Navigating the Career Labyrinth with SVM and XGBoost: A Comprehensive Career and Salary Prediction System

Ananya Singhal (20BCI0273), G Joshita Reddy (20BDS0178),
Keerthika Reddy (20BCE0919), and Sai Ruthvik Athota (20BDS0075)

***Abstract*—The growing number of job opportunities poses a challenging decision for students when it comes to choosing a career path. Moreover, it's crucial for them to assess if the resources they receive match their expertise and knowledge. To tackle this, we created a mechanism that suggests a suitable career path to students based on their proficiency in various fields and areas of study. By assessing their performance, we can determine the specific skills necessary for the job. Additionally, we're working on a salary estimation tool to help students gauge their future earnings based on their current expertise.**

***Index Terms*—svm, xgboost, career prediction, salary prediction**

I. Introduction

Choosing a career path can be a daunting task for students today, given the growing number of job opportunities and options available to them. While students receive guidance from various sources, including parents, friends, and educational institutions, ultimately, it is the student who must decide on their course of action. In addition to this, students need to assess if the job package offered is proportionate to their skills and expertise, as companies may underestimate their abilities and offer a sub-optimal package.

To help address these challenges, we have proposed a career prediction mechanism that recommends a suitable career path for students based on their proficiency in various domains and subjects. The approach focuses on the ability to forecast the careers of applicants, with a particular emphasis on engineering and computer science. These fields have seen rapid changes and advancements due to emerging technologies, and the system can help identify the specific job functions that best suit the student's skills.

The system works by assessing the student's performance in various areas of study and then providing a recommended career path based on their expertise. We have also introduced a salary prediction tool that can help students estimate their future earnings based on their current level of expertise.

Overall, our study offers a promising solution to the challenges students face when choosing a career path. By leveraging the latest technology and approaches, the proposed system can help students make informed decisions about their future, maximizing their potential and ensuring they receive an optimum package for their skills and expertise.

II. Literature Survey

Wang, Yuan, et.al., [1] aimed to predict college students' career choices using eXtreme Gradient Boosting (XGBoost), a machine learning technique, and to analyze the importance of individual features using SHAP in their study. The study used a dataset of 18,000 graduates from a college in Beijing and found that XGBoost was able to predict students' career choices with high accuracy. The study also explored the interaction of features among four different choices of students and found that educational features, especially differences in grade point average (GPA) during college, had a relatively larger impact on the final choice of career. Specifically, the study found that the score of the college entrance examination plays an important role in predicting graduates' career choice, with students with high scores tending to choose further education in China. The total amount of scholarships was also found to have an important impact on the final academic direction, with students with higher amounts tending to choose domestic postgraduate education rather than employment. Additionally, GPA in the first semester was found to have a vital impact on students' future choices, with most students with low GPA in the first semester not considering further education or studying abroad. The study's results can be useful in the planning, design, and implementation of higher educational institutions' events.

Nie, Min, et.al., [2] proposed a model called Approach Cluster Centers Based On XGBOOST (ACCBOX) to predict college students' career choices based on their behavioral data. The model aims to use the most remarkable characteristics of classes reflected by the main samples of a category to forecast students' career choices, and it uses a prototypical cluster center generation approach to leverage a priori information from each college. The results of experiments conducted on 13 million behavioral data of over four thousand students demonstrate that the ACCBOX model outperforms existing state-of-the-art techniques in predicting students' career choices. The authors suggest future directions for research, such as discovering cluster

centers in a more precise way, using multimodal data, and improving the model to provide career planning advice. The study concludes that using students' behavioral data can be a promising approach to predicting their career choices.

Song, Q. Chelsea, et.al., [3] aimed to improve the accuracy of interest inventory-based career choice prediction through the application of machine learning (ML) in their study. The researchers compared the accuracy of a traditional interest profile method to a new ML-augmented method in predicting occupational membership and vocational aspirations. Results showed that the ML-augmented method yielded higher predictive accuracy than the traditional profile method in predicting both types of career choices. However, the profile method outperformed the ML-augmented method in predicting career choices in occupational areas with low base rates. The study suggests that ML can enhance the predictive validity of vocational interests and provides important contributions to theory and practice, including employee development, recruitment, job placement, and retention. The researchers provide access to their analysis code and detailed instructions to facilitate the implementation of the new ML-augmented method in both research and applied contexts. Overall, the study highlights the potential of ML to improve the accuracy of interest inventory-based career choice prediction and calls for future research to explore the optimal application of ML methods for predicting career choices.

The process of selecting the right career path can be challenging, especially for students who may lack knowledge and maturity in the field. Choosing the wrong career path can result in job dissatisfaction or a lack of knowledge in the subject area. To help students make informed decisions, B. Harsha, et.al., [4] in their study proposed a machine learning-based approach to predicting the most suitable career path based on a survey of the student's inherent talents and attributes. The study utilizes three machine learning classifiers to train and test data, with the Support Vector Machine (SVM) algorithm providing the highest accuracy at 90.3%. The study suggests developing more powerful web applications that collect student parameters through various assessments and exams to provide a more accurate prediction of the student's appropriate profession. This could be achieved by creating technical, analytical, logical, memory, psychometric, and general cognitive, interest, and skill-based assessments. The system could also cover multiple disciplines and provide technical, logical, memory-based, and skill-based testing to achieve better results. The proposed concept encourages students to select a career based on their inherent skills and attributes, rather than solely on interests. Students with excellent skills and qualities in a particular subject can choose that subject, and those lacking skills can remove the subject and select a different one. The study provides important recommendations for refining the model, such as creating websites and mobile applications to help students understand their strengths and weaknesses, completing quizzes in their area of interest, and analyzing their knowledge of the field. Overall, their study highlights the importance of choosing the appropriate career path and proposes a machine learning-based approach to predicting the most suitable career path for students based on their inherent talents and attributes. The proposed system could be developed further to cover multiple disciplines and provide technical, logical, memory-based, and skill-based testing to achieve better results.

Al-Dossari, H., et.al., [5] proposed a recommendation system called CareerRec, which utilizes machine learning algorithms to help IT graduates select a career path based on their skills. The system was trained and tested on a dataset of 2255 employees in the IT sector in Saudi Arabia, and the XGBoost algorithm was found to be the most accurate in predicting the best-suited career path among the three classes. However, the low accuracy achieved in the experiments can be attributed to the small dataset and the method of data collection, which involved directly asking employees about their skill levels. The paper suggests that collecting more data about IT employees in the Saudi market and using alternative approaches to measure employees' skills, such as obtaining technical skills from academic records or acquiring skills indirectly from colleagues and supervisors, can improve the system's performance and reliability. The proposed system not only benefits IT graduates but can also assist employers in determining an applicant's suitability for a specific job position. The paper concludes by highlighting the challenges of collecting and ensuring the credibility of data and suggests exploring more algorithms, such as deep learning models, to improve the accuracy of the system.

As the world becomes increasingly reliant on the internet, cybersecurity has become a crucial factor to prevent malicious cyber threats. This survey reviews the use of machine learning (ML) techniques to detect potential cybersecurity risks, including fraud, intrusion, spam, and malware. Shaukat, Kamran, et.al., [6] provide a comprehensive comparison of commonly used ML models based on performance and time complexity. They also discuss the challenges and limitations of using ML techniques in cybersecurity. This study focuses on the application of ML models on both the attacker and defender sides. The former uses ML to find new ways to evade security systems, while the latter uses it to prevent illegal penetration and unauthorized access. The authors compare six ML models, including random forest, support vector machine, naive Bayes, decision tree, artificial neural network, and deep belief network. They also compare the models on sub-domains of cyber threats such as anomaly-based, signature-based, and hybrid-based for intrusion detection, static detection, dynamic detection, and hybrid detection for malware detection, and classification of spam mediums like images, videos, emails, SMS, and calls. The authors have provided a detailed discussion on the performance of ML models and the challenges of using these models in

cybersecurity. Despite being a pivotal element in the cyber world, cybersecurity has its limitations and constraints. The authors conclude that while ML models have become an integral part of modern cybersecurity, their application is limited by challenges such as data quality, privacy concerns, and the adversarial nature of cyber threats.

Student dropout is a major issue in higher education that incurs economic and social costs. Maldonado, Sebastián, et.al., [7] proposed a data-driven approach for predicting student dropout by adopting a profit-driven perspective. A novel performance measure is designed that quantifies the net savings of a retention campaign, enabling the identification and selection of students to optimally allocate resources for preventing dropout and maximizing resulting savings. Experiments using data from three bachelor's programs of a higher education institution were conducted, and the proposed metric yielded tangible savings for the institution. The presented approach and experimental results highlight pathways to design tailored student retention programs. The paper argues that the performance of classification methods should not be assessed solely based on statistical measures but also consider the benefits and costs of the decision-making process. By adopting a profit-driven perspective, the efficiency and effectiveness of investments in preventing student dropout can be improved. The proposed metric enables a better choice of the prediction model and classification threshold, leading to tangible savings for the institution.

B. M. D. E. Bannaka, et.al., [8] presented a study based on the IT industry in Sri Lanka having a large workforce with many job opportunities available for fresh graduates. However, employees change careers frequently, and employers struggle to retain employees. This research aims to develop a career mentoring system that predicts career suitability, progression, and attrition to help IT employees achieve their career goals. Data was collected from IT employees, and several classification algorithms were implemented to predict career suitability, initial salary, career progression, salary progression, professional courses, and employee attrition. XGBoost resulted in higher accuracies for career suitability, initial salary, career progression, and salary progression, while Random Forest had higher accuracies for professional courses and employee attrition. The goal of this research is to guide IT graduates and employees towards better performance and assist them in embracing responsibilities throughout their career life.

Wang, Ping Liao, et.al., [9] in their paper analyzed the determinants of the high salary of graduates from a financial and economic university in 2020 and establishes a logistic regression model using R. The study shows that academic qualifications, professional disciplines, employment regions, employment industries, the nature of employment units, gender, and whether they have served as student cadres have a significant impact on whether graduates can get "high salaries." The main factors affecting the starting salary of graduates are the accumulation of human capital and social capital, but the segmentation of the labor market is also the main reason affecting the starting salary of graduates. The paper uses five machine learning methods to predict whether graduates can get a high starting salary, and the XGBoost model is the best. The paper suggests that college students should improve their human capital, find employment in innovative enterprises in new first-tier cities, and improve their self-efficacy. The paper has practical guiding significance for college students and their career development.

Employee recruitment and salary are crucial components of a business's success. A salary prediction system has been proposed to help college students understand their expected salary after graduation and identify the necessary skills to achieve their professional goals. Pansare, Dr.Jayshree Lunawa, et.al., [10] proposed a system that uses data mining techniques to compare the profiles of students and graduates, providing a more accurate prediction. The study demonstrates that data mining techniques perform well and has conducted experiments on a student dataset using 10-fold cross-validation. The proposed system not only benefits the students by increasing their motivation but also provides better assistance to higher education institutions. The system's goal is to improve the smooth conduction of companies' production and competitiveness, making it a significant activity for any business.

III. Method

A. Student career prediction ML model

- The dataset for the ML model is chosen from the Kaggle dataset for career prediction.
- Student career prediction is identified as a classification problem, as there are different classes of potential job roles that will be predicted as an output.
- Data preprocessing of the dataset is to be performed by-Identifying and handling missing values by substituting with mean, most frequent values, etc. Encode the categorical data. Split the dataset. Perform feature scaling to ensure that all the values are in the same range.
- There are 5 classification algorithms in ML. Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Tree, and Support Vector Machines. Each of the algorithms is to be performed to determine the algorithm that provides the best accuracy.
- The algorithm chosen from the previous step will be chosen to test sample data and predict a student's career.

B. Student salary prediction ML model

- The dataset for the ML model is chosen from the Kaggle dataset for salary prediction.
- Student salary prediction is identified as a regression problem, as each student will get a unique salary based on the user input.
- Data preprocessing of the dataset is to be performed by-Identifying and handling missing values by substituting

with mean, most frequent values, etc. Encode the categorical data. Split the dataset. Perform feature scaling to ensure that all the values are in the same range.
- There are 4 regression algorithms that will be used. Linear Regression, Logistic Regression, Polynomial Regression, and Bayesian Linear Regression. Each of the algorithms is to be performed to determine the algorithm that provides the best accuracy.
- The algorithm chosen from the previous step will be chosen to test sample data and predict a student's expected salary.

C. Website Frontend

- The website front-end will be coded in HTML, and CSS by following the design made in Figma.
- EJS(Embedded Javascript) will be used to display the dynamic results like salary and career based on the form input.

D. Website Backend

- The website backend is coded in node-js and express-js.
- To implement the backend we need to require all of the libraries.
- Define the required GET, POST, PUT, and DELETE functions based on the requirements and call the ML model APIs.
- Define the port on which the server listens.

E. Integration

- The previously written ML models need to be converted in the form of a Python object into a character stream using pickling.
- We need to build an API using Flask. We hosted the API on pythonanywhere platform.
- The output after hosting the API will be 2 links of each of the ML models which can be integrated with our website backend.
- The links are put in the required GET routes to access the prediction models.
- The functionality of the website is now complete and the website can be hosted on the Cyclic platform to make it accessible to everyone.

F. Testing

- After the completion of the website testing needs to be performed to ensure that the platform works as per the requirements
- The different testing methodologies that will be employed are:
  - Functionality testing: This step ensures that the functionalities of a web application are properly functioning or not.
  - Usability testing: Usability testing involves several parameters such as UI design, speed, navigability, content readability, and accessibility.
  - Interface testing: This testing method ensures that the three main components of a web application which are the web server, web browser, and database are running harmoniously.
  - Compatibility testing: This testing methodology ensures that a particular web application is compatible with all browsers.
  - Performance testing: The web application is tested in terms of how better it can perform under stress conditions and heavy load.
  - Security testing: To ensure the security of the data provided by users is safely stored.

IV. Result and Discussions

Two discrete models were trained, namely a classification model and a regression model. Both models were utilised to formulate predictions. The study employed a dataset sourced from Kaggle, containing career prediction data, to construct a classification model. The efficacy of various classification algorithms, including logistic regression, decision tree, random forest, and support vector classifier, was subsequently evaluated. The support vector classifier demonstrated superior performance on the testing set, attaining an accuracy of 86 and an F1 score of 80. The support vector classifier was identified as the optimal model. This methodology can accurately predict the professional trajectories of students by analysing their current skill sets.

The study aimed to evaluate the effectiveness of various regression techniques, namely linear regression, polynomial regression, and XGBoost, through the utilisation of a salary dataset in constructing the regression model. The XGBoost model exhibited superior performance compared to other models on the testing set, as evidenced by its R-squared value of 0.85. This methodology enables a job seeker to acquire a precise estimation of the potential income they could receive. Both models exhibited commendable performance on their respective datasets and possess the potential to be valuable in forecasting forthcoming career trajectories and remunerations. The website's users possess the capability to input relevant characteristics and obtain anticipated results derived from the model, thereby aiding in the process of making well-informed decisions.

V. Conclusion

On this website, we have presented a classification model predicting the student's career and a regression model that predicts the expected salary that a student should expect. The models were trained and evaluated using accuracy and R squared metrics. Our results indicate that the model can precisely predict the salary and the career of the student based on the input features. Students can fill out the forms and get career or salary predictions, which can help them make informed decisions about their professional life. The website is designed to provide maximum information to students which will help them analyze all aspects in a better way. As part

of our future works, more algorithms, such as deep learning models, can be considered to increase the accuracy of the suggested system. Further research is required to determine how better career choice predictions relate to job outcomes, particularly those measured over time.

REFERENCES

[1] Wang, Yuan, Liping Yang, Jun Wu, Zisheng Song, and Li Shi. 2022. "Mining Campus Big Data: Prediction of Career Choice Using Interpretable Machine Learning Method" Mathematics 10, no. 8: 1289. doi: https://doi.org/10.3390/math10081289

[2] Nie, Min, Zhaohui Xiong, Ruiyang Zhong, Wei Deng, and Guowu Yang. 2020. "Career Choice Prediction Based on Campus Big Data—Mining the Potential Behavior of College Students" Applied Sciences 10, no. 8: 2841. doi: https://doi.org/10.3390/app10082841

[3] Song, Q. Chelsea, Hyun Joo Shin, Chen Tang, Alexis Hanna, and Tara Behrend. "Investigating machine learning's capacity to enhance the prediction of career choices." Personnel Psychology (2022). doi: https://doi.org/10.1111/peps.12529

[4] B. Harsha, N. Sravanthi, N. Sankeerthana and M. Suneetha, "Career Choice Using Machine Learning Algorithms," 2022 International Conference on Inventive Computation Technologies (ICICT), Nepal, 2022, pp. 171-176. doi: https://doi.org/10.1109/ICICT54344.2022.9850697

[5] Al-Dossari, H., Nughaymish, F. A., Al-Qahtani, Z., Alkahlifah, M. and Alqahtani, A. (2020) "A Machine Learning Approach to Career Path Choice for Information Technology Graduates", Engineering, Technology Applied Science Research. Greece, 10(6), pp. 6589–6596. doi: https://doi.org/10.48084/etasr.3821

[6] Shaukat, Kamran, Suhuai Luo, Vijay Varadharajan, Ibrahim A. Hameed, Shan Chen, Dongxi Liu, and Jiaming Li. 2020. "Performance Comparison and Current Challenges of Using Machine Learning Techniques in Cybersecurity" Energies 13, no. 10: 2509. doi: https://doi.org/10.3390/en13102509

[7] Maldonado, Sebastián, Jaime Miranda, Diego Olaya, Jonathan Vásquez, and Wouter Verbeke. "Redefining profit metrics for boosting student retention in higher education." Decision support systems 143 (2021): 113493. doi: https://doi.org/10.1016/j.dss.2021.113493

[8] B. M. D. E. Bannaka, D. M. H. S. G. Dhanasekara, M. K. Sheena, A. Karunasena and N. Pemadasa, "Machine learning approach for predicting career suitability, career progression and attrition of IT graduates," 2021 21st International Conference on Advances in ICT for Emerging Regions (ICter), Colombo, Sri Lanka, 2021, pp. 42-48. doi: https://doi.org/10.1109/ICter53630.2021.9774825

[9] M. Y. Arafath, M. Saifuzzaman, S. Ahmed and S. A. Hossain, "Predicting Career Using Data Mining," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, 2018, pp. 889-894. doi: https://doi.org/10.1109/GUCON.2018.8674995

[10] Roy, K. Roopkanth, K. Teja, V. Bhavana, V. Priyanka, J.. (2018). Student Career Prediction Using Advanced Machine Learning Techniques. International Journal of Engineering and Technology(UAE). 7. 26-29. doi: https://doi.org/10.14419/ijet.v7i2.20.11738

[11] H, Manoj A, Sufiyan L, Vinay P, Abhinav Reddy, Veerendra. (2022). STUDENT CAREER GUIDANCE. International Research Journal of Computer Science. 9. 312-315. doi: https://doi.org/10.26562/irjcs.2022.v0908.30

[12] S. Vignesh, C. Shivani Priyanka, H. Shree Manju and K. Mythili, "An Intelligent Career Guidance System using Machine Learning," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 987-990, doi: https://doi.org/10.1109/ICACCS51430.2021.9441978

[13] Qutub, Aseel, Asmaa Al-Mehmadi, Munirah Al-Hssan, Ruyan Aljohani, and Hanan S. Alghamdi. "Prediction of employee attrition using machine learning and ensemble methods." Int. J. Mach. Learn. Comput 11, no. 2 (2021): 110-114. doi: https://doi.org/10.18178/ijmlc.2021.11.2.1022

[14] Parate, Poonam Barapatre, Mukesh Kalode, Harish Harkut, Shruti Kamdi, Shivam Alne, Arpit. (2021). Machine Learning base Student Future Orientation and Recommendation System. International Journal of Advanced Research in Science, Communication and Technology. 482-487. doi: https://doi.org/10.48175/IJARSCT-1049

[15] Wang, Ping Liao, Wensheng Zhao, Zhongping Miu, Feng. (2022). Prediction of Factors Influencing the Starting Salary of College Graduates Based on Machine Learning. Wireless Communications and Mobile Computing. 2022. 1-14. doi: https://doi.org/10.1155/2022/7845545

[16] Saeed, Ashty Abdullah, Pavel Tahir, Avin. (2023). Salary Prediction for Computer Engineering Positions in India. Journal of Applied Science and Technology Trends. 4. 13-18. doi: https://doi.org/10.38094/jastt401140

[17] Pansare, Dr.Jayshree Lunawa, Sakshi Shivatare, Janhavi Oswal, Jhanavi Mehta, Krupa. (2022). SALARY ESTIMATOR: A LITERATURE REVIEW. International Research Journal of Computer Science. 9. 101-105. doi: https://doi.org/10.26562/irjcs.2022.v0905.001

[18] Kolhe, Hrugved Chaturvedi, Ruchi Chandore, Shruti Sakarkar, Gopal Sharma, Gopal. (2023). Career Path Prediction System Using Supervised Learning Based on Users' Profile. doi: https://doi.org/10.1007/978-981-19-7346-850

[19] Maaliw III, Renato Quing, Karen Anne Lagman, Ace Ballera, Melvin Ugalde, Bernard Ligayo, Michael Angelo. (2022). Employability Prediction of Engineering Graduates Using Ensemble Classification Modeling. doi: https://doi.org/10.1109/CCWC54503.2022.9720783

[20] Ge, Chunmian Kankanhalli, Atreyi Huang, Ke-Wei. (2015). Investigating the Determinants of Starting Salary of IT Graduates. ACM SIGMIS Database. 46. 9-25. doi: https://doi.org/10.1145/2843824.2843826