ACKNOWLEDGEMENT

A project is always a result of the amalgamation of various ideas and support from countless people. We would like to acknowledge the contribution made by them.

I express my sincere thanks to the Principal **Dr. Koel Roychoudhury**, Also express my thanks to Ms. Anu Thomas, Professor and Head Department of Information Technology for constant support.

I would like to express my gratitude to my guide, **Ms. Arti Bansode**, for her encouragement and guidance. She not only helped me suggest the project topic but also assisted me in conducting extensive research. Through her support, I have gained knowledge about numerous new things.

Last but not the least I would like to thank my Friends who have helped me a lot.

Zainab Shaikh M.Sc. IT (Part II)
Vignesh Nadar M.Sc. IT (Part II)
SIES (Nerul) College of Arts, Science &
Commerce, Mumbai University

Index

Abstract:

In response to the expanding data volume and complexity in astronomy, this documentation will help in understanding how we used machine learning (ML) and deep learning (DL) approaches for classifying stars, galaxies, and quasars. We evaluated a range of ML algorithms, including logistic regression, decision trees, random forest, SVM, HDBSCAN, XGBoost, and KNN, alongside DL methods like CNNs, ANNs, and transfer learning from Resnet50, Xception, and EfficientNetB2, the study utilizes datasets from prominent surveys such as SDSS, LAMOST, and WISE. This documentation meticulously curates and selects datasets, underscoring their pivotal role in assessing classification model robustness. The choice of the Random Forest algorithm is particularly notable, given its demonstrated effectiveness in handling imbalanced datasets, a common challenge in astronomy, and its interpretability, offering insights into feature importance for a deeper understanding of the underlying data. Systematic analysis of results, encompassing accuracy, precision, and efficiency metrics, informs a comprehensive synthesis of findings, detailing the strengths and weaknesses of the selected approach. With discussions on various classification algorithms and datasets, the documentation provides a valuable resource for researchers and practitioners engaged in celestial object classification, navigating the evolving landscape of data-driven astronomy.

Keywords: Machine Learning, Random Forest Classifier, SDSS, SkySurveys, Celestial Objects.

1. Introduction:

The cosmos, an expansive tapestry of wonder and intricacy, has been a perennial source of fascination and exploration for astronomers throughout the annals of human history. As our understanding of the universe has deepened, so too has the imperative for systematic classification methods that can distill meaning from the vast array of celestial objects that populate the mesmerizing night sky. Consider the stars, those incandescent spheres of hydrogen and helium, whose formation unfolds over millions of years, serving as cosmic chronicles that unravel the intricate processes shaping the cosmos. In parallel, galaxies, sprawling realms of stars, planets, and elusive dark matter, beckon astronomers to decipher their compositions and unravel the cosmic drama written in their celestial structures. Meanwhile, quasars, born from the colossal gravitational dance around supermassive black holes, emit luminous signals that carry the tales of the universe's early epochs, offering a unique glimpse into the unfolding drama of cosmic evolution.

The historical reliance on visual observations, where early astronomers discerned stars based on brightness and color, has given way to the transformative era of spectroscopy. This sophisticated technique enables the dissection of light emitted or absorbed by celestial objects, heralding a new epoch in celestial classification. As telescopes have grown in size and capability, the need for nuanced classification schemes has become more pronounced. In this context, wide-angle sky surveys, such as the Sloan Digital Sky Survey and Gaia, have emerged as pivotal endeavors, aiming to map the large-scale structure of the universe and derive cosmological insights.

Amidst this celestial odyssey, the quest for accurate classification has taken on new significance. In the intricate dance of stars, galaxies, and quasars, accurate classification forms the bedrock for comprehending fundamental cosmic processes, from the evolution of individual stars to the dynamics of entire galactic structures. The emergence of machine learning, data science, and IT technologies adds a new dimension to this pursuit. As we stand on the cusp of a new era in astronomy, the convergence of these disciplines holds promise for automating and refining celestial classification, unlocking new realms of understanding in the ever-expanding cosmos. Thus, the exploration of the night sky, now intertwined with the tools of modern technology, continues to captivate our collective imagination, beckoning us to unravel the cosmic mysteries that have intrigued humanity for centuries.

Stars, those brilliant beacons of cosmic illumination, represent the fundamental building blocks of the universe, each a burning hot sphere of elemental magnificence primarily composed of hydrogen and helium. Their genesis is a cosmic symphony that unfolds over vast stretches of time, a narrative scripted in the turbulence of colossal gas and dust clouds. These celestial cradles, agitated by cosmic forces, intricately gather and coalesce, a mesmerizing dance choreographed by the unseen hands of gravity.

The genesis of a star, a process extending across millions of years, commences within these turbulent clouds. Gravity, the cosmic orchestrator, assumes the role of conductor, compelling particles of gas and dust to gather, initiating a gravitational ballet that steadily evolves into a cosmic masterpiece. As these cosmic constituents amass in increasingly dense clusters, the gravitational forces intensify, reaching a crescendo that culminates in a magnificent collapse under the weight of their mass and gravity.

It is in this moment of celestial birth that a star emerges, a radiant entity borne from the crucible of gravitational collapse. The intense pressures and temperatures within the core ignite a nuclear fusion furnace, unleashing an effulgence that traverses the cosmic expanse. This luminous manifestation, a testament to the transformative power of celestial alchemy, radiates brilliance across the cosmic tapestry.

Analyzing stars transcends mere astronomical curiosity; it is an endeavor integral to unraveling the intricate processes and inner workings of celestial objects. Stars, in their various stages of evolution, serve as cosmic laboratories, where the alchemical processes of fusion and fission sculpt the destiny of the cosmos. Each stellar lifecycle, a narrative etched in the pulsating glow of hydrogen fusion, provides invaluable insights into the grand design of the universe.

Furthermore, the study of stars holds a profound significance in our quest to comprehend the solar system's central luminary—the Sun. As a star of medium size and luminosity, the Sun's secrets are entwined with those of its cosmic brethren. By peering into the depths of stellar atmospheres and deciphering the spectral signatures of distant suns, astronomers unravel the intricacies of solar dynamics, shedding light on the enigmatic processes that govern our home star.

In essence, the exploration of stars transcends the boundaries of astronomical inquiry; it is a cosmic pilgrimage that unravels the mysteries of creation and sheds light on the celestial processes shaping the universe. Each star, a radiant testament to the cosmic ballet of gravity, gas, and dust, beckons astronomers to decipher the universal language written in their luminous glow, offering a glimpse into the celestial symphony that resonates throughout the cosmos.

Galaxies, those cosmic tapestries of immense proportions, beckon astronomers to embark on a journey through the vast pools of stars, planets, galactic dust, gas, and the enigmatic shroud of dark matter that collectively shape the celestial panorama. These colossal agglomerations of celestial wonders transcend the individuality of stars, weaving a narrative of cosmic evolution and unveiling the profound mysteries of the universe.

At the heart of galactic exploration lies the quest to unravel the intricacies of their composition—a symphony of cosmic elements orchestrating the celestial dance. Stars, like luminescent jewels, adorn the sprawling expanse of galaxies, their brilliance contributing to the cosmic radiance that captivates the imagination. Interspersed among them, planets chart their celestial orbits, and galactic dust particles, like ethereal brushstrokes, paint the cosmic canvas with hues of celestial beauty.

Yet, within the cosmic realms of galaxies, an unseen force holds sway—the elusive dark matter, a mysterious entity that defies direct observation yet exerts gravitational influence, shaping the very fabric of galactic structures. This enigmatic substance, though unseen, is an integral player in the cosmic drama, influencing the motion of stars and sculpting the grand architecture of galaxies.

The study of galaxies transcends the boundaries of mere astronomical inquiry; it is a cosmic pilgrimage that spans the eons of time, unveiling the story of cosmic evolution etched in the celestial archives. Understanding how galaxies have formed and developed over the epochs offers a cosmic lens through which astronomers glean insights into the functions, processes, and timeline of the universe on a monumental scale.

The cosmic chronicles imprinted in the annals of galaxies provide astronomers with a unique opportunity to fathom the vast expanse of the universe on both grand and intricate scales. From the formation of colossal galactic structures to the birth and demise of individual stars within their folds, galaxies serve as cosmic textbooks, elucidating the laws of celestial dynamics and the interconnectedness of cosmic elements.

Moreover, the study of galaxies offers a unique vantage point to better comprehend our celestial abode—the Milky Way. This spiral arm of cosmic splendor, where our solar system resides, becomes more comprehensible as astronomers decode the galactic symphony echoing across the cosmos. By unraveling the cosmic ballet within our own galactic neighborhood, astronomers gain crucial insights into the intricate tapestry of stellar life cycles, galactic interactions, and the broader cosmic narrative.

In essence, the exploration of galaxies is a cosmic odyssey that transcends the boundaries of spatial and temporal scales. It is a profound journey into the heart of the universe, where the study of celestial ensembles unveils the secrets of cosmic evolution, offering a panoramic vista into the celestial mechanics that govern the vast expanse of the cosmos. As astronomers navigate the cosmic seas of galaxies, they chart a course towards a deeper understanding of the universe's grand design—a celestial voyage that continues to illuminate the cosmic wonders that have inspired awe and curiosity throughout the ages.

In the vast cosmic expanse, quasars emerge as celestial marvels, challenging our understanding of the universe and inviting astronomers into the heart of galactic mysteries. These quasi-stellar objects, aptly named quasars, are cosmic luminaries whose brilliance is intimately tied to the gravitational symphony orchestrated by supermassive black holes residing in the galactic centers. In the cosmic theater, the stage is set by a disk-like cloud of dust and gas known as the accretion disk, swirling in a mesmerizing cosmic dance around the colossal gravitational well of the supermassive black hole.

The interplay of forces within this gravitational ballet becomes a cosmic drama, with the accretion disk spiraling ever closer to the voracious appetite of the black hole. The immense gravitational pressure at the galactic nucleus, coupled with the relentless gravitational pull, transforms this swirling cosmic cloud into a cauldron of extreme temperatures and pressures. The friction generated in this celestial crucible produces extraordinary amounts of heat and light, catapulting quasars into the cosmic limelight as some of the brightest objects in the universe. The luminous intensity of quasars pierces through the cosmic darkness, illuminating the vast cosmic landscapes.

What adds an additional layer of fascination to quasars is their temporal origins. These celestial beacons began their cosmic journey at a time when the universe was in its infancy. Their formation is intricately woven into the fabric of the early universe, making them cosmic time capsules that carry within them the echoes of an ancient era. The light and electromagnetic radiations emitted by quasars become cosmic messengers, transporting astronomers across billions of years and offering a unique window into the universe's evolutionary history.

As astronomers gaze upon the brilliance of quasars, they embark on a transformative journey through time and space. The study of these luminous entities is not merely an astronomical pursuit; it is an odyssey that unravels the cosmic narrative, unveiling the secrets of galactic evolution and the cosmic symphony that has played out over eons. The brilliance of quasars becomes a celestial storyteller, narrating tales of the universe's formative stages and providing invaluable insights into the interplay of cosmic forces.

In essence, the study of quasars transcends the boundaries of conventional astronomy; it becomes a cosmic archaeology, delving into the ancient epochs of the universe. The radiations emitted by quasars carry the imprints of cosmic processes, offering a unique perspective on the formation of galaxies, the evolution of black holes, and the broader celestial dynamics that have shaped the cosmic landscape. As astronomers peer into the luminous glow of quasars, they are not only studying celestial phenomena but also traversing the depths of cosmic time, unlocking the mysteries that have captivated humanity's curiosity since the dawn of stargazing. Quasars, with their brilliance and cosmic significance, stand as beacons illuminating the path toward a deeper understanding of the universe's grand design—a journey that continues to unfold as we explore the cosmic wonders embedded in these extraordinary quasi-stellar objects.

1.1 Background

The historical progression of astronomical observation traces a fascinating trajectory from the rudimentary visual analyses conducted by early astronomers to the sophisticated methodologies employed in modern celestial classification. In the nascent stages of astronomical exploration, stars were scrutinized based on fundamental properties such as brightness and color, while galaxies presented themselves as mysterious smudges on photographic plates, shrouded in cosmic ambiguity.

However, as telescopic capabilities evolved, reaching unprecedented sizes and technological sophistication, the demands for more refined classification schemes became increasingly apparent. This pivotal shift marked a turning point in astronomical methodologies, propelling the field into a new era characterized by enhanced precision and nuanced analyses.

The revolutionary advent of spectroscopy emerged as a transformative tool, offering astronomers the ability to dissect and scrutinize the light emitted or absorbed by celestial objects. Spectroscopy became the metaphorical scalpel, allowing scientists to delve into the intricate details of the electromagnetic spectrum and extract invaluable information about the composition, temperature, and evolutionary stage of celestial entities.

Stars, galaxies, and quasars, though distinct in their cosmic roles, became intricately interwoven components of the vast cosmic tapestry. The application of spectroscopic techniques enabled astronomers to transition from broad categorizations based on visual observations to a more nuanced and insightful classification based on spectral characteristics. This shift not only refined our understanding of individual stars but also unlocked the secrets embedded within galaxies and quasars, illuminating their unique contributions to the cosmic narrative.

In the realm of stellar classification, the analysis of spectral characteristics became paramount. Stars were no longer merely points of light in the night sky; they became cosmic laboratories, each spectrum providing a wealth of information about the elements present, temperatures reached, and stages of evolution undergone. The spectral signatures etched across the cosmic canvas became a language, allowing astronomers to decipher the cosmic stories written in the stellar glow.

Galaxies, once elusive smudges, emerged from the shadows of ambiguity through the lens of spectroscopy. The diverse compositions and evolutionary paths of galaxies unfolded as astronomers scrutinized the intricacies of their spectral imprints. The hitherto mysterious formations of galactic structures became clearer, contributing to a deeper understanding of the vast cosmic landscapes that extend beyond individual stars.

Quasars, as celestial enigmas, revealed their secrets under the scrutiny of spectroscopic analysis. The emissions from these luminous entities, forged in the gravitational cauldron of supermassive black holes, carried distinctive spectral signatures. Spectroscopy allowed astronomers to not only comprehend the intense processes occurring within quasars but also to trace their cosmic origins back to the early epochs of the universe.

This need for nuanced classification, underscored by the power of spectroscopy, underscores the intricate connections between stars, galaxies, and quasars. These celestial entities, once disparate and enigmatic, now stand as interconnected threads in the cosmic fabric. As we navigate the vastness of the universe, armed with advanced spectroscopic tools, the cosmic tapestry unravels before our eyes, revealing the intricate details and profound narratives woven into the very essence of the cosmos. The journey from visual observations to spectroscopic analyses mirrors humanity's evolving quest to comprehend the celestial wonders that have captivated our collective curiosity throughout the ages.

1.2 Goal

The ongoing and planned wide-angle sky surveys represent a monumental undertaking in the realm of astronomy, driven by the overarching goal of mapping the large-scale structure (LSS) of the universe. These ambitious projects employ a diverse array of observational techniques, ranging from galaxy clustering to gravitational lensing, to derive cosmological constraints and illuminate the intricate tapestry of the cosmos.

Among these groundbreaking surveys, the Sloan Digital Sky Survey (SDSS) occupies a prominent position, having left an indelible mark on astronomical research since its inception. The SDSS, with its comprehensive multicolor imaging covering approximately one-third of the celestial sphere, has provided astronomers with an unprecedented panoramic view of the cosmos.

The high-resolution spectra obtained for millions of Galactic and extragalactic objects have proven to be a treasure trove of information, enabling researchers to delve into the composition, temperature, and evolutionary stages of celestial entities.

Moving forward into the realm of systematic exploration of the variable sky, Pan-STARRS and the Zwicky Transient Facility (ZTF) emerge as pioneers. These surveys are engineered to provide a detailed time-series analysis of the dynamic celestial phenomena that grace the cosmic stage. From the continuous monitoring of asteroids to the study of variable stars, supernovae, and active galactic nuclei, Pan-STARRS and ZTF capture the transient nature of cosmic events, offering a unique glimpse into the ever-changing celestial landscape.

In the realm of charting the Milky Way's cosmic geography, the Gaia mission stands as a beacon of precision and scale. Gaia's mission to measure the three-dimensional positions and radial velocities of over a billion stars within the Milky Way and the Local Group promises to revolutionize our understanding of galactic dynamics. The intricate dance of stars and the distribution of stellar populations within our galaxy are poised to be unveiled with unprecedented accuracy.

Looking ahead to the future of observational astronomy, the landscape is set to undergo transformative changes with projects such as the Dark Energy Spectroscopic Instrument (DESI), the Square Kilometer Array (SKA), and the Large Synoptic Survey Telescope (LSST). DESI's focus on dark energy studies will contribute to unraveling the cosmic mysteries behind the universe's accelerated expansion. SKA, with its colossal radio telescope array, is poised to probe the cosmic dawn, unlocking the secrets of the universe's infancy. LSST, as a wide-field survey telescope, is slated to conduct a systematic scan of the southern hemisphere, providing a dynamic and comprehensive view of the cosmic panorama.

In essence, the ongoing and forthcoming wide-angle sky surveys mark a collaborative effort of unprecedented scale and ambition within the global astronomical community. From the foundational contributions of SDSS to the temporal explorations of Pan-STARRS and ZTF, the precision mapping by Gaia, and the transformative promise of future surveys like DESI, SKA, and LSST, these endeavors collectively shape the trajectory of observational astronomy. The vastness of the cosmic web and the underlying principles governing the universe's large-scale structure stand as the focal points of exploration, underscoring humanity's collective endeavor to unravel the mysteries of the cosmos and chart the course for the next era of astronomical discovery.

As we navigate the cosmic seas of discovery, it becomes imperative to appreciate the profound impact of each survey on our understanding of the universe. The Sloan Digital Sky Survey, with its pioneering approach, set the stage for large-scale observational efforts, providing a wealth of data that continues to shape our understanding of galactic and extragalactic phenomena. The multicolor imaging spanning a significant portion of the sky has not only broadened our cosmic perspective but has also laid the groundwork for subsequent surveys seeking to explore the dynamic and transient aspects of the cosmos.

Pan-STARRS and the Zwicky Transient Facility represent the next phase in our quest to understand the variable sky. These surveys, with their systematic exploration and time-series analysis, offer a unique opportunity to capture the ephemeral nature of celestial events. From the trajectories of asteroids to the explosive brilliance of supernovae, these initiatives enhance our capacity to monitor and comprehend the dynamic processes unfolding in the cosmic theater. The transient nature of these phenomena provides critical insights into the evolution and life cycles of celestial objects, contributing to our understanding of the broader cosmic narrative.

The Gaia mission, with its ambitious goal of charting the three-dimensional map of the Milky Way, represents a monumental leap in precision astrometry. By measuring the positions and velocities of an unprecedented number of stars, Gaia not only unravels the intricate structure of our galaxy but also offers a glimpse into the gravitational interactions shaping stellar orbits. The detailed mapping of stellar populations enhances our comprehension of galactic dynamics, providing a holistic view of the Milky Way's history and evolution.

As we gaze towards the future, the promise of transformative surveys such as DESI, SKA, and LSST ushers in a new era of observational capabilities. DESI, with its emphasis on dark energy studies, is poised to illuminate the cosmic phenomena driving the accelerated expansion of the universe. The Square Kilometer Array, a colossal radio telescope project, represents an unprecedented leap in radio astronomy, opening avenues to explore the cosmic dawn and probe the fundamental properties of the universe. The Large Synoptic Survey Telescope, with its wide-field survey capabilities, is set to conduct a systematic exploration of the southern hemisphere, capturing a comprehensive view of the dynamic and evolving cosmic landscape.

In conclusion, the ongoing and future wide-angle sky surveys stand as testament to humanity's insatiable curiosity and relentless pursuit of knowledge. Each survey, with its unique focus and observational techniques, contributes to a mosaic of understanding that collectively deepens our comprehension of the cosmos. From the foundational contributions of SDSS to the temporal insights provided by Pan-STARRS and ZTF, the precision mapping by Gaia, and the transformative potential of DESI, SKA, and LSST, these surveys weave together a narrative that unravels the mysteries of the universe. As we stand on the precipice of a new era in observational astronomy, these endeavors not only expand the frontiers of human knowledge but also inspire future generations to continue the exploration of the cosmic realms that beckon beyond the celestial horizon.

The accurate classification of stars, galaxies, and quasars represents a cornerstone in the edifice of astronomical exploration, transcending mere categorization to unlock the profound secrets of the cosmos. This meticulous classification serves as the keystone for comprehending the fundamental processes that govern the vast tapestry of the universe, ranging from the intricate dance of stellar evolution to the dynamic choreography of galactic structures.

At the heart of this pursuit lies the imperative to decipher the celestial language encoded in the characteristics of stars. Accurate classification of stars is not merely a taxonomic endeavor; it is a journey into the life cycles of these luminous entities that dot the cosmic expanse. Each star, with its unique spectral signature and luminosity, becomes a cosmic storyteller, narrating tales of its birth, evolution, and eventual fate. The nuances embedded in the classification of stars offer a portal to understanding the elemental composition, temperatures, and stages of development, unraveling the celestial mysteries written in the stellar spectra.

Galaxies, vast cosmic islands containing billions of stars, present a complex mosaic of structures and interactions. Accurate classification of galaxies is a linchpin for unraveling the larger cosmic narrative, providing insights into the formation and evolution of these colossal cosmic entities. The diverse compositions, shapes, and sizes of galaxies encapsulate the evolutionary history of the universe itself. Classifying galaxies allows astronomers to discern the underlying principles that have shaped the cosmic landscape over billions of years, offering a panoramic view of the universe's unfolding drama.

Quasars, enigmatic quasi-stellar objects powered by supermassive black holes, add an extra layer of complexity to the cosmic puzzle. Accurate classification of quasars is not only a scientific imperative but also a key to unlocking the cosmic past. These luminous entities, with their intense emissions and gravitational dynamics, serve as cosmic time capsules. The light and electromagnetic radiations emanating from quasars carry within them the echoes of the early universe, offering a unique opportunity to unravel the mysteries of cosmic infancy and the emergence of supermassive black holes.

Beyond individual classifications, the broader significance lies in the interconnectedness of stars, galaxies, and quasars in shaping the cosmic tapestry. Accurate classification becomes a linchpin for deciphering the intricate web of relationships between celestial objects and their roles in the grand cosmic narrative. It enables astronomers to trace the threads of influence that connect the stellar nurseries to the sprawling galactic structures and the luminous beacons of quasars.

Moreover, this meticulous classification fosters a deeper understanding of the dynamic interplay between various celestial components. It provides a roadmap to comprehend how stars contribute to the formation of galaxies and how galaxies, in turn, influence the cosmic environment. Accurate classification becomes a tool for unraveling the cosmic symphony, where each celestial object plays a distinct note, contributing to the harmonious composition of the universe.

In conclusion, the accurate classification of stars, galaxies, and quasars is not merely a scientific endeavor; it is a voyage into the heart of the cosmos. It forms the bedrock for understanding the intricate processes that govern the celestial realm, offering a lens to peer into the evolutionary stories written across the cosmic canvas. This pursuit goes beyond taxonomy; it is a quest to decipher the cosmic narrative, where stars, galaxies, and quasars converge in a symphony of celestial phenomena that has shaped the universe over eons. Accurate classification becomes the guiding light, illuminating the path toward a deeper comprehension of the cosmos and our place within its vast and awe-inspiring expanse.

2. Literature Review:

The exploration of classifying stars, galaxies, and quasars has indeed been a persistent challenge in astronomy, and the evolution of methodologies in this pursuit is fascinating. A comprehensive review of the existing literature reveals a rich tapestry of approaches, with researchers leveraging advancements in technology, particularly the integration of machine learning and data science, to unravel the complexities of these celestial entities.

One cornerstone in this journey is the utilization of the Sloan Digital Sky Survey (SDSS) dataset, a pivotal resource for researchers in the field. The integration of photometric methods marked a significant stride forward, where the measurement of brightness across different wavelength bands became a key tool for characterizing stars, galaxies, and quasars. Landmark studies, such as [1], have played a pivotal role in shaping the trajectory of these methodologies. The introduction of color indices in this seminal work facilitated the differentiation of stars based on their spectral characteristics, contributing substantially to the field.

In the comprehensive study detailed in [2], the authors embarked on a journey to leverage the power of support vector machines (SVM) for the automated classification of celestial objects within the DR6 release of the Sloan Digital Sky Survey (SDSS). This pioneering effort aimed to harness the potential of machine learning techniques to efficiently categorize stars, galaxies, and quasars based on their multicolor characteristics. The initial phase of this endeavor involved the training of the SVM classifier on a meticulously curated dataset. This dataset comprised the u—g, g—r, r—i, and i—z colors of 47,401 stars, 415,634 galaxies, and 71,031 quasars, all of which possessed well-defined spectral classifications. The inclusion of spectral classifications in the training set provided a robust foundation for the SVM to learn intricate patterns and relationships within the four-dimensional color space. The results of the analysis of the classifier's performance were nothing short of remarkable. The total classification error was revealed to be a mere 3.80%, underscoring the efficacy of the SVM in navigating non-linear, four-dimensional class boundaries with exceptional precision. This demonstrated the capacity of the SVM to discern subtle variations and relationships within the multicolor features of celestial objects, showcasing its adaptability to the complexities inherent in the SDSS data. With the SVM classifier finely tuned and validated, the study progressed to the application phase. The SVM was deployed to predict class membership probabilities for an expansive dataset comprising 12,362,179 objects in DR6 that lacked spectral information. The selected subset of objects within the inner 90% of the training color space, coupled with magnitude errors below 10%, constituted the focus of this prediction phase. The SVM predictions yielded a classification distribution of 11,012,775 stars, 1,088,862 galaxies, and 260,542 quasars. Notably, the higher number of galaxies in the predicted distribution was explained by specific constraints imposed on color and magnitude errors. This nuanced approach ensured a more accurate classification in scenarios where galaxies exhibit diverse colors and potential errors in magnitude measurements. To validate the robustness of the SVM predictions, a cross-match was conducted against external surveys, namely the FIRST and USNO-B surveys. In the cross-match with FIRST, a radio survey, 8,666 radio sources were identified. The SVM classifier exhibited a remarkable accuracy of 94.8% in predicting these radio sources to be galaxies or quasars, aligning seamlessly with the anticipated outcomes. This alignment with external surveys served as a powerful validation of the SVM's capability to generalize its learned patterns to new, unseen data, affirming its reliability in the classification of celestial objects. In summary, the study showcased the prowess of support vector machines in automating the classification of celestial objects within the SDSS DR6 release. From the meticulous training on a dataset with spectral classifications to the efficient prediction of class membership probabilities for a vast dataset lacking spectra, the SVM emerged as a robust tool for navigating the complexities of the cosmic landscape. The success of the SVM in accurately classifying objects, even in the absence of spectral information, paves the way for enhanced efficiency and scalability in the automated categorization of celestial phenomena, marking a significant stride in the synergy between machine learning and astronomy.

In their notable paper [3], Yulun Wu embarked on a pivotal quest with the primary objective of developing a predictive model for the classification of stars, galaxies, and quasars. Grounded in the hypothesis that leveraging astronomical data from the Sloan Digital Sky Survey (SDSS) could yield a potent machine learning model, the investigation unfolded as a systematic exploration of various classification models to discern the most effective approach. The initial focal point of the study centered on the implementation of a multinomial logistic regression model. This model was meticulously trained and rigorously tested to ascertain its efficacy in classifying celestial entities. The results were noteworthy, with the logistic regression model achieving an accuracy of 0.87. Correspondingly, the weighted average precision, recall, and f-1 score values echoed this high level of accuracy. Moreover, the cross-validation accuracy score, serving as a litmus test for the model's generalizability, stood at a commendable 0.8664. This initial phase laid the groundwork for a comprehensive understanding of the logistic regression model's performance in discerning stars, galaxies, and quasars. The investigation took a significant turn with the introduction of a decision tree model into the analysis. The results garnered from this addition were nothing short of robust, as the decision tree model exhibited an accuracy that soared to an impressive 0.99. The weighted average precision, recall, and f-1 score, all reflective of the model's performance, matched this exceptional accuracy level at 0.99. Cross-validation accuracy scores further underscored the superiority of the decision tree model, reaching an outstanding 0.99 and 0.9858. This stark contrast in performance between the logistic regression and decision tree models emphasized the decisiveness of the decision tree in accurate celestial object classification. Yulun Wu's findings not only showcased the effectiveness of the decision tree model but also highlighted its superiority over the logistic regression model across key performance metrics. The decision tree, with its exceptional accuracy and robustness, emerged as a formidable classifier for stars, galaxies, and quasars. This success not only validated the initial hypothesis but also demonstrated the feasibility of training a machine learning model on SDSS astronomical data for precise and reliable celestial object classification. In essence, Yulun Wu's research represented a significant stride in the intersection of machine learning and astronomy. The study not only contributed a robust decision tree model for celestial object classification but also provided valuable insights into the comparative performance of different classification approaches. The success of the decision tree model in this context opens avenues for further exploration and applications of machine learning techniques in the realm of astronomical data analysis, marking a noteworthy advancement in our ability to understand and categorize celestial phenomena with unprecedented precision.

Zhuliang Qi's paper [4] stands as a significant contribution to the field of astronomical data analysis, emphasizing the critical importance of improved classification processes in the context of the Sloan Digital Sky Survey (SDSS). Qi's research delves into the intricacies of leveraging machine learning models for the classification of celestial objects, shedding light on the challenges posed by imbalanced datasets and presenting a nuanced evaluation of decision tree, random forest, and support vector machine (SVM) models. The foundation of Qi's study lies in a carefully curated training dataset derived from the SDSS. This dataset encompasses spectral features and consists of 100,000 star samples, each characterized by 17 features. Within this rich dataset, a significant imbalance in sample numbers among galaxies, quasars, and stars is identified, presenting a potential hurdle for effective machine learning due to this imbalance. Qi addresses this challenge by employing the Synthetic Minority Over-sampling Technique (SMOTE) function from the imbalanced-learn library in Python for oversampling, ensuring a more balanced representation of the different celestial classes. The evaluation of the performance of decision tree, random forest, and SVM models yields intriguing insights. The decision tree model emerges as a standout performer among the three, showcasing high accuracy and commendable operational efficiency. Notably, the decision tree model processes information at the fastest speed, making it an efficient choice for rapid classification tasks. Additionally, it excels in star classification, underscoring its versatility and precision in discerning specific celestial objects. The random forest model, while slightly trailing the decision tree in terms of overall accuracy, takes the lead in precision rates. This indicates its effectiveness in minimizing false positives, a crucial aspect in astronomical classification where precision is paramount. The combination of high precision and operational efficiency positions the random forest model as a robust contender, particularly excelling in handling imbalanced datasets—a common challenge in astronomical data. The SVM model, while achieving an impressive overall accuracy of 97%, stands out for its extended training time. The meticulous training process takes 6 minutes and 46.9 seconds, marking the longest duration among the three models. Despite this, the SVM model maintains a high accuracy rate, underscoring its efficacy in making correct predictions. The longer training time, however, does impact its operational efficiency, highlighting a trade-off between accuracy and processing speed. In summary, Zhuliang Qi's research not only underscores the importance of addressing imbalanced datasets in astronomical classification but also provides a nuanced comparison of decision tree, random forest, and SVM models. The decision tree's speed and versatility, the random forest's precision and efficiency in handling imbalanced data, and the SVM's accuracy despite longer training times collectively contribute to a comprehensive understanding of the strengths and trade-offs associated with each model. Qi's work opens avenues for optimizing classification processes in astronomy, paving the way for more accurate and efficient analysis of celestial objects in large-scale surveys like the SDSS.

The study conducted by the author [5] addresses a crucial challenge in the realm of astronomical data analysis – the creation of training sets for classification in supervised learning, a process often demanding significant time and resources. Instead of following the traditional supervised learning approach, the author proposed an innovative method leveraging unsupervised machine learning for the categorization of stars, galaxies, and QSOs based on photometric data. The primary methodology employed in this research involves the application of Hierarchical Density-Based Spatial Clustering of Applications with Noise (hdbscan) to identify clusters in a color space that correspond to stars, galaxies, and QSOs. The author undertook a meticulous process of fine-tuning hyperparameters and selecting input attributes through three separate hdbscan runs, each specifically targeting a distinct object class. The results from these runs were treated as binary classifiers, and their outputs were combined to derive final classifications, optimizing the process based on F1 scores. As part of the data preparation phase, the author explored the utilization of Random Forest and Principal Component Analysis (PCA) for feature selection and data simplification. With a dataset comprising approximately 50,000 spectroscopically labeled objects, the author achieved remarkable F1 scores of 98.9, 98.9, and 93.13 for star, galaxy, and QSO identification, respectively, using the unsupervised hdbscan method. This high level of accuracy underscores the effectiveness of the proposed approach in accurately categorizing celestial objects based solely on photometric data. The study places particular emphasis on the importance of careful attribute selection for achieving accurate classification results using hdbscan. The innovative application of unsupervised learning in this context not only streamlines the classification process but also demonstrates the potential for correcting misclassified spectra. This aspect is particularly valuable for large-scale projects like the Dark Energy Spectroscopic Instrument (DESI) and 4-metre Multi-Object Spectroscopic Telescope (4MOST), where accurate spectral classification is crucial. Furthermore, the author extended the application of this classification method to a subset of the Sloan Digital Sky Survey (SDSS) spectroscopic catalog, showcasing its potential in rectifying misclassifications within existing datasets. The creation of a catalog containing 2.7 million sources from various surveys, complete with classifications and photometric redshifts, further emphasizes the scalability and versatility of the proposed approach. In conclusion, the study not only introduces an innovative and efficient method for unsupervised classification of celestial objects based on photometric data but also highlights its potential applications in correcting misclassifications within existing datasets. The results presented in the research showcase the feasibility of achieving high accuracy in object classification without the need for extensive labeled training data, paving the way for more streamlined and resource-efficient approaches in the field of astronomical data analysis.

The paper [6] presents a pioneering approach in the field of astronomical object classification by proposing a deep convolutional neural network (CNN) architecture specifically tailored for galaxy morphology classification. This research marks a significant stride in integrating deep learning methodologies into the landscape of astronomical classification, offering a novel system that demonstrates the capabilities of machine learning in automating and improving the efficiency of large-scale astronomical studies. The core of the proposed system lies in the deployment of two distinct models, each serving a unique purpose within the classification framework. The first model focuses on predicting the types of celestial objects, distinguishing between galaxies, stars, and quasars. The second model is designed to forecast the specific shapes of galaxies, providing a more granular understanding of the morphological characteristics of these cosmic entities. To arrive at this innovative system, the authors conducted a thorough comparison of various machine learning models. Notably, Random Forest and K-Nearest Neighbors (KNN) emerged as the top-performing models. The Random Forest algorithm exhibited particular effectiveness in class prediction, showcasing high accuracy in discerning the types of celestial objects. On the other hand, Convolutional Neural Networks (CNNs), a specialized form of deep learning, demonstrated superior capabilities in predicting galaxy morphology, capturing intricate features that might be challenging for traditional machine learning models. The proposed system's efficacy is underscored by its high accuracy in both predicting the types of celestial objects and determining the shapes of galaxies. The strategic combination of Random Forest for class prediction and CNNs for galaxy morphology prediction contributes to a holistic and accurate classification approach. A notable aspect of the study is the integration of data from different sources, including the Sloan Digital Sky Survey (SDSS) and the LAMOST survey. This multi-source data fusion enhances the accuracy of predictions, highlighting the potential benefits of leveraging diverse datasets for comprehensive astronomical studies. The utilization of machine learning techniques in this context not only automates the classification process but also facilitates the extraction of intricate features, paving the way for a more efficient and accurate analysis of astronomical objects on a large scale. In essence, the paper reinforces the potential of machine learning, specifically deep learning with CNNs, in revolutionizing the classification of astronomical objects. The combination of Random Forest and CNNs, along with the integration of diverse data sources, exemplifies a powerful and versatile methodology for extracting meaningful insights from large-scale astronomical surveys. This work signifies a paradigm shift in astronomical data analysis, opening avenues for more sophisticated and automated approaches to understanding the diverse and complex nature of celestial objects in our universe.

The exploration of ensemble learning is evident in [7], where the efficacy of asymmetric AdaBoost is highlighted for photometric data classification. The comparative analyses presented in various papers, including [8], [9], and [10], demonstrate the meticulous efforts of researchers in fine-tuning machine learning models for the classification of Star, Galaxies, and Quasars (SGQs).

Photometric data emerges as a focal point in several studies, with researchers like Shashank Shetye Saudagar [9] suggesting its efficiency for celestial object classification. Machine learning models are scrutinized based on their performances in [10], where Idel R. Waisberg discusses the application of machine learning algorithms on 5-band photometric data. Additionally, the utilization of multiple neural networks for analyzing photometric data is proposed in [11], reflecting a commitment to achieving higher accuracies.

Deep learning makes its mark in the classification landscape, as evident in [12], where it is employed for classifying, detecting, and predicting SGQs. The integration of classical machine learning algorithms and artificial neural networks for tabular and image data, respectively, is explored in [13], showcasing a multi-modal approach for enhanced classification results.

Transfer learning and feature selection also find a place in this exploration. Clarke [14] discusses the implications of transfer learning, while [15], by Miguel Garcia-Torres, provides a valuable comparison of five different feature selection algorithms for both photometric and spectroscopic data. This emphasis on efficient feature selection underscores the importance of optimizing resources for accurate and timely results.

In conclusion, the landscape of classifying stars, galaxies, and quasars is vibrant and dynamic, with researchers at the intersection of IT technologies and astronomy continually pushing boundaries. The integration of machine learning, data science, and advanced algorithms has not only expanded our understanding of the cosmos but also paved the way for more sophisticated and efficient classification methodologies. This rich tapestry of research contributes significantly to the exploration and development of projects at the convergence of IT and astronomy, fostering a collaborative and innovative spirit in the scientific community.

3. Methodology:

3.1 Data

Throughout the paper, we have used the Sloan Digital Sky Survey (SDSS) dataset for all the analysis and classification. The SDSS is a pioneering astronomical survey that has significantly contributed to our understanding of the universe.

The SDSS project was initiated in the late 1990s with the primary goal of creating a comprehensive 3D map of the universe. It officially began operations in the year 2000 and has since undergone multiple phases of data collection. The project was a collaboration between several institutions, including the Astrophysical Research Consortium, the University of Chicago, the Institute for Advanced Study, and others.

The survey utilized a dedicated 2.5-meter wide-angle optical telescope located at Apache Point Observatory in New Mexico, equipped with a unique imaging camera and spectrographs. The telescope's wide field of view enabled the efficient mapping of large regions of the sky. The survey captured data in multiple wavelength bands, providing detailed information about celestial objects.

One of the key components of the SDSS dataset is its photometric catalog. This catalog includes precise measurements of the brightness of celestial objects in different filters, allowing astronomers to study the spectral energy distribution of stars, galaxies, and quasars. The spectroscopic data, on the other hand, provides detailed information about the composition and redshift of these objects.

The SDSS dataset has been instrumental in advancing various areas of astronomical research. Here are some notable applications:

1. **Cosmology Studies:** SDSS data has been crucial in constraining the cosmological parameters, contributing to our understanding of the large-scale structure of the universe, including the distribution of galaxies and the cosmic web.

2. **Galaxy Morphology and Evolution:** The detailed imaging data from SDSS has allowed astronomers to study the morphology and evolution of galaxies. This includes identifying and classifying different types of galaxies based on their shapes, sizes, and color properties.

3. **Quasar Research:** SDSS has discovered and characterized a vast number of quasars—highly luminous and active galactic nuclei. These findings have shed light on the early universe, providing insights into the conditions prevailing during its infancy.

4. **Stellar Astrophysics:** The survey has contributed to the study of stellar populations, enabling researchers to investigate the properties and distribution of stars in our Milky Way and nearby galaxies.

An important aspect of the SDSS project is its commitment to open science. The data collected, including images, spectra, and catalogs, are made publicly available. This has empowered astronomers worldwide to conduct their research using this rich dataset. In addition to its scientific contributions, SDSS has engaged in public outreach activities, making its findings accessible to a broader audience. Educational initiatives, citizen science projects, and interactive tools have been developed to involve the public in exploring the wonders of the universe revealed by the survey. The legacy of the SDSS continues with ongoing and upcoming projects building upon its success. The next-generation surveys, such as the Dark Energy Spectroscopic Instrument (DESI), aim to push the boundaries further by mapping a larger volume of the universe with increased precision.

In conclusion, the SDSS dataset stands as a monumental achievement in modern astronomy, providing a wealth of information that has fueled numerous discoveries and advancements. Its impact extends beyond the realm of astronomy, influencing fields like machine learning and data science, where the analysis of such vast datasets poses unique challenges and opportunities. Despite its tremendous value, working with the SDSS dataset also presents challenges. The data is heterogeneous, with variations in quality and completeness across different regions of the sky. Dealing with such complexities requires careful consideration and innovative solutions.

1. Quality Assurance: Ensuring the accuracy and reliability of the data is a continuous challenge. Machine learning models need to account for potential biases and errors in the

observational data, and efforts are ongoing to improve data quality through calibration and validation processes.

2. Scalability: As datasets in astronomy continue to grow, scalability becomes a significant concern. Machine learning models need to be designed to handle the increasing volume of data efficiently, requiring advancements in parallel computing and distributed processing.

3. Interdisciplinary Collaboration: Bridging the gap between astronomy and machine learning requires interdisciplinary collaboration. Engaging astronomers, data scientists, and domain experts in collaborative projects is essential for developing effective machine learning solutions tailored to the specific challenges of astronomical data.

Integration with Data Science Techniques:

Beyond machine learning, the SDSS dataset offers a rich playground for various data science techniques. Exploratory data analysis, statistical modeling, and data visualization play crucial roles in uncovering patterns, trends, and outliers within the dataset.

1. Statistical Analysis: Statistical methods are applied to infer properties of the overall population of celestial objects based on the observed sample. Confidence intervals, hypothesis testing, and regression analysis are employed to draw meaningful conclusions from the data.

2. Data Visualization: Visualizing complex astronomical data is essential for understanding patterns and trends. Techniques such as heatmaps, scatter plots, and 3D visualizations enable researchers to explore the relationships between different parameters and identify interesting correlations.

3. Time-Series Analysis: For studying transient phenomena, time-series analysis techniques are employed. This includes identifying periodic signals, characterizing variability, and detecting short-lived events captured by the SDSS observations.

The convergence of IT technologies, machine learning, and data science with astronomy opens up exciting possibilities for collaborative projects. Initiatives that bring together experts from both fields can drive innovation and address challenges more effectively.

1. **Community Challenges:** Organizing community challenges and competitions focused on specific tasks within the SDSS dataset encourages collaboration and the development of novel algorithms. Platforms like Kaggle have hosted competitions related to astronomical data analysis, fostering a global community of data scientists and astronomers.

2. **Open Source Tools and Libraries:** Developing and maintaining open-source tools and libraries specifically designed for analyzing astronomical data, including the SDSS dataset, facilitates knowledge sharing and collaboration. This can include software for

data preprocessing, machine learning, and visualization tailored to the unique characteristics of astronomical datasets.

In summary, the SDSS dataset serves as a remarkable resource for the convergence of machine learning, data science, and astronomy. Leveraging these technologies not only enhances our understanding of the universe but also presents exciting opportunities for collaboration and innovation at the intersection of these fields. As the field continues to evolve, it is crucial to foster a collaborative and interdisciplinary approach to unlock the full potential of the SDSS dataset and similar astronomical surveys.

3.1.1 Data Acquisition

Acquiring data from the Sloan Digital Sky Survey (SDSS) involves accessing the extensive dataset through dedicated platforms and interfaces. Here, we'll explore the various ways to acquire SDSS data, including the source site, available tools, and different methods for obtaining the dataset.

1. SDSS Data Access:

The primary source for accessing SDSS data is the official SDSS website, which provides user-friendly interfaces and tools for retrieving data. The website is maintained by the SDSS collaboration and offers several avenues for accessing different types of data, including images, spectra, and catalogs.

The SDSS website is designed to cater to both astronomers and the general public. It features a user-friendly interface with navigation menus, search functionalities, and detailed documentation to assist users in navigating the vast dataset. Here are some key sections of the SDSS website relevant to data access:

- **Data Access Menu:** The website typically includes a dedicated section or menu for data access. This is where users can find links to tools, interfaces, and documentation related to obtaining SDSS data.
- **Data Release Information:** SDSS releases its data in stages known as data releases. Each release includes new and updated data. The website provides detailed information about each release, including release notes, data formats, and data quality assessments.
- **SkyServer Interface:** SkyServer is a web portal hosted by SDSS that allows users to interactively explore and query the SDSS database. It provides a convenient way for users to visualize and retrieve data based on specific criteria. This will be further explained.

There are multiple ways to download SDSS data, catering to different user preferences and needs. Here are some common methods:

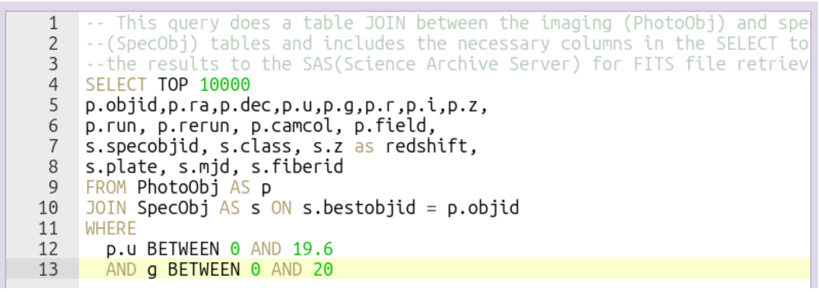
- **Bulk Download:** SDSS offers bulk download options for users who need to retrieve large amounts of data. This is particularly useful for researchers working with extensive datasets or conducting analyses that require a local copy of the data.

- **APIs (Application Programming Interfaces):** SDSS provides APIs that allow users to programmatically access and retrieve data. This is beneficial for users who wish to integrate SDSS data directly into their analysis pipelines or custom applications.
- **Virtual Observatory Tools:** The SDSS dataset is often accessible through Virtual Observatory (VO) tools and protocols. VO tools enable interoperability between different astronomical datasets and can facilitate complex queries and cross-referencing with other surveys.

SkyServer, mentioned earlier, is a powerful tool for exploring the SDSS dataset interactively. It provides a web interface for constructing SQL queries, visualizing sky coverage, and accessing various data products.

- **SQL Queries:** Researchers can write SQL queries to extract specific subsets of data based on criteria such as celestial coordinates, object types, and redshift ranges. SkyServer provides a SQL tutorial and a query tool for constructing and executing queries.

The following SQL code snippet was used to fetch the data:



```sql
-- This query does a table JOIN between the imaging (PhotoObj) and spe
--(SpecObj) tables and includes the necessary columns in the SELECT to
--the results to the SAS(Science Archive Server) for FITS file retriev
SELECT TOP 10000
p.objid,p.ra,p.dec,p.u,p.g,p.r,p.i,p.z,
p.run, p.rerun, p.camcol, p.field,
s.specobjid, s.class, s.z as redshift,
s.plate, s.mjd, s.fiberid
FROM PhotoObj AS p
JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
  p.u BETWEEN 0 AND 19.6
  AND g BETWEEN 0 AND 20
```

Explanation:

SELECT Clause: SELECT TOP 10000: This part of the query instructs the database to retrieve the top 10,000 records that match the specified conditions. The TOP keyword is used in SQL Server to limit the number of rows returned. The subsequent lines list the specific columns to be included in the result set. These columns are selected from two tables: PhotoObj (aliased as p) and SpecObj (aliased as s).

FROM Clause: FROM PhotoObj AS p: Specifies the primary table (PhotoObj) from which the columns will be retrieved. The table is aliased as p for brevity in the query.

JOIN SpecObj AS s ON s.bestobjid = p.objid: This part of the query performs an inner join between the PhotoObj table and the SpecObj table. The condition for the join is that the bestobjid column in the SpecObj table must match the objid column in the PhotoObj table.

WHERE Clause:

- **p.u BETWEEN 0 AND 19.6:** specifies a condition on the u column in the PhotoObj table. It filters the results to include only those, where the u value is between 0 and 19.6.
- **g BETWEEN 0 AND 20:** Adds another condition, filtering the results to include only those where the g value is between 0 and 20.

The columns selected in the result set include various parameters related to astronomical observations. Here is a short explanation of each of the column selected:

- ra, dec — right ascension and declination respectively
- u, g, r, i, z — filter bands (a.k.a. photometric system or astronomical magnitudes)
- run, rerun, camcol, field — descriptors of fields (i.e. 2048 x 1489 pixels) within image
- redshift — increase in wavelength due to motion of astronomical object
- plate — plate number
- mjd — modified Julian date of observation
- fiberid — optic fiber ID

This SQL query retrieves information from the SDSS database about celestial objects based on specific criteria related to the magnitudes in the u and g filters. The use of the JOIN clause combines information from both the PhotoObj and SpecObj tables, providing a comprehensive set of parameters for the selected objects. The result is limited to the top 10,000 records that meet the specified conditions.

In conclusion, acquiring data from the Sloan Digital Sky Survey involves navigating the official SDSS website, utilizing tools like SkyServer and CasJobs, and choosing the appropriate download method based on the user's specific requirements. The availability of various interfaces, APIs, and bulk download options makes SDSS data accessible to a wide range of users, from astronomers conducting in-depth research to enthusiasts exploring the wonders of the universe.

3.1.2 Data Processing

Throughout the paper we use the SDSS DR18 as our raw data. The raw data obtained from the Sloan Digital Sky Survey (SDSS) represents a vast collection of astronomical observations, comprising imaging and spectroscopic data. The processing of this data involves several steps aimed at cleaning, transforming, and extracting relevant information to facilitate meaningful analysis.

Raw astronomical data often contains artifacts, outliers, and imperfections that can affect the accuracy of subsequent analyses. To ensure the reliability of our results, a rigorous quality control process is applied. Data points with missing or anomalous values are identified and either corrected or removed. The SQL conditions, such as filtering based on magnitude ranges in the u and g bands, serve as an initial step in this quality control process, ensuring that the selected data points meet predefined criteria.

First, we do some basic EDA in order to understand the data. EDA helps us to see the types of data present in the data, the available features in it and how the features are correlated with each other, missing values or any kind of inconsistencies in the data.

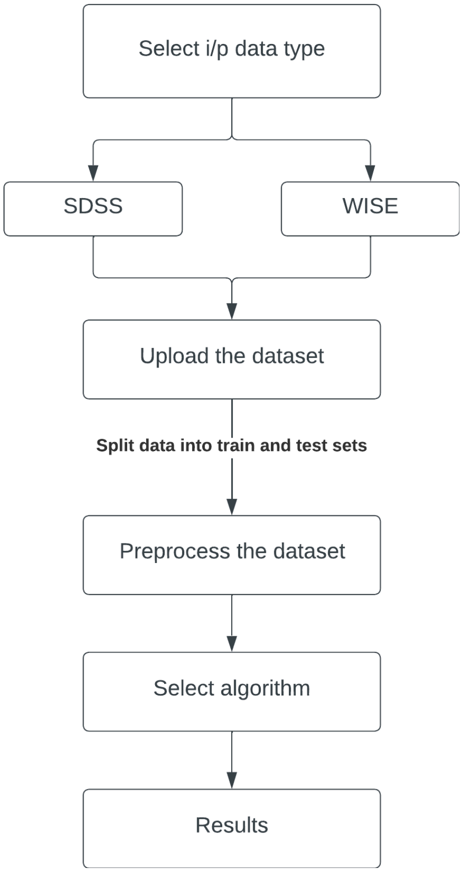
EDA allows us to meticulously explore the features present in the dataset. By examining each feature individually, we can uncover the range and distribution of values, providing insights into the variability and characteristics of the data. This step is particularly crucial for comprehending the diversity of information encapsulated within the dataset.

3.2 Algorithm Selection

The XGBoost algorithm emerges as a compelling choice for your research in astronomical data analysis. Its notable effectiveness in class prediction, as demonstrated in various studies, ensures accurate categorization of celestial objects. The algorithm's ability to handle imbalanced datasets, a common challenge in astronomy, makes it particularly well-suited for tasks involving stars, galaxies, and quasars. Moreover, XGBoost excels in scenarios where interpretability is crucial, offering insights into feature importance for a deeper understanding of the underlying data. The algorithm's versatility, demonstrated across different studies and datasets, positions it as a robust tool for automating and enhancing the efficiency of large-scale astronomical studies, making it a valuable asset for your research endeavors at the intersection of machine learning and astronomy.

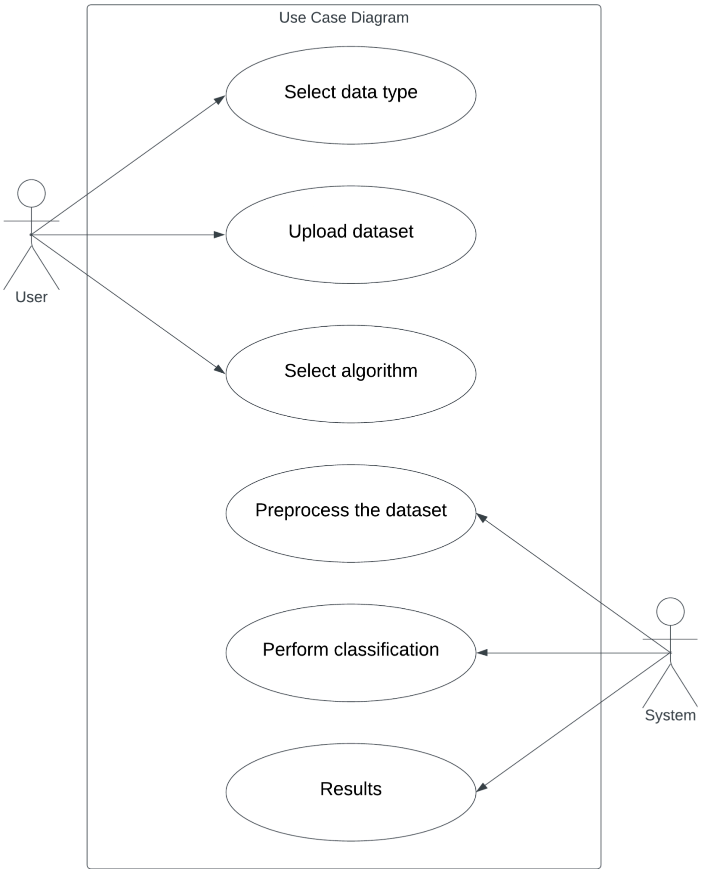
4. Designing:

4.1 Data Flow Diagram:



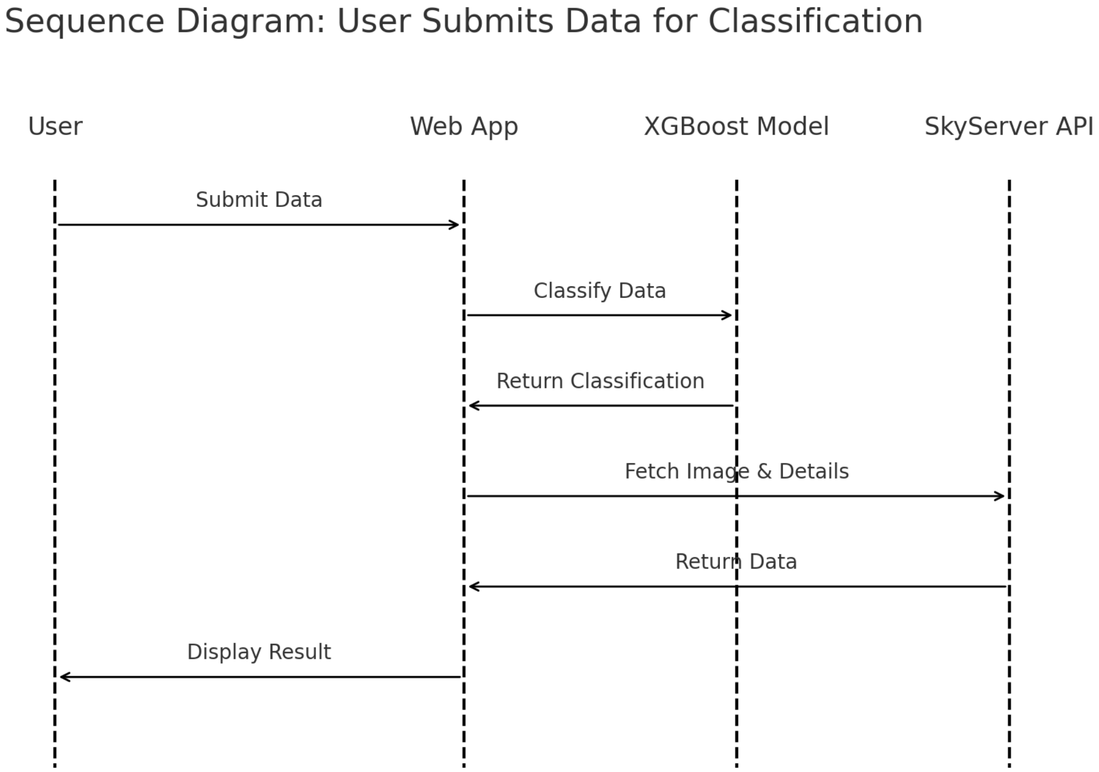
The first step is to select the type of data that will be used for implementing the algorithm. Once the data type has been selected, the dataset is uploaded to a local storage. The dataset is then split into two sets: a train set and a test set. The train set is used to train the algorithm, and the test set is used to evaluate its performance. The train set is preprocessed to improve the algorithm's performance. Finally, a machine learning algorithm is selected and trained on the train set. The algorithm is then evaluated on the test set, and the results are analyzed.

4.2 Use Case diagram:



This is a Use-Case diagram for the classification application. The system has six main components: the system itself and the user. The use case diagram shows the different actions that the user can take within the system, including uploading a dataset, selecting a data type, selecting an algorithm, preprocessing the dataset, performing classification, and viewing results. The diagram also shows the relationships between the user and the different components of the system. Specifically, the user interacts with the system by uploading a dataset, selecting an algorithm, and viewing results. The system interacts with the user by providing the results of the classification. Finally, the system interacts with the dataset by preprocessing it before classification and provides the users with the acquired results.

4.3 Sequence Diagram:



The diagram illustrates a workflow for processing and analyzing datasets. The user selects the dataset type and uploads it to the system, which stores it in local storage. The user then requests the dataset and selects an algorithm to apply to the processed dataset. The processed dataset is stored and the selected algorithm is applied to it. The results are then displayed with visualization and evaluation of the classified output. The processed dataset can also be accessed directly from storage. Overall, the workflow provides a clear and structured approach on how the data is being collected from the user and how its processed and analyzed to provide the required classification with some visualizations.

5. Implementation

5.1 Overview of Implementation

The implementation phase of the 'Stellar Classification' project involved the development of a web application designed to provide users with valuable insights into the Sloan Digital Sky Survey (SDSS) dataset. The primary objectives were to create an interactive user interface for exploratory data analysis (EDA) and to build a robust classification model capable of identifying whether input astronomical data corresponds to stars, galaxies, or quasars. This chapter details the various components and technologies utilized in the project, including the frontend and backend implementations, data handling, and the machine learning model.

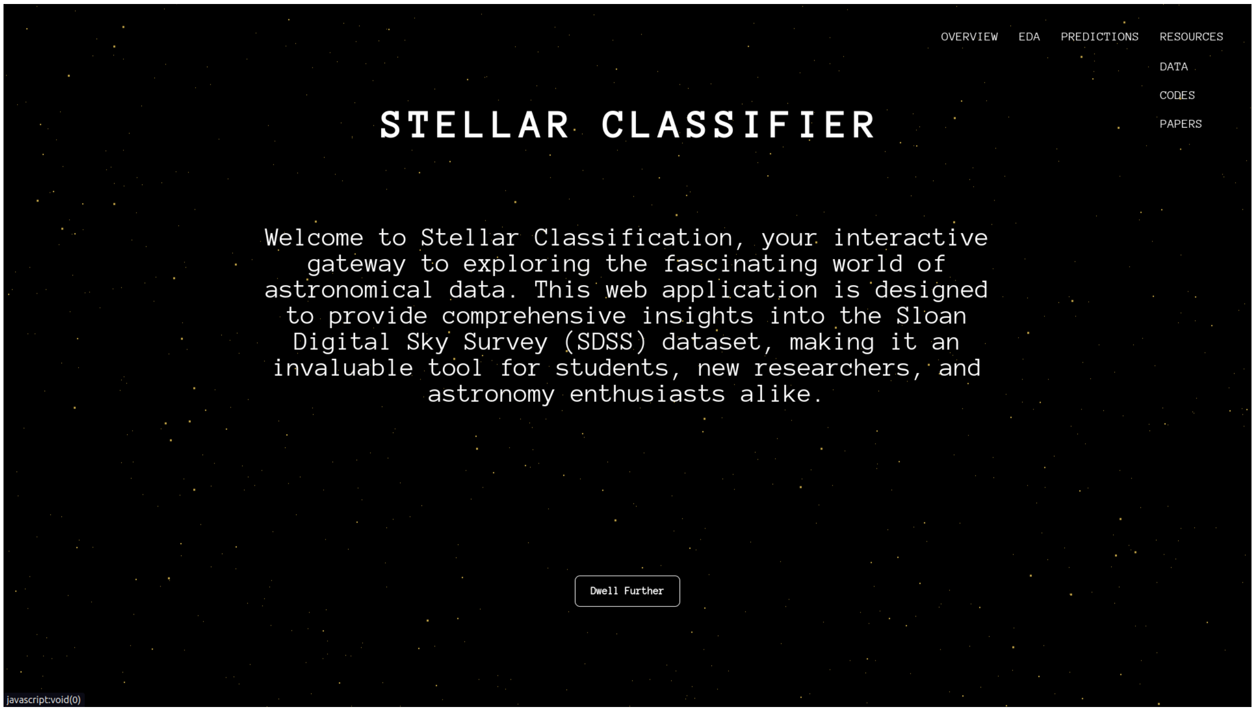
5.2 System Architecture

The system architecture of the 'Stellar Classification' project was meticulously designed to ensure seamless interaction between the frontend, backend, and data processing components. The architecture consists of several key components: a Flask web application serving as the primary user interface, a Streamlit page dedicated to displaying various EDA charts and plots, a highly accurate XGBoost classification model, a database for managing the astronomical data, and APIs for retrieving images and additional details from the SkyServer portal.

The Flask web application is the core component that ties together all the functionalities, facilitating navigation through different sections such as the homepage, overview, EDA, predictions, and resources. The Streamlit page provides an interactive platform for users to visualize the data through a variety of charts and plots. The XGBoost model, which was meticulously trained and optimized, serves as the backbone of the classification feature. The database, structured as a '.db' file, stores the data collected from various SDSS data releases, making it accessible for analysis and visualization. Finally, the APIs enhance the application by sourcing relevant images and details from the SkyServer portal based on the classification results.

5.3 Frontend Implementation

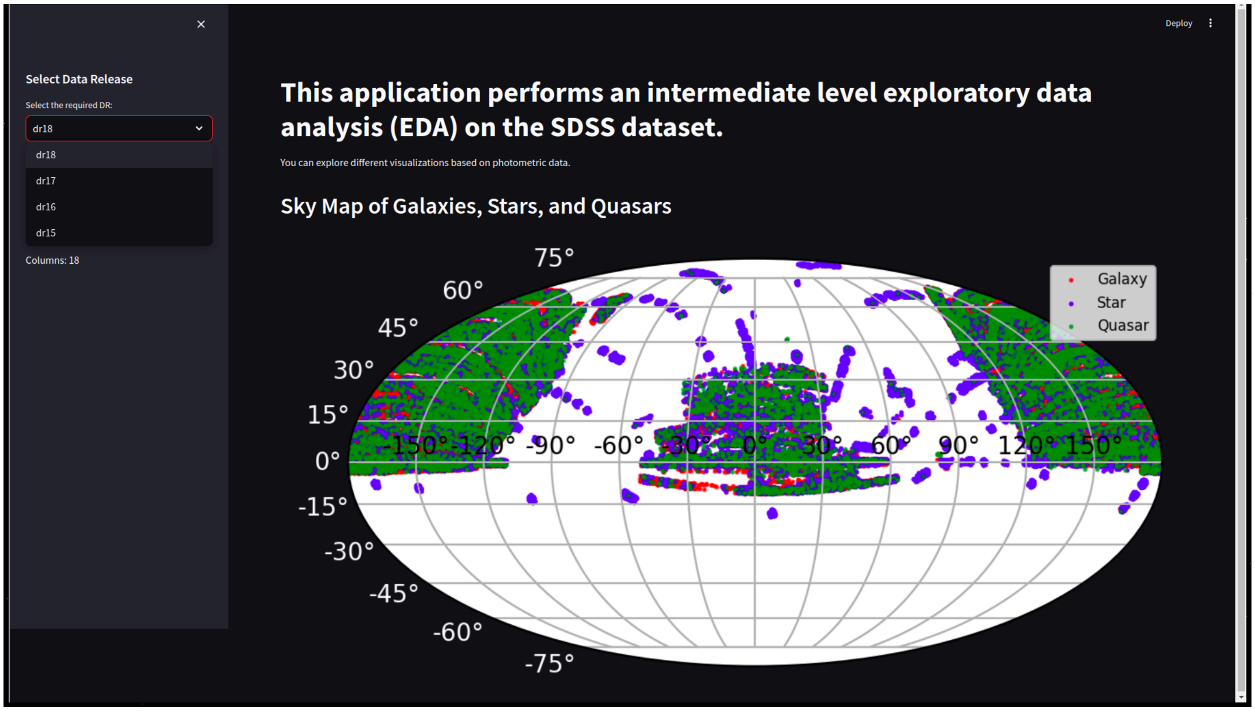
The frontend of the 'Stellar Classification' project was developed using Flask, HTML, CSS, and JavaScript, aiming to create a user-friendly and intuitive interface. The homepage features a welcoming text introducing the project and its purpose, along with links to various sections of the web application. The overview section provides introductory information about the SDSS dataset, organized in ascending order by years to showcase the evolution of data releases.



STELLAR CLASSIFIER

Welcome to Stellar Classification, your interactive gateway to exploring the fascinating world of astronomical data. This web application is designed to provide comprehensive insights into the Sloan Digital Sky Survey (SDSS) dataset, making it an invaluable tool for students, new researchers, and astronomy enthusiasts alike.

The EDA section allows users to perform exploratory data analysis on different SDSS data releases, offering a dropdown menu for selecting specific releases and various graphs for visualizing the data. This section includes interactive elements that enable users to explore the dataset in detail, enhancing their understanding of the astronomical data.



Select Data Release

Select the required DR: dr18

Columns: 18

This application performs an intermediate level exploratory data analysis (EDA) on the SDSS dataset.

You can explore different visualizations based on photometric data.

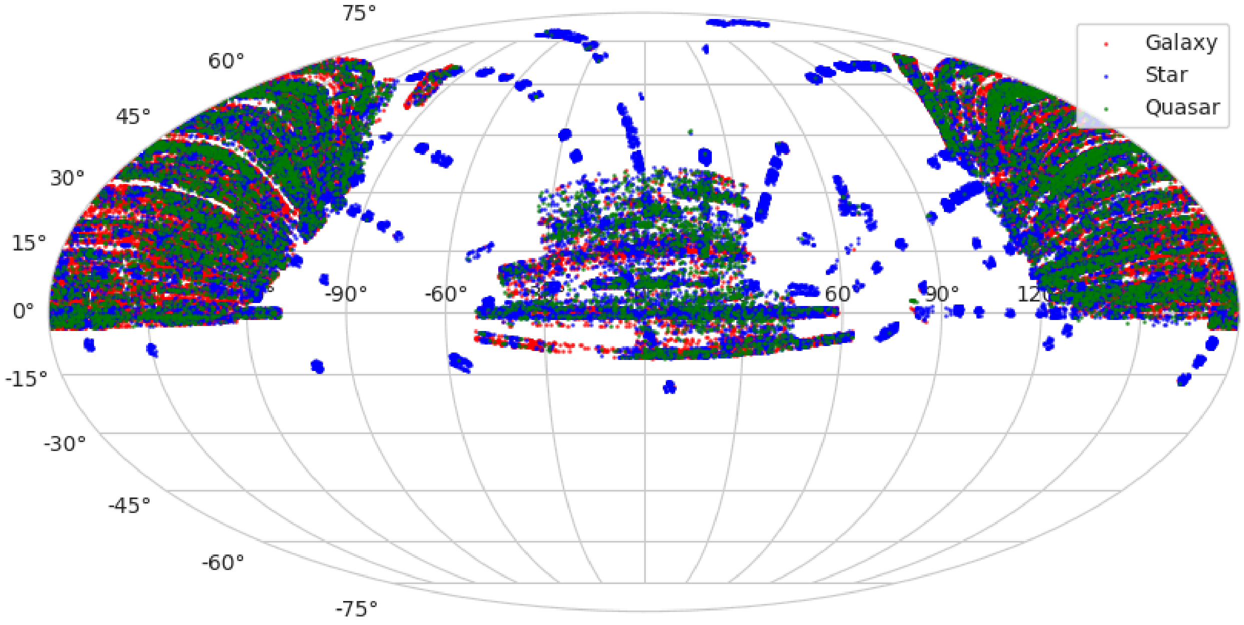
Sky Map of Galaxies, Stars, and Quasars

The prediction section is designed to take user inputs in the form of values such as dec, u, g, r, i, z, and redshift, and provide a classification of whether these values correspond to a star, galaxy, or quasar. Upon receiving the classification results, the application displays relevant images and additional information about the identified celestial object. The resources section, accessible via a dropdown menu, links to relevant code, research papers, and the data source, providing users with a comprehensive set of resources for further exploration.

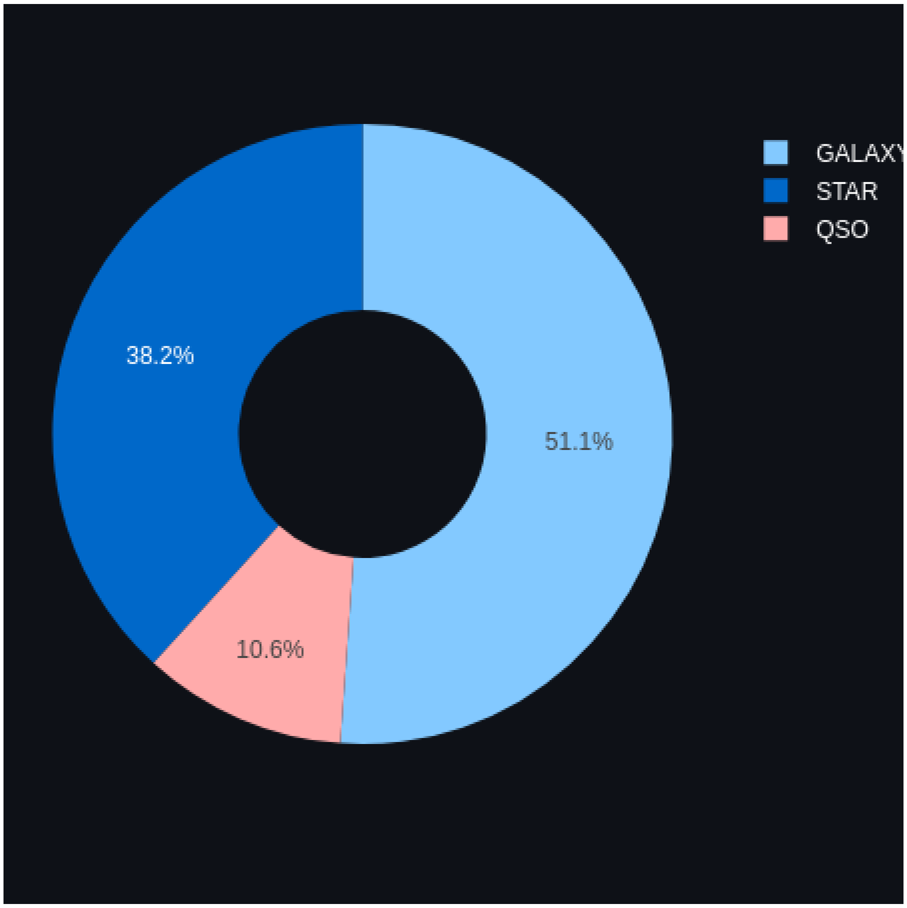
5.4 Exploratory Data Analysis (EDA) Implementation

The EDA component of the project is implemented using a Python script that leverages various data visualization libraries, including Altair, Plotly, Matplotlib, and Seaborn. This script accesses different SDSS data releases (DR18 to DR15) and generates a series of charts and plots to help users visualize the data.

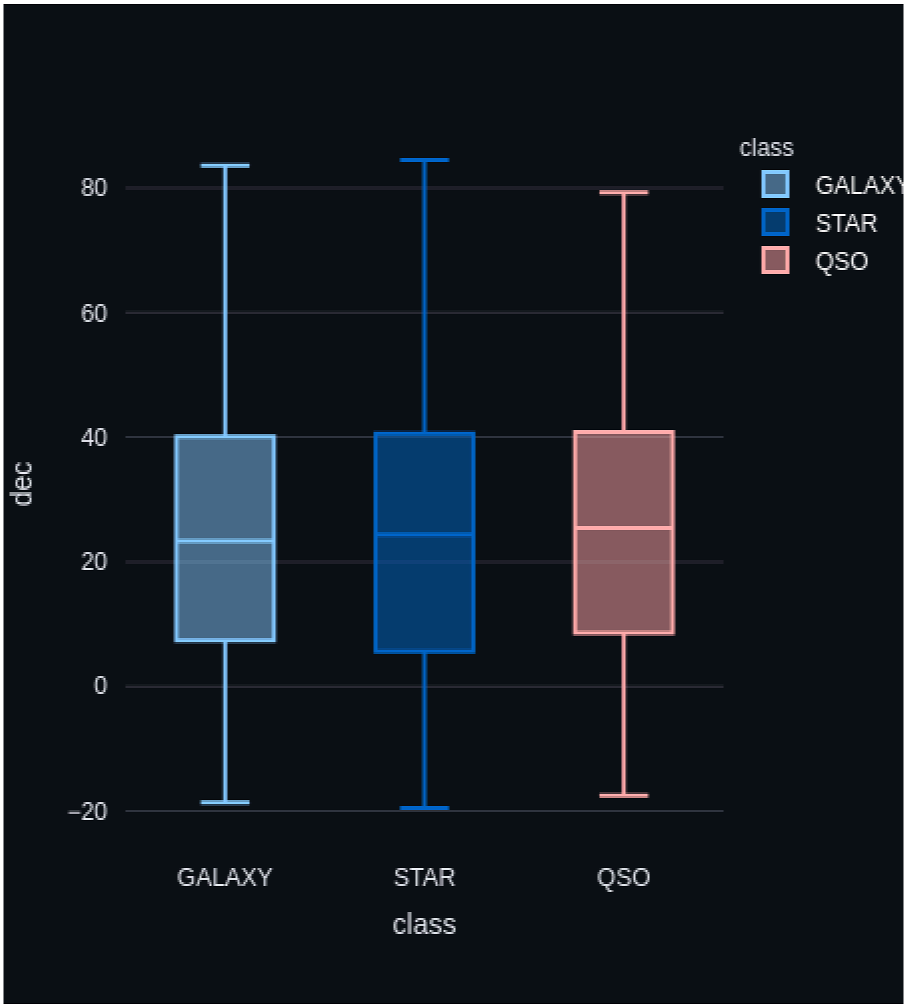
Users can select the desired data release from a dropdown menu, after which the script loads the corresponding data from the '.db' file. The script generates several types of graphs to visualize the data, such as a coordinate distribution graph that shows how stars, galaxies, and quasars are distributed based on their celestial coordinates.



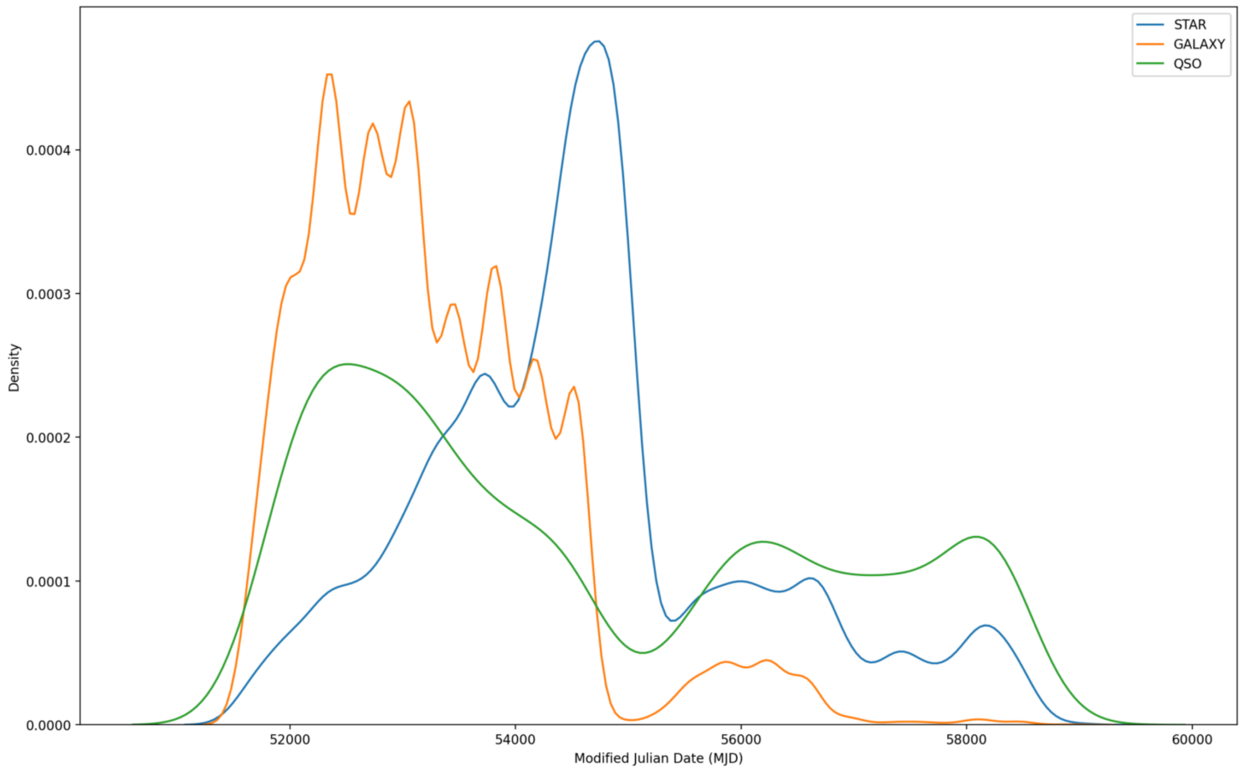
Another graph, the donut chart, visualizes the percentage distribution of these celestial objects, highlighting their relative abundance within the dataset.



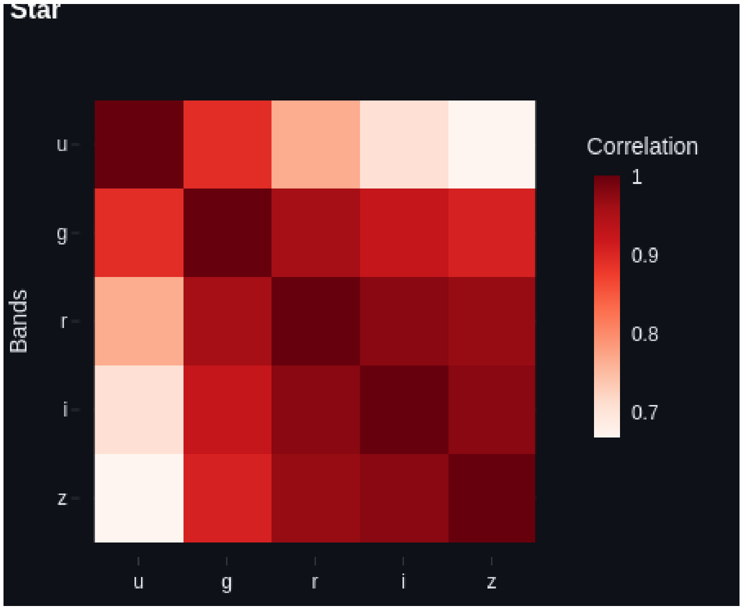
Additionally, the box plot illustrates the distribution of stars, galaxies, and quasars based on the 'dec' value, helping users understand how these objects vary with declination.

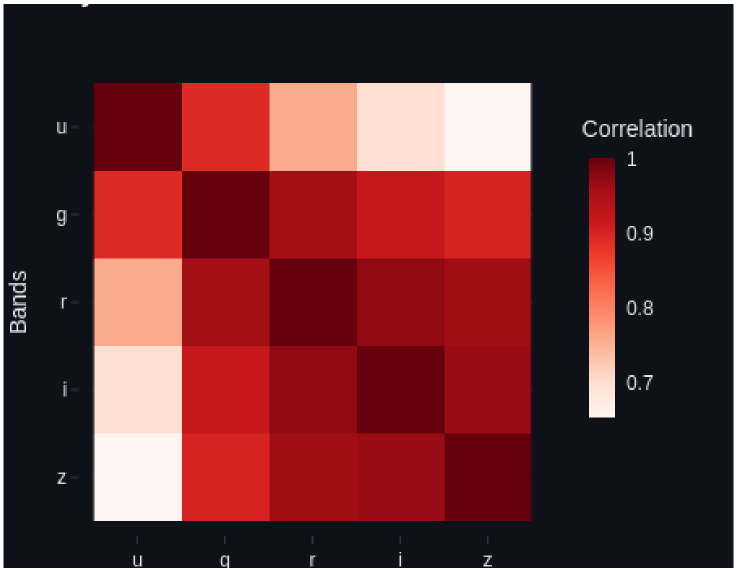


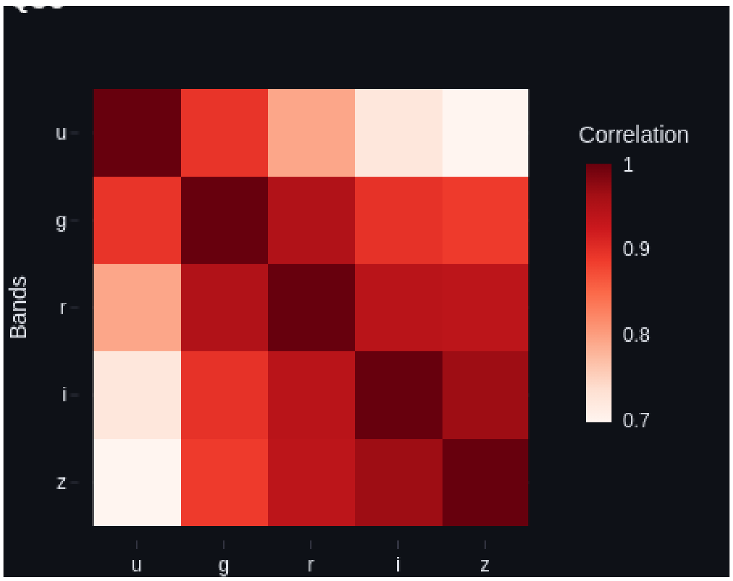
The stars, galaxies and quasars distribution based on Modified Julian Date (MJD) chart offers insights into how stars, galaxies, and quasars were discovered in different periods of time, indicating how these stars, galaxies and quasars densities differ with time. This shows the periods were how at a certain point of time a particular celestial body was discovered a lot compared to the other two.



Finally, the correlation graphs reveal the relationships between various features such as 'u', 'g', 'r', 'i', and 'z' for stars, galaxies, and quasars, enabling users to identify significant correlations within the dataset.







The implementation of these visualizations not only aids in understanding the dataset but also provides a solid foundation for further analysis and research.

5.5 Database Implementation

The database component is crucial for managing and accessing the large volumes of data collected from the SkyServer portal. The data from SDSS data releases (DR18 to DR15) was collected using SQL queries and subsequently converted into a '.db' file using a Python script. This '.db' file contains four tables, each corresponding to a different data release, and includes relevant columns such as 'dec', 'u', 'g', 'r', 'i', 'z', 'redshift', and 'class' (star, galaxy, quasar).

The data collection process involved writing SQL queries to extract the necessary information from the SkyServer portal. The data was then structured and stored in a database format that could be easily accessed and utilized by the EDA script. This approach ensured that the data was organized and ready for analysis, facilitating the generation of various visualizations and the training of the classification model.

By maintaining a structured and comprehensive database, the project ensures that users can efficiently explore and analyze the SDSS dataset, gaining valuable insights into the distribution and characteristics of celestial objects.

5.6 Classification Model Implementation

The classification model is a key component of the 'Stellar Classification' project, responsible for predicting whether input astronomical data corresponds to stars, galaxies, or quasars. The XGBoost model was chosen for its high accuracy and performance in classification tasks. The model was trained and tested on the latest SDSS data release (DR18), following extensive data preprocessing steps such as normalization and feature selection to improve its performance.

The classification model is a key component of the 'Stellar Classification' project, responsible for predicting whether input astronomical data corresponds to stars, galaxies, or quasars. The XGBoost model was chosen for its high accuracy and performance in classification tasks. This section will delve into the data preprocessing steps, the reasoning behind using XGBoost, the training process, and the evaluation of the model.

5.6.1 Data Preprocessing

Data preprocessing is a crucial step in any machine learning project as it prepares the raw data for the model, ensuring better performance and accuracy. For this project, several preprocessing steps were undertaken. Initially, the target variable 'class', which categorizes the celestial objects into stars, galaxies, and quasars, was encoded using LabelEncoder. This conversion is necessary as machine learning algorithms require numerical input. Following this, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data while retaining most of the variance present in the original features. The features 'u', 'g', 'r', 'i', and 'z' were transformed into three principal components (PCA_1, PCA_2, PCA_3). This step not only simplifies the model but also helps in mitigating multicollinearity, thus enhancing the model's performance.

Feature scaling was the next preprocessing step, achieved using MinMaxScaler to normalize the data. Scaling ensures that all features contribute equally to the model's predictions, preventing features with larger magnitudes from dominating the model. These preprocessing steps collectively ensure that the data is clean, normalized, and suitable for training the XGBoost model, setting a solid foundation for accurate and reliable predictions.

5.6.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique used to emphasize variation and capture strong patterns in a dataset. It transforms the data into a new coordinate system, reducing the number of dimensions without significant loss of information. By applying PCA, the original features 'u', 'g', 'r', 'i', and 'z' were reduced to three principal components. This reduction simplifies the dataset by reducing the number of features, making the model less complex and faster to train. Moreover, PCA retains most of the important information (variance) from the original dataset, ensuring that the model's performance is not compromised. By transforming the features, PCA also helps in reducing the multicollinearity among the input variables, leading to more stable and reliable model predictions. The application of PCA in this project effectively balances the trade-off between dimensionality reduction and information retention, contributing to a more efficient and accurate classification model.

5.6.3 Benefits of Using XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful and scalable machine learning algorithm based on gradient boosting. It is widely recognized for its efficiency and performance in classification and regression tasks. The decision to use XGBoost for this project was influenced by several factors. Firstly, XGBoost is known for its superior performance in terms of accuracy and speed, making it an ideal choice for classification tasks. The algorithm's robustness to various data inconsistencies, such as missing values and outliers, is another significant advantage, as these issues are common in real-world datasets. Additionally, XGBoost includes regularization parameters that help in preventing overfitting, ensuring that the model generalizes well to unseen data. The implementation of parallel processing in XGBoost significantly speeds up the training process, making it suitable for large datasets. Furthermore, XGBoost supports various objective functions and evaluation metrics, providing flexibility to fine-tune the model according to the specific requirements of the project. These benefits make XGBoost a powerful tool for developing an accurate and reliable classification model for astronomical data.

5.6.4 Training and Evaluation Process

The training process of the XGBoost model involves several key steps. The preprocessed data was first split into training and testing sets using a 67-33 split, ensuring that the model is trained on a substantial portion of the data while retaining a significant portion for evaluation. The XGBoost model was then instantiated with 100 estimators. The training process was timed to measure the efficiency of the model, and the model was subsequently trained on the training data (X_train, y_train), learning the patterns and relationships within the dataset.

After training, the model's performance was evaluated on the testing set (X_test). Predictions were made, and various metrics such as accuracy, precision, recall, and F1-score were calculated. The evaluation process also included timing the prediction process to assess the model's speed. The XGBoost model achieved a remarkable classification accuracy of 99.03%. The training time and prediction time were also recorded, demonstrating the model's efficiency. The high accuracy and quick prediction time underscore the effectiveness of the model in classifying celestial objects.

5.6.5 Detailed Analysis of Results

The results obtained from the XGBoost model were impressive, with a classification accuracy of 99.03%. This high accuracy indicates that the model is highly effective in distinguishing between stars, galaxies, and quasars based on the input features. The use of PCA and feature scaling contributed significantly to this performance by ensuring that the input data was well-prepared and optimized for the model. The evaluation metrics provide a comprehensive view of the model's performance: accuracy indicates the overall correctness of the model's predictions, precision measures the proportion of true positive predictions among all positive predictions, recall measures the proportion of true positive predictions among all actual positive cases, and F1-score provides a balanced measure of precision and recall, especially useful in cases where there is a class imbalance.

The high values of these metrics reflect the robustness and reliability of the XGBoost model. Additionally, the training and prediction times demonstrate the efficiency of the model, making it suitable for real-time applications where quick predictions are crucial. In conclusion, the implementation of the XGBoost model in the 'Stellar Classification' project has proven to be highly successful. The model's high accuracy, combined with efficient data preprocessing and dimensionality reduction techniques, ensures that the web application provides reliable and accurate classifications of celestial objects. This comprehensive approach to model implementation and evaluation highlights the project's potential for educational and research applications in the field of astronomy.

The training process involved splitting the dataset into training and testing sets to evaluate the model's accuracy and robustness. Hyperparameter tuning was conducted to optimize parameters such as learning rate, max depth, and the number of estimators, resulting in a model that achieved a classification accuracy of 99.03%. This high accuracy indicates the model's effectiveness in distinguishing between stars, galaxies, and quasars based on their input features.

To integrate the model with the Flask application, it was converted into a pickle file, allowing for easy loading and use for predictions whenever users input data in the prediction section. This seamless integration ensures that users receive immediate and accurate classification results, enhancing the overall functionality of the web application.

5.7 API Implementation

APIs were developed to enhance the user experience by sourcing relevant images and additional details from the SkyServer portal. An API was created to retrieve images of celestial objects based on the classification results, providing users with visual representations of the identified objects. For instance, if the classification result is a star, an image of a star from the SDSS dataset is displayed.

Another API was developed to provide additional details about the classified celestial object, such as coordinates, magnitudes, and redshift values. These details offer users a deeper understanding of the classified objects and enrich the overall user experience.

By incorporating these APIs, the web application not only provides accurate classification results but also delivers a comprehensive set of information and visual aids that enhance the user's understanding of the astronomical data.

6. Results and Discussions

6.1 Overview of Results

The results of the 'Stellar Classification' project demonstrate the effectiveness of the developed web application in providing insightful exploratory data analysis (EDA) and accurate classification of astronomical objects. The primary outcomes of the project include the successful implementation of an interactive user interface, comprehensive EDA charts, and a highly accurate classification model. The main objectives of creating a user-friendly platform for exploring the SDSS dataset and developing a robust model for classifying stars, galaxies, and quasars were achieved through meticulous implementation and thorough testing. This has resulted in a web application that meets the needs of students and new researchers interested in astronomical data.

6.2 Detailed Results

The Exploratory Data Analysis (EDA) component offers a visual representation of various aspects of the SDSS dataset, allowing users to select different data releases (DR18 to DR15) and explore the data through interactive charts and plots. The coordinate distribution graph illustrates how stars, galaxies, and quasars are spatially distributed, providing insights into their celestial coordinates. The donut chart visualizes the percentage distribution of these celestial objects, highlighting their relative abundance within the dataset. Additionally, the box plot displays the distribution of stars, galaxies, and quasars based on the 'dec' value, facilitating an understanding of how these objects vary with declination. The redshift distribution chart offers a view into how stars, galaxies, and quasars are distributed based on their redshift values, indicating their distance and movement. Finally, the correlation graphs reveal the relationships between various features such as 'u', 'g', 'r', 'i', 'z' for stars, galaxies, and quasars, allowing users to identify significant correlations within the dataset.

The XGBoost classification model, which forms a crucial part of the project, was trained on the latest SDSS data release (DR18) and achieved a remarkable classification accuracy of 99.03%. This high accuracy underscores the model's effectiveness in distinguishing between stars, galaxies, and quasars based on their input features.

Models	**Prediction accuracy for train data (%)**	**Prediction accuracy for test data (%)**	**Time consumed for training (seconds)**	**Time consumed for prediction (seconds)**
XGBoost	99.91	99.03	4.526	0.35491
Random Forest Classifier	99.90	99.02	3.025	0.11499

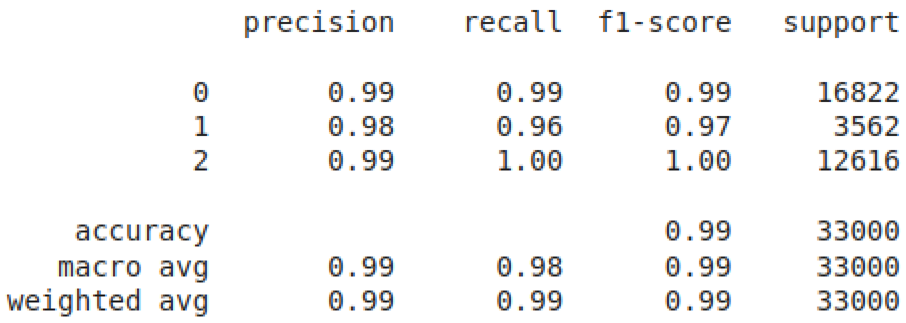
After applying cross validation on XGB and RFC, we found that they were indeed performing really well and their accuracies didn’t change much at all.

Models	**Mean of the Scores (%)**	**Standard Deviation**
XGBoost	0.9916	0.000884
Random Forest Classifier	0.9909	0.000773

Accuracy after optimizatiation/hypertuning has been applied to the XGBoost Classifier model.

Model	**Prediction accuracy WITH optimal hyperparameters (%)**
Random Forest Classifier	98.91
XGBoost	99.01

The model's performance was further evaluated using metrics such as precision, recall, and F1-score, all of which demonstrated its robustness and reliability.



```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99     16822
           1       0.98      0.96      0.97      3562
           2       0.99      1.00      1.00     12616

    accuracy                           0.99     33000
   macro avg       0.99      0.98      0.99     33000
weighted avg       0.99      0.99      0.99     33000
```

The classification results are seamlessly integrated into the web application, enabling users to input astronomical data and receive immediate predictions along with relevant images and details. This interactive feature significantly enhances the user experience and provides practical insights into the classification process.

6.3 Implications of Results

The implications of the results obtained from the 'Stellar Classification' project are substantial for both educational and research purposes. From an educational standpoint, the web application serves as a valuable tool for students and new researchers by offering a hands-on platform to explore astronomical data and learn about data analysis and classification techniques. In the realm of research, the high accuracy of the classification model signifies its potential for scientific studies where precise identification of celestial objects is essential. The project also establishes a foundation for future enhancements, such as incorporating additional features, improving the user interface, and updating the dataset to include more recent data releases. These improvements could further expand the application's utility and relevance in the field of astronomy.

6.4 Limitations

Despite the successful outcomes, the project has certain limitations that should be acknowledged. One notable limitation is the reliance on the SDSS data releases (DR18 to DR15), which may not encompass the most recent astronomical data. To address this, future work could involve updating the dataset to ensure more comprehensive coverage.

Another limitation pertains to the generalization of the model; while the XGBoost model achieved high accuracy with the current dataset, its performance may vary with different datasets. Additional testing and validation with diverse datasets are necessary to confirm its generalizability.

Lastly, the feature selection process could be expanded to include more variables, which might improve the model's performance further. Incorporating domain-specific knowledge could also provide deeper insights and enhance the classification accuracy.

7. Conclusion

7. 1 Summary of Findings

The 'Stellar Classification' project successfully achieved its objectives of developing a web application for exploratory data analysis (EDA) and classification of astronomical objects. The project culminated in the creation of an interactive web application that allows users to explore the SDSS dataset through various charts and plots, offering a comprehensive view of the data. The high-accuracy XGBoost classification model, capable of distinguishing between stars, galaxies, and quasars, demonstrated exceptional performance and reliability. This combination of EDA and classification capabilities provides a powerful tool for educational and research purposes, fulfilling the project's goals and addressing the needs of its intended audience. Several valuable lessons emerged from the implementation of the 'Stellar Classification' project. First and foremost is the importance of data quality and consistency. Ensuring that the dataset is properly preprocessed and validated is crucial for achieving accurate and reliable results. This includes steps such as normalization and feature selection, which are essential for effective model training. Another significant lesson is the critical role of model optimization. Hyperparameter tuning and iterative testing are necessary to enhance the performance of machine learning models, as evidenced by the high accuracy achieved by the XGBoost model. Additionally, designing an intuitive and interactive user interface greatly enhances the application's usability. User feedback and testing are vital for identifying areas of improvement and ensuring that the application meets the users' needs effectively.

7.2 Future Work

The 'Stellar Classification' project has laid the groundwork for further development and improvement. One potential avenue for future work is the incorporation of new data, either by updating the dataset with more recent SDSS data releases or by including additional astronomical data sources. This would enhance the comprehensiveness and accuracy of the analysis. Another area for expansion is the addition of new features to the web application. Advanced filtering options, more detailed visualizations, and the inclusion of additional classification models could provide users with a richer experience and more valuable insights. Furthermore, exploring different machine learning algorithms, ensemble methods, and deep learning techniques could improve the accuracy and robustness of the classification model, ensuring its applicability across various datasets and contexts.

7.3 Final Thoughts

The 'Stellar Classification' project represents a significant achievement in the field of astronomical data analysis. By providing an interactive platform for exploratory data analysis and a highly accurate classification model, the project offers valuable insights and tools for students, researchers, and educators. The lessons learned throughout the project and the potential for future work underscore its ongoing relevance and importance in advancing our understanding of the universe. The project's success not only highlights the effectiveness of the implemented solutions but also opens the door for continuous improvement and innovation in the field of astronomy.

8. References:

1. Johnson, H. L., Morgan, W. W. (1953) Fundamental stellar photometry for standards of spectral type on the revised system of the Yerkes spectral atlas.
2. Photometric Classification of Stars, Galaxies and Quasars in the Sloan Digital Sky Survey DR6 Using Support Vector Machines C. Elting, C. A. L. Bailer‑Jones, and K. W. Smith
3. Machine learning classification of stars, galaxies and quasars. Logistic, decision
4. Stellar Classification by Machine Learning. random forest support vector
5. Unsupervised star, galaxy, QSO classification Application of HDBSCAN
6. Classification and Feature Prediction of Star, Galaxies, Quasars, and Galaxy Morphologies Using Machine Learning. XGBoost, navies bayes
7. Machine Learning in Astronomy: A Case Study in Quasar-Star Classification.
8. Machine learning-based photometric classification of galaxies, quasars, emission-line galaxies, and stars
9. Star-Galaxy Classification of Photometric Data – A Comparative Study of Machine Learning Algorithmic Models
10. Astronomical Point Source Classification through Machine Learning
11. Photometric identification of compact galaxies, stars, and quasars using multiple neural networks
12. Deep learning Approach for Classifying, Detecting and Predicting Photometric Redshifts of Quasars
13. Classification of quasars, galaxies, and stars using Multi-Modal Deep Learning

14. Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra
15. Feature Selection Applied to Data from the Sloan Digital Sky Survey
16. Machine Learning in Astronomy: A practical overview
17. Machine learning based catalogs of quasars and galaxies for cosmological studies
18. Automated physical classification in the SDSS DR10. A catalogue of candidate Quasars.