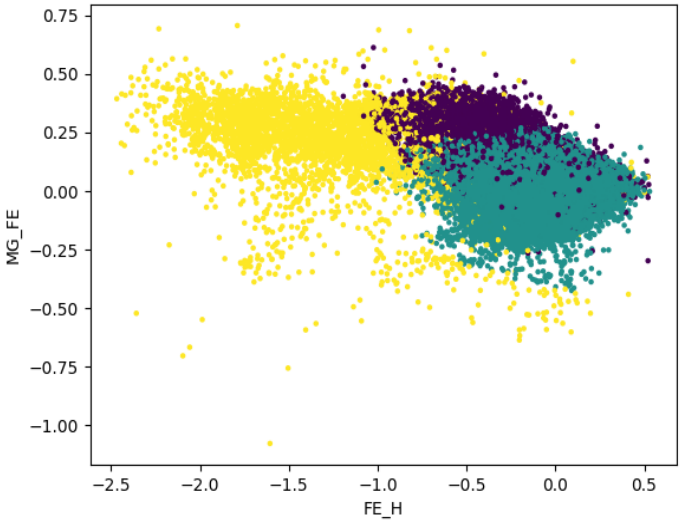

Figure 6. Scatterplot of clusters generated by Gaussian Mixture Model for $n = 3$.

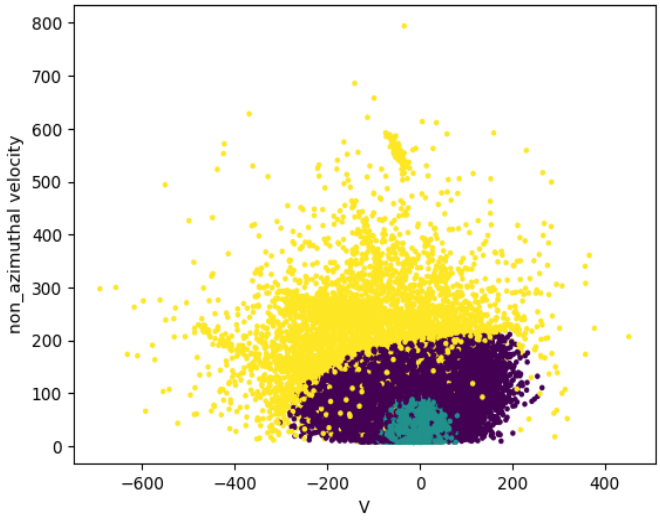

Figure 7. Toomre diagram with clusters generated by Gaussian Mixture Model for $n = 3$.

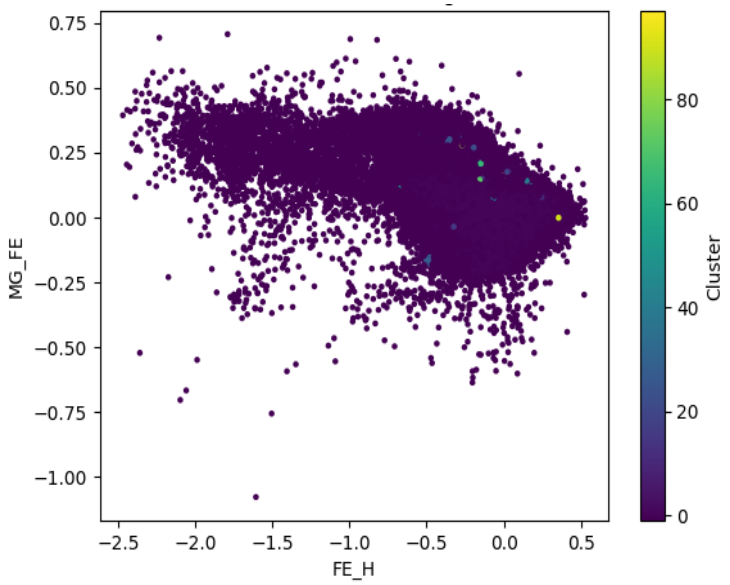

Figure 8. Scatterplot of clusters generated by DBSCAN model.

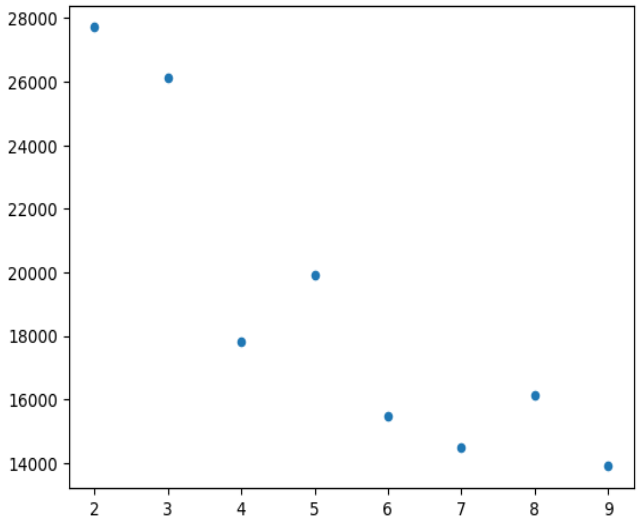

Figure 9. Scatterplot of C-H scores (y-axis) of Gaussian Mixture Models for $n = x$.

Table 1. Silhouette, C-H, and D-B scores for Gaussian Mixture Models from n =2 to n = 9, rounded to 2-3 significant figures.

	2	3	4	5	6	7	8	9
Silhouette	0.545	0.335	0.225	0.116	0.062	0.038	0.030	-0.008
C-H	27700	26100	17800	19900	15500	14500	16100	13900

D-B	1.64	1.91	2.22	1.79	2.26	2.29	2.78	2.89

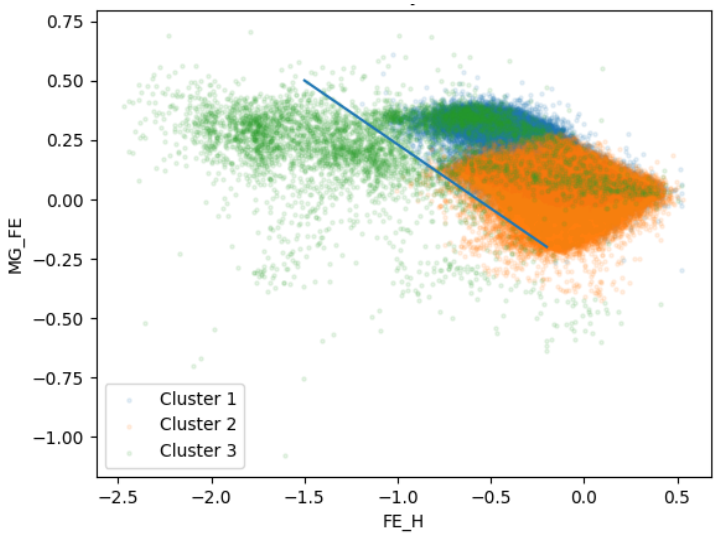

Figure 10. Line drawn over scatterplot of Gaussian Mixture Model for $n = 3$. Cluster 1 is the thin disk, cluster 2 is the thick disk, and cluster 3 is the halo.

Table 2. Table of the line's accuracies for the thin disk and the thick disk.

	Thin Disk	Thick Disk
Accuracy	99.9%	98.5%

Table 3. Number of stars belonging to each cluster (thin disk, thick disk and halo).

	Thin Disk	Thick Disk	Halo
Number	29247	66542	5303

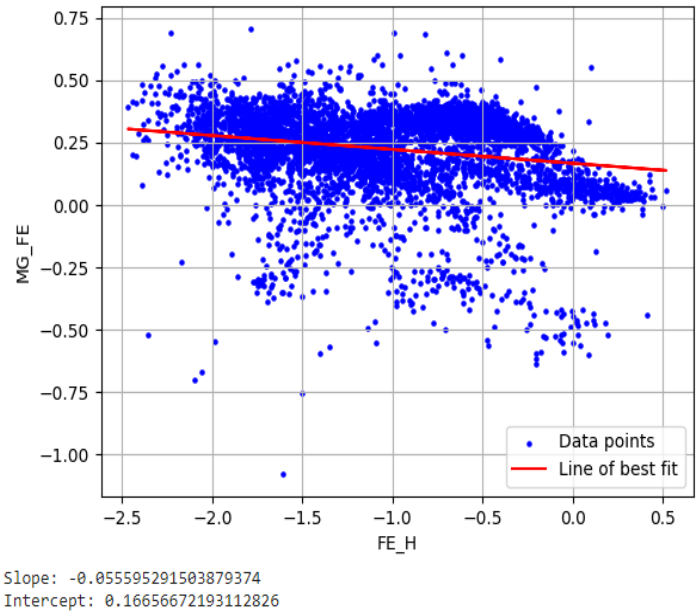

Figure 11. Best line of fit for the scatterplot of Fe/H on x-axis and Mg/Fe on y-axis.

Table 4. Mean absolute error with one line of best fit, two lines of best fit split at x= -0.4, and thirteen lines of best fit.

	1 Line	2 Lines	13 Lines
Mean Absolute Error	0.122	0.114	0.101

Discussion

The results from the study have significant impacts in understanding the composition of the Milky Way. Firstly, the shapes of the Fe/H and Mg/Fe plot and the Toomre diagram both resemble the plots shown in various other papers (Mackereth et al., 2018). This means that the data was accurate. For the Fe/H and Mg/Fe plot, the Gaussian Mixture Model came up with clusters that roughly resembled the clusters shown in various other papers (Mackereth et al., 2018). DBSCAN, on the other hand, only found one significant cluster, likely representing the halo stars and the disk stars, but it didn't resemble the clusters from other papers well. Additionally, while silhouette scores were calculated, it was quickly realized that these would be ineffective, since they measure the average distances between the points. Since all the points were relatively close together, density based model metrics may be better suited. As a result, C-H score and D-B score was computed after for later samples. The Decision Tree algorithm only came up with one cluster, so it was clear that this algorithm would be ineffective for the data.

After creating several models for the Fe/H and Mg/Fe plot, more parameters were included to try and increase the model's performance on a wider scale. At first, non-azimuthal velocity, total velocity, and effective temperature were added. However, this actually slightly reduced the silhouette scores, meaning that the addition of these variables was unnecessary.

Instead, Fe/H, Mg/Fe, V, and non-azimuthal velocity were used. Using these parameters resulted in a higher silhouette score. Additionally, it was identified that the Gaussian Mixture Model was doing the best at representing the data, since DBSCAN continued to only find one cluster. Furthermore, it was decided to look predominantly at C-H scores and D-B scores since they are density-based and would likely represent the model's

efficacy better. The GMM model for $n = 2$ had the best C-H and D-B scores, but this may be because there is a more clear distinction between halo stars and disk stars than the halo, thin disk, and thick disk. The GMM model for $n = 3$ had only a slightly worse C-H and D-B scores, but a major drop came with the GMM model for $n = 4$ (Figure 9, Table 1). This indicates that the model is significantly less effective with four clusters than three clusters. Since there are known to be three clusters in the MIlky Way—the thin disk, thick disk, and halo—the GMM model for $n = 3$ is the best representation of the clusters. Additionally, the clusters developed by this model visually appear to be the most similar to prior papers (Mackereth et al., 2018).

The approximate line used in prior papers to differentiate between halo stars and disk stars was graphed on the Fe/H and Mg/Fe plot as shown in Figure 11 (Mackereth et al., 2018). 99.9% of the thin disk stars were on the right side of the line, along with 98.5% of the thick disk stars (Table 2). This meant that the line was 98.9% accurate, since 98.9% of the stars in the model were on the correct side of the line. Thus, the model supports the line being an accurate separation point for disk stars and halo stars.

The linear regression models for the thirteen individual sections of the data performed the best since they had the lowest average mean absolute error (Table 4). The linear regression models for the two individual sections performed the second best, while the overall model performed the worst. The general declining trend in the slope may indicate an increase in the number of supernovas taking place due to the increasing proportion of Fe/H to Mg/Fe. This is also where a lot of the halo stars may lie.

Using the Initial Mass Function, the upper and lower limits of the number of neutron stars in the Milky Way was found. The upper limit is the total number of halo stars in the Milky Way, while the lower limit uses the Initial Mass Function and assumes that neutron stars only form when they have the mass of 8 to 15 solar masses. The upper limit was found to be $3.73 * 10^{10}$ (this is also the approximate number of halo stars in the Milky Way) and the lower limit was found to be $9.53 * 10^{7}$ (both rounded to 3 significant figures). This means that there should be between $9.53 * 10^{7}$ and $3.73 * 10^{10}$ neutron stars in the Milky Way.

Conclusion

The Gaussian Mixture Model for $n = 3$ worked the best in clustering the data; it took Fe/H, Mg/Fe, V, and non-azimuthal velocity as parameters. Density-based metrics were used to gauge the efficacy of the model, since data points were very close together.

The approximate line used to define the border between halo stars and disk stars was very accurate. 99.9% of the thin disk stars were on the right side of the line, along with 98.5% of the thick disk stars. This meant that the line was 98.9% accurate, since 98.9% of the stars in the model were on the correct side of the line.

Additionally, the linear regression models of the thirteen sections of the Fe/H and Mg/Fe plot were the most accurate. The general declining trend in the slope may indicate an increase in the number of supernovas taking place due to the increasing proportion of Fe/H to Mg/Fe.

Finally, the Initial Mass Function was used to approximate the total number of neutron stars in the Milky Way. It was found that there should be between $9.53 * 10^{7}$ and $3.73 * 10^{10}$ neutron stars in the Milky Way.

Future studies may include experimenting with different types of clustering models, as well as using different metallicity ratios as parameters.

Acknowledgments

I would like to thank Mr. Antonio Rodriguez for being my mentor and guiding me towards the completion of this project. I would also like to thank the Inspirit team for their continued support and cooperation.

References

APOGEE | SDSS. (n.d.). Www.sdss4.org. https://www.sdss4.org/dr17/irspec/

Bonaca, A., Conroy, C., Wetzel, A., Hopkins, P. F., & Kereš, D. (2017). Gaia reveals a metal-rich, in situ component of the local stellar halo. The Astrophysical Journal, 845(2), 101.

Blancato, K., Ness, M., Johnston, K. V., Rybizki, J., & Bedell, M. (2019). Variations in α-element Ratios Trace the Chemical Evolution of the Disk. The Astrophysical Journal, 883(1), 34.

Chabrier, G. (2005). The initial mass function: from Salpeter 1955 to 2005. The Initial Mass Function 50 Years Later, 41-50.

Cooper, Keith. (2022, December 20). Strange arrangement of Milky Way's groupie galaxies may undermine dark matter. Space.com. https://www.space.com/milky-way-dwarf-galaxies-alignment-dark-matter

DBSCAN Clustering Algorithm Demystified. (n.d.). Built In. Retrieved May 29, 2024, from https://builtin.com/articles/dbscan#:~:text=Its%20effective%20at%20identifying%20and

EarthSky | History of Milky Way mergers revealed in Gaia data. (2022, February 17). Earthsky.org. https://earthsky.org/space/history-of-milky-way-mergers-revealed-in-gaia-data/

Gaia overview. (n.d.). Www.esa.int. https://www.esa.int/Science_Exploration/Space_Science/Gaia_overview

Gaussian Mixture Model Explained. (n.d.). Built In. https://builtin.com/articles/gaussian-mixture-model#:~:text=A%20Gaussian%20mixture%20model%20is%20a%20soft%20clustering%20technique%20used

Kilic, M., Munn, J. A., Harris, H. C., von Hippel, T., Liebert, J. W., Williams, K. A., ... & DeGennaro, S. (2017). The ages of the thin disk, thick disk, and the halo from nearby white dwarfs. The Astrophysical Journal, 837(2), 162.

Lai, D. (2001). Neutron star kicks and asymmetric supernovae. In Physics of Neutron Star Interiors (pp. 424-439). Berlin, Heidelberg: Springer Berlin Heidelberg.

Mackereth, J. T., Schiavon, R. P., Pfeffer, J., Hayes, C. R., Bovy, J., Anguiano, B., ... & Fernández-Trincado, J. G. (2018). The origin of accreted stellar halo populations in the Milky Way using APOGEE, $\textit {Gaia} $, and the EAGLE simulations. arXiv preprint arXiv:1808.00968.

Spectroscopy | Center for Astrophysics. (n.d.). Www.cfa.harvard.edu. https://www.cfa.harvard.edu/research/topic/spectroscopy

Zhao, Y., Gandhi, P., Dashwood Brown, C., Knigge, C., Charles, P. A., Maccarone, T. J., & Nuchvanichakul, P. (2023). Evidence for mass-dependent peculiar velocities in compact object binaries: towards better constraints on natal kicks. Monthly Notices of the Royal Astronomical Society, 525(1), 1498-1519.