Galaxy Classification Using Discriminants

---

by
SANTTOSH A/L MUNIYANDY

SCHOOL OF PHYSICS
UNIVERSITI SAINS MALAYSIA

2024

Galaxy Classification Using Discriminants

---

by
SANTTOSH A/L MUNIYANDY

Thesis submitted in fulfilment of the requirements
for the degree of
Bachelor of Science (Hons.) Physics

July 2024

Author:
Santtosh a/l Muniyandy

Supervisors:
Dr. John Soo Yue Han

Institute:
Universiti Sains Malaysia, Penang, Malaysia

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

One of the biggest challenges in the field of astronomy involves accurately classifying stars and galaxies including distinguishing between galaxy types. While many discriminants have been provided, their performance with different variable transformation have not been properly assessed. This study addresses this gap by effectively analyzing the efficiency of several discriminants in the classification of stars and galaxies. This study identifies the most efficient discriminants to improve the accuracy of astronomical classifications by comparing the performance of different discriminants in classifying stars and galaxies and additionally between rounded galaxies and galaxies with no bulges. The study was carried out by analyzing the performance of multiple discriminants, such as Boosted Decision Trees (BDT), Fisher, Mahalanobis, *H*-Matrix, Artificial Neural Networks (ANN), linear discriminants, and in addition of applying different variable transformations (none, N, U, G, D, and PCA) on the discriminants. The discriminants were evaluated and assessed by utilizing metrics and statistical calculations. Our findings revealed that ANNs consistently showed the highest efficiency, followed by BDT which validates the strong classification of these two discriminants. Moreover, Fisher, Mahalanobis, and linear discriminants exhibited significant performance with the variable transformation U in part A of our study where these three discriminants achieved the highest accuracy (0.991) while in part B, which focused on distinguishing between rounded and no bulge galaxies, the same discriminants showed the highest accuracy (0.9306) utilizing the U transformation. The potential of Fisher, Mahalanobis, and linear discriminants in terms of classification is thoroughly analyzed from this study which makes it worth noting for future research in the field of cosmological classification.

ABSTRAK

Salah satu cabaran terbesar dalam bidang astronomi melibatkan klasifikasi bintang dan dan galaksi termasuk membezakan antara jenis-jenis galaksi. Walaupun banyak pembeza layan telah disediakan, prestasi mereka dengan menggunakan transformasi pembolehubah yang berbeza tidak dinilai secara menyeluruh. Kajian ini menangani jurang ini dengan menganalisis kecekapan secara berkesan menggunakan beberapa pembeza layan dalam klasifikasi daripada bintang dan galaksi. Kajian ini mengenal pasti pembeza layan yang paling berkesan untuk meningkatkan ketepatan klasifikasi dengan membandingkan prestasi daripada pembeza layan yang berbeza dalam mengklasifikasikan bintang dan galaksi dan di antara galaksi bulat dan galaksi tanpa bonjolan. Kajian telah dijalankan dengan menganalisis prestasi pelbagai pembeza layan, seperti Boosted Decision Tree (BDT), Fisher, Mahalanobis, *H*-Matrix, Artificial Neural Network (ANN) dan pembeza layan linear, dan di samping menggunakan transformasi pembolehubah yang berbeza (tiada, N, U, G, D, dan PCA) pada pembeza layan. Pembeza layan telah dinilai dan dinilai dengan menggunakan kaedah metrik dan statistik pengiraan. Penemuan kami mendedahkan bahawa ANN menunjukkan kecekapan tertinggi secara konsisten dan diikuti oleh BDT yang mengesahkan pengelasan kukuh bagi dua pembeza layan ini, tambahan pula, Fisher, Mahalanobis, dan pembeza layan linear mempamerkan prestasi yang ketara dengan transformasi pembolehubah U dalam bahagian A daripada kajian kami di mana ketiga-tiga pembaza layan ini mencapai ketepatan tertinggi (0.991) manakala di bahagian B, yang memberi tumpuan kepada membezakan antara galaksi bulat dan galaksi tanpa bonjolan, pembeza layan yang sama menunjukkan ketepatan tertinggi (0.9306) menggunakan transformasi U. Potensi Fisher, Mahalanobis, dan pembeza layan linear dari segi klasifikasi dianalisis dengan teliti daripada kajian ini dan patut diberi perhatian untuk penyelidikan masa depan dalam bidang kosmologi pengelasan.

STATEMENT OF AUTHORSHIP

I, Santtosh a/l Muniyandy, born October 19, 2002, declare that this thesis titled *Galaxy Classification Using Discriminants* and the work presented in it are the result of my own research and work. I confirm that this work was done mainly while in candidature for degree at Universiti Sains Malaysia.

Santtosh a/l Muniyandy 2024

ACKNOWLEDGMENT

I want to express my gratitude to my supervisor, Dr. John Soo Yue Han, for his knowledge, guidance, and encouragement throughout this study. His feedbacks and suggestions helped to improve the quality and depth of this thesis. Dr. John provided the perfect combination of flexibility and direction which allow us to explore our project while utilizing his helpful feedbacks. His understanding of astronomical and machine learning knowledge was the source of inspiration and his ability to share his knowledge assisted me in overcoming issues throughout this study. His dedication to my academic development has had a long-lasting influence, and I am grateful for his guidance and support.

I want to express my gratitude to my research partner, Sazatul Nadhilah Binti Zakaria, for her continuous support and participation throughout this research project. Her dedication, effort, and focus to this project were crucial for the successful outcomes from this study. Her creative ideas and ability to tackle challenges from multiple perspectives really enhanced our work and overcome many obstacles we encountered.

Finally, I would like to dedicate my deepest gratitude to my mother for her support and encouragement throughout my time at university. Her love, understanding, and sacrifices served as the foundation for my achievement. She has always been a source of strength for me, giving me the inspiration and courage to achieve my goals. Her confidence in my interests has inspired me to overcome various obstacles and strive for perfection. Without her emotional and moral support, this achievement would not have been possible. I will be forever thankful for her presence in my life and everything she has done to help me accomplish my goals.

1

INTRODUCTION

1.1 COSMOLOGY

1.1.1 *Introduction*

The origins and evolution of the universe are among the most significant mysteries. The emergence of cosmological topics as well as questions about their origins and structures, have been long overdue with just a handful of answers in this era of rapid technological advancement. Cosmology is the science of discovering the mysteries of the universe. The term 'cosmos' is derived from the Greek word 'universe', whereas 'logos' means 'study of'. Thus, the term "cosmology" denotes the study of the cosmos [20].

Historical cosmological models play an important role in explaining how the state of the universe works along with what it represents. The 'Cosmic Egg', which appeared in India between the 15th and 12th centuries B.C., was one of the first models developed to describe the cyclic universe expanding and collapsing infinitely [20]. However, the 'cosmic egg' or 'world egg' was widely accepted as a mythological design of various cultures and civilizations, rather than a cosmological model for scientific purposes. In other words, the 'Cosmic Egg' is an illustration that represents the idea of the universe expanding from the Big Bang and is used metaphorically to understand the principles of science. It's fascinating that so many ancient philosophers, including Anaximander, Anaxagoras, Democritus, Aristotle, and Aristarchus, contributed to primitive models early on. Vesto Slipher, Edwin Hubble, Alexander Friedmann, Carl Wilhelm Wirtz, Knut E. Lundmark, Willem de Sitter, and Georges H. Lemaître, discovered that galaxies were moving apart in the early 1930s, giving the foundation for the theory of the expanding universe [83].

This implies that it calls to question Albert Einstein's concept of a static universe, which he shared with other astronomers. Einstein's cosmological constant, Λ, was revoked due to evidence of an expanding universe. However, it has made an appearance in modern cosmology in the form of dark energy, which is believed to accelerate the expansion of the universe [20][29]. Thus, the impact of observational and theoretical physics on modern understanding of the universe is represented by the mythological theories of the 'Cosmic Egg' and various philosophical models. Throughout the history of astronomy, human curiosity and questions fueled our understanding of the universe, and we are reminded of how remarkable the human mind is at deciphering the vastness of the universe.

Cosmology uses a variety of distinct units and standards to correctly describe and quantify the universe's scale. The distance and standard units used by astronomers to calculate distances between objects is the astronomical unit (AU), $1\,\mathrm{AU} = 1.50 \times 10^{11}$ m which represents the mean distance between the Earth and Sun. The astronomical unit was later found to be incapable of measuring distances beyond our Solar System to nearby stars. To compare and calculate these far distances, we use the parsec (pc) where 1 pc $= 3.09 \times 10^{16}$ m. Going further than the distance between stars and objects, the parsec is not sufficient to measure the distances between galaxies. The use of megaparsec (Mpc) is needed, where $1\,\mathrm{Mpc} = 10^6\,\mathrm{pc}$ or 3.09×10^{22} m.

The Sun was used as the standard unit of power in astronomy, such as the standard unit of mass ($M_\odot$) where the Sun's mass, $1\ M_\odot = 1.99 \times 10^{30}$ kg. Similarly, the Sun provides a standard measurement of luminosity, $1\ L_\odot = 3.83 \times 10^{26}$ W. With these given standard units, it was given estimations to the mass of our galaxy, $M_{gal} \approx 10^{12}\ M_\odot$ and luminosity of our galaxy, $L_{gal} \approx 3 \times 10^{10}\, L_\odot$. Particle physicists were known to play a part in cosmology, as there were certain particle physics events occurring in the Galaxy. Thus, they had introduced their sets of units, to measure the energy in electron volts (eV), $1\ \mathrm{eV} = 1.60 \times 10^{-19}$ J. However, these introductions of the units are biased to only human capacities and capabilities, which brings to a means of a new universal system. A less biased system was introduced in the Planck system , the Newtonian gravitational constant, $G = 6.67 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$, the speed of light, $c = 3.00 \times 10^8\,\mathrm{ms^{-1}}$ and Planck constant, $h = 1.05 \times 10^{-34}\,\mathrm{J}\ s$ or 6.58×10^{-16} eVs [54]. Max Planck established the Planck system

in 1899, which is a set of units derived from the basic constants that describe the universe as a whole. He discovered a series of physical constants by combining the Planck constant h, gravitational constant G, and speed of light c. These physical constants, Planck mass $m_p \approx 2.17651 \times 10^{-8}$ kg, Planck energy $E_p \approx 1.9561 \times 10^{9}$ J, Planck time $t_p \approx 5.39106 \times 10^{-44}$ s and Planck length $l_p \approx 1.61619 \times 10^{-35}$ m [84]. Planck units provide appropriate scales for length, time, mass, temperature, and other measurements. These scales does not depend on creations of humans, and hence provide a more simple understanding of the the universe. The Planck units removed unpredictability of traditional measuring methods, for example, the measurement of meters is based on human history, but the Planck length is derived from basic features of the universe. Changing to natural units can result in a better understanding of physical limitations [85].

1.1.2 *Stars*

The birth of stars is a fascinating process that begins in massive clouds of dust and gas known as nebulae. These nebulae are the birthplaces of stars, providing all the required conditions for their formation. When a portion of the cloud collects enough mass through gravitational force, it begins to collapse inwards and shrink. As the cloud collapses into smaller bits, these particles could form a new star [15]. Throughout their lives, stars have a major impact on their surroundings, such as fostering the development of new stars in nebulae and generating ideal circumstances for planet formation. In their middle ages, like as our Sun, they act as catalysts to produce a suitable environment for our planetary systems. Stars are like the universe's lighthouses, illuminating our wide night sky. Each star is a hot, compact ball of gas composed of hydrogen and helium that emits an immense amount of energy, allowing it to burn brilliantly for billions of year [49]. When a star's life cycle finishes, rather than fading away, it explodes as a supernova, and the scattered stardust becomes material for new planets, stars, or other celestial bodies. Stars make significant contributions to the universe's narrative throughout their lives and deaths, from dust to capturing light for their sole purpose of the continuous cycle of cosmic development.

1.1.3 *Galaxies*

The universe's building blocks are vast and complex systems, including intergalactic gases, objects, stars, and celestial bodies. These systems are held together by gravity and exist in galaxies of various sizes, shapes, luminosities, and other galactic properties. Astronomers can classify or derive crucial information about galaxies using properties such as spectral lines, velocities, wavelength emission, and morphological features. In addition to that being stated, galaxies range in size from dwarf to supermassive, and they can be clustered together in groups. The spaces between galaxies are filled with interstellar gases and dusts, which not only serve as a womb for the birth of new stars and bodies, but also give and change the structures of the galaxies. One of the most fascinating aspects of galaxies is dark matter, a substance that scientists and physicists continue to strive to fully comprehend with respect to its properties and key forces at work in the universe. These unknown forms of matter do not emit or interact with light, but they do possess gravitational energy. The mass-to-light ratio ($\frac{M}{L}$) is used to measure the contribution of dark matter on a scale [5]. Every breakthrough of unlocking and uncovering the nature of galaxies is only possible when we delve deeper into the study of fundamental components of galaxies.

1.1.4 *Problem Statement*

In the vast field of astronomy, the classification of celestial bodies such as stars and galaxies are critical to understanding the distribution and composition of the cosmos. As the volume of data generated by cosmological surveys continues to grow, existing classification methods such as manual inspection of telescope images and human skill will become insufficient. The ongoing surveys that made many important contributions include the extended Baryon Oscillation Spectroscopic Sky Survey (eBoss: https://www.sdss4.org/surveys/eboss/), the Kilo-Degree Survey (KiDs: https://kids.strw.leidenuniv.nl/) and LOFAR Two-meter Sky Survey (LoTSS: https://lofar-surveys.org/), meanwhile for the upcoming surveys include the Legacy Survey of Space and Time (LSST: https://rubinobservatory.org/) and Euclid satellite mission (https://www.euclid-ec.org/public/data/surveys/) [86]. Thus, the problem is to analyze these data efficiently and accurately to extract valuable insights into the

characteristics and distribution of stars and galaxies. As such, we use machine learning to automate the classification process. However, applying machine learning to classify stars and galaxies requires the use of different algorithms or models that consider the significant variability and complex dimension of the data. Furthermore, building a balanced dataset from either vast or limited sources presents major difficulties for developing an effective and balanced training procedure.

Ultimately, this thesis attempts to overcome these issues by implementing a variety of machine learning algorithms that are specifically designed for both star and galaxy classification and galaxy classification by type. The goal is to increase classification accuracy by combining numerous methods and configurations in various algorithms.

1.1.5 *Thesis Objective*

This thesis aims to improve and extract information regarding the classification of stars and galaxies using machine learning techniques and different algorithms. The objectives of this study are structured upon one another such that it addresses progressively the classification process. They are as follows:

- Objective 1: To identify the most effective discriminant for separating stars from galaxies. This will be accomplished by comparing the performance of six different machine learning discriminants, with the goal of determining which method maximizes both accuracy and efficiency in classification.
- Objective 2: To investigate the influence of these variable transformations on the different discriminants used to process and classify the dataset, therefore assessing the possibility and potential for improving the classification process.
- Objective 3: To accurately classify disk-type galaxies using the Galaxy Zoo 2 Decision Tree by correctly categorizing between boxy, rounded and no bulge galaxies.

1.1.6 *Thesis Layout*

Chapter 2 focuses on reviewing important and relevant literature related to the research topic. The topics that are discussed are Edwin Hubble's contribution to galaxy classification, morphology of galaxies, the importance of redshift in light measurement, photometry, and comoving distance, which explain and summarize each key theme that contributes to the study.

Chapter 3 describes the approaches and procedures used to conduct this research topic. This chapter describes how the research will be designed to achieve the main objectives of the overall research topic, as well as how to address these goals. We will also explain the method used to extract data for classification, the types of algorithms used for classification, the changes made to input parameters, and the metric scores of each algorithm using the evaluation techniques.

Chapter 4 presents the research results and findings. The findings in this chapter will contribute to and address the scope of our goals, as well as provide discussion for each section of the study and research questions.

Chapter 5 focuses on the thesis's conclusion, where the discussion of the results and summarization of key findings will highlight the contributions of this research to future study.

2

LITERATURE REVIEW

2.1 GALAXY CLASSIFICATION

2.1.1 *Edwin Hubble's Galaxy Classification*

Edwin Powell Hubble (1889-1953), an important figure in making significant contributions to early classification, influenced modern astronomical discoveries. He was born in Missouri and attended the University of Chicago, where he studied mathematics and astronomy [43]. The most significant impact of his journey was the discovery of billions of galaxies beyond the Milky Way. Initially, astronomers believed that the Milky Way was the only galaxy in the vast universe. However, Hubble's work concept, using the telescope at Mount Wilson Observatory, revealed distant nebulae that were observed as independent galaxies. The astronomers' view of the Milky Way as a single galaxy was eliminated when Hubble established a relationship between the light emitted by celestial bodies and their distances from a specific reference point. Hubble calculated the distance between celestial bodies by measuring Cepheid periods. Cepheid periods are characteristics of Cepheid variable stars, which are pulsating (vibrating) variable stars with fixed periodic changes in their brightness. The term period refers to the total time it takes for a star to complete one cycle of pulsation [66][43].

This concept is related to Henrietta Swan Leavitt's early 20^{th} century discoveries about the Cepheid variables. She discovered a direct relationship between the period of the Cepheid variable and the brightness of the source, with the Cepheid with longer periods of pulses appearing brighter [66]. According to this theory, there is a correlation between a star's brightness and the time it takes for that star to dim from its observed brightness.

This established relationship was widely used in astronomy to determine the distances between celestial bodies and galaxies. The working methodology for measuring the observed galaxy using Cepheid variables is as follows, through the known luminosity value of the Cepheid variable, astronomers will compare the Cepheid variable's intrinsic luminosity to the luminosity observed from Earth.

Hubble used the method developed by Henrietta Swan Leavitt, which resulted in the discovery of millions of galaxies outside the Milky Way, each with its own collection of stars and gases. This leads to Hubble's Law and the expanding universe theory [70]. Hubble discovered the reddening of spectrum in his analysis, which is known as redshift. This indicates that the redshift is a value for knowing the recessions of galaxies from us. He established the relationship that the greater the redshift, the farther it moves away. Hubble's Law is expressed as $V = H_0 D$ where V is the recessional velocity, D the distance of the object from Earth and H_0 the Hubble's constant. Furthermore, by observing differences in galaxy structures, Hubble developed a classification system that resembles a 'tuning fork' diagram, categorizing galaxies by different shapes and structures, such as lenticular galaxies, spiral galaxies, and elliptical galaxies [17][70].

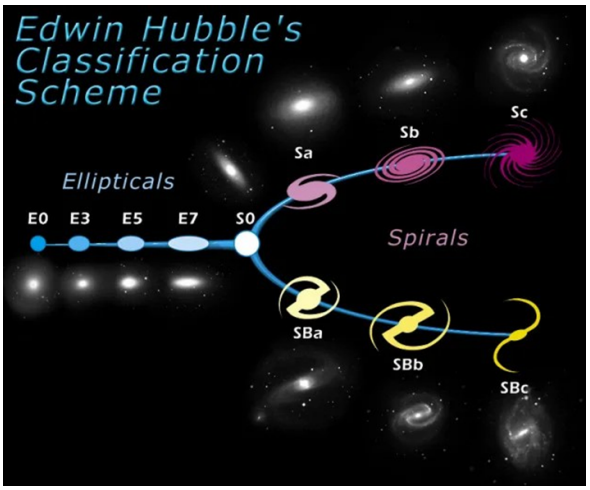

Figure 2.1

Figure 2.1 shows Edwin Hubble's galaxy classification (tuning fork diagram) which was obtained from https://esahubble.org/images/heic9902o/. In addition, several advances have been made to improve Edwin Hubble's initial galaxy classification. First, the de Vaucouleurs classification system helps to understand various galaxy morphological structures by categorizing galaxies into ellipticals, irregulars, and spirals [56]. Ellipticals

(E) range from ball-shaped galaxies (E0) to flat-like discs (E7). Spirals (S) are classified based on their central bulge and arms; spiral galaxies with a large central bulge and tight wound arms are denoted Sa, while galaxies with a small central bulge and loose wound arms are denoted Sc and the spiral galaxy located between those classes are denoted Sb.

There are also spiral galaxies that resemble a bar filled with stars in the center, known as Barred Spirals (SB), which are classified from spirals that are difficult to see at the end of the bar (SBa) to spirals that are easy to see (SBc). Finally, irregular galaxies (Irr) have an irregular shape and are difficult to classify using the tuning fork diagram [56][57][31].

Other classifications worth mentioning include Van Den Bergh's and Morgan's classifications. Van Den Bergh discovered that the spiral arms play an important role in the absolute luminosity of galaxies. He proved that the galaxy with the highest luminosity has the longest and most developed spiral arms [11]. Morgan's classification is intended for distinguishing and evaluate the contents of the spiral arm and central bulge. Morgan introduced a new system called Hubble-de Vaucouleurs color-class notation to better understand the variety of appearances of galaxies [56][11].

2.1.2 *Elliptical Galaxies*

Elliptical galaxies are distinct through their ellipsoidal, smooth and elongated spheroidal shapes with weak spiral arms or discs structures. The brightness of elliptical galaxy can be represented as

$$I = I_o\left(\frac{r}{a} + 1\right)^{-2}, \tag{2.1}$$

where a is a normalising constant, I and I_o are the intensity and central intensity of light, respectively. Hubble classified the elliptical galaxies based on their apparent ellipticity, E_n, where $n = 10\left(1 - \frac{b}{a}\right)$, $\frac{b}{a}$ is the apparent flattening (b and a are the minor and major axes of the ellipse). Thus, we are able to classify nearly spherical galaxies (E0) to highly elongated galaxies (E7). The additional feature of late ellipticals, E^+ was implemented to de Vaucouleurs classification to categorize elliptical galaxies that fit in standard E0 to E7

[11].

Moreover, a Sersic profile of $r^{\frac{1}{n}}$ allows us to differentiate the structures of elliptical galaxies where the n represents concentration of light. Larger and more luminous galaxies exhibits $n > 4$ in which it shows a higher concentrated core, while on other hand, for $n < 4$, it represents smaller and lesser luminous ellipticals with lesser concentrated core [31][56]. Figure 2.2 shows the elliptical galaxies which was obtained from https://ned.ipac.caltech.edu/level5/Sept11/Buta/Buta5.html.

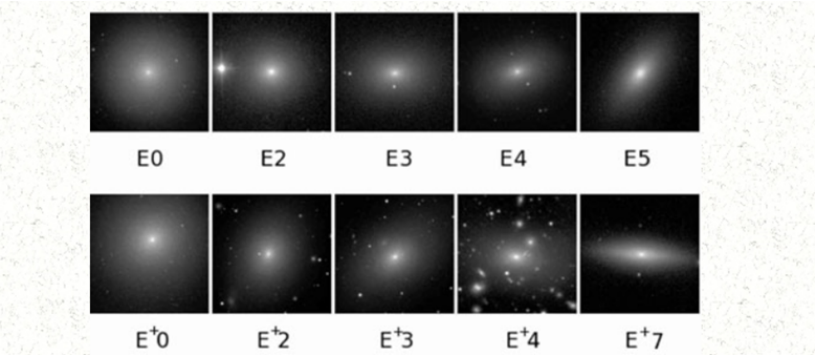

Figure 2.2

2.1.3 *Spiral Galaxies*

The spiral arms of spiral galaxies are easily distinguishable and include different star-bearing regions that emerge from the centre of the galaxy. The nucleus, disc, corona, spiral arms, and central bulge which is the core of the galaxy where stars are frequently concentrated in high densities are the typical components of spiral galaxies. Figure 2.3 shows the regions and structure of a spiral galaxy which was obtained from https://www.researchgate.net/figure/Components-of-Milky-Way-galaxy-18_fig2_333882064.

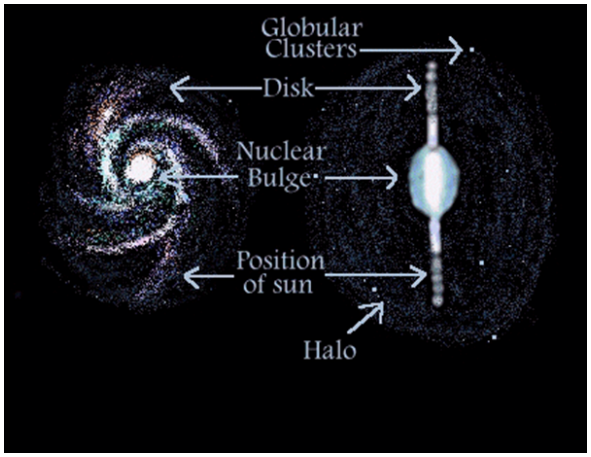

Figure 2.3

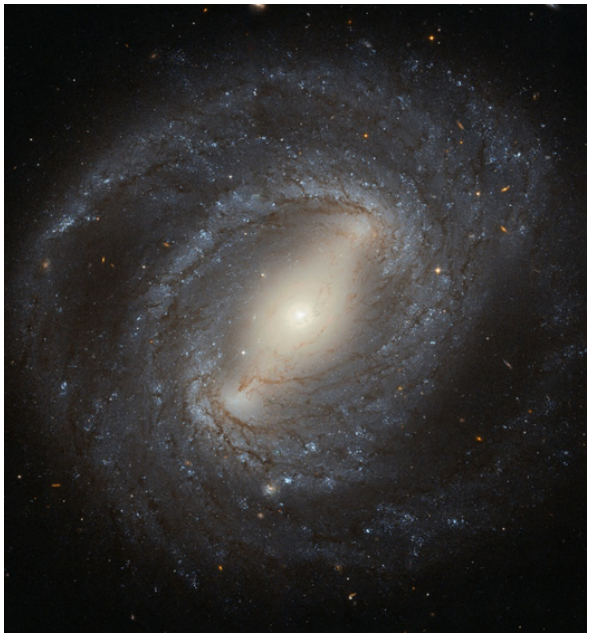

Figure 2.4

Figure 2.4 shows the spiral galaxy which was obtained from https://en.m.wikipedia.org/wiki/File:Hubble_spies_NGC_4394.jpg. The spiral galaxy is divided into two classes of normal spirals and barred spirals in which the normal spirals consist of spiral arms emit from the centre while the barred spirals consist of a bar that connects the ends of the galaxy. Hubble classified the normal spirals by (S) and barred spirals by (SB) [24]. It is crucial to note that there are galaxies that lie between spirals and ellipticals. These are known as lenticular galaxies (S0), and they are identified by their disc shape without spiral arms.

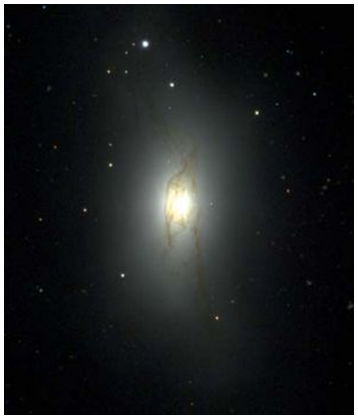

Figure 2.5

Figure 2.5 shows the example of S0 galaxy which was obtained from [24]. In the classification of Hubble-Sandage, the possibility to classify S0 was enabled such that in terms of developmental stage of spiral formation [11]. As such, early ($S0^{-}$) represents little to minimal features of discs with no evidence of spiral arms, intermediate ($S0^{o}$), represents moderate features of discs with similar lack of spiral arms and late ($S0^{+}$), represents minimal to very subtle features of discs which marks the beginning of spiral formation [24][31]. Figure 2.6 shows the examples of barred and non-barred S0 galaxies which was obtained from [11]. Figure 2.7 shows the spiral galaxies that have been classified as bar types which was obtain from [11].

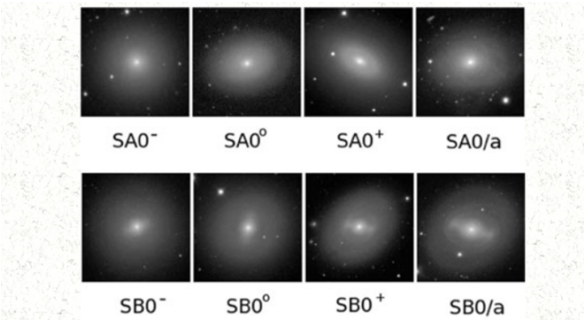

Figure 2.6

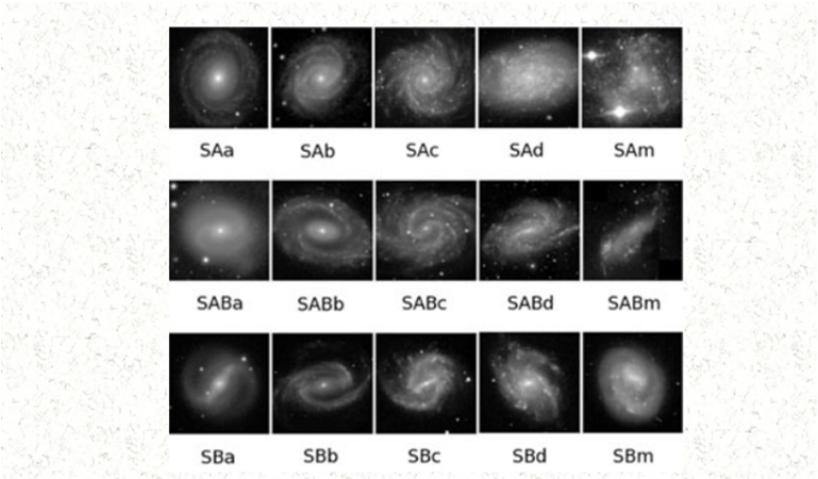

Figure 2.7

2.1.4 *Irregular Galaxies*

Observations of spiral and elliptical galaxies with rhythm or a standard structure allowed them to be classified using discs, bars, and other structures. Irregular galaxies are outliers among these two types of galaxies because they lack a fixed and definite structure or shape that distinguishes them from the others. The randomized and asymmetrical shapes include clumped or distorted versions of stars and gases. Irregular galaxies are divided into two categories: Irr I and Irr II [11]. Irr I is regarded as an extension for spiral classes, describing an Sc galaxy that lacks spiral structures. Irr I also contains many stars and intergalactic objects. Irr I holds the well-defined spiral galaxy better than Irr II. Irr II galaxies are groups of randomly structured galaxies for which it appears impossible to categorize the galaxy using any kind of characteristic. There is very little to no symmetry with stars or clusters in Irr II [31][24]. Figure 2.8 shows the optimized de Vaucouleurs diagram which was obtained from https://en.wikipedia.org/wiki/File:Hubble_-_de_Vaucouleurs_Galaxy_Morphology_Diagram.png. Figure 2.9 shows the example of irregular galaxy which was obtained from https://www.gizbot.com/news/nasa-hubble-telescope-snaps-an-irregular-galaxy-060732.html.

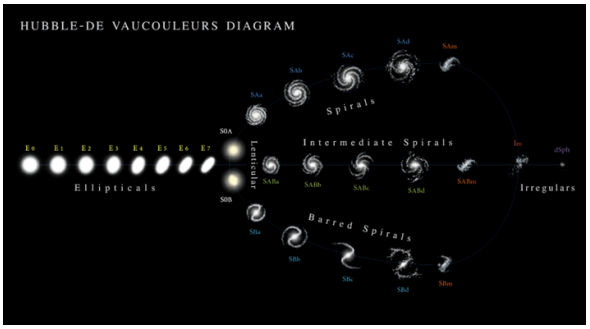

Figure 2.8

Figure 2.9

2.1.5 *Machine Learning Galaxy Classification*

Machine learning is a form of artificial intelligence that uses computers to learn from raw or organized data and make predictions without relying on extensive human labor or explicit programming. Machine learning provides fundamental concepts and has the potential to address a wide range of complex issues and challenges in fields such as medicine, economics, politics, and education [65]. In today's world, the abundance and overabundance of data has become a critical source of information in almost every aspect of the modern world's capabilities.

Data is a collection of information gathered through various methods including observations, measurements, research, and analysis. Data includes numbers, names, pictures, figures, time, and other information that can be stored for later use. The volume of data accumulated had exceeded the limitations, making it difficult for data scientists and data analysts to understand and make appropriate decisions for their respective reasons. There

are two types of data: qualitative and quantitative [13].

Qualitative describes information in words, for example, describing the quality of a shirt, a person's religion, or the type of shoes. Quantitative refers to numerical information such as a person's weight, the price of a shoe, or any person's phone number. Data comes in a variety of forms and formats, and it can be structured or unstructured, with each having its own distinct characteristics. Structured data is classified as quantitative data and is typically stored in relational databases and spreadsheets, making it easier to manage [50].

Structured data is widely used in industries because it is well organized and can be easily understood by machine. The data produced or organized will be of higher quality and consistency, reducing the risk of data compromise. Unstructured data is classified as qualitative data, which is difficult to analyze because it is non-relational, or lacks a proper categorization. As previously mentioned, qualitative data could include the name of a restaurant, file names, and much more. However, it does not imply that unstructured data is difficult to work with; rather, this type of data can be used as a technique for learning comparisons in trends, which aids in informed decision making [38].

Starting from the historical perspective, in the 20^{th} century, Alan Turing and Arthur Samuel focused on creating algorithms which could learn from datasets and improve independently periodically [75]. This leads to developments of various algorithms such as neural networks, decision trees, naïve bayes, support vector machines and etc.

Supervised learning is entirely based on training from labeled data, which means that the input data is accompanied by theoretical output. It trained the model using the training dataset, resulting in the desired output, which is provided as labeled data. This method allows the algorithm to learn and compare to the actual output, after which it optimizes and adjusts until the error is minimized. The most common application of supervised learning methods is to use historical data to statistically predict future events.

Supervised machine learning can be divided into two types: classification and regression. Classification is the process of predicting a categorical class and assigning the test dataset to specific categories. It recognizes trends or patterns as early as the learning stage, assisting in the classification of unlabeled data. Regression predicts a continuous numerical value, and the fundamental goal is to understand the relationship between dependent and independent variables. The output may include numerical values within a specified range to indicate the likelihood of an event occurring [42][44][40].

Unsupervised learning learns from unlabeled data that lacks target output and allows for the extraction of insights without the guidance provided by supervised learning. It must generate inferences independently because they are not given a fixed output to train with, so they must discover patterns on their own. Clustering is one of the most used forms of unsupervised learning. Clustering requires no prior knowledge of data classes, and the cluster algorithm attempts to familiarize data points using traits, characteristics, or features. This greatly helps in identifying hidden patterns and trends. Clustering examples include hierarchical clustering, K-means, and the OPTICS algorithm [63][76][44].

A powerful data analysis tool has emerged because of recent advances in computational techniques for classifying galaxies based on their shapes, features, and traits. The increase in astronomical data from surveys such as the Sloan Digital Sky Survey (SDSS), Galaxy Zoo, and Hubble Space Telescope, which are used to analyze and classify different types of galaxies by extracting patterns and trends using various algorithms [27]. The key feature in using machine learning to classify galaxies is based on features in galaxy images such as size, shape, color, and brightness. Thus, inputs, or photometric attributes, are critical features for improving the accuracy of machine learning classifications. Photometric attributes include color indices, spectral energy distributions, brightness, and so on. Most machine learning galaxy classification algorithms use parameters such as luminosity, apparent magnitudes, color indices ranging from B-V to U-B, and flux ratios to produce a high metric score.

2.2 DISTANCE MEASUREMENTS IN GALAXY

2.2.1 *Cosmological Redshift*

Redshift, an observed phenomenon involving the light emitted by galactic objects, is frequently used as a tool to aid in galaxy measurements and universe expansion. Hubble observed and concluded that emitted light from galaxies shifted towards longer wavelengths, implying that galaxies were moving away from the source. His analysis stated that the farther away an object is in the galaxy, the longer the shift of wavelength of light. The fundamentals of Doppler principles are relevant to this matter, as it demonstrates that the universe is expanding by stretching the redshift received [26][10]. Multiple attempts to determine distances between galaxies have been made and proven successful using redshift calculations. Various maps of galaxies can be plotted using applications or websites such as SDSS and the Galaxy Redshift Survey, both of which contain numerous integrated redshift datasets [10][45].

The cosmological redshift,

$$z = \frac{\lambda_{\text{observed}} - \lambda_{\text{emitted}}}{\lambda_{\text{emitted}}}, \tag{2.2}$$

where the expansion of the universe is related with the elongated (redder) wavelength. $\lambda_{\text{observed}}$ is the observed wavelength while λ_{emitted} is emitted wavelength.

The estimation distance used in cosmology follows the Hubble's Law which incorporates the relationship between redshift, z and the recessional velocity, v which is

$$v = H_0 \cdot D, \tag{2.3}$$

D represents the galaxy's distance from the source in which this case is Earth, and H_0 is Hubble's constant where H_0 = 74 (km/s) Mpc^{-1}. This principle briefly explains how fast and further the galaxy is moving away. Similarly, we could use the scale factor to represent cosmological redshift,

$$z = \frac{a_0}{a(t)} - 1, \tag{2.4}$$

where a_0 and $a(t)$ represent scale factors during present and time t respectively. As the universe expands periodically, the scale factor increases which leads to higher redshifts from the objects [71].

Redshifts are commonly used in observational industries and sectors, such as spectroscopy and photometry. Industrial tools such as the Multi-Object Spectrograph (MOS) [78] are effective at receiving and observing celestial object redshift. In the context of photometry, more sensitive and redshifting objects necessitate more precise and accurate measurements, which are frequently assisted by detectors such as charge-coupled devices (CCDs) [53].

The redshift data obtained is critical in categorizing galaxies into three types: spiral, elliptical, and irregular. The results of measuring the redshift reveal important information about galaxies, such as their sizes, speeds, shapes, and structural composition. For example, most spiral galaxies have lower redshifts and a slow recession speed, whereas elliptical galaxies exhibit the opposite.

2.2.2 *Comoving Distance*

The comoving distance, d_C shows the distance between two points which are not affected by the principle of expansion of the universe. In the case of a spatially flat universe and an epoch where the distance between two objects is constant (a specific time period in the evolution of the universe) with the expansion of the universe (Hubble Flow), the comoving distance is given by

$$d_C = \frac{c}{H_0} \int_0^z \frac{1}{\sqrt{\Omega_M (1+z')^3 + \Omega_k (1+z') + \Omega_\Lambda}} dz', \tag{2.5}$$

c represents the speed of light, Ω_M and Ω_Λ are the density parameters of matter and dark energy, meanwhile $\Omega_k = 1 - \Omega_r - \Omega_M - \Omega_\Lambda$ where it determines the curvature (Ω_r is radiation energy density). In other terms, it is called as 'proper distance' where the position of an object is determined at a specific moment of cosmological time and changes

periodically as to the Hubble Flow.

The measurement of the physical size of an object to its angular size from a telescope is known as the angular diameter distance, d_A where

$$d_A = \frac{d_M}{1+z}. \tag{2.6}$$

From the derivation of angular diameter distance, d_A, we get to estimate the apparent brightness of celestial objects,

$$d_L = (1+z) \cdot d_M = (1+z) \cdot d_A, \tag{2.7}$$

where d_L is the luminosity distance. This implies that as expansion of the universe occurs, the wavelength of light emitted from the object will get stretched and appears more dimmer when compared to its original luminosity.

Moreover, due to redshift, astronomers often observe celestial objects at different observed wavelengths from their sources' emitted wavelength. This could be factored from extinction where it is a phenomenon when the light emitted from the celestial object travels through the medium in space such as intergalactic gases and dusts. This occurs due to processes of absorption and scattering. Absorption occurs when the photons in the travelling light are absorbed from the dust when the light travels through them. Meanwhile scattering occurs due to the light being scattered from the dust particles. To account for this, astronomers utilize K-correction,

$$K = -2.5 \log \left[\frac{\int_0^\infty f_\lambda \left(\frac{\lambda}{1+z}\right) R(\lambda) \frac{\lambda}{1+z} \, d\lambda}{\int_0^\infty f_\lambda(\lambda) R(\lambda) \lambda \, d\lambda} \right] \tag{2.8}$$

to the observed luminosity to compare the difference between observed and emitted wavelengths. Breaking down this formula, the term $f_\lambda \left(\frac{\lambda}{1+z}\right)$ accounts for how the object's spectrum appears to be stretched towards longer wavelength when observe. $\frac{\lambda}{1+z}$ is the stretched wavelength from emitted light wavelength, λ due to redshift where $f_\lambda(\lambda)$ is the flux per unit wavelength. $R(\lambda)$ is a term representing the system of observational equipment to intergrate the observed flux correctly [79][80]. From the obtained luminosity

distance, d_L, we could also represent the difference between absolute magnitude, M (constant brightness of standard candle regardless of distance from Earth) and the apparent magnitude, m (brightness of standard candle depending on its distance from Earth) by using $m = M + \mu + K$ where

$$\mu = 5 \log \left(\frac{d_L}{10 \text{ parsec}} \right) \tag{2.9}$$

is distance modulus and K is k-correction. So,

$$m - M = 5 \log \left(\frac{d_L}{10 \text{ p}} \right) + 2.5 \log \left[\frac{\int_0^\infty f_\lambda \left(\frac{\lambda}{1+z} \right) R(\lambda) \frac{\lambda}{1+z} \, d\lambda}{\int_0^\infty f_\lambda(\lambda) R(\lambda) \lambda \, d\lambda} \right]. \tag{2.10}$$

The relationship established between apparent and absolute magnitudes is widely used in determining distances to objects [32][28][71].

2.3 MEASUREMENT OF LIGHT

2.3.1 *Photometry*

The Greek terms "photo" and "metry," which translate to "light" and "measure," are frequently used to describe the visible light spectrum. "Photometry" is the study of interpreting electromagnetic radiation, or light, emitted by stars, galaxies, or other objects in order to derive properties and useful information for light measurement and analysis. Hipparchus, a Greek astronomer, divided the stars into groups according to brightness in 130 BCE [8]. The logarithmic scale used for this analysis has six categories: the first magnitude denotes the brightest stars, while the sixth and final magnitude denotes the faintest stars. Despite being the foundational system, the logarithmic scale evolved over time due to technological growth and advancement. When the magnitude increases by 1, it corresponds to a 2.512-fold decrease in light intensity, applied in the opposite direction. Since photometry primarily focuses on measuring light that can be seen by human eyes, it is limited to visible light of the electromagnetic spectrum. Visible light is a portion of the electromagnetic spectrum, which ranges from lowest frequency (radio waves) to highest frequency (gamma rays) [2][47].

A key aspect of photometry, luminous intensity,

$$I_V = \frac{d\Phi}{d\Omega}, \tag{2.11}$$

is the luminous flux, $d\Phi$ from a source emitted per unit solid angle, $d\Omega$ (coverage of three-dimensional space as viewed from a fixed point) which can be explain in simple terms as how intense a light source appears to the observer. In 1729, Pierre Bouguer created a system of wax candles which was later widely used as the international standard of theoretical point sources where one candela emits 540 $\times 10^{12}$ Hz [2]. Luminous intensity was measured with one unit of candela. Luminous flux,

$$\Phi_V = K_m \int_\lambda \Phi_{e,\lambda} V(\lambda)\, d\lambda, \tag{2.12}$$

where $K_m = 683 \frac{\text{lm}}{\text{W}}$ (efficiency of light source emitting light compared to total radiant energy emitted) and $\Phi_{e,\lambda}$ is the radiant energy of light emitted at each wavelength, λ, is the total amount of light emitted by a source in all directions [47]. We must take into account of the spread of the light emitted from the source which produces into measurement of intensity in a particular direction. The spread in one particular direction is denoted as a plane angle which is the spread of light over an area, this is the concept of a solid angle,

$$\Omega = \frac{A}{r^2}. \tag{2.13}$$

Similar to the opening of two intersecting planes in 2$^{\text{nd}}$-dimensional space, solid angles measure the spread in the region of 3$^{\text{rd}}$-dimensional space when viewed from a point [81]. Figure 2.10 shows the conical solid angle which was obtained from [81].

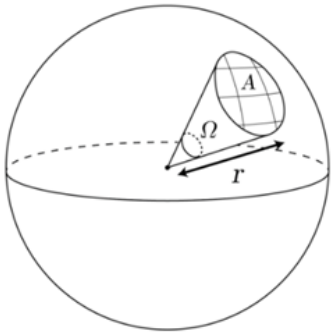

Figure 2.10

Illuminance,

$$E_v = \frac{d\Phi}{dA} \tag{2.14}$$

where $d\Phi$ is luminous flux and dA surface area is an important concept alongside luminance,

$$L_V = \frac{d^2\Phi}{d\Omega.dA.\cos\theta}, \tag{2.15}$$

where θ is the angle between normal to the surface and direction of beam, represents the total amount of light emitted onto a surface; while luminance represents the brightness of light emitted from the source. This concept enables astronomers to analyse the difference of wavelengths from celestial objects between blue and visible light filters, where it is utilized by optical filters such as the Ultraviolet, Blue and Visual filters (UBV photometric system) [6].

Colour index is defined as the magnitudes between two colours, $B - V = m_B - m_V$ where this shows the differences between blue and visible light colour filters. We can also represent this in terms of flux, f_B (flux for B filter) and f_V (flux for V filter),

$$B - V = m_B - m_V = -2.5 \cdot \log_{10}\left(\frac{f_B}{f_V}\right) + \text{constant}. \tag{2.16}$$

The colour filters can be related as the more negative the differences between two magnitudes, the bluer the appearance in the observed colour [53]. Figure 2.11 shows the illuminance which was is obtained from [47]. Figure 2.12 shows the luminance which was is obtained from [47].

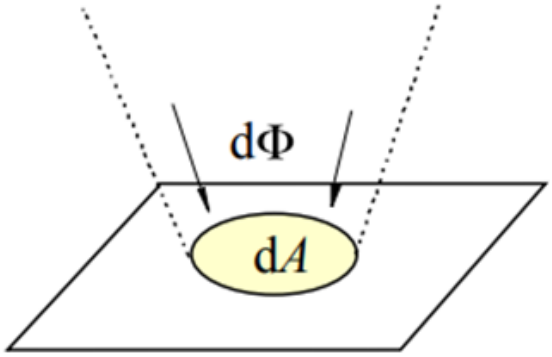

Figure 2.11

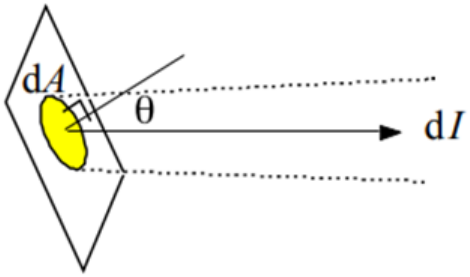

Figure 2.12

Astronomers use CCDs as digital imaging tools to capture light emissions and transform them into electronic signals. The ordered grid pattern in CCDs serves as a light-sensitive pixel structure. It also has a sensor that divides the image into pixels, which are then converted to electrical charges, within the pixel region, which transforms light particles (photons) into electrical charges. After that, the charges are directed in the direction of the output node, where they are amplified and ultimately transformed into voltage. To produce a digital image, these voltages are processed and digitalized [62]. The Point Spread Function (PSF) is the method used to accurately measure the flux of celestial objects in high density object maps. Typically, the environment and external influences distort or prevent the light source's components from reaching the detector. The goal of the PSF invention is to accurately determine the light spectrum. The PSF model, or PSF fitting in this particular case, serves as a guide for distributing the light source across the appropriate detector parameters (width, shape, or area), taking into consideration the challenges posed by outside disturbances. This is crucial in separating the densely mapped objects into a single image so that measurements can be made with greater accuracy. Other PSF models include theoretical models that employ mathematical concepts to predict how light will interact with observational equipment and practical models that analyze how point sources appear from various regions of the image. By measuring the differences in brightness between the same observed star in a bigger aperture and an observed object in a smaller aperture, one can detect fluctuations caused by atmospheric conditions and external disturbances (e.g., when a big aperture is used, an excessive amount of sky light

will be entered for measurements in aperture along with the observed star), this is known as aperture correction,

$$\Delta m = m_{\text{bigger aperture}} - m_{\text{smaller aperture}}, \tag{2.17}$$

where Δm represents differential magnitude. If the differential magnitude, Δm usually results in negative, it simply means that the light is denser in larger aperture than smaller aperture [53].

3

METHODOLOGY

3.1 INTRODUCTION

The summary of methodology for galaxy classification using machine learning begins with the use and leveraging of various algorithms such as Artificial Neural Networks (ANNs), Linear discriminants (LD), Fisher discriminants, *H*-Matrix discriminants, and Boosted Decision Trees. Machine learning can learn and classify the morphological diversity of various types of galaxies, as well as identify patterns for more efficient classification. The process begins with data collection from the Sloan Digital Sky Survey (SDSS) via a user-friendly query methods (CasJobs, see section 3.2.1). Next, by utilizing programs like Jupyter Notebook and Microsoft Excel, we then train the processed dataset using the machine learning algorithms. Finally, algorithm evaluation measures each algorithm's performance and its ability to generalize to previously unknown datasets.

3.2 DATA

A machine learning model's ability to adapt and generalize against previously unseen data for future use is heavily reliant on the quality, reliability, and quantity of the data. For this project, we used two different types of data: the first part consisted of data acquisition of stars and galaxies, and the second part consisted of boxy, no bulge, and rounded datasets. However, later in the project, we discovered that the overall boxy dataset is less than 20, which has a significant impact on the training aspect of machine learning, so we decided to proceed with no bulge and a rounded dataset.

3.2.1 *Sloan Digital Sky Survey (SDSS)*

We obtained the dataset from a website that contains all of the relevant and large-scale structure of the universe, as well as the properties of galaxies, stars, quasars, and other cosmic components. In 2000, the SDSS began using the Apache Point Observatory's 2.5 m telescope to capture high resolution images and a wide range of wavelength spectra [82]. SDSS is also equipped with cameras that capture images in five different wavelength bands: u, g, r, i and z which determine the properties of objects in the universe such as mass, age, temperature, and so on. SDSS includes four different survey operations: SDSS-I (2000-2005), SDSS-II (2005-2008), SDSS-III (2008-2014), and SDSS-IV (2014-2020), each of which contributed significantly to the discovery of the universe. SDSS data consists of both image and spectroscopy data, with the image data representing detailed optical images and the spectroscopy data representing the properties of the objects. The data then flows through data processing, which involves calibrating and correcting for atmospheric resistances, noise, and other external effects to produce accurate results. Thus, SDSS includes data releases from DR1 to the most recent DR18. These data releases are frequently updated with additional and more recent enhanced image and spectroscopic data [51][22].

Catalog Archive Server Jobs System (CasJobs: https://skyserver.sdss.org/casjobs/) is an easy-to-use tool for managing and manipulating SDSS datasets. CasJobs allows us to run Structured Query Language (SQL) on the server with a simple interface and little to no delays or timeouts, without overloading the system. We can use CasJobs to extract the dataset we need for our machine learning project by submitting queries and storing the results in our own private space called 'MyDB'. CasJobs' 'MyDB' section allows us to extract the dataset in a variety of formats, including csv, XML, GZIP, and others [37][60]. Figure 3.1 shows the sample query language used to extract rounded galaxy data from SDSS. Query process was done in https://skyserver.sdss.org/casjobs/. Figure 3.2 shows the different formats available to extract data in CasJobs which was taken from https://skyserver.sdss.org/casjobs/.

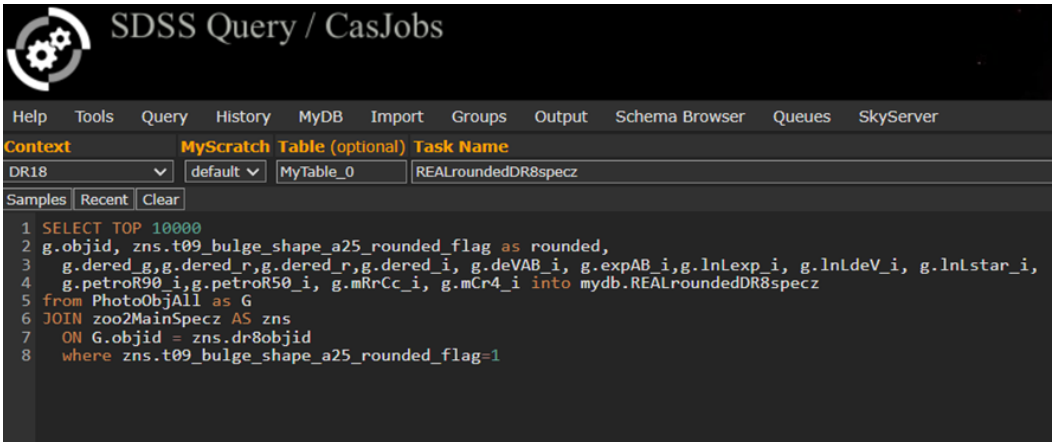

Figure 3.1

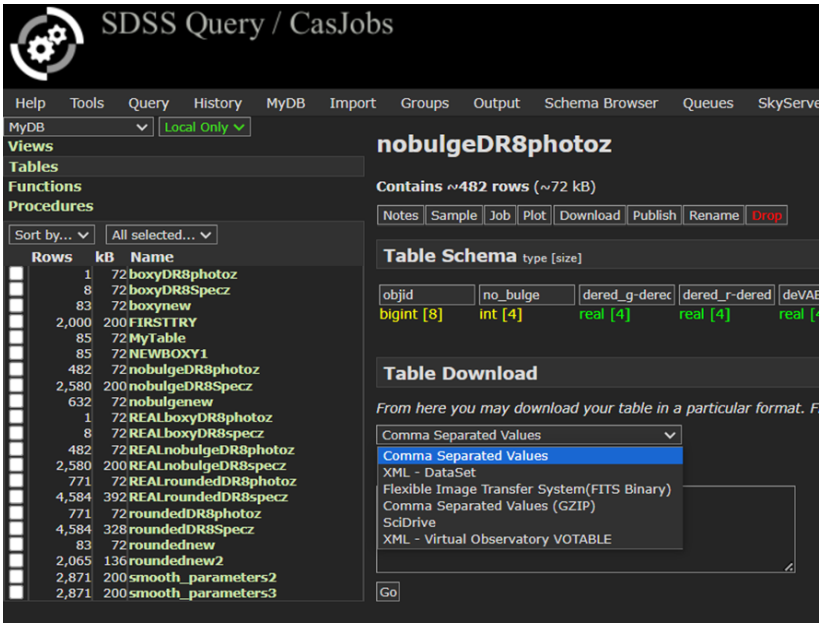

Figure 3.2

3.2.2 *Galaxy Zoo 2*

Galaxy Zoo (GZ2: https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/) was founded in 2007 to encourage volunteers worldwide to classify large number of galaxies to determine different patterns or aspects of the objects [87]. Galaxy Zoo 2 is an expansion of the original project that asks more in-depth questions about galaxy characteristics. For example, the number of spiral arms, the shape of the central bulge (boxy, rounded, or without a bulge), and the presence of bars. Galaxy Zoo 2 offers an online interface with prepared questions and SDSS images of galaxies. The survey follows a decision tree

known as the Galaxy Zoo Decision Trees. In GZ1 (Galaxy Zoo), online survey users are only asked one question, whereas in the newly established Galaxy Zoo Decision Tree, users can choose from a variety of survey types, including GZ 2, GZ 3 Hubble, GZ 4 Sloan, GZ 4 CANDELS, GZ 4 UKIDSS, GZ 4 Ferengi, GZ 4 DECaLs, GZ 4 Illustris, and GZ GAMA-KiDS [88].

For this study, we focus on the GZ 2 survey, in which participants start at the top of the tree of the first question and work their way down depending on their response. The questions are color-coded, as shown at the bottom left of Figure 3.3. GZ 2 had a significant impact on tens of thousands of volunteers as an educational tool for fostering interest in galaxies among students and the general public [74][25].

For the second part of our study, we will primarily focus on the type of bulge that appears in the centre of the galaxy, whether it is boxy, rounded, or does not have a bulge (T08). Figure 3.3 shows the Galaxy Zoo Decision Tree which was obtained from [88].

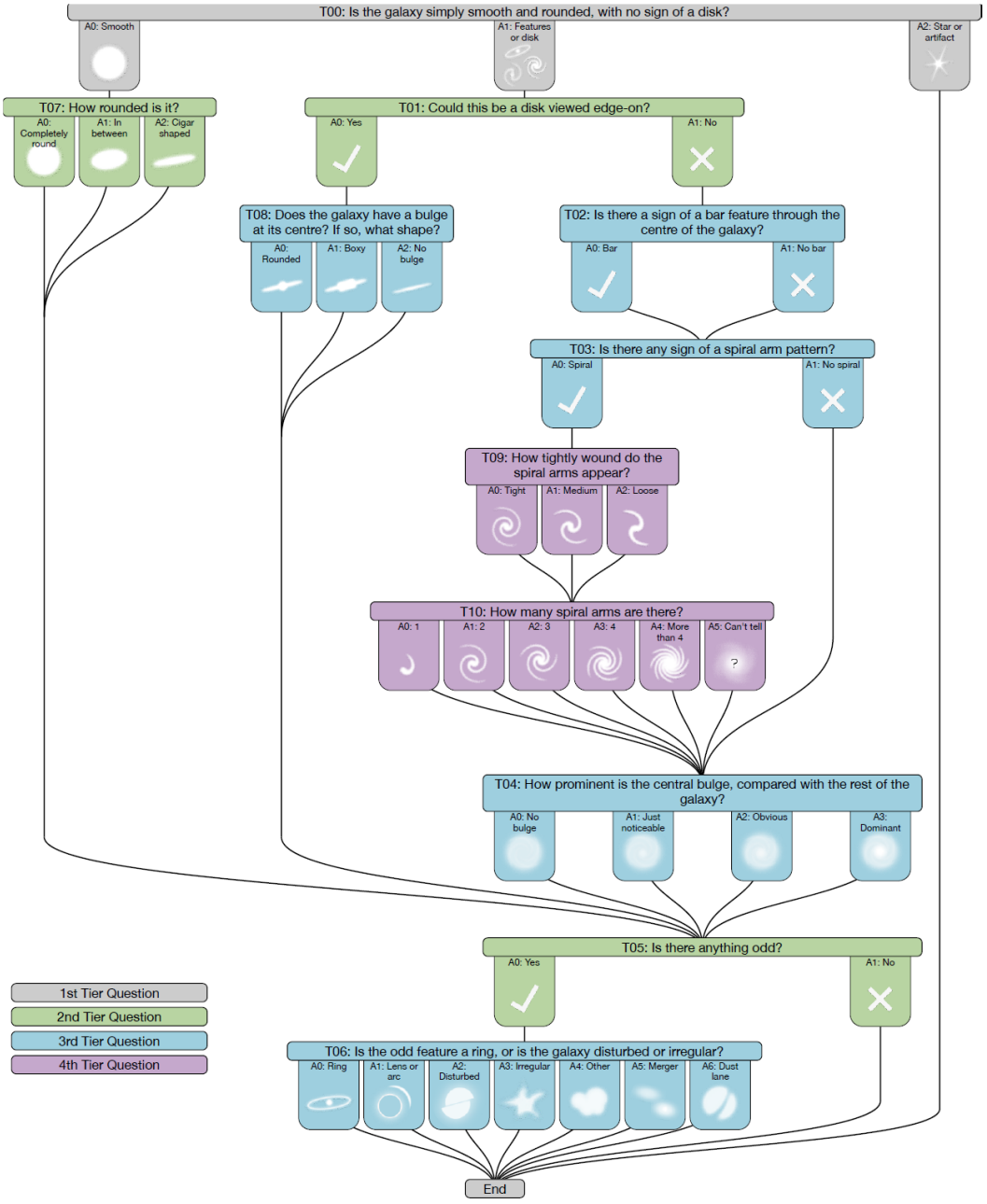

Figure 3.3

3.2.3 *Input Parameters*

Input parameters are critical attributes for determining and training machine learning models efficiently. Thus, selecting appropriate sets of input parameters is critical for influencing the model's ability to learn and make accurate predictions. For both parts of our study, we decided to use the same parameters in order to prioritize our objective when comparing our machine learning algorithms. The SDSS Sky Server's Schema Browser allows us

to select any set of variables from a variety of tables, including `PhotoObjAll`, `SpecObjAll`, `PhotoObjDR7`, and others, ranging from u-band to z-band. The first part of this project will use `PhotoPrimary`, while the second part will use `PhotoPrimary` and `Zoo2MainSpecz` in the *r*-band. Table 3.1 shows the input variables [58][59].

TABLE	INPUT PARAMETER	DESCRIPTION
PhotoPrimary	`psfMag_r`	The magnitude of stars is measured using the Gaussian Point Spread Function in the *r*-band. A single image of a star is adjusted until it is perfectly isolated in the image. The model will be used to estimate and quantify the amount of light that spreads from the star until it reaches the telescope.
PhotoPrimary	`fiberMag_r`	A magnitude which measures the amount of light emitted from an object using a spectrograph, with the lig ht entering through an aperture. The aperture size is important because it determines how much light can enter and be captured. For fiberMag, it was assumed that the aperture would be 3 inches in diameter to gather light from the object.
PhotoPrimary	`petroMag_r`	The petrosian magnitude which is used to generalize the brightness of a galaxy across different wavelengths, and it is calculated as total light in a predefined aperture set.

TABLE	INPUT PARAMETER	DESCRIPTION
PhotoPrimary	`modelMag_r`	The magnitudes of a galaxy measured using matched aperture approach to determine the characteristics of the galaxies or objects. It is used as a matched aperture to calculate flux in all bands. This helps to determine the colour of the galaxy.
PhotoPrimary	`petroR50_r` `petroR90_r`	A Petrosian system that calculates total flux based on radii measurements (`petroR50` and `petroR90`). `PetroR50` is the radius of the galaxy's 50% total Petrosian flux, whereas `PetroR90` is the radius of 90% total Petrosian flux.

PhotoPrimary	`mE1_r`
	`mE2_r`
	`mRrCc_r`
		Adaptive moments used to quantify the size and shape of an object by calculating its light distribution. We can improve the signal-to-noise ratio by measuring the ellipticity and size of the object. `mRrCc` is calculated as a second moments to determine the spread of an object's light in a specific direction. `mE1` and `mE2` are ellipticity components that represent the elongation and orientation of an object.

TABLE	INPUT PARAMETER	DESCRIPTION

PhotoPrimary	`lnLStar_r`
	`lnLExp_r`
	`lnLDeV_r`
		Likelihoods are used to analyze and classify the light of astronomical objects by fitting different models and calculating the likelihood of how well these models represent the object. `DeV_L` is typically used for elliptical galaxies to determine how well it matches the de Vaucouleurs profile, whereas `exp_L` is used for spiral galaxies to assess the accuracy of an object's light in matching an exponential decay model. Finally, `star_L` represents a star or point-like source, which corresponds PSF.

PhotoPrimary And *self-added	`type`	For section 4.1, `type=3` denotes `GALAXY`, while `type=6` denotes `STAR`. For section 4.2, `type=1` denotes `NO_BULGE`, while `type=0` denotes `ROUNDED`. *`BOXY` was not initialized or included because it had a very small total dataset.
SpecObjAll	`class`	Divided into GALAXY and STAR
*self-added	`wgt`	A weight used to balance the asymmetrical overall data. For example, `ROUNDED` has an estimated 1170 total training data, whereas `NO_BULGE` has approximately 360 total training data. This could cause an imbalance in the training/testing process, so we implemented a weighted column to balance it.

Table 3.1

3.3 ARTIFICIAL NEURAL NETWORK 2 (ANNZ2)

The software package we used to run our machine learning algorithms to predict our `STAR/GALAXY` and `ROUNDED/NO_BULGE` was Artificial Neural Network 2, which is an upgraded version of Collister and Lahav's original ANNz1 code [55]. This software can be used to predict redshifts, classify galaxies, stars, and quasars, depending on our project's regression or classification needs, both of which are supported by ANNz2. The machine learning algorithms in ANNz2 are implemented via the Toolkit for Multivariate Analysis (TMVA section 3.3.1) package, which includes multivariate analysis techniques described later in the TMVA section. For our project, we employ the *H*-Matrix discriminant, Fisher discriminant, linear discriminant analysis, Artificial Neural Network, and Boosted Decision Trees. The software package is downloadable via `https://github.com/IftachSadeh/ANNZ` [55].

3.3.1 *TMVA*

The Toolkit for Multivariate Data Analysis with ROOT was created as part of the ROOT framework, which is a machine learning library intended to handle analysis techniques primarily in the field of high energy physics. With the integration of TMVA with ROOT, where the ROOT framework was known to exhibit highly efficient processes for analyzing large datasets and even handling higher dimensional data [72]. TMVA provides a variety of multivariate techniques, including classification and regression methods. Artificial Neural Networks (ANNs), Support Vector Machines (SVM), Boosted Decision Trees (BDT), *H*-Matrix discriminant, k-Nearest Neighbours, Fisher discriminant, rectangular cut optimisation, and other methods are used to distinguish between signal and background events in classification.

In regression methods, they are used to predict continuous variables using algorithms such as linear regression, Boosted Regression Trees (BRT), and ANNs. Both classifier and multivariate regression are implemented through training, testing, and performance evaluation, with user-friendly interfaces to assist new users with minimal knowledge of this framework. TMVA allows users to use scripts in both C++ and Python bindings,

promoting the comfort of a flexible interface. Users can freely manipulate these discrimination techniques, such as parameter fitting, weighted, and variable transforms, and each discriminant has its own modification that we can experiment with. TMVA is an open-source product and can be downloaded through TMVA Toolkit for Multi Variate Analysis - Browse Files at SourceForge.net [67][72].

Data preprocessing is critical in the machine learning workflow because it allows TMVA to improve algorithm performance, transform the input dataset as needed, and ensure accurate results in data analysis. For our project, we used variable normalization (N), decorrelation via square-root of covariance matrix (D), principal component decomposition (PCA), uniform (U), and Gaussian (G) distributions, as described in Table 3.2. Table 3.2 shows the types of variable transforms.

Variable Transform	Description
Normalisation (N)	Also known as min-max normalization, the features are rescaled to values between 0 to 1. The equation of max-min normalization is $X_{\text{norm}} = \frac{X-\min(X)}{\max(X)-\min(X)}$. Normalisation has been shown to produce small standard deviations; however, if there are outliers in a specific feature, normalization will not be able to handle them well because it scales the data to a small interval [72][41].

Variable Transform	Description
Decorrelation (D)	Decorrelation converts the original variable into new variables that are not linearly correlated, implying that there are no relations between them. Decorrelation only works under two conditions: the variables are linearly correlated and they follow the Gaussian distributions. If the condition is not met, the decorrelation process may be ineffective, reducing the model's performance. This occurs because, if the original variables are not linearly correlated, the decorrelation will increase the degree of nonlinearity between the variables. By decorrelating the variables, the features contribute independently to the classification, reducing overfitting in models that do not rely heavily on a set of correlated features [72][3].

Variable Transform	Description
Principal Component Analysis (PCA)	Principal Component Analysis is frequently used to reduce the high-dimensionality of complex and large datasets by focusing on the first few principal components or largest eigenvalues while ignoring the rest of the variables. Reducing the number of variables reduces computation time while retaining important information. The model can now deal with lower-dimensional and less-correlated data, which helps to prevent overfitting. PCA can also reveal variables that have a significant impact on data but were previously invisible. PCA begins by calculating the covariance, which is the linear relationship between variables. The covariance matrix will then be decomposed to yield eigenvectors and eigenvalues. The original data points will be added to the new eigenvectors. These new transformations then ranked coordinates, with the first indicating the largest variance (first principal component), the second indicating the second largest variance (second principal component), and so on. Lower-order principal components have more variance than higher-order components, so retaining these lower-order ones allows us to capture important information in the original dataset while using fewer dimensions [72][33].

Variable Transform	Description
Uniform (U)	Uniformisation transforms the given variable into a uniform distribution interval of [0,1] using the cumulative distribution function. If a variable X has distribution F, the cumulative distribution function at a given point x is defined as the probability that the variable X will have a value equal to or less than x. The variable's value will then be transformed into a new value that is uniformly distributed between 0 and 1. Uniform distributions produce equal probability distribution outcomes. To facilitate effective transformation of the Gaussian distribution, the uniform transform is often performed first [72][14][23].
Gaussian (G)	The Gaussian transform causes the variable to follow the Gaussian distribution, which controls outliers. The symmetric distribution aids in data centering, for example, values considered outliers are brought closer to the mean to reduce their impact on model training [72].

Table 3.2

3.3.2 *H-Matrix Discriminant*

H-Matrix discriminant is an algorithm for classification which distinguish the differences between classes of data by the distribution of feature vectors. A feature vector consists of multiple measurements where each of these measurements is Gaussian distributed and the inverse covariance matrix of the feature vectors is known as the *H*-Matrix. It takes accounts for the relationship of different variables in the feature vector. *H*-Matrix utilizes a multivariate, χ^2 estimator to measure the deviation of a respective feature vector from expected values under one class against another. The final score which indicates whether that feature vector lies in signal or background class depends on the differences of mean

values between two classes alongside the inverse covariance matrix. Initially, for each measurement or an observation, the discriminant score for either signal (S) or background (B) is calculated with,

$$\chi_{\mathrm{U}}{}^{2}(i) = \sum_{k,l=1}^{n_{\mathrm{var}}} (x_k(i) - \overline{x}_{U,k}) C_{U,kl}{}^{-1} (x_l(i) - \overline{x}_{U,l}), \tag{3.1}$$

where $U = \text{Signal (S)}, \text{ Background (B)}, \text{C}$

U^{-1} is the inverse covariance matrices (H-Matrix), and $\overline{x}_{S(B),k}$ is the sample means from training data. Then, the discriminant value is determined through,

$$y_H(i) = \frac{\chi^2{}_B(i) - \chi^2{}_S(i)}{\chi^2{}_B(i) + \chi^2{}_S(i)} \tag{3.2}$$

in which it indicates the likelihood of the feature vector belonging to either signal or background class. It was stated that the performance of H-Matrix underperforms compared to Fisher discriminant for similar types of problems. There are no specific options or configurations available for H-Matrix [72].

3.3.3 *Fisher Discriminant*

The Fisher discriminant focuses on the linear combination of features which works the best in separating the different classes by projecting a line that maximizes the separation between different classes while the variation within those classes is minimized. The working concept of the Fisher discriminant is that the covariance matrix, C which is distributed into the sum of within-class matrix, W and between-class matrix, B where $C_{kl} = W_{kl} + B_{kl}$. Within-class matrix represents the variability within each class, how far the measurements in respective classes deviate from their class mean; meanwhile between-class matrix represents variability between class, how far the classes deviate from overall mean. The within-class matrix is given,

$$W_{kl} = \sum_{U=S,B} \langle U_{U,k} - \overline{x}_{U,k} \rangle \langle x_{U,l} - \overline{x}_{U,l} \rangle = C_{S,kl} + C_{B,kl}. \tag{3.3}$$

The Fisher coefficients are calculated,

$$F_k = \frac{\sqrt{N_S N_B}}{N_S + N_B} \sum_{l=1}^{n_{\text{var}}} {W_{kl}}^{-1} (\overline{x}_{S,l} - \overline{x}_{B,l}), \tag{3.4}$$

where $N_{S(B)}$ is the number of signal (background) events obtained from training sample. The discriminant for each event is measured through the Fisher discriminant,

$$(y_{Fi}(i) = (F_0 + \sum_{k=1}^{n_{\text{var}}} F_k x_k(i), \tag{3.5}$$

where F_0 is determined so that the average of discriminant values for all events remain zero. For between-class matrix,

$$B_{kl} = \frac{1}{2} \sum_{U=S,B} (\overline{x}_{U,k} - \overline{x}_k)(\overline{x}_{U,l} - \overline{x}_l), \tag{3.6}$$

where $\overline{x}_{S(B),k}$ represents average variable x_k for signal (background) samples while $\overline{x}_k$ represents the average of the entire sample.

There is also another method called the Mahalanobis variant which determines the Fisher coefficient,

$$F_k = \frac{\sqrt{N_S N_B}}{N_S + N_B} \sum_{l=1}^{n_{var}} {C_{kl}}^{-1} (\overline{x}_{S,l} - \overline{x}_{B,l}). \tag{3.7}$$

The Fisher discriminant pushes the projection of signal and background classes as far as possible while keeping the data points of the same class close together. The performance of Fisher discriminant is found to degrade if the signal and background are too similar [72].

3.3.4 *Linear Discriminant Analysis (LD)*

The Linear discriminant, like the Fisher discriminant, maximizes the distance between the means of two classes while minimizing variance within each class using a linear model. The linear model's discriminant function is $y(x) = x^{\mathrm{T}}\beta + \beta_0$ where β_0 represents a bias term. The bias term is adjusted so that $y(x) \geq 0$ indicates signal and $y(x) \leq 0$ represents the background class. Data is organized as such vector Y represents the target val-

ues and vector X represents bias term β_0 which is absorbed into vector β . This can be expressed as $Y = X\ \beta$. From applying the least-squares method, we obtain parameters, $\beta = (X^T X)^{-1} X^T Y$ where the term $(X^T X)^{-1} X^T$ can be modified if the events are weighted which leads to $\beta = (X^T W X)^{-1} X^T W Y$. The linear discriminant ranks the input variables from their respective values of coefficients, β which shows the influential features that acts in separating the classes. If the feature variables have similar mean values for different classes, linear discriminant would struggle to separate the classes as it relies heavily on the differences in mean values to establish a line boundary between the classes. There is no configuration option for LD [72].

3.3.5 *Other Machine Learning Methods*

Other machine learning methods worth mentioning include artificial neural networks and boosted decision trees, as we want to see how H-Matrix, LD, and Fisher discriminants compare to the popularly used classifiers in ANNz2.

Artificial Neural Networks (ANNs) are interconnected sets of artificial neurons inspired by the human brain, with the goal of distinguishing between various signal and background inputs. The working concept of ANN is that it begins at the input layer, which receives input data for processing, with each neuron representing a feature of the input data. The hidden layers perform the computational work in their neurons using a weighted sum and pass the result to the next layer. Before that, the result will be propagated and compared to true or false values (error junction). The network will propagate backwards to learn from the computed error, allowing it to modify weights and biases. This is accomplished via the neuron activation function. The output layer generates the final output, which may be a final classification or a continuous value for regression [72][64][77].

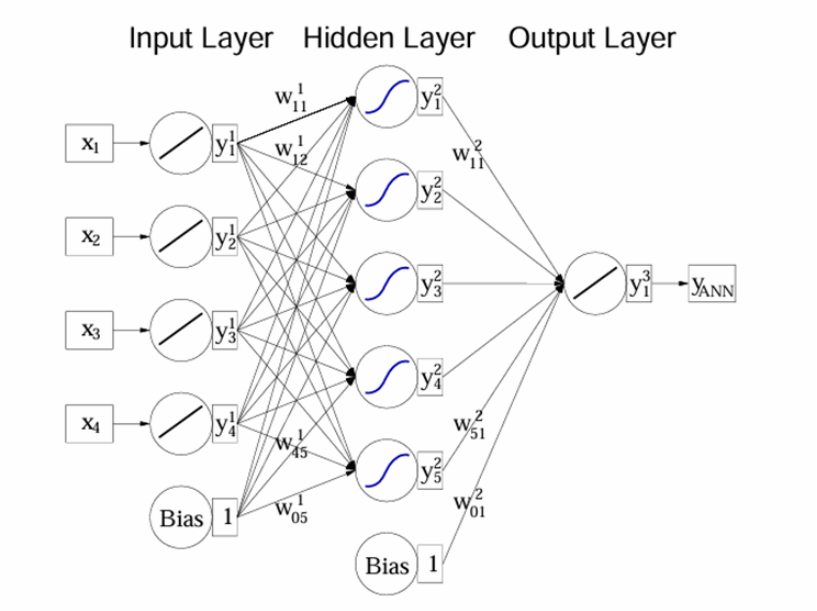

Figure 3.4

Figure 3.4 shows the structure of a multilayer perceptron with only one hidden layer which was obtained from [72]. Multilayer perceptron consists of multiple layers of neurons which are organized into input layers, hidden layers and output layers. Input layers are the neurons that received input features which then passed onto hidden layers that will apply weights or biases to analyse and capture the complex patterns in the data. Finally, the output layer consists the neurons which will output the final predictions [89]. By using a tanh activation function, the output of a network is

$$y_{\text{ANN}} = \sum_{j=1}^{n_h} y_j{}^{(2)} w_{j1}{}^{(2)} = \sum_{j=1}^{n_h} \tanh\left(\sum_{i=1}^{n_{\text{var}}} x_i w_{ij}{}^{(1)}\right).w_{j1}^{(2)}, \tag{3.8}$$

where n_{var} represents the number of neurons in the input layer while n_h represents the number of neurons in hidden layers. $w_{ij}{}^{(1)}$ is the weight for input-layer neuron i and hidden-layer neuron j. $w_{i1}{}^{(2)}$ is the weight for hidden-layer neuron j and output neuron. The error junction is measured,

$$E\left(x_1, \ldots x_N | w\right) = \sum_{a=1}^{N} E_a\left(x_a | w\right) = \sum_{a=1}^{N} \frac{1}{2}(y_{\text{ANN},a} - \widehat{y}_a), \tag{3.9}$$

where w represents the weights in the network. For the error function, E to decrease thus the weight will be updated in output layer,

$$\Delta w_{j1}^{(2)} = -\eta \sum_{a=1}^{N} \frac{\partial E_a}{\partial w_{ij}{}^{(2)}} = -\eta \sum_{a=1}^{N} (y_{ANN,a} - \widehat{y}_a) y_{\mathrm{j,a}}^{(2)}, \tag{3.10}$$

while for the hidden layers,

$$\Delta w_{ij}^{(1)} = -\eta \sum_{a=1}^{N} \frac{\partial E_a}{\partial w_{ij}^{(1)}} = -\eta \sum_{a=1}^{N} (y_{ANN,a} - \widehat{y}_a) y_{j,a}{}^{(2)} (1 - y_{j,a}^{(2)}) w_{j1}^{(2)} x_{i,a}, \tag{3.11}$$

For ANNs, TMVA provides three types of networks which are MLP, ANN from ROOT networks and the Clermont-Ferrand neural network [72]. Clermont-Ferrand shows the worst performance among the three types thus being excluded in our study. We have carried out our test using MLP networks. MLP exhibits variable ranking where we get to determine which variables have the most influence on the network's prediction. This is given by the importance,

$$I_i = \overline{x}_i^2 \sum_{j=1}^{n_h} (w_{ij}^{(1)}), \tag{3.12}$$

$i = 1, \ldots n_{var}$ where $\overline{x}_i^2$ is the mean of the sample of input variable i. A higher sum of the weights-squared indicates a greater influence on the input variables. Some of the important configurations that was utilized were `TrainingMethod`, `SamplingTraining`, `NeuronType`, `UseRegulator`, `HiddenLayers` and `RandomSeed`. For more configurations regarding ANNs, refer to [72].

The final algorithm is boosted decision trees (BDTs). A decision tree is a model that makes decisions along a sequential and hierarchical pathway, with data divided into branches at each node. Edges are the paths that connect nodes, and leaf nodes are the terminal nodes that produce the outcome. Boosting techniques are applied to decision trees in such a way that they create a forest of trees, with each tree correcting its error through learning. The final outcome or prediction is calculated by taking the weighted average of all tree series predictions.

Individual decision trees have a higher bias because the number of splits is limited, which is known as weak classifiers. New trees are added in an iterative process because each tree generated focuses on making accurate predictions while the previous ones were inaccurate. The new iterative process of creating trees is carried out by reweighting the data

and learning from the inaccurate ones. Thus, for classification, the final output is the sum of the predictions from all individual trees. Overfitting is reduced with the addition of trees. BDTs also rank the input variables that influence the splits in decision trees [72][16][39].

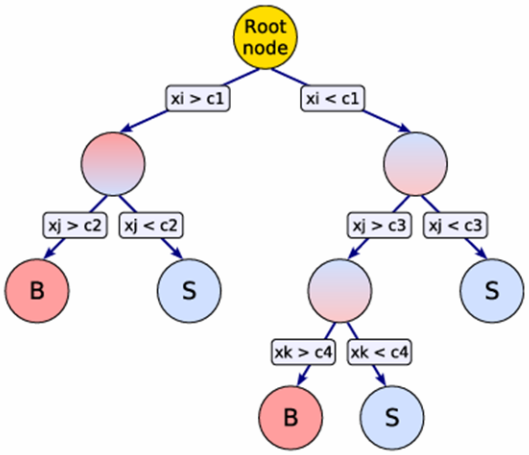

Figure 3.5

Figure 3.5 shows the Architecture of a decision tree which was obtained from [72]. BDTs configuration options that were utilized were `NTrees`, `BoostType`, `BaggedSampleFraction`, `nCuts` and `MinNodeSize`. For more configuration options, refer to [72].

3.4 EVALUATION METRICS

After the models and algorithms have completed their predictions, we need to perform model validation using evaluation metrics like confusion matrix, accuracy, precision, recall, and F1 score to compare which models are the most effective and appropriate for the classification. We also learn more about which parameters or configuration options cause the chosen model to change, either improving or decreasing in accuracy. To evaluate the models, we have written our own Python script. For Star/Galaxy separation, refer to Appendix 1; while for Rounded/No_Bulge separation, refer to Appendix 2.

3.4.1 *Confusion Matrix*

A confusion matrix provides a visual summary of the model's performance by comparing the actual outcomes with predictions made by the model. The instances made in actual

class and predicted class are tabulated in a 2×2 table, as shown in Table 3.3. Table 3.3 shows the structure of a confusion matrix.

- True Positive (TP): Model correctly predicts the positive (signal) class.
- True Negative (TN): Model correctly predicts the negative (background) class.
- False Positive (FP): Model incorrectly predicts the actual negative instance as positive class.
- False Negative (FN): Model incorrectly predicts the actual positive instance as a negative class.

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Table 3.3

As an illustration of class imbalance, the discriminant we used classifies the signal class with good accuracy but predicts the background class poorly. The confusion matrix thus indicates the overall proportions as well as the ratio of predicted instances in both classes [7].

3.4.2 *Accuracy*

Accuracy is used to evaluate the overall proportion of both true positives and true negatives for its performance in classification,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.13}$$

However, accuracy is not optimal when the dataset is imbalanced. For example, if TP has 5 values while TN has 90, this indicates that the accuracy is more likely to classify the signal than the background [4].

3.4.3 *Precision*

Precision measures the accuracy of how the model classifies the positive predictions of both true and false positives. A good model classifier should have precision near or equal to 1,

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{3.14}$$

For instance, if the FP is known to have a high value, the precision will undoubtedly decrease, indicating that the model frequently misclassifies the signal as background [4].

3.4.4 *Recall*

Recall guarantees the model's sensitivity in identifying positive occurrences, whereas precision seeks to determine the accuracy of positive predictions [4]. We need to determine whether the model can classify the signal better than the background or inversely,

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3.15}$$

3.4.5 *F1 Score*

A high F1 score guarantees that the model must function well in terms of both recall and precision. The proportion and balance between recall and precision can be seen in the F1 score [35].

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{3.16}$$

3.4.6 *Gaussian Kernel Density Estimation graph*

For this project, I will employ Gaussian Kernel Density Estimation (KDE) and scatter plot graphs for each discriminant's performance to visualize how well each instance is able to

distribute across the classification. Gaussian KDE plots each data point to visualize the distribution of the data and produces a smooth curve of the probability density function for the total data. We may learn about the characteristics of each discriminant and how biased or effective they are at classifying both positive and negative instances by displaying the spreads, densities, and overlaps of curves. The region where the discriminant is uncertain about their ability to distinguish between the classes is represented by the overlap of curves. I have chosen a decision threshold of 0.5 for this study, which is a vertical line that establishes the classification cutoff. In order words, this threshold is the transition line of one class to another [21]. Refer to Appendix 3 for more information on this.

3.4.7 *Statistical Techniques and Calculation*

Lastly, the purpose of this section is to measure the deviation of the predicted values from theoretical values and evaluating how well each discriminant and variable transform classifies both positive and negative instances equally. We will use statistical analysis, such as the standard deviation, to evaluate how well or how imbalanced each discriminant is in terms of classification in both cases. A low standard deviation denotes a high degree of consistency in the discriminant's performance throughout that particular instance, whereas a high standard deviation suggests a notable imbalance in the classification. Refer to appendix 4 for more information on this.

3.5 METHODOLOGY

Our goals in this study were based on how well different classification models could categorize stars, galaxies, and various galaxy morphologies. The two main components of our project's approach, each focusing on a distinct aspect of classification and purpose, are meant to help us achieve our goals. In Part A, the classification of stars and galaxies is the main topic. Various discriminants' performance to distinguish between stars and galaxies are examined; ANN, BDT, *H*-matrix, Fisher, Mahalanobis and linear discriminant, along with how different variable transforms impact the discriminants' performance. In order to compare the effectiveness of discriminants in terms of differentiating galaxy

morphology, Part B focuses on `NO_BULGE/ROUNDED` classification. We will use the following 11 input parameters for the training process for each part: `lnLStar_r`, `lnLExp_r`, `lnDeV_r`, `psfMag_r`, `fiberMag_r`, `petroMag_r`, `modelMag_r`, `petroRad_r`, `petroR50_r`, `petroR90_r`, and `Tmath::Log(pow(mE1_r,2))`. `Tmath` is briefly summarized as sets of mathematical functions and computations lies in the framework of ROOT which usually used in data analytic tasks or scripts. Thus, `Tmath::Log(pow(mE1_r,2))` is defined as $\log_2(\text{mE1\_r})$

3.5.1 *Part A: Comparative Analysis of Discriminants in Star-Galaxy Classification*

ANNz2 includes pre-made scripts and datasets for distinguishing stars and galaxies. We used all of the methods mentioned above to complete the star-galaxy classification task for this part. We then used our Jupyter notebook to execute our evaluation Python script on the output file and observe the key metrics, which included the confusion matrix, accuracy, precision, recall, and F1 score. We can identify which model is best suited to the classification task and compete against others depending on key indicators. We will also investigate how the variable transform might impact the computing efficiency of the discriminants. We will use Normalisation (N), Gaussianisation (G), Uniformization (U), Decorrelation (D), and Principal Component Analysis (PCA) to alter the variables. The dataset we used for this part is the star-galaxy dataset which is balanced where for both galaxy and star contain the same amount of data in training (2000), testing (2000) and evaluation files (2000).

3.5.2 *Part B: Comparative Analysis of Discriminants in No_bulge-Rounded Classification*

Similar to Part A, we ran the algorithms with few modifications in the python script for `No_Bulge-Rounded` classification. Initially, we were trying to classify for disk galaxies in the GZ2 decision tree which contains three types: boxy, rounded and no bulge galaxies. Nevertheless, we only obtained a very small amount of data for boxy galaxies using the CasJobs in Skyserver SDSS. Because of the excessive data imbalance caused by boxy galaxies, we decided to only classify rounded galaxies and not any bulge galaxies. We created a new, weighted input parameter specifically for this section. Given that the rounded and no bulge datasets are not perfectly balanced; rounded has 1170 total data, while the no

bulge dataset has 360 data, we chose to test whether the weighted dataset would affect the computational efficiency of all the algorithms when compared to the non-weighted dataset. For the weightings, we applied 0.33 for rounded galaxies and 1.08 for no bulge galaxies respectively. Next, we use our evaluation script to run the final output and compare the algorithms.

4

RESULTS AND DISCUSSION

4.1 PART A

A range of models, including Artificial Neural Networks (ANNs), Boosted Decision Trees (BDT), *H*-Matrix discriminants, linear discriminants, and Fisher discriminants using various techniques (Fisher and Mahalanobis), are among those used in the data analysis for the classification of stars and galaxies. Both scatter plots and graphs with the Gaussian distribution will be shown in advance to show how well each discriminant can categorize the star and galaxy data points. Then, an evaluation table containing statistical calculations such standard deviations of both positive and negative instances will be presented to show how consistent each discriminant is able to classify . The metrics include accuracy, precision, recall, F1 score, and confusion matrix.

4.1.1 *No Variable Transform*

Table 4.1: Performance assessment of each discriminant using no variable transform.

Discriminants	Accuracy	Precision	Recall	F1 Score	Confusion Matrix
ANN	0.9930	0.9950	0.9910	0.9930	$\begin{bmatrix} 998 & 5 \\ 9 & 988 \end{bmatrix}$
BDT	0.9905	0.9970	0.9840	0.9904	$\begin{bmatrix} 1000 & 3 \\ 16 & 981 \end{bmatrix}$
H-Matrix	0.9475	0.9980	0.9980	0.9500	$\begin{bmatrix} 900 & 103 \\ 2 & 995 \end{bmatrix}$
Fisher	0.9630	1.0000	0.9258	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$
Mahalanobis	0.9630	1.0000	0.9258	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$
Linear Discriminant	0.9630	1.0000	0.9258	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$

Table 4.1 shows the overall performance metrics of all six types of discriminants under the variable transform 'None' condition. ANN has the highest accuracy (0.993) among the discriminants, allowing it to efficiently classify both true positives and true negatives across the dataset. The number of false negatives and true negatives is relatively low, according to the confusion matrix. Figure 4.1 shows that ANN exhibits minimal deviation, with the majority of them packed together in their respective diagonal predictions. The discriminant with the highest F1 Score is also ANN, (0.993) which maintains the balance of identifying positives while keeping both true negatives and false positives low.

In terms of precision, both the Fisher discriminant and the Linear discriminant have a precision of (1), indicating that they can completely classify all of the galaxies without

false positives, implying that these three discriminants completely avoided misclassifying stars as galaxies. Figure 4.1 shows that none of the data points for stars are outliers because they are all within the respective threshold boundaries. However, the distribution of data points for the galaxy in these three discriminants is described as highly scattered. As a result, these discriminants are more accurate when they classify stars rather than galaxy.

H-Matrix has the highest recall (0.998), allowing it to minimize misclassifying galaxies as stars. In contrast to Fisher and Linear discriminants, *H*-Matrix can accurately classify the majority of the galaxy but performs poorly for stars.

Thus, we assert that ANN is the best classifier for the no variable transform. Among the four discriminants, Fisher and Linear discriminant has the best accuracy, making it a superior classifier than *H*-Matrix.

Table 4.2: Standard deviation of the instances for no variable transform.

Discriminants	Standard deviation of negative instances	Standard deviation of positive instances
ANN	2.0	5.0
BDT	6.5	9.5
H-Matrix	50.5	47.5
Fisher	37.0	40.0
Mahalanobis	37.0	40.0
Linear Discriminant	37.0	40.0

Table 4.2 shows that ANN has the lowest standard deviation in both negative (2.0) and positive instances (5.0), implying that it is more accurate and consistent in classifying stars and galaxies than other discriminants. However, when focusing on the four discriminants, both the Fisher and Linear discriminant methods produce lower standard deviations for negative instances (37.0) and positive instances (40.0) than the *H*-matrix. This implies that

Fisher and Linear discriminants are more efficient than *H*-Matrix, despite their poor and inconsistent performance when compared to other discriminants.

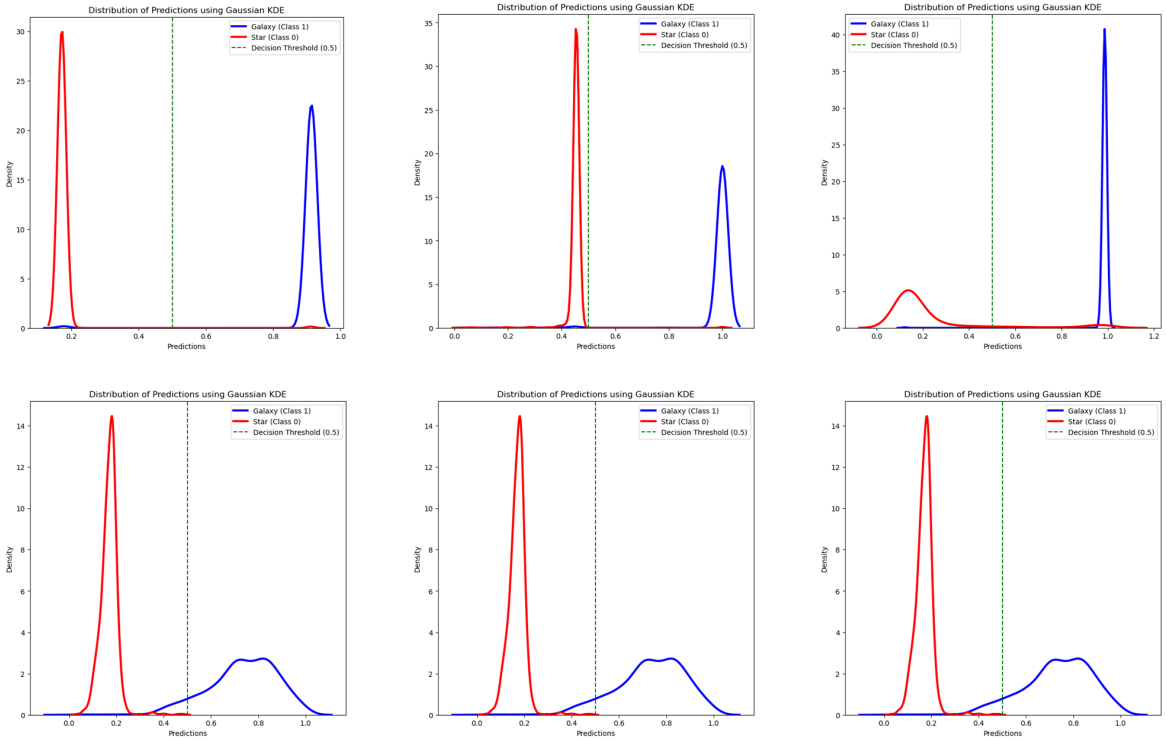

Figure 4.1: Gaussian graph for ANN (top left), BDT (top middle), H-Matrix (top right), Fisher (bottom left), Mahalanobis (bottom middle) and linear discriminant (bottom right) using no variable transform.

4.1.2 *Variable transform N*

Table 4.3: Performance assessment of each discriminant using variable transform N.

Discriminants	Accuracy	Precision	Recall	F1 Score	Confusion Matrix
ANN	0.9965	0.9980	0.9950	0.9965	$\begin{bmatrix} 1001 & 2 \\ 5 & 992 \end{bmatrix}$
BDT	0.9905	0.9970	0.9840	0.9904	$\begin{bmatrix} 1000 & 3 \\ 16 & 981 \end{bmatrix}$
H-Matrix	0.9745	0.9062	0.9980	0.9499	$\begin{bmatrix} 900 & 103 \\ 2 & 995 \end{bmatrix}$
Fisher	0.9630	1.0000	0.9258	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$
Mahalanobis	0.9630	1.0000	0.9258	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$
Linear Discriminant	0.9630	1.0000	0.9258	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$

According to Table 4.3, ANN has the highest accuracy (0.9965), with false positive and false negative instances of 2 and 5, respectively, implying that ANN is highly accurate and consistent when classifying galaxies and stars, as well as the highest F1 Score (0.9965), indicating a balanced precision and recall. In terms of precision, both Fisher and Linear discriminants have a maximum of 1 and a false positive of 0. The *H*-Matrix discriminant has the highest recall (0.9499). Thus, for the variable transform 'N', we claim ANN to be the best classifier, whereas *H*-Matrix outperforms Fisher and linear discriminants in terms of accuracy.

Table 4.4: Standard deviation of the instances for variable transform N.

Discriminants	Standard deviation of negative instances	Standard deviation of positive instances
ANN	1.5	4.5
BDT	6.5	9.5
H-Matrix	50.5	47.5
Fisher	37.0	40.0
Mahalanobis	37.0	40.0
Linear discriminant	37.0	40.0

Table 4.4 shows that ANN has the lowest standard deviation of negative instances (1.5) and positive instances (4.5) when compared to other discriminants, making it the most consistent classifier. Among the four discriminants, *H*-Matrix is the most inconsistent, having the largest standard deviation for both positive (47.5) and negative (50.5) instances.

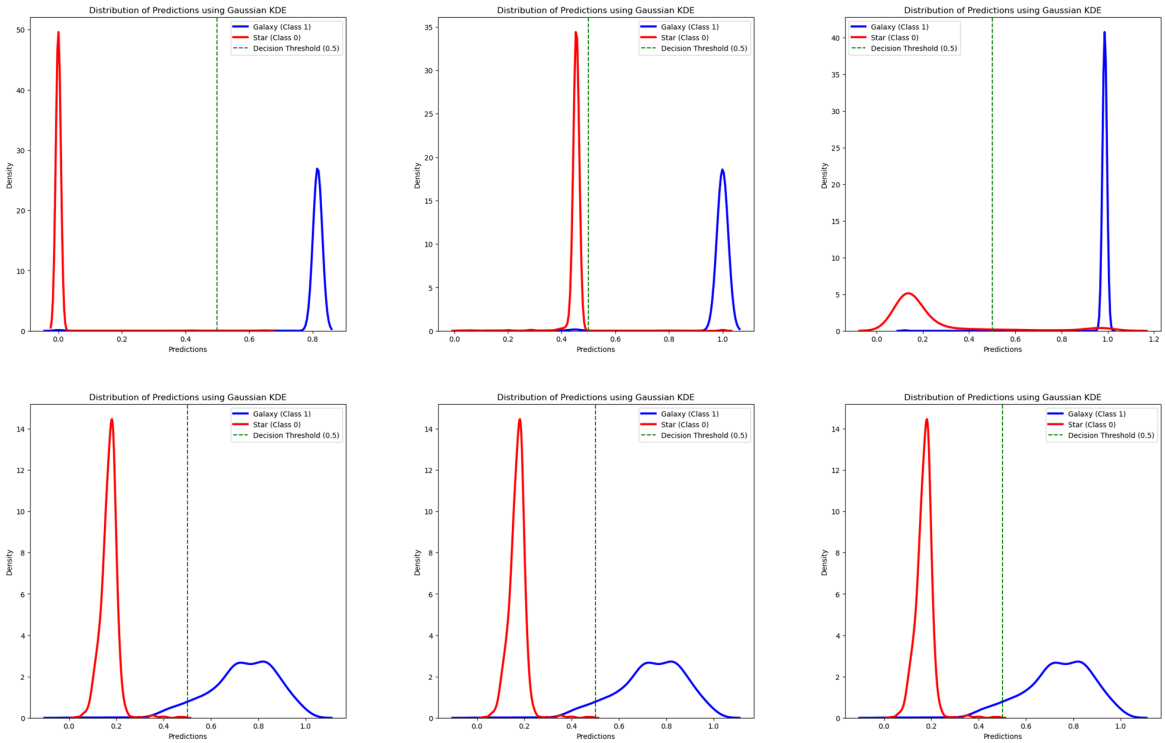

Figure 4.2: Gaussian graph for ANN (top left), BDT (top middle), H-Matrix (top right), Fisher (bottom left), Mahalanobis (bottom middle) and linear discriminant (bottom right) using N variable transform.

4.1.3 *Variable transform U*

Table 4.5: Performance assessment of each discriminant using variable transform U.

Discriminants	Accuracy	Precision	Recall	F1 Score	Confusion Matrix
ANN	0.9930	0.9920	0.9940	0.9930	$\begin{bmatrix} 995 & 8 \\ 6 & 991 \end{bmatrix}$
BDT	0.9905	0.9979	0.9910	0.9904	$\begin{bmatrix} 1001 & 2 \\ 17 & 980 \end{bmatrix}$
H-Matrix	0.9810	0.9669	0.9970	0.9812	$\begin{bmatrix} 968 & 35 \\ 3 & 994 \end{bmatrix}$
Fisher	0.9910	0.9870	0.9950	0.9910	$\begin{bmatrix} 990 & 13 \\ 5 & 992 \end{bmatrix}$
Mahalanobis	0.9910	0.9870	0.9950	0.9910	$\begin{bmatrix} 990 & 13 \\ 5 & 992 \end{bmatrix}$
Linear Discriminant	0.9910	0.9870	0.9950	0.9910	$\begin{bmatrix} 990 & 13 \\ 5 & 992 \end{bmatrix}$

According to Table 4.5, ANN has the best accuracy (0.993) and F1 score (0.993). BDT has the best precision (0.9979) and the lowest false positive rate by 2. *H*-Matrix has the highest recall (0.997) and the fewest false negatives by 3. In conclusion, ANN has been perceived the best classifier for variable transform 'U', whereas Fisher and Linear discriminant are more suited for classification than *H*-Matrix. Surprisingly, both Fisher discriminants and linear discriminants outperform *H*-Matrix and even BDT as classifiers, with high accuracy (0.991), which is closely aligned with ANN.

Table 4.6: Standard deviation of the instances for variable transform U.

Discriminants	Standard deviation of negative instances	Standard deviation of positive instances
ANN	1.0	2.0
BDT	7.5	10.5
H-MATRIX	16.0	13.0
Fisher	4.0	1.0
Mahalanobis	4.0	1.0
Linear Discriminant	4.0	1.0

According to Table 4.6, ANN has the lowest standard deviation of instances, generating the greatest F1 Score among other discriminants. Both Fisher and Linear discriminants have proved extremely consistent in providing both positive and negative consistent instances, with standard deviations of 1.0 and 4.0, respectively. These measurements show that, while these discriminants are less consistent than ANN, they outperform BDT and *H*-Matrix discriminants as reliable classifiers.

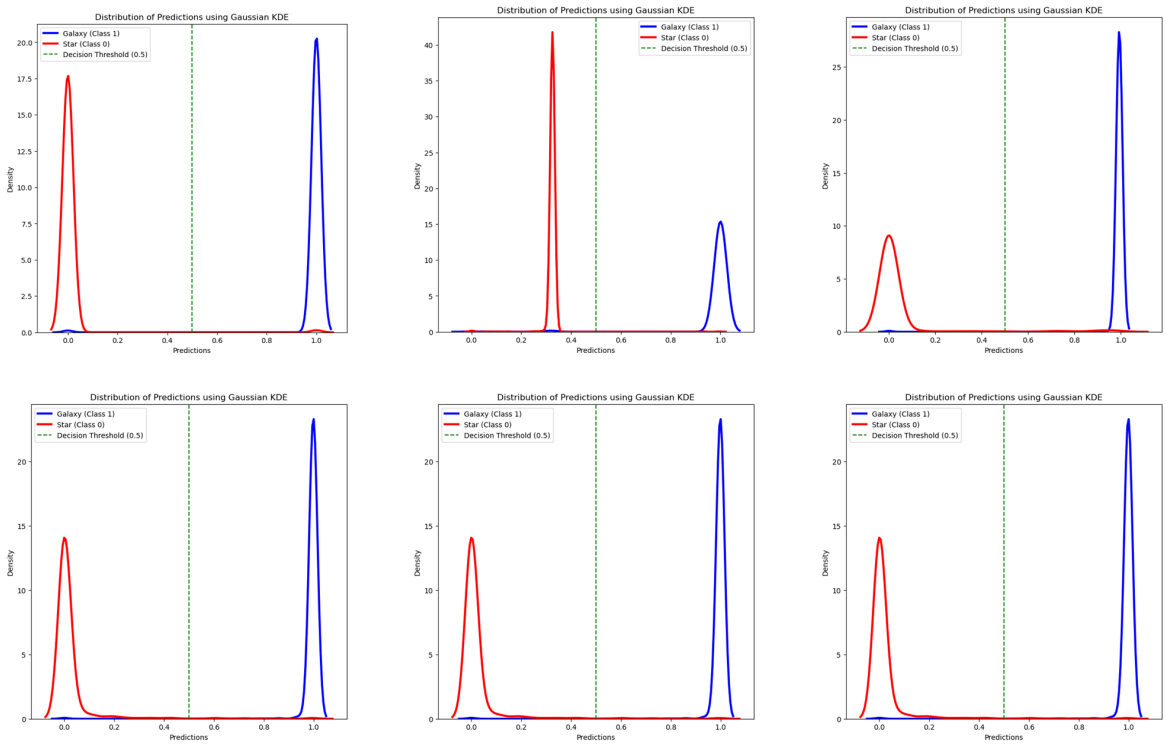

Figure 4.3: Gaussian graph for ANN (top left), BDT (top middle), H-Matrix (top right), Fisher (bottom left), Mahalanobis (bottom middle) and linear discriminant (bottom right) using U variable transform.

4.1.4 *Variable transform G*

Table 4.7: Performance assessment of each discriminant using variable transform G.

Discriminants	Accuracy	Precision	Recall	F1 Score	Confusion Matrix
ANN	0.9915	0.9930	0.9900	0.9915	$\begin{bmatrix} 996 & 7 \\ 10 & 987 \end{bmatrix}$
BDT	0.9915	0.9980	0.9850	0.9914	$\begin{bmatrix} 1001 & 2 \\ 15 & 982 \end{bmatrix}$
H-Matrix	0.9900	0.9841	0.9960	0.9900	$\begin{bmatrix} 987 & 16 \\ 4 & 993 \end{bmatrix}$
Fisher	0.9900	0.9870	0.9930	0.9900	$\begin{bmatrix} 990 & 13 \\ 7 & 990 \end{bmatrix}$
Mahalanobis	0.9900	0.9870	0.9930	0.9900	$\begin{bmatrix} 990 & 13 \\ 7 & 990 \end{bmatrix}$
Linear Discriminant	0.990	0.9870	0.9930	0.9900	$\begin{bmatrix} 990 & 13 \\ 7 & 990 \end{bmatrix}$

Table 4.7 demonstrates that both ANN and BDT achieved the highest accuracy (0.9915). If we look at the confusion matrix, we can see that ANN has a consistent number of false positives and false negatives, 7 and 10, respectively, implying that it is balanced in terms of producing errors. Meanwhile, BDT has more false negatives than false positives, indicating an imbalance and bias when classifying stars more accurately than galaxies. This also contributes to ANN having the highest F1 score (0.9915). Because BDT has just 2 false positives, it has the highest precision (0.998), whereas *H*-Matrix has only 4 false negatives, implying that it has the highest recall among other discriminants (0.996). Thus, we declare ANN as the most suitable classifier for variable transform G.

Table 4.8: standard deviation of the instances for variable transform G

Discriminants	Standard deviation of negative instances	Standard deviation of positive instances
ANN	1.5	4.5
BDT	6.5	9.5
H-MATRIX	6.0	3.0
Fisher	3.0	0.0
Mahalanobis	3.0	0.0
Linear Discriminant	3.0	0.0

Based on Table 4.8, ANN has the lowest standard deviation of instances and when compared to other discriminants. Other discriminants, such as Fisher, Mahalanobis and linear discriminants, can classify both star and galaxy consistently due to their very low standard deviation of positive instances by 0. This can be proven also by observing the confusion matrix in table 4.7 where it shows the true positives and true negatives contains 990 each equally. However, when observing the figure 4.4, the galaxy is not consistently classified at a denser specific data point compared to the star where the concentration of classification is denser between 0 and 0.2. Nevertheless, Fisher, Mahalanobis and linear discriminants are the subsequent best classifiers for variable transform 'G', following ANN.

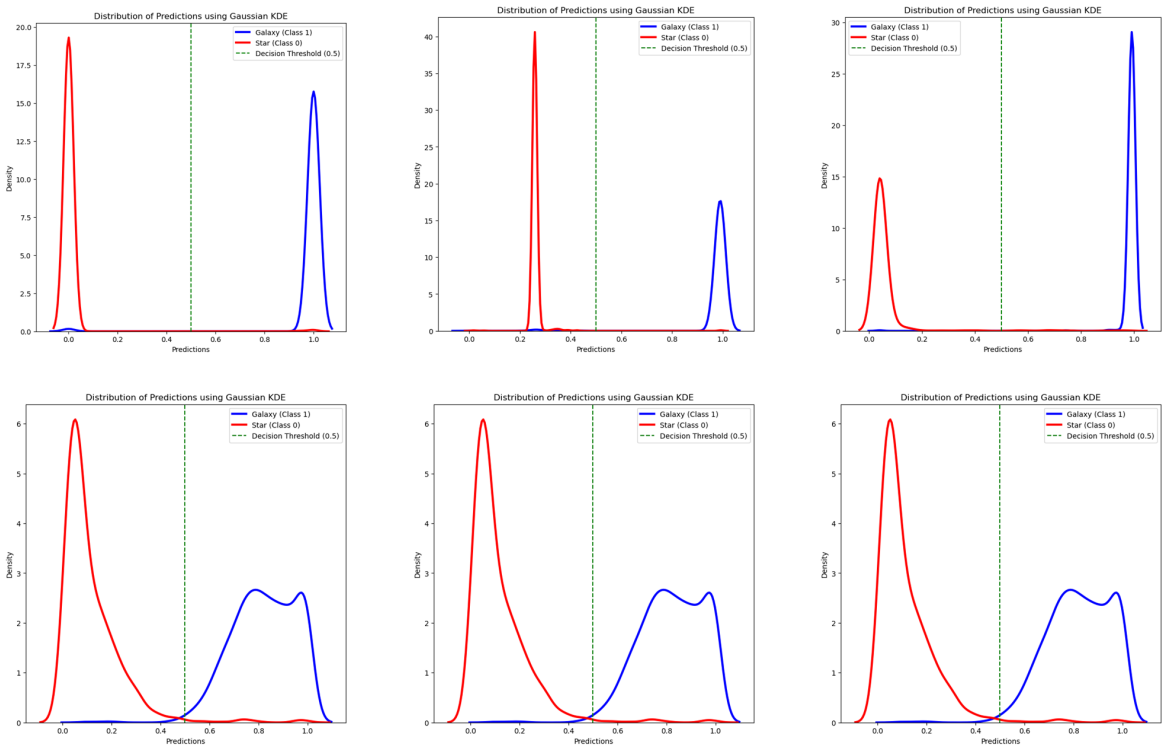

Figure 4.4: Gaussian graph for ANN (top left), BDT (top middle), H-Matrix (top right), Fisher (bottom left), Mahalanobis (bottom middle) and linear discriminant (bottom right) using G variable transform.

4.1.5 *Variable transform D*

Table 4.9: Performance assessment of each discriminant using variable transform D.

Discriminants	Accuracy	Precision	Recall	F1 Score	Confusion Matrix
ANN	0.9930	0.9920	0.9940	0.9930	$\begin{bmatrix} 995 & 8 \\ 6 & 991 \end{bmatrix}$
BDT	0.9945	0.9980	0.9910	0.9945	$\begin{bmatrix} 1001 & 2 \\ 9 & 988 \end{bmatrix}$
H-Matrix	0.9475	0.9062	0.9980	0.9500	$\begin{bmatrix} 900 & 103 \\ 2 & 995 \end{bmatrix}$
Fisher	0.9630	1.0000	0.9260	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$
Mahalanobis	0.9630	1.0000	0.9260	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$
Linear Discriminant	0.9630	1.0000	0.9260	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$

According to Table 4.9, the BDT has the best accuracy (0.9945), and also has the greatest F1 Score (0.9945) suggesting that it generates balanced, low-error results. Fisher discriminants and Linear discriminants both have high precision (1). *H*-Matrix has the highest recall (0.998). We define BDT as the best classifier for variable transform 'D', with Fisher discriminant and Linear discriminant outperforming *H*-Matrix.

Table 4.10: Standard deviation of the instances for variable transform D.

Discriminants	Standard deviation of negative instances	Standard deviation of positive instances
ANN	1.0	2.0
BDT	3.5	6.5
H-Matrix	50.5	47.5
Fisher	37.0	40.0
Mahalanobis	37.0	40.0
Linear Discriminant	37.0	40.0

Based on Table 4.10, BDT is inconsistent in producing positive (3.5) and negative instances (6.5) due to its higher standard deviation than ANN. When comparing the four discriminants, Fisher discriminants and linear discriminants have lower standard deviation than *H*-Matrix, implying that they are more consistent than *H*-Matrix.

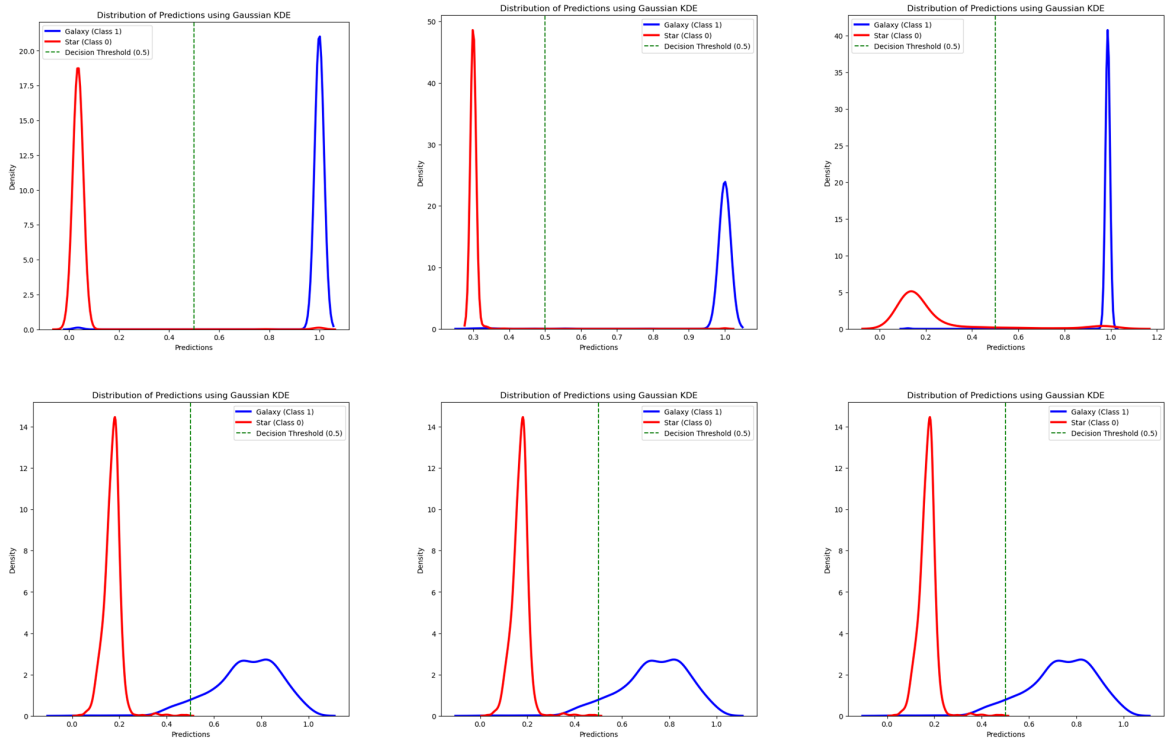

Figure 4.5: Gaussian graph for ANN (top left), BDT (top middle), H-Matrix (top right), Fisher (bottom left), Mahalanobis (bottom middle) and linear discriminant (bottom right) using D variable transform.

4.1.6 *Variable transform PCA*

Table 4.11: Performance assessment of each discriminant using variable transform PCA.

Discriminants	Accuracy	Precision	Recall	F1 Score	Confusion Matrix
ANN	0.9885	0.9959	0.9810	0.9884	$\begin{bmatrix} 1000 & 3 \\ 15 & 982 \end{bmatrix}$
BDT	0.9950	0.9970	0.9930	0.9950	$\begin{bmatrix} 1000 & 3 \\ 7 & 990 \end{bmatrix}$
H-Matrix	0.9475	0.9062	0.9980	0.9499	$\begin{bmatrix} 900 & 103 \\ 2 & 995 \end{bmatrix}$
Fisher	0.9630	1.0000	0.9258	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$
Mahalanobis	0.9630	1.0000	0.9258	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$
Linear Discriminant	0.9630	1.0000	0.9258	0.9614	$\begin{bmatrix} 1003 & 0 \\ 74 & 923 \end{bmatrix}$

According to Table 4.11, BDT has the greatest accuracy (0.995) and F1 score (0.995). Fisher and linear discriminants obtained the maximum accuracy (1), while *H*-Matrix got the highest recall (0.998). Thus, BDT is the best classifier for variable transform 'PCA', whereas Fisher discriminant and linear discriminant are better classifiers than *H*-Matrix due to their greater accuracy (0.963).

Table 4.12: Standard deviation of the instances for variable transform PCA.

Discriminants	Standard deviation of negative instances	Standard deviation of positive instances
ANN	6.0	9.0
BDT	2.0	5.0
H-Matrix	50.5	47.5
Fisher	37.0	40.0
Mahalanobis	37.0	40.0
Linear Discriminant	37.0	40.0

According to Table 4.12, BDT is the most appropriate classifier due to its low standard deviations (2.0) for negative instances and (5.0) for positive instances. When compared to Fisher and linear discriminants, *H*-Matrix performs the worst, with the highest standard deviation errors of (50.5) and (47.5) for negative and positive instances, respectively. This suggests that Fisher and linear discriminants are better suited than *H*-Matrix.

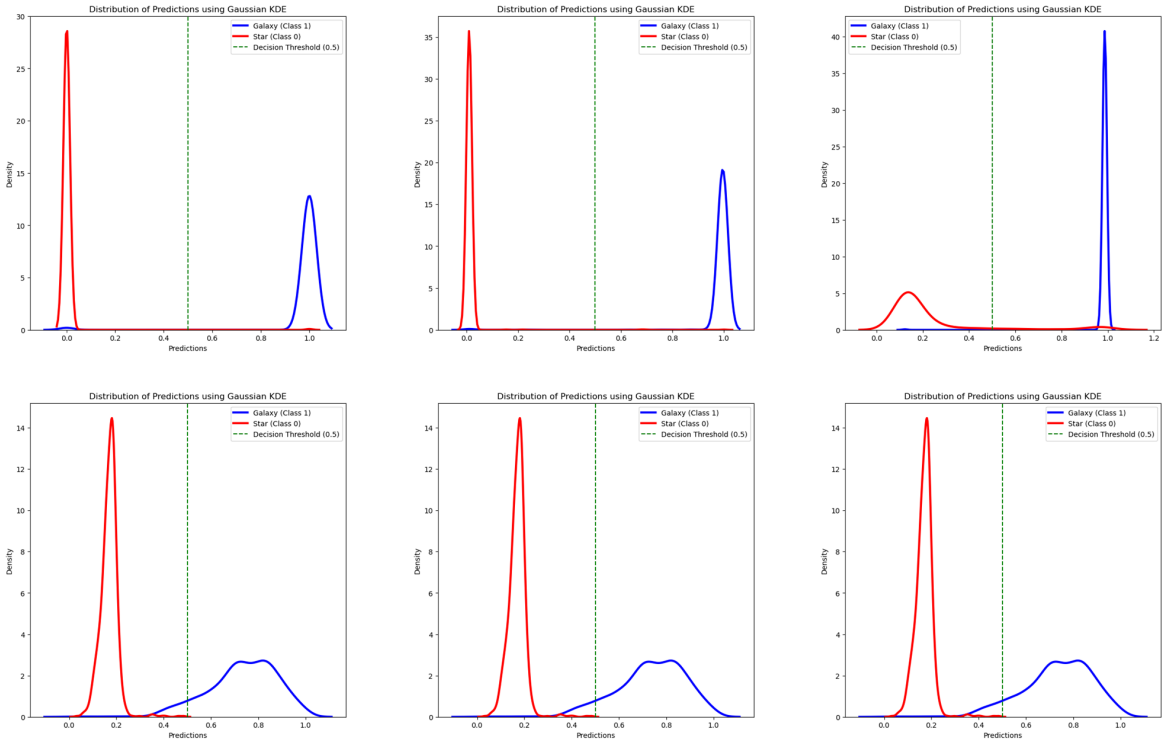

Figure 4.6: Gaussian graph for ANN (top left), BDT (top middle), H-Matrix (top right), Fisher (bottom left), Mahalanobis (bottom middle) and linear discriminant (bottom right) using PCA variable transform.

4.1.7 *Part A Discussion*

From the analysis made in part A for all ranges of variable transforms, we had decided variable transform U to be the most suitable configuration as it is able to output the highest accuracy among the chosen machine learning algorithms; *H*-Matrix, Fisher, Mahalanobis and linear discriminant. We chose these four methods as ANN and BDT are not our main focus in this study. The highest accuracy was produced by Fisher, Mahalanobis and linear discriminant (0.991).

Other important aspects that we can take away from this section include the fact that nearly every variable transform, *H*-Matrix, is stated to have the highest recall among other discriminants. By definition this means that *H*-Matrix able to avoid the issue of misclassifying actual galaxy data as predicted star which implies low false negatives (FN). Although *H*-Matrix excels in recall, its precision which is its ability to avoid falsely classify actual stars as galaxies, is observed to be inconsistent throughout different variable transformation options. In different scope of objective such as one's goal is to ensure every galaxy is classified correctly even at expense of incorrectly classifying some stars as galaxies and with confident to fully avoid misclassifying galaxy as star, they are able to fully utilize *H*-Matrix with its significantly high bias towards galaxy classification. The significant bias trait of *H*-Matrix suggests that it is a suitable option for users that aims to prioritize complete detection of galaxies over the classification of stars.

Inversely, we had observed that Fisher and linear discriminants were able to produce a precision of 1 which exhibits the perfect ability to avoid misclassifying actual stars as predicted galaxies, implying its 0 false positives. However, both of these discriminants output inconsistent recall throughout different variable transforms. Such scenario that aims to map only stars, one is able to utilize both Fisher or linear discriminants as they exhibit bias traits towards stars classification.

There was a noticeable "RandomSeed=300000" setting for discriminants ANN to prevent the error shown in Figure 4.7. We suspected that the reason for this could be "Random-Seed," which is one of the configurations for Artificial Neural Network, as we use this

method in part B as well to avoid errors in ANN. We would not go into detail to prioritize or elaborate on these specific issues because their importance is not central to our project's scope. Figure 4.43 shows the ANN error in Part A.

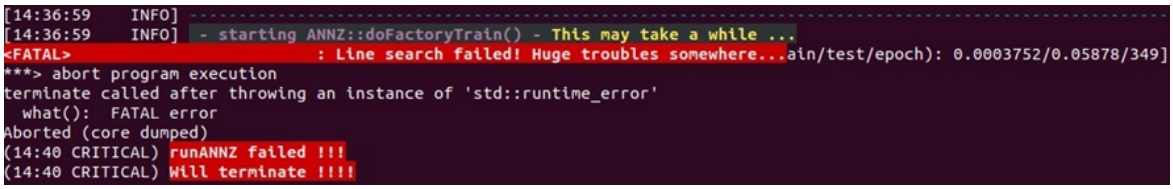

Figure 4.7

4.2 PART B

Part B will employ the same discriminants as part A to categorize rounded and no bulge galaxies. This section will experiment with three different types of datasets due to differences in the data for rounded and no bulge, where they contain 1170 and 360, respectively:

- Dataset A: dataset (360:360) non-weighted
- Dataset B: all data (1170:360) non-weighted
- Dataset C: all data (1170:360) weighted

Variable transform U will be used for all of the dataset. Note that the "RandomSeed" configuration option for ANN was set to 1000 to avoid the error shown in Figure 4.7. Due to the imbalance in the dataset, F1 score is suitable for measuring the discriminants' ability to identify a certain class, whereas accuracy is adequate for a well-balanced dataset. As a result for the parts utilizing all the dataset, we will assess and make the final decision with the highest F1-score.

4.2.1 *Dataset A*

Table 4.13: Performance assessment of each discriminant for dataset A.

Discriminants	Accuracy	Precision	Recall	F1 Score	Confusion Matrix
ANN	0.9236	0.9296	0.9167	0.9231	$\begin{bmatrix} 67 & 5 \\ 6 & 66 \end{bmatrix}$
BDT	0.9097	0.8933	0.9306	0.9116	$\begin{bmatrix} 64 & 8 \\ 5 & 67 \end{bmatrix}$
H-Matrix	0.8958	0.9130	0.8750	0.8936	$\begin{bmatrix} 66 & 6 \\ 9 & 63 \end{bmatrix}$
Fisher	0.9306	0.9559	0.9028	0.9286	$\begin{bmatrix} 69 & 3 \\ 7 & 65 \end{bmatrix}$
Mahalanobis	0.9306	0.9559	0.9028	0.9286	$\begin{bmatrix} 69 & 3 \\ 7 & 65 \end{bmatrix}$
Linear Discriminant	0.9306	0.9559	0.9028	0.9286	$\begin{bmatrix} 69 & 3 \\ 7 & 65 \end{bmatrix}$

Figure 4.13 shows that methods of Fisher, Mahalanobis and linear discriminants acquired the highest in accuracy (0.9306), F1 score (0.9286) and precision (0.9559). In terms of recall, BDT (0.9306) is declared as the highest among the discriminants. Thus, Fisher, Mahalanobis and linear discriminants will be chosen as the suitable classifier for dataset A due to its highest accuracy.

Table 4.14: Standard deviation of the instances for dataset A.

Discriminants	Standard deviation of negative instances	Standard deviation of positive instances
ANN	0.5	0.5
BDT	1.5	1.5
H-Matrix	1.5	1.5
Fisher	2.0	2.0
Mahalanobis	2.0	2.0
Linear Discriminant	2.0	2.0

Based on Table 4.14, ANN is proven to be the most consistent to classify both positive and negative instances while Fisher, Mahalanobis and linear discriminants are proven to the most inconsistent due to the highest standard deviation of the instances compared to other discriminants.

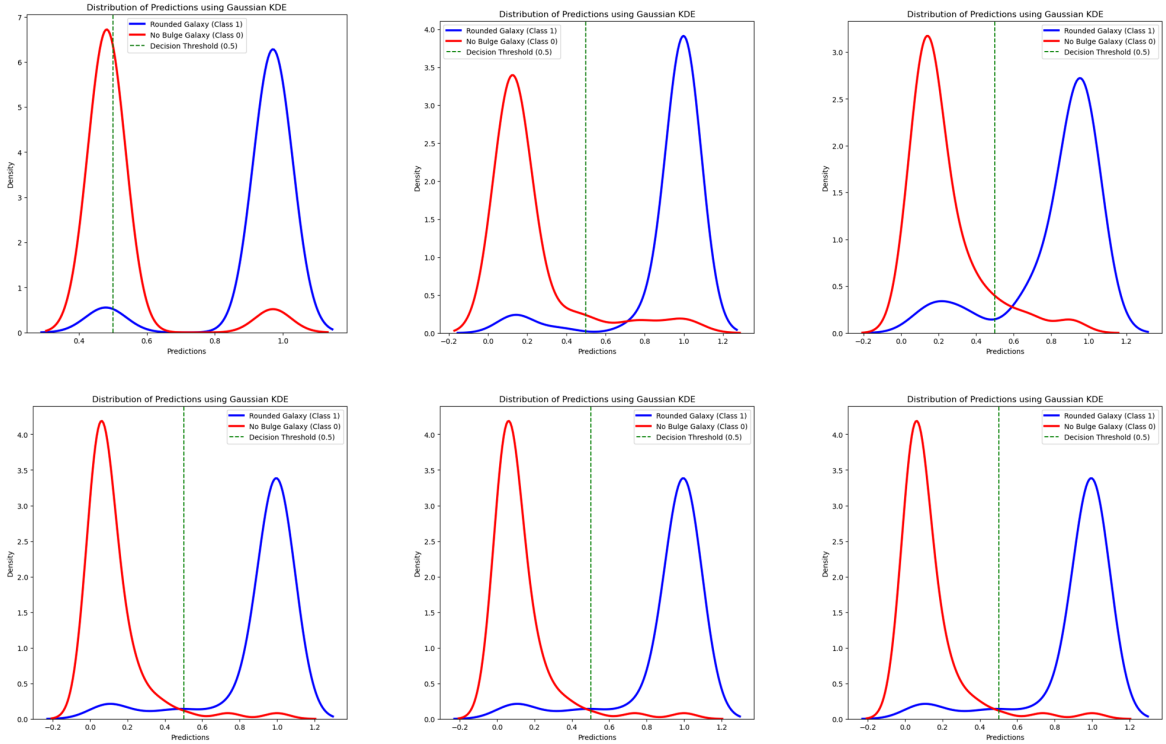

Figure 4.8: Gaussian graph for ANN (top left), BDT (top middle), H-matrix (top right), Fisher (bottom left), Mahalanobis (bottom middle) and linear discriminant (bottom right) for dataset A.

4.2.2 *Dataset B*

Table 4.15: Performance assessment of each discriminant for dataset B.

Discriminants	Accuracy	Precision	Recall	F1 Score	Confusion Matrix
ANN	0.9380	0.9574	0.9615	0.9595	$\begin{bmatrix} 62 & 10 \\ 19 & 215 \end{bmatrix}$
BDT	0.9020	0.9766	0.8932	0.9330	$\begin{bmatrix} 67 & 5 \\ 25 & 209 \end{bmatrix}$
H-Matrix	0.8856	0.9585	0.8889	0.9224	$\begin{bmatrix} 63 & 9 \\ 26 & 208 \end{bmatrix}$
Fisher	0.9052	0.9952	0.8803	0.9342	$\begin{bmatrix} 71 & 1 \\ 28 & 206 \end{bmatrix}$
Mahalanobis	0.9052	0.9952	0.8803	0.9342	$\begin{bmatrix} 71 & 1 \\ 28 & 206 \end{bmatrix}$
Linear discriminant	0.9052	0.9952	0.8803	0.9342	$\begin{bmatrix} 71 & 1 \\ 28 & 206 \end{bmatrix}$

According to Table 4.15, ANN has the highest accuracy (0.9380), recall (0.9615), and F1 score (0.9595), making it the most effective classifier for the dataset B. Fisher, Mahalanobis and linear discriminants have a single false positive, resulting in the greatest precision (0.9952) and subsequent highest F1-Score (0.9342) next to ANN, giving it a superior alternative than *H*-Matrix as a classifier below ANN.

Table 4.16: Standard deviation of the instances for dataset B non-weighted.

Discriminants	Standard deviation of negative instances	Standard deviation of positive instances
ANN	4.5	76.5
BDT	10.0	71.0
H-Matrix	8.5	72.5
Fisher	13.5	67.5
Mahalanobis	13.5	67.5
Linear discriminant	13.5	67.5

Due to the imbalance dataset, we will prioritizing standard deviation of negative instances only, on how well each discriminant are consistent in terms of false positive and false negative. ANN exhibits the lowest standard deviation of negative instances (4.5) while Fisher, Mahalanobis and linear discriminants are the most inconsistent with standard deviation (13.5)

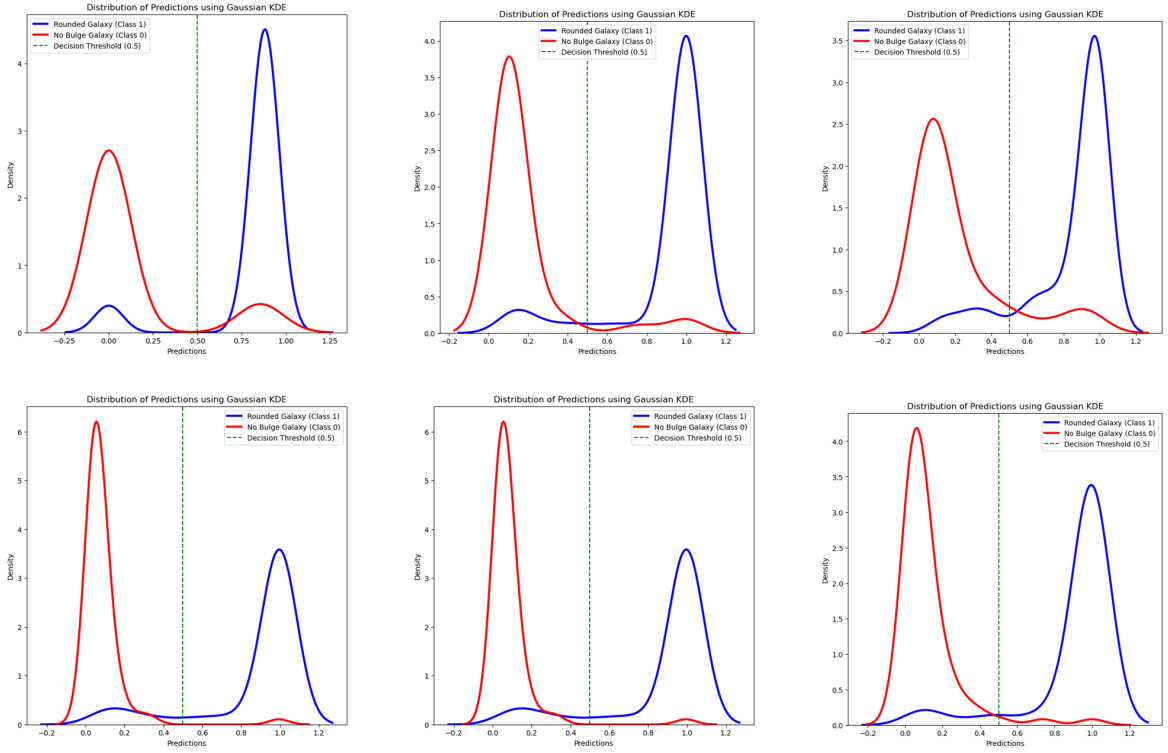

Figure 4.9: Gaussian graph for ANN (top left), BDT (top middle), H-matrix (top right), Fisher (bottom left), Mahalanobis (bottom middle) and linear discriminant (bottom right) for dataset B.

4.2.3 *Dataset C*

Table 4.17: Performance assessment of each discriminant for dataset C.

Discriminants	Accuracy	Precision	Recall	F1 Score	Confusion Matrix
ANN	0.9013	0.7595	0.8451	0.8000	$\begin{bmatrix} 214 & 11 \\ 19 & 60 \end{bmatrix}$
BDT	0.9013	0.7156	0.9577	0.8193	$\begin{bmatrix} 206 & 3 \\ 27 & 68 \end{bmatrix}$
H-Matrix	0.8586	0.6373	0.9155	0.7514	$\begin{bmatrix} 196 & 6 \\ 37 & 65 \end{bmatrix}$
Fisher	0.8783	0.6604	0.9859	0.7910	$\begin{bmatrix} 197 & 1 \\ 36 & 70 \end{bmatrix}$
Mahalanobis	0.8783	0.6604	0.9859	0.7910	$\begin{bmatrix} 197 & 1 \\ 36 & 70 \end{bmatrix}$
Linear discriminant	0.8783	0.6604	0.9859	0.7910	$\begin{bmatrix} 197 & 1 \\ 36 & 70 \end{bmatrix}$

From Table 4.17, ANN and BDT exhibit same accuracy (0.9013), however BDT has higher F1 score (0.8193), declaring it as the best classifier for dataset C. Fisher, Mahalanobis and linear discriminants have the highest recall (0.9859) where they only produced 1 false negative while ANN has the highest precision (0.7595) with the lowest false positive of 19 among the discriminants.

Table 4.18: Standard deviation of the instances for dataset C, weighted.

Discriminants	Standard deviation of negative instances	Standard deviation of positive instances
ANN	4.0	77.0
BDT	12.0	69.0
H-Matrix	15.5	65.5
Fisher	17.5	63.5
Mahalanobis	17.5	63.5
Linear discriminant	17.5	63.5

Table 4.18 shows that Fisher, Mahalanobis and linear discriminants are the most inconsistent with the highest standard deviation (17.5) while ANN is the most consistent discriminant to produce errors by exhibiting lowest standard deviation (4.0).

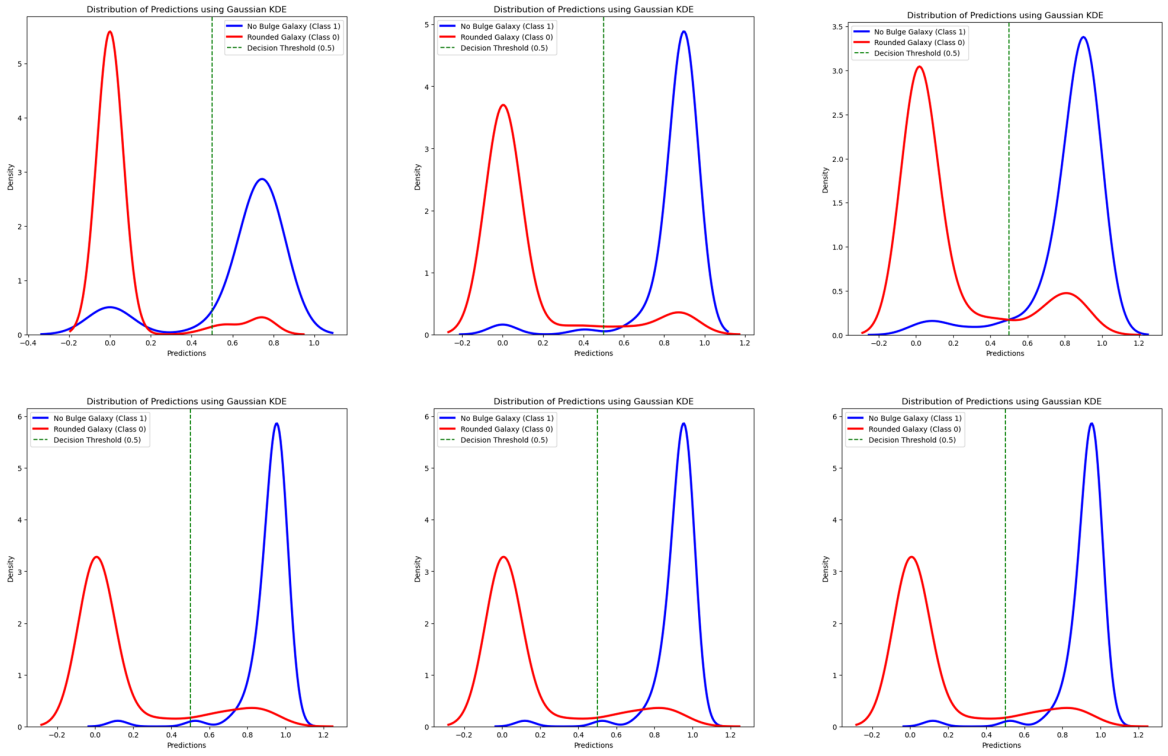

Figure 4.10: Gaussian graph for ANN (top left), BDT (top middle), H-matrix (top right), Fisher (bottom left), Mahalanobis (bottom middle) and linear discriminant (bottom right) for dataset C.

4.2.4 *Part B Discussion*

Based on the findings presented in Part B, we can infer that Fisher, Mahalanobis and linear discriminants had the greatest accuracy (0.9306), precision (0.9559), and F1 score (0.9286) for the dataset A. This shows that not only are both of these discriminants capable of consistently and reliably distinguishing rounded galaxies and no-bulge galaxies, but they also effectively categorize the majority of no-bulge galaxies as their actual value, signifying an extremely low rate of false positives. Another major finding from this investigation is that BDT has the highest recall (0.9306) and is effectively able to avoid incorrectly classifying rounded galaxies as no bulge galaxies.

In terms of the entire dataset, we discovered that when weighting was implemented with the sole goal of balancing the dataset's training process, the overall metrics performance degraded. However, we discovered that in the entire dataset with weighted, Fisher, Mahalanobis and linear discriminants produced the greatest recall (0.9859) among the three types of datasets. This implies that, despite their degraded precision, these discriminants can still accurately classify actual positive instances (no bulge galaxies), demonstrating a bias toward the classification of no bulge galaxies only, even if it means misclassifying rounded galaxies as no bulge galaxies. In terms of all dataset non-weighted, ANN has the highest recall (0.9595) and precision (0.9615). This means that, even with an imbalanced dataset, ANN can classify both rounded and no bulge galaxies accurately.

In conclusion, we determined that for imbalance dataset (dataset B), ANN is the most efficient classifier for rounded galaxies and no bulge galaxies, with the greatest output F1 score (0.9595) while Fisher, Mahalanobis and linear discriminants is the second best classifier for equal dataset from dataset A with second highest accuracy overall (0.9306).

5

CONCLUSION

5.1 CONCLUSION

In this thesis, we evaluated the effectiveness of several discriminants; *ANN*, *BDT*, *Fisher*, *Mahalanobis H-Matrix*, and *linear discriminants* for star-galaxy classification and the rounded-no bulge galaxy classification. The study we conducted compared these discriminants across several variable transformations, namely none, N, U, G, D, and PCA.

The results we obtained from part A showed that *ANN* consistently outperformed all the evaluation parameters, including accuracy, precision, F1 score, and recall, across the discriminants. *BDT* followed closely, exhibiting its effectiveness as a reliable classifier. *Fisher*, *H-Matrix*, *Mahalanobis* and *linear discriminants*, while not as productive as *ANN* and *BDT*, showed significant promises and nearly achieving the level of the top classifiers.

Our focus was to compare the performance of four discriminants; *Fisher*, *Mahalanobis*, *H-Matrix*, and *linear dscriminants* in classifying star-galaxy datasets. In particular, we determined the variable transform U as the most optimal configuration, since it produced the maximum accuracy (0.991) particularly when using *Fisher*, *Mahalanobis* and *linear discriminants*. Additionally, these three discriminants also demonstrated their ability to effectively categorize stars without misclassifying them as galaxies, as proven by the highest output of precision (1). However, their recall were inconsistent indicating its value in situations when avoiding false positives is crucial, even if it meant compromising their recall.

Part B of the study concludes that by utilizing variable transform U from part A, *Fisher*, *Mahalanobis* and *linear discriminants* performed highest in terms of accuracy (0.9306), precision (0.9559), and F1 score (0.9286) for dataset A. This demonstrates their consistent ability in distinguishing rounded galaxies from no bulge galaxies.

In conclusion, our study shows that, while *ANN* is the most efficient classifier for both parts of classifications, other discriminants such as *Fisher*, *Mahalanobis*, *H-Matrix*, and *linear discriminants* are also able to perform effectively in certain settings. These findings are useful for identifying appropriate discriminants depending on specific classification aims and dataset features in astronomical analysis of data.

5.2 FUTURE RESEARCH AND OUTLOOK

- While this study explored multiple variable transformations, future research could investigate at different methods or combination of variable transformations that would enhance classification performance even further.
- Investigating which input variables contribute significantly to classification fulfillment can reveal insights about the important features that distinguish stars from galaxies and various galaxy types.
- Exploring various configuration options for discriminants, such as ANN and BDT, may enhanced classifiers' efficiency as well as flexibility in astronomy classification.
- Exploring into weighting methods to balance imbalanced datasets may enhance classification performance, as our experiment discovered that using weighting methods impacted results badly. Future research should focus on designing and evaluating weighting techniques to improve model balance and accuracy.
- Classifying a wider range of galaxy types from the Galaxy Zoo Decision Tree 2, in besides rounded and no bulge galaxies, might give additional information and increase the effectiveness of classification models. Future research should try extending the classification to include more galaxy morphologies to improve the models' generalizability in astronomical classification.

APPENDIX 1

A.1 APPENDIX: CODE LISTING

Listing 1: Python code for evaluating a model: Star/Galaxy Classification.

```
1
import pandas as pd
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
    confusion_matrix
#This is where we put our csv file. As a sample I will import from the directory
    where I keep my output files 'bnnz/output/test_singleCls_quick/clasif/eval/ANNZ_
    singleCls_0000.csv'
#In this case i'm using jupyter notebook, so i will import the csv file into the same
     file where im doing this notebook
6 # Load your CSV file
file_path = 'H2_N.csv'
data = pd.read_csv(file_path)
predicted = data['F :ANNZ_0']
#(I:type)<= i added this in the eval output var so i can see it in the csv file
11 # Convert true labels where the (galaxy==3) to 1 and (star ==6) to 0
data['type'] = data['type'].replace({3: 1, 6: 0})
true_labels = data['type']
threshold = 0.5
binary_predictions = (predicted >= threshold).astype(int)
16 accuracy = accuracy_score(true_labels, binary_predictions)
precision = precision_score(true_labels, binary_predictions)
recall = recall_score(true_labels, binary_predictions)
```

```
f1_score = f1_score(true_labels, binary_predictions)
confusion_matrix=confusion_matrix(true_labels,binary_predictions)
print("Model Metrics:")
print(f"Accuracy: {accuracy}")
print(f"Precision: {precision}")
print(f"Recall: {recall}")
print(f"F1 Score: {f1_score}")
print(f"Confusion_Matrix:\n{confusion_matrix}")
```

Listing 2: Python code for evaluating a model: Rounded/No_Bulge Classification.

```
import pandas as pd
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
    confusion_matrix
#This is where we put our csv file from the directory where we keep our output files
# Load your CSV file
file_path = 'annh2_eq_nowgt.csv'
data = pd.read_csv(file_path)
predicted = data['F :ANNZ_0']
#1:rounded and 0:no bulge
data['type'] = data['type'].replace({1: 1, 3: 0})
true_labels = data['type']
threshold = 0.5
binary_predictions = (predicted >= threshold).astype(int)
accuracy = accuracy_score(true_labels, binary_predictions)
precision = precision_score(true_labels, binary_predictions)
recall = recall_score(true_labels, binary_predictions)
f1_score = f1_score(true_labels, binary_predictions)
confusion_matrix=confusion_matrix(true_labels,binary_predictions)
print("Model Metrics:")
print(f"Accuracy: {accuracy}")
print(f"Precision: {precision}")
print(f"Recall: {recall}")
print(f"F1 Score: {f1_score}")
print(f"Confusion_Matrix:\n{confusion_matrix}")
```

Listing 3: Python code for visualization of Gaussian Kernel Density Estimation (KDE).

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
plt.figure(figsize=(8,8))
sns.kdeplot(data[true_labels == 1]['F :ANNZ_0'], color='blue', label='Rounded Galaxy
    (Class 1)', lw=3)##lw is line width
sns.kdeplot(data[true_labels == 0]['F :ANNZ_0'], color='red', label='No Bulge Galaxy
    (Class 0)', lw=3)
plt.axvline(x=0.5, color='green', linestyle='--', label='Decision Threshold (0.5)')
plt.title(f'Distribution of Predictions using Gaussian KDE')
plt.xlabel('Predictions')
plt.ylabel('Density')
plt.legend()
plt.show()
```

Listing 4: Python code for statistical calculation.

```
#Statistics Calculation
from sklearn.metrics import confusion_matrix, mean_absolute_error
import numpy as np
tn, fp, fn, tp = confusion_matrix(true_labels,binary_predictions).ravel() #ravel() is
     something like .items() but this one is for dictionary : key,value
# ravel() makes it into 1d array and we can unpack
errors = [fp,fn]
positive=[tn,tp]
# Standard deviation
std_dev = np.std(errors)
std_dev_p=np.std(positive)
# Mean absolute error
print (f"tn:{tn}")
print(f"fp:{fp}")
print(f"fn:{fn}")
print(f"tp:{tp}")
print(f"Standard Deviation of Errors: {std_dev}")
```

```
print(f"Standard Deviation of Positives: {std_dev_p}")
```

B

REFERENCES

1. Andreas, H. (2010). 2 Spectroscopic techniques: I Spectrophotometric techniques. https://research-repository.griffith.edu.au/bitstream/handle/10072/34561/62679_1.pdf

2. Ashdown, I. (1998). Photometry and Radiometry. ResearchGate. https://www.researchgate.net/publication/2711215_Photometry_and_Radiometry

3. Aydoğdu, Ç. (2021). Decorrelating Gaussian Vectors. Medium. https://medium.com/@cagriaydogdu2334/decorrelating-gaussian-vectors-e7982c9a28e6

4. Harikrishnan, B. (2020). Confusion Matrix, Accuracy, Precision, Recall, F1 Score. Medium. https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

5. Bahcall, N. A., Lubin, L. M., & Dorman, V. (1995). Where is the Dark Matter? ArXiv.org. https://arxiv.org/abs/astro-ph/9506041

6. Bessell, M. S. (2005). Standard Photometric Systems. Annual Review of Astronomy and Astrophysics, 43(1), 293–336.https://doi.org/10.1146/annurev.astro.41.082801.100251

7. Bhandari, A. (2020). Confusion Matrix for Machine Learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/#:~:text=A%20Confusion%20matrix%20is%20an

8. Britannica, T. Editors of Encyclopaedia. (2017). photometry. Brittanica. https://www.britannica.com/science/photometry-astronomy

9. Britannica, The Editors of Encyclopaedia. (2014). radiometer. Brittanica. https://www.britannica.com/technology/radiometer

10. Bunn, E. F., & Hogg, D. W. (2009). The kinematic origin of the cosmological redshift. American Journal of Physics, 77(8), 688–694. https://doi.org/10.1119/1.3129103

11. Buta, R. (2011). Galaxy Morphology. http://carina.fcaglp.unlp.edu.ar/extragalactica/Bibliografia/Galaxy_Morphology.pdf

12. Celso Luís Levada. "Edwin Hubble and The Expansion of The Universe" American Journal of Engineering Research (AJER), vol.8, no.03, 2019, pp.288-292

13. Chaudhary, B. (2023). Importance of Data (A Term Paper). ResearchGate. https://www.researchgate.net/publication/374544786_Importance_of_Data_A_Term_Paper

14. Chen, J. (2021). Uniform Distribution Definition. Investopedia. https://www.investopedia.com/terms/u/uniform-distribution.asp

15. Choi, C. Q. (2017). Star Facts: The Basics of Star Names and Stellar Evolution. Space.com; Space.com. https://www.space.com/57-stars-formation-classification-and-constellations.html

16. Cravit, R. (2019). What Is a Decision Tree and How to Make One [Templates + Examples] - Venngage. Venngage. https://venngage.com/blog/what-is-a-decision-tree/

17. Cui, Q. (2022). A novel model for the interpretation of cosmological redshift. ResearchGate. https://www.researchgate.net/publication/359561006_A_novel_model_for_the_interpretation_of_cosmological_redshift

18. Dadi, M. (2022). Spectroscopy and Spectrophotometry: Principles and Applications for Colorimetric and Related Other Analysis. ResearchGate. https://www.researchgate.net/publication/362202612_Spectroscopy_and_Spectrophotometry_Principles_and_Applications_for_Colorimetric_and_Related_Other_Analysis

19. Datla, R. (2005). 1. Introduction to Optical Radiometry. ResearchGate. https://www.researchgate.net/publication/260691090_1_Introduction_to_Optical_Radiometry

20. de Haro, J., & Elizalde, E. (2022). Topics in Cosmology—Clearly Explained by Means of Simple Examples. Universe, 8(3), 166. https://doi.org/10.3390/universe8030166

21. Drapala, J. (2023). Kernel Density Estimation explained step by step. Medium. https://towardsdatascience.com/kernel-density-estimation-explained-step-by-step-7cc5b5bc4517

22. eoPortal. (2020). SDSS (Sloan Digital Sky Survey) - eoPortal. Www.eoportal.org. https://www.eoportal.org/other-space-activities/sdss#events-and-science-results

23. Frost , J. (n.d.). Cumulative Distribution Function (CDF): Uses, Graphs& vs PDF. Statistics by Jim. https://statisticsbyjim.com/probability/cumulative-distribution-function-cdf/

24. Gotame, R. C. (2020). Classification of Galaxies. Physics Feed. https://physicsfeed.com/post/classification-galaxies/

25. Gray, R. (2017). Galaxy Zoo: Citizen science trailblazer marks tenth birthday. Www.bbc.com. https://www.bbc.com/news/science-environment-40558759

26. Gray, R., & Dunning-Davies, J. (2008). A review of redshift and its interpretation in cosmology and astrophysics. https://arxiv.org/vc/arxiv/papers/0806/0806.4085v1.pdf

27. Guruprasad, A. (2023). Data Mining Project Report Galaxy Classification: A machine learning approach for classifying shapes using numerical data. https://arxiv.org/ftp/arxiv/papers/2312/2312.00184.pdf

28. HandWiki. (2022). Distance Measures (Cosmology). Encyclopedia.pub. https://encyclopedia.pub/entry/30206#:~:text=The%20comoving%20distance%20should%20be

29. Harrison, E. (2003). Masks of the Universe: Changing Ideas on the Nature of the Cosmos. In Google Books. Cambridge University Press. https://books.google.com.my/books?id=tSowGCP0kMIC&q=ylem+supersymmetrie&pg=PA224&redir_esc=y#v=snippet&q=ylem%20supersymmetrie&f=false

30. Harvey, D. (n.d.). Spectroscopic Methods. https://resources.saylor.org/wwwresources/archived/site/wp-content/uploads/2012/07/Chapter1011.pdf

31. Hodge, P. W. (2024). galaxy. Brittanica. https://www.britannica.com/science/galaxy

32. Hogg, D. (2000).Distance measures in cosmology. https://arxiv.org/pdf/astro-ph/9905116.pdf

33. Jaadi, Z. (2019). A Step by Step Explanation of Principal Component Analysis. Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

34. Johnson, B. C. (2014). Principles of Optical Radiometry and Measurement Uncertainty. ResearchGate. https://www.researchgate.net/publication/285988310_Principles_of_Optical_Radiometry_and_Measurement_Uncertainty

35. Koo. Ping Shung. (2018). Accuracy, Precision, Recall or F1? Towards Data Science; Towards Data Science. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

36. Liew, Anthony. (2007). Understanding Data, Information, Knowledge And Their Inter-Relationships. Journal of Knowledge Management Practice. Vol. 7.

37. Li, N., Thakar, A. (2008). CasJobs and MyDB - A batch query workbench. http://www.sdss.jhu.edu/~thakar/pubs/cise08/casjobs.pdf

38. Liew, A. (2007). Understanding Data, Information, Knowledge And Their Inter-Relationships. ResearchGate. https://www.researchgate.net/publication/224937037_Understanding_Data_Information_Knowledge_And_Their_Inter-Relationships

39. likebupt. (2022). Boosted Decision Tree Regression: Component Reference - Azure Machine Learning. Learn.microsoft.com. https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/boosted-decision-tree-regression?view=azureml-api-2

40. Lindholm, A., Wahlström, N., Lindsten, F., & Schön, T. (2019). Supervised Machine Learning Lecture notes for the Statistical Machine Learning course. https://mwns.co/blog/wp-content/uploads/2020/01/Supervised-Machine-Learning.pdf

41. Liu, C. (2022). Data Transformation: Standardization vs Normalization. KDnuggets. https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html

42. Liu, Q. (2012). Supervised Learning. ResearchGate. https://www.researchgate.net/publication/229031588_Supervised_Learning

43. Luís Levada, C. (2019). Edwin Hubble and the Expansion of the Universe. American Journal of Engineering Research (AJER). https://www.ajer.org/papers/Vol-8-issue-3/ZZI0803288292.pdf

44. Mahesh, B. (2019). Machine Learning Algorithms -A Review. ResearchGate. https://www.researchgate.net/publication/344717762_Machine_Learning_Algorithms_-A_Review

45. Marshall, O., Tojeiro, R., Weijmans, A.-M. (2022). Demonstrating Cosmological and Doppler Redshift in the Classroom. https://upcommons.upc.edu/bitstream/handle/2117/370780/SSEA_2022_233.pdf?sequence=1&isAllowed=y

46. matplotlib. (n.d.). matplotlib.pyplot.scatter — Matplotlib 3.5.1 documentation. Matplotlib.org. https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.scatter.html

47. Ohno, Y. (1999). OSA Handbook of Optics, Volume III Visual Optics and Vision Chapter for Photometry and Radiometry. https://www.physics.muni.cz/~jancely/PPL/Texty/IntegracniKoule/Photometry%20and%20Radiometry.pdf

48. OP-TEC. (n.d.). Basics of Spectroscopy Photonics-Enabled Technologies OPTICS AND PHOTONICS SERIES STEP (Scientific and Technological Education in Photonics), an NSF ATE Project.

49. Pearson, E. (2021). What are stars? Www.skyatnightmagazine.com. https://www.skyatnightmagazine.com/space-science/beginners-guide-stars

50. Pickell, D. (2023). Structured vs. Unstructured Data: What the Difference? https://www.g2.com/articles/structured-vs-unstructured-data

51. Princeton University. (n.d.). The SDSS Data System and Data Products. Www.astro.princeton.edu. https://www.astro.princeton.edu/PBOOK/datasys/datasys.htm

52. Quincey, P. (2020). Solid angles in perspective. Physics Education, 55(5), 055003. https://doi.org/10.1088/1361-6552/ab9323

53. Romanishin, W. (2002). An Introduction to Astronomical Photometry Using CCDs. https://www1.phys.vt.edu/~jhs/phys3154/CCDPhotometryBook.pdf

54. Ryden, B. (2017). Introduction to Cosmology. In Google Books. Cambridge University Press. https://books.google.com.my/books?hl=en&lr=&id=07WSDQAAQBAJ&oi=fnd&pg=PA1&dq=explaining+cosmology&ots=wDS3960oXb&sig=RsVTyIBQdn0nroc59o7GIvMBIdM#v=onepage&q=explaining%20cosmology&f=false

55. Sadeh, I. (2014). ANNz2 - Photometric redshift and probability density function estimation using machine-learning. Proceedings of the International Astronomical Union, 10(S306), 316–318. https://doi.org/10.1017/s1743921314010849

56. Sandage, A. (1975). Classification & Stellar Content of Galaxies -A. Sandage CLASSIFICATION AND STELLAR CONTENT OF GALAXIES OBTAINED FROM DIRECT PHOTOGRAPHY. https://ned.ipac.caltech.edu/level5/Sandage/paper.pdf

57. Sandage, A., Wilson, M., Observatories, P.,& Carnegie. (1961). THE HUBBLE ATLAS OF GALAXIES 1961. https://www.cs.utexas.edu/users/mitra/csSummer2019/hsra/lectures/Hubble_Atlas.pdf

58. SciServer. (n.d.). Schema Browser - SkyserverSDSS. Skyserver.sdss.org. Retrieved June 9, 2024, from https://skyserver.sdss.org/dr18/MoreTools/browser?&&history=description+zoo2MainPhotoz+U

59. SDSS. (n.d.). Glossary of SDSS-IV Terminology | SDSS. Live-Sdss4org-Dr13.Pantheonsite.io. Retrieved June 9, 2024, from https://live-sdss4org-dr13.pantheonsite.io/help/glossary/#

60. SDSS Query CasJobs. (n.d.). CasJobs Guide. Skyserver.sdss.org. https://skyserver.sdss.org/casjobs/Guide.aspx

61. Seaborn. (2012). seaborn.kdeplot — seaborn 0.9.0 documentation. Pydata.org. https://seaborn.pydata.org/generated/seaborn.kdeplot.html

62. Sheldon, R. (2021). What is a charge-coupled device (CCD)? SearchStorage. https://www.techtarget.com/searchstorage/definition/charge-coupled-device

63. Siadati, S. (2018). What is unsupervised Learning. ResearchGate. https://www.researchgate.net/publication/342121950_What_is_unsupervised_Learning

64. Singh, G. (2021). Introduction to Artificial Neural Networks. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/introduction-to-artificial-neural-networks/

65. Smola, A. S., & Vishwanathan, S. V. N. (2010). INTRODUCTION TO MACHINE LEARNING. https://alex.smola.org/drafts/thebook.pdf

66. Soares, D. (2021). Henrietta Leavitt, the woman who discovered a cosmic ruler. https://lilith.fisica.ufmg.br/~dsoares/reino/hleavitt-e.pdf

67. Speckmayer, P., Höcker, A., Stelzer, J.,& Voss, H. (2010). The toolkit for multivariate data analysis, TMVA 4. Journal of Physics: Conference Series, 219(3), 032057. https://doi.org/10.1088/1742-6596/219/3/032057

68. Stoll , M.-P. (2000). Principles Of Radiometry. SpringerLink. https://link.springer.com/article/10.1023/A:1006758614317

69. Stoll, MP. (2000). Principles Of Radiometry. Surveys in Geophysics 21, 133–146. https://doi.org/10.1023/A:1006758614317

70. University of Western Australia. (2014). Evidence for the Big Bang. In Evidence for the Big Bang (pp. 1–4). https://www.uwa.edu.au/study/-/media/Faculties/Science/Docs/Evidence-for-the-Big-Bang.pdf

71. Vavrycuk, V. (2021). Expansion of the Universe and cosmological redshift. ResearchGate. https://www.researchgate.net/publication/355861318_Expansion_of_the_Universe_and_cosmological_redshift

72. Voss, H. (2009). TMVA: Toolkit for Multivariate Data Analysis with ROOT. ResearchGate. https://www.researchgate.net/publication/323023374_TMVA_Toolkit_for_Multivariate_Data_Analysis_with_ROOT

73. Willett, Kyle W. (2013). Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. Astrophysics Data System. https://ui.adsabs.harvard.edu/abs/2013MNRAS.435.2835W/abstract

74. Willett, K. (2015). Visualizing the decision trees for Galaxy Zoo. Galaxy Zoo. https://blog.galaxyzoo.org/2015/04/06/visualizing-the-decision-trees-for-galaxy-zoo/

75. Wilpon, J., Thomson, D., Bangalore, S., Haffner, P.,& Johnston, M. (2019). THE FUNDAMENTALS OF MACHINE LEARNING. https://www.interactions.com/wp-content/uploads/2017/06/machine_learning_wp-5.pdf

76. Wu, W. (2022). Unsupervised Learning. https://na.uni-tuebingen.de/ex/ml_seminar_ss2022/Unsupervised_Learning%20Final.pdf

77. Yasar, K. (2023). What is an Artificial Neural Network (ANN)? SearchEnterpriseAI. https://www.techtarget.com/searchenterpriseai/definition/neural-network

78. JWST User Documentation. (2018). JWST Multi-Object Spectroscopy - JWST User Documentation. Jwst-Docs.stsci.edu. https://jwst-docs.stsci.edu/methods-and-roadmaps/jwst-multi-object-spectroscopy#gsc.tab=0

79. Hogg, David W. (2002). The K Correction. https://arxiv.org/abs/astro-ph/0210394

80. Soo, J. Y. H. (2018). Enhancing photometric redshifts for the era of precision cosmology. In Doctoral thesis, UCL (University College London). https://discovery.ucl.ac.uk/id/eprint/10055277/1/My_Thesis_10p.pdf

81. Quincey, P. (2020). Solid angles in perspective. Physics Education, 55(5), 055003. https://arxiv.org/ftp/arxiv/papers/2108/2108.05226.pdf

82. The Sloan Digital Sky Survey: Technical Summary. (n.d.). https://classic.sdss.org/dr4/instruments/technicalPaper/index.php

83. Grøn, Ø. (2018). The Discovery of the Expansion of the Universe. Galaxies, 6(4), 132. https://doi.org/10.3390/galaxies6040132

84. Wang, X. (2018). New Discovery on Planck Units and Physical Dimension in Cosmic Continuum Theory. Journal of Modern Physics. https://doi.org/10.4236/jmp.2018.914153

85. PRMJ. (2024). Planck Units. Medium. https://medium.com/@prmj2187/plank-units-0cca79974424

86. Ferreira, P. (n.d.). Cosmological Surveys | University of Oxford Department of Physics. https://www.physics.ox.ac.uk/research/group/beecroft-institute-particle-astrophysics-and-cosmology/research/cosmological-surveys

87. Wikipedia. (2022). Galaxy Zoo. Wikipedia. https://en.wikipedia.org/wiki/Galaxy_Zoo

88. Galaxy Zoo. (2018). Galaxy Zoo Decision Trees. https://data.galaxyzoo.org/gz_trees/gz_trees.html

89. Jalswal, S. (2024). Multilayer Perceptrons in Machine Learning: A Comprehensive Guide. https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning