Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 4 Issue 2 (2024)

The Impact of Machine Learning on Enhancing Diversity and Inclusion through Advanced Recommence Screening Techniques

Sandip Bhattacharjee,
HOD and Assistant Professor, Department of Multimedia, Brainware University, Kolkata, E-Mail: iamsandipin2007@gmail.com

Dr. R Naveenkumar,
Associate Professor, Department of Computer Science and Engineering, Chandigarh Group of Colleges, Jhanjeri, Mohali, E-Mail: naveen.j3390@cgc.ac.in

Rahul Singha,
Assistant Professor, Department of Multimedia, Brainware University, Kolkata, E-Mail: rsingha907@gmail.com

Somnath Mullick,
Assistant Professor, Department of Multimedia, Brainware University, Kolkata, E-Mail: somnath.office8@gmail.com

Rubi Sarkar,
Assistant Professor, Department of Computer Science and Engineering, Chandigarh Group of Colleges, Jhanjeri, Mohali, E-Mail: rubi.j3387@cgc.ac.in

Abstract:

In the era of big data, a plethora of data is accumulating, spanning a wide range of domains. Among them, the realm of Data Science unfolds as individuals armed with a potent arsenal of data analysis, statistical modeling, and machine learning techniques embark on a quest to unravel the mysteries hidden within vast troves of data. Embedded within are myriad coding paradigms, architectural blueprints, and optimization strategies, offering aspiring developers a fertile ground for honing their craft, troubleshooting challenges, and fostering innovation within the Java ecosystem. In addition, the enigmatic world of Data Scientists unfolds, encompassing a diverse array of datasets spanning domains as varied as finance, healthcare, marketing, and beyond. Curated datasets preserve the rich tapestry of human creativity and serve as catalysts for artistic innovation and expression, empowering individuals to push the boundaries of their creativity. Through meticulous analysis of such datasets, individuals glean invaluable insights into prevailing trends, user inclinations, and design best practices, thus empowering them to craft compelling digital experiences that resonate with audiences at a profound level. The vital function of Human Resources (HR) unfolds in parallel as organizations seek to cultivate vibrant, inclusive workplaces where talent flourishes and organizational objectives thrive. While categorizing resumes into specific roles offers numerous advantages in streamlining recruitment processes and identifying qualified candidates, it's crucial to acknowledge and address the potential drawbacks.

Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 4 Issue 2 (2024)

Keywords: Kappa statistics, ZeroR, Rule OneR, Rule PART, and Matthews Correlation Coefficient

Introduction:

In the fast-paced world of job hunting, your resume acts as your ambassador, speaking on your behalf to potential employers. It's not just a piece of paper; it's your ticket to new opportunities. However, with the job market becoming increasingly competitive, standing out from the crowd is no easy feat. Enter the Resume Screening model, a game-changer in the hiring process. Think of it as your assistant, equipped with cutting-edge technology like Machine Learning and Natural Language Processing. Its job? To sift through the mountain of resumes flooding HR departments and connect the dots between candidates and job openings [1]. By analyzing your resume with a keen eye, the model identifies your strengths, experiences, and unique talents. It then matches these with the requirements of various job postings, effectively streamlining the hiring process [2]. This not only saves valuable time and resources for companies but also ensures that you, as a candidate, are considered for roles that align with your skills and aspirations. So, while you polish up your resume to showcase your best self, rest assured that the Resume Screening model is working tirelessly in the background, opening doors to your next career move. With its help, finding the perfect job fit becomes less of a daunting task and more of an exciting opportunity for growth and success [3].

Data Warehouse is Subject-Oriented

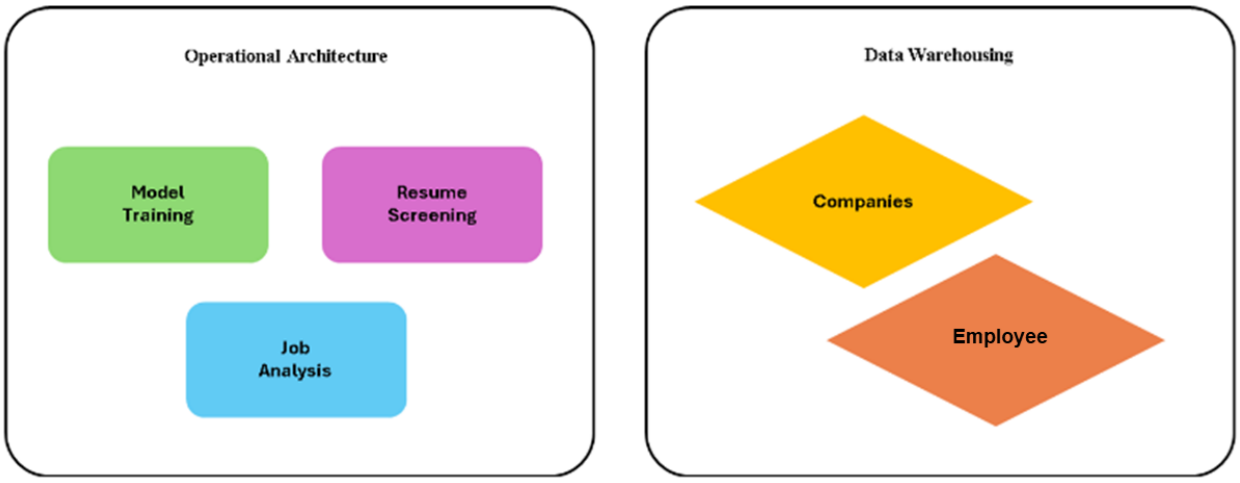

Fig 1: Data Warehouse is Subject-Oriented

In the world of Data Warehousing, we focus on organizing information around specific topics, like companies and candidates [4]. Let's look at how this works simply: Imagine a big company receiving lots of resumes from job seekers. They use a system to help them go through all these resumes quickly and find the best matches for their job openings [5].

- **Model Training:** First, they teach their system to recognize important things in resumes, like skills and experiences. It's like teaching a computer to understand what makes a good candidate [6].

- **Resume Screening:** Once the system is trained, it starts screening resumes. It reads through each one and picks out the most relevant information. This helps the company see which candidates might be a good fit for their jobs.
- **Job Analysis:** After screening, the company looks at the results and matches candidates to suitable jobs. They analyze each candidate's skills and experiences to find the best fit.

For candidates, understanding this process can be helpful. It's like knowing the steps of a game – you can play better if you know how it works [7]. By knowing how companies screen resumes, candidates can make sure their resumes highlight the right skills and experiences. Overall, this system helps companies find the right people for their jobs quickly and helps candidates understand what companies are looking for. It's like a matchmaking service for jobs and candidates!

Data source:

In the ever-evolving landscape of data analysis, the pursuit of the perfect dataset is akin to embarking on a captivating journey, traversing through diverse terrains across continents and industries [8]. From the early epochs of data scarcity to the contemporary era of abundance, the trajectory of data science has been characterized by a myriad of challenges and opportunities. In this narrative, repositories like Kaggle emerge as guiding lights, offering a rich repository of meticulously curated datasets and nurturing collaborative communities where insights are exchanged, and innovative solutions take root [9]. Reflecting on this evolutionary trajectory, one can't help but marvel at the profound transformation in opportunities afforded to data scientists. In the yesteryears, the scarcity of resources and data presented formidable obstacles to those venturing into the realm of data science. However, as the digital age unfurled its wings, a paradigm shift ensued, with enterprises embracing machine learning and data-driven approaches with unbridled enthusiasm. This monumental shift not only underscored the paramount importance of high-quality datasets but also galvanized concerted efforts to amass data tailored to specific problem domains.In the narrative of data exploration, personal projects serve as crucibles wherein aspiring data scientists refine their skills and put their abilities to the test. Consider, for instance, a recent endeavor centered around resume analysis, where a dataset sourced from Kaggle emerged as the canvas for exploration [10]. Curated by eminent contributors like GAURAV DUTTA, this dataset unfurled two pivotal columns: 'Category' and 'Resume'. Within the confines of these columns lies the potential to predict job classifications solely based on the content of resumes—an eloquent testimony to the transformative power of machine learning in the realm of talent acquisition.Venturing into specialized domains like Web Design, Java Development, or Data Science, one unravels distinct realms of expertise and creativity. In the realm of Web Design, artisans blend aesthetics with technical prowess, crafting digital experiences that captivate and engage. Similarly, Java Developers navigate the intricate landscape of software engineering, wielding their mastery to build robust applications that stand the test of time. Meanwhile, Data Scientists unravel the mysteries hidden within vast troves of data, wielding algorithms and tools to distill actionable insights from the chaos.Delving deeper into specialized domains such as Web Design, Java Development, or Data Science, one unravels distinct realms of expertise and creativity, each offering its own unique challenges and opportunities. In the realm of Web Design, artisans merge aesthetics with technical prowess, fashioning digital experiences that not only captivate but also engage users. Armed with a

discerning eye for design aesthetics and adeptness in HTML, CSS, JavaScript, and a myriad of design tools like Adobe Photoshop and Sketch, these creatives weave together digital marvels that leave lasting impressions on audiences. A curated dataset tailored to this domain serves as a treasure trove, presenting a diverse tapestry of web design projects replete with varying styles, interfaces, and interactive elements. Through meticulous analysis of such datasets, aspiring web designers glean invaluable insights into prevailing trends, user inclinations, and design best practices, thus empowering them to craft compelling digital experiences that resonate with audiences on a profound level. In the realm of Java Development, seasoned professionals harness their mastery in crafting robust and scalable software applications using the versatile Java programming language. Proficient in a plethora of Java frameworks and libraries, and well-versed in agile methodologies like Scrum and Kanban, these developers navigate the intricate landscape of software engineering with finesse [11]. A meticulously curated dataset for Java Development beckons with a cornucopia of projects, spanning from enterprise-grade applications to nimble web solutions and mobile apps. Embedded within are myriad coding paradigms, architectural blueprints, and optimization strategies, offering aspiring developers a fertile ground for honing their craft, troubleshooting challenges, and fostering innovation within the Java ecosystem.Meanwhile, the enigmatic world of Data Science unfolds as individuals armed with a potent arsenal of data analysis, statistical modeling, and machine learning techniques embark on a quest to unravel the mysteries concealed within vast troves of data [12]. Proficient in programming languages like Python and R, and wielding tools such as Pandas and NumPy with aplomb, these modern-day alchemists transmute raw data into actionable insights that fuel informed decision-making. A bespoke dataset tailored to the realm of Data Science serves as a veritable goldmine, encompassing a diverse array of datasets spanning domains as varied as finance, healthcare, marketing, and beyond. Within these datasets lie structured, semi-structured, and unstructured data, each presenting its own set of challenges, from data cleansing to feature engineering and model selection [13]. Through meticulous exploration of these datasets, aspiring Data Scientists refine their analytical acumen, experiment with cutting-edge machine learning algorithms, and craft predictive models that illuminate pathways to tangible solutions for real-world problems.Meanwhile, the enigmatic world of Data Science unfolds as individuals equipped with a potent arsenal of data analysis, statistical modeling, and machine learning techniques embark on a quest to unravel the mysteries hidden within vast troves of data. Proficient in programming languages like Python and R, and wielding tools such as Pandas and NumPy with aplomb, these modern-day alchemists transmute raw data into actionable insights that fuel informed decision-making. A bespoke dataset tailored to the realm of Data Science serves as a veritable goldmine, encompassing a diverse array of datasets spanning domains as varied as finance, healthcare, marketing, and beyond. Within these datasets lie structured, semi-structured, and unstructured data, each presenting its own set of challenges, from data cleansing to feature engineering and model selection. Through meticulous exploration of these datasets, aspiring Data Scientists refine their analytical acumen, experiment with cutting-edge machine learning algorithms, and craft predictive models that illuminate pathways to tangible solutions for real-world problems.In parallel, the vital function of Human Resources (HR) unfolds as organizations seek to cultivate vibrant, inclusive workplaces where talent flourishes and organizational objectives thrive. The role of HR professionals extends far beyond merely staffing positions; it encompasses the holistic management of human capital, encompassing recruitment, training, employee relations, performance

management, and organizational development.Successfully categorizing resumes into the realm of HR professionals empowers recruiters to discern candidates proficient in the delicate art of talent acquisition, employee relations, and organizational development. With an astute eye for HR processes like recruitment, onboarding, and performance management, these individuals navigate the intricate tapestry of human interactions with finesse. By delving into resumes, recruiters evaluate candidates' interpersonal prowess, communication finesse, and conflict resolution techniques, vital for nurturing a harmonious work environment and resolving workplace discord.Insights derived from resume analysis not only inform targeted recruitment strategies but also enable organizations to develop tailored approaches for talent development and retention. By identifying candidates whose skills and experiences align with the organization's culture and values, recruiters can curate a workforce that is not only highly competent but also deeply invested in the company's mission and vision.Furthermore, if a candidate exhibits multifaceted capabilities beyond traditional HR domains, such as data analysis or project management, recruiters may explore avenues like HR analytics, workforce planning, or strategic HR consulting. In leveraging the diverse skill set of such candidates, organizations can drive innovation and excellence in HR practices, thereby enhancing overall organizational performance and competitiveness in the marketplace.Simultaneously, the realm of legal advocacy beckons as Advocates stand as stalwarts in the defense of justice, wielding their expertise in legal advocacy, litigation, and astute legal research. Emboldened by a wealth of experience and fortified by a formidable arsenal of legal acumen, these defenders of justice navigate the labyrinthine corridors of law with dexterity, serving as beacons of hope for those seeking redress and fairness within the legal system.A meticulously curated dataset tailored for Advocates offers a rich tapestry of legal expertise, brimming with insights into educational backgrounds, professional trajectories, and specialized domains within the legal sphere. Delving into resumes, recruiters evaluate candidates' prowess in legal writing, oral advocacy, and negotiation skills—essential attributes for championing clients' interests in legal arenas. Through this process, recruiters glean valuable insights into a candidate's ability to navigate complex legal landscapes, analyze intricate legal issues, and provide strategic counsel to clients embroiled in legal disputes.Moreover, if a candidate showcases additional proficiencies such as public speaking, policy analysis, or dispute resolution, recruiters may explore avenues beyond traditional legal practice. This expansion of focus opens doors to roles in legal consulting, corporate governance, or regulatory compliance, where diverse skill sets can be leveraged to surmount multifaceted legal challenges and advance organizational objectives.In essence, the recruitment of skilled advocates is not merely about filling positions; it is about assembling a team of legal professionals who embody integrity, empathy, and unwavering commitment to justice. By harnessing the insights derived from resume analysis, organizations can attract and retain advocates who not only possess the requisite legal expertise but also demonstrate the versatility and adaptability needed to thrive in an ever-evolving legal landscape.In the realm of Arts, individuals harness their creativity and expression to bring forth works that inspire, provoke thought, and evoke emotions, thereby enriching the human experience and shaping cultural narratives. Whether through visual arts, performing arts, or literary endeavors, artists imbue their creations with a unique blend of imagination, skill, and personal perspective, contributing to the diverse tapestry of human creativity.A curated dataset tailored to the Arts domain offers a rich tapestry of artistic endeavors, encompassing a myriad of mediums, genres, and styles. Within such datasets lie a treasure trove of paintings, sculptures, music compositions, theatrical performances, literary works,

and more, each reflecting the artist's individuality and creative vision. By delving into these datasets, aspiring artists gain access to a wealth of inspiration and knowledge, immersing themselves in the vast spectrum of artistic expression.Through the exploration of curated datasets, aspiring artists can discern prevailing artistic trends, innovative techniques, and the evolution of creative processes across different time periods and cultural contexts. They can study the works of renowned masters, analyze the intricacies of composition, color theory, narrative structure, and gain valuable insights into the artistic journey.Moreover, curated datasets serve as invaluable resources for artists seeking to refine their craft, experiment with new styles, and find their unique voice. By studying the works of their peers and predecessors, aspiring artists can identify patterns, uncover hidden connections, and draw inspiration from a diverse range of sources. This process of exploration and discovery fosters artistic growth and empowers individuals to push the boundaries of their creativity.In essence, curated datasets not only preserve the rich tapestry of human creativity but also serve as catalysts for artistic innovation and expression. By providing a platform for exploration, learning, and collaboration, these datasets nurture a vibrant ecosystem where artists can thrive, evolve, and contribute meaningfully to the cultural landscape.In the domain of Mechanical Engineering, professionals are the architects of innovation, utilizing the foundational principles of physics, mathematics, and material science to conceive, analyze, and refine mechanical systems and components. Armed with specialized knowledge in disciplines like thermodynamics, fluid mechanics, and mechanical design, these engineers serve as catalysts for progress across diverse industries, including automotive, aerospace, manufacturing, and robotics.A curated dataset tailored for Mechanical Engineering represents a goldmine of project data, encapsulating a wealth of design specifications, simulation outcomes, and performance metrics pertaining to a multitude of mechanical systems and components. Within these datasets lie the blueprints of innovation, offering aspiring mechanical engineers a roadmap to navigate the complexities of real-world engineering challenges.By immersing themselves in such datasets, aspiring mechanical engineers gain invaluable insights into industry best practices, emerging technologies, and the nuances of engineering design. They dissect the intricacies of design optimization, explore the dynamics of fluid flow, and unravel the mysteries of structural integrity, all while honing their problem-solving skills and analytical acumen.Moreover, curated datasets serve as repositories of collective wisdom, encapsulating the cumulative knowledge of seasoned professionals and industry pioneers. Through meticulous analysis and experimentation, aspiring mechanical engineers glean insights into the forefront of technological advancement, discovering novel solutions and pushing the boundaries of innovation within the field.In essence, a tailored dataset for Mechanical Engineering serves as a beacon of knowledge, guiding aspiring engineers on their journey towards mastery and excellence. By leveraging the insights gleaned from such datasets, they equip themselves with the tools, techniques, and understanding needed to tackle the engineering challenges of tomorrow and propel the industry forward into a future defined by ingenuity and progress.In the dynamic sphere of Sales, professionals are adept navigators of the ever-changing currents of market dynamics, customer relations, and revenue generation. Equipped with a potent blend of persuasive communication skills, keen market insights, and a results-oriented mindset, sales professionals embark on a journey to forge meaningful connections with clients, unearth lucrative opportunities, and propel revenue growth for their organizations.A meticulously curated dataset tailored to the Sales domain serves as a treasure trove of invaluable sales data, comprising a diverse array of customer interactions, sales performance

metrics, and prevailing market trends spanning various industries and markets. Within these datasets lie the keys to unlocking success in the competitive landscape of sales and business development.By delving into such datasets, aspiring sales professionals glean actionable insights into effective sales strategies, discerning customer behavior patterns, and deploying sophisticated market segmentation techniques. Armed with this knowledge, they fine-tune their sales acumen, refining their approach to client engagement, and optimizing their strategies to maximize revenue generation and achieve sustainable growth.Moreover, curated datasets serve as virtual classrooms, offering a platform for continuous learning and skill development. Aspiring sales professionals immerse themselves in the wealth of data, extracting valuable lessons from past successes and failures, and leveraging this knowledge to navigate the complexities of the sales landscape with confidence and finesse.In essence, a tailored dataset for the Sales domain is more than just a repository of numbers—it is a strategic asset, empowering aspiring sales professionals to chart their course towards success in the competitive world of sales and business development. By harnessing the insights gleaned from such datasets, they equip themselves with the tools, techniques, and understanding needed to thrive in a dynamic and ever-evolving marketplace, driving growth, and prosperity for themselves and their organizations.Apart from that there are another ten or twelve categories and all that job titles or the category can has its own advantages and can be helpful by that. The main advantages that it have are -**Streamlines the recruitment process:** Categorizing resumes into specific roles enables recruiters to quickly pinpoint candidates with the necessary qualifications and experience for particular positions. This targeted approach expedites the initial screening process, saving valuable time and effort.**Saves time and resources:** By focusing recruitment efforts on candidates closely aligned with job requirements, manual review of resumes is minimized. This efficient allocation of resources ensures that recruiting teams invest their time where it matters most, reducing wasted effort on unsuitable candidates.**Facilitates targeted talent acquisition strategies:** Categorized resumes empower recruiters to tailor outreach and recruitment strategies to attract candidates with specific skill sets and relevant experience. This customized approach enhances the likelihood of attracting qualified candidates who are a perfect fit for the organization.**Provides valuable insights into candidate qualifications:** Categorized resumes offer recruiters valuable insights into candidates' qualifications, skills, and experiences. Armed with this information, recruiters can make more informed hiring decisions, ensuring they select candidates who best meet the job requirements.**Enhances organizational efficiency:** Optimal recruitment processes, facilitated by categorized resumes, contribute to improved overall efficiency in hiring. Streamlined procedures enable recruiters to fill vacancies promptly, reducing time-to-hire and ensuring key positions are filled promptly to meet organizational demands.Balancing its advantages, the practice of categorizing resumes into specific roles also harbors certain disadvantages.**Overlooking potential talent:** Strict categorization may inadvertently exclude candidates with transferable skills or diverse experiences that could benefit the organization. Focusing solely on predefined roles risks overlooking valuable talent that doesn't fit neatly into established categories.**Bias and stereotyping:** Categorization based solely on job titles or keywords may perpetuate bias and stereotypes, potentially resulting in the exclusion of qualified candidates from underrepresented groups or unconventional backgrounds.

Missing out on innovation: Rigid categorization may stifle creativity and innovation by favoring candidates who conform to traditional molds. This approach risks overlooking individuals who bring

fresh perspectives and unconventional approaches to problem-solving.**Narrow focus:** Excessive focus on predefined roles may limit the organization's ability to adapt to evolving needs or recognize emerging talent with skills spanning multiple domains. A rigid approach to categorization may hinder flexibility and agility in responding to changing market demands.**Limited diversity:** Categorizing resumes strictly based on specific roles may inadvertently lead to a lack of diversity in the workforce. Candidates from diverse backgrounds or with non-traditional career paths may be overlooked, resulting in a less diverse and inclusive workplace.

Difficulty in matching candidates: Categorizing resumes into predefined roles may sometimes result in mismatches between candidates' actual skills and the requirements of the roles. This can lead to inefficient hiring processes or poor job fit, ultimately hampering organizational effectiveness.In conclusion, while categorizing resumes into specific roles offers numerous advantages in streamlining recruitment processes and identifying qualified candidates, it's crucial to acknowledge and address the potential drawbacks. Striking a balance between efficiency and inclusivity is essential to ensure that organizations not only attract top talent but also foster diversity, innovation, and adaptability within their workforce. By remaining mindful of these considerations, organizations can optimize their recruitment strategies and cultivate a dynamic and inclusive workplace culture poised for sustained success.

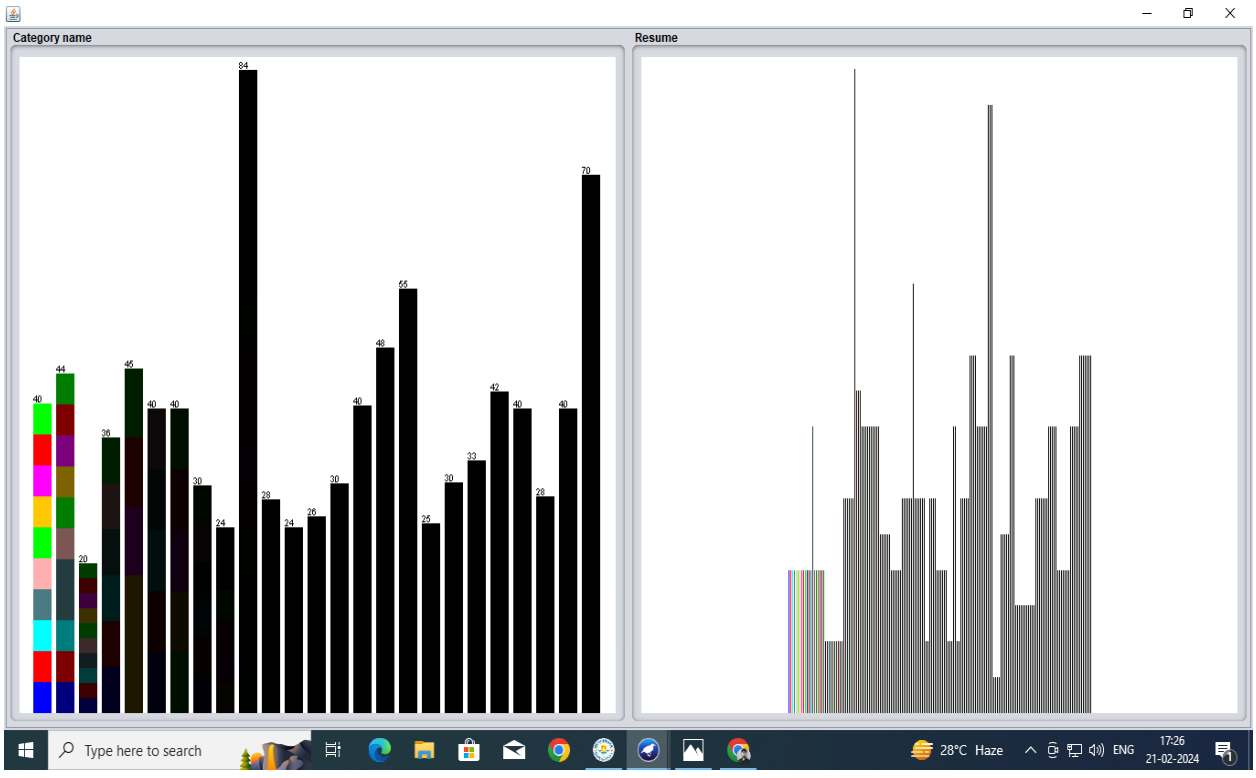

Fig 2: Resume dataset (Category & Resume)

After successfully uploading file in Weka:

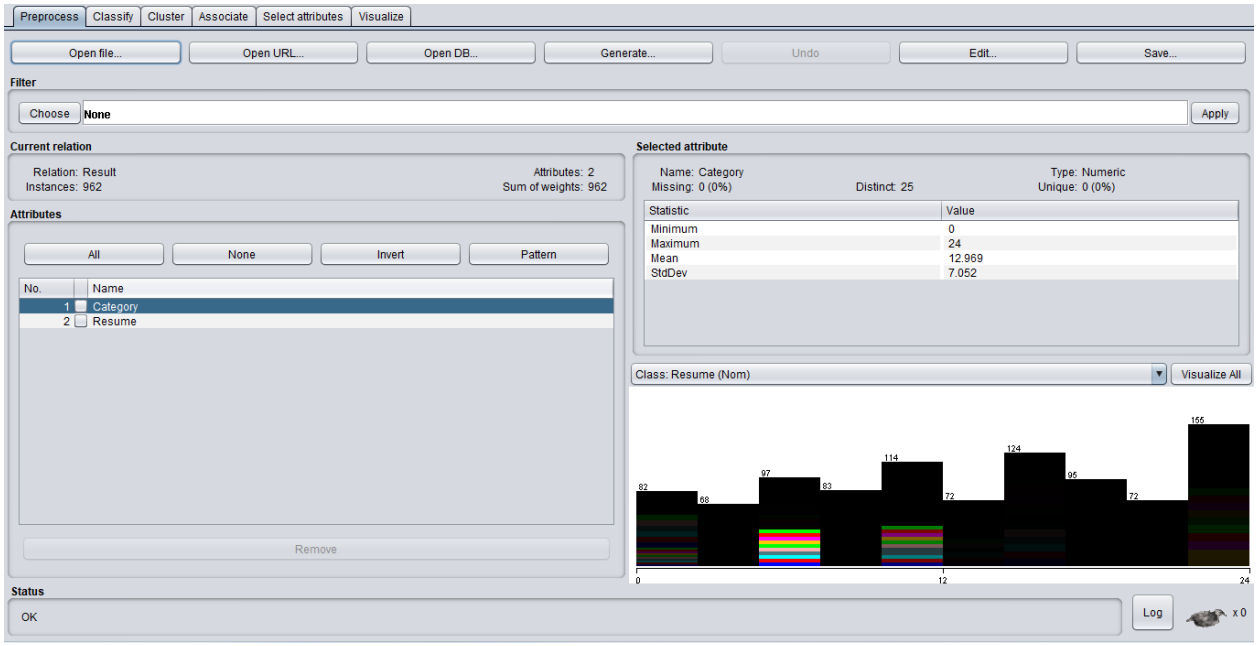

Fig 3: Resume dataset

After applying the discretized method to filter the dataset, an examination was conducted on two attributes: 'Category' and 'Resume'. The dataset comprises 962 instances, with a total weight equal to 962.

For the 'Category' attribute, only one missing value was identified, accounting for 0% of the total dataset. Within this attribute, 25 distinct categories were found, all of which are represented more than once. Notably, no unique categories were observed. This attribute is of a numeric type, indicating its likely association with numerical classifications or groupings within the dataset. Noteworthy roles within this category include that of a Data Scientist, which embodies proficiency in statistical analysis, machine learning, and domain expertise. Data Scientists play a crucial role in extracting actionable insights from complex datasets, utilizing advanced analytical techniques and programming languages like Python or R to inform strategic decisions and drive innovation across various industries.

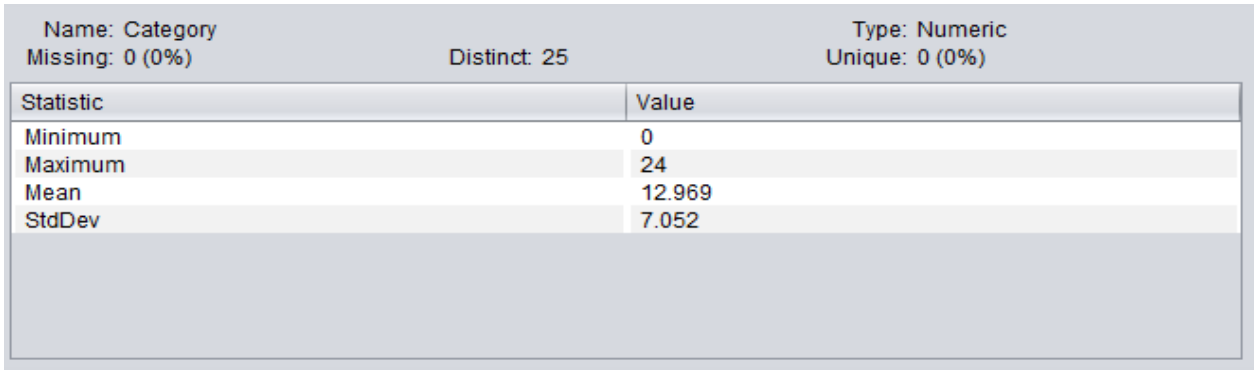

Fig 4: After filtering (Category)

Regarding the 'Resume' attribute, no missing values were encountered, accounting for 0% of the total dataset. A total of 166 distinct resumes were identified, each representing a unique professional profile or individual. However, it's notable that within these resumes, four unique categories were discerned. This suggests potential clusters or groupings based on similarities in experience, skills, or career trajectory. Classified as nominal, this attribute pertains to qualitative distinctions rather than quantitative measurements, highlighting the diverse and multifaceted backgrounds represented within the dataset.

Name: Resume
Missing: 0 (0%)
Distinct: 166
Type: Nominal
Unique: 4 (0%)

No.	Label	Count	Weight
1	Skills Programming Language...	4	4.0
2	Education Details May 2013 to ...	4	4.0
3	Areas of Interest Deep Learnin...	4	4.0
4	Skills R Python SAP HANA Tab...	4	4.0
5	Education Details MCA YMCA...	4	4.0
6	SKILLS C Basics IOT Python M...	4	4.0
7	Skills Python Tableau Data Vis...	4	4.0
8	Education Details B Tech Ray...	4	4.0

Fig 5: After filtering (Resume)

Visual Representation of Category and Resume Attributes:

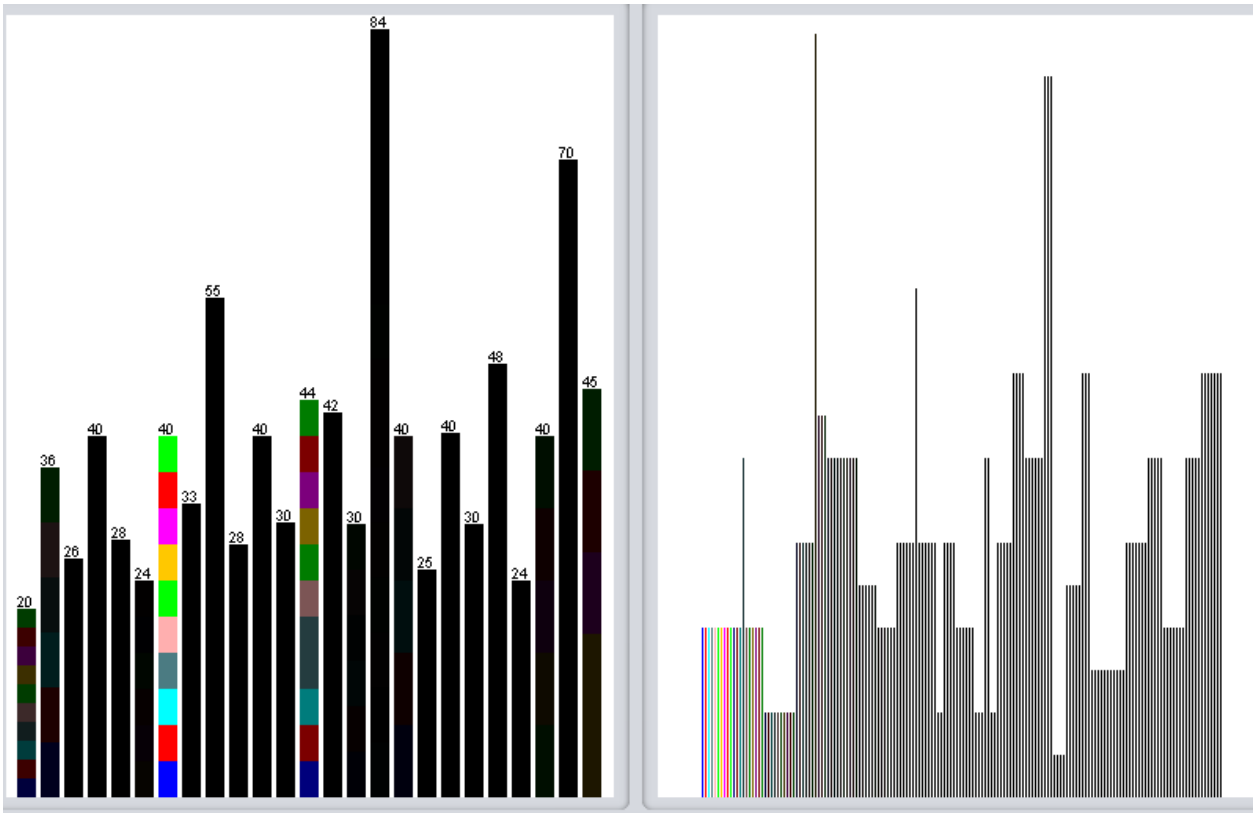

Fig 6: Visual Representation of Category and Resume Attributes

To test the model I split 10.0% train, the remainder test. I have 962 instances.

Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 4 Issue 2 (2024)

Rule ZeroR:

For rule ZeroR, Correctly Classified Instances are 12 which is 1.3857 % of the dataset, and Incorrectly Classified Instances are 854 which is 98.6143 % of the dataset. Kappa statistics is 0, Mean absolute error is 0.012, Root mean squared error is 0.0774, Relative absolute error is 100%, Root relative squared error is 100%, Relative absolute error is 100% and Total Number of Instances is 866. Time taken to build a model: 0 seconds and Time taken to test the model on training split: 0.01 seconds.

Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.000	0.000	0.000	0.000	0.000	0.198	0.004	Skills Programming Languages Python pandas numpy scip
	0.000	0.000	0.000	0.000	0.000	0.000	0.198	0.004	Education Details May 2013 to May 2017 B E UIT RGPV D
	0.000	0.000	0.000	0.000	0.000	0.000	0.198	0.003	Areas of Interest Deep Learning Control System Design
	0.000	0.000	0.000	0.000	0.000	0.000	0.199	0.003	Skills R Python SAP HANA Tableau SAP HANA SQL SAP HAN
	0.000	0.000	0.000	0.000	0.000	0.000	0.199	0.003	Education Details MCA YMCAUST Faridabad Haryana Data
	0.000	0.000	0.000	0.000	0.000	0.000	0.198	0.004	SKILLS C Basics IOT Python MATLAB Data Science Machin
	0.000	0.000	0.000	0.000	0.000	0.000	0.198	0.004	Skills Python Tableau Data Visualization R Studio Mac
	0.000	0.000	0.000	0.000	0.000	0.000	0.198	0.004	Education Details B Tech Rayat and Bahra Institute of
	0.000	0.000	0.000	0.000	0.000	0.000	0.198	0.004	Personal Skills Ability to quickly grasp technical as
	0.000	0.000	0.000	0.000	0.000	0.000	0.398	0.008	SKILLS Bitcoin Ethereum Solidity Hyperledger Beginner
	0.000	0.000	0.000	0.000	0.000	0.000	0.499	0.010	Good logical and analytical skills Positive attitude
	0.000	0.000	0.000	0.000	0.000	0.000	0.499	0.010	COMPUTER PROFICIENCY Basic MS Office PowerPoint word C
	0.000	0.000	0.000	0.000	0.000	0.000	0.499	0.010	Computer Skills Proficient in MS office Word Basic Exc
	0.000	0.000	0.000	0.000	0.000	0.000	0.499	0.010	Willingness to a ept the challenges Positive thinking
	0.000	0.000	0.000	0.000	0.000	0.000	0.499	0.010	PERSONAL SKILLS Quick learner Eagerness to learn new t
	0.000	0.000	0.000	0.000	0.000	0.000	0.499	0.010	COMPUTER SKILLS SOFTWARE KNOWLEDGE MS Power Point MS C
	0.000	0.000	0.000	0.000	0.000	0.000	0.499	0.010	Skill Set OS Windows XP 7 8 8 1 10 Database MYSQL sql
Weighted Avg.	0.019	0.019	0.000	0.019	0.001	0.000	0.336	0.007	

Fig 7: Accuracy class (ZeroR)

Confusion matrix:

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	ab	ac	ad	ae	af	ag	ah	ai	aj	ak	al	am	an	ao	ap	aq	ar	as	at	au
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0

Fig 8: Confusion matrix (ZeroR)

Rule PART:

Correctly Classified Instances are 105(12.1247 %) and Incorrectly Classified Instances 761 (87.8753 %). Kappa statistics is 0.1157, Mean absolute error is 0.0105, Root mean squared error is 0.0813, Relative absolute error is 100%, Root relative squared error is 87.5668 %, Relative absolute error is 105.0189 %, and Total Number of Instances 866. Time taken to build a model: 0.05 seconds and Time taken to test the model on training split: 0.01 seconds.

Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 4 Issue 2 (2024)

Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.020	0.000	0.000	0.000	-0.009	0.970	0.078	Skills Programming Languages Python pandas numpy scip
	0.000	0.009	0.000	0.000	0.000	-0.006	0.970	0.078	Education Details May 2013 to May 2017 B E UIT RGPV D
	0.000	0.006	0.000	0.000	0.000	-0.005	0.968	0.069	Areas of Interest Deep Learning Control System Design
	0.000	0.006	0.000	0.000	0.000	-0.005	0.965	0.078	Skills R Python SAP HANA Tableau SAP HANA SQL SAP HAN
	0.000	0.000	0.000	0.000	0.000	0.000	0.966	0.067	Education Details MCA YMCAUST Faridabad Haryana Data
	0.000	0.000	0.000	0.000	0.000	0.000	0.970	0.078	SKILLS C Basics IOT Python MATLAB Data Science Machin
	0.000	0.000	0.000	0.000	0.000	0.000	0.970	0.078	Skills Python Tableau Data Visualization R Studio Mac
	0.000	0.000	0.000	0.000	0.000	0.000	0.966	0.067	Education Details B Tech Rayat and Bahra Institute of
	0.000	0.000	0.000	0.000	0.000	0.000	0.968	0.069	Personal Skills Ability to quickly grasp technical as
	0.000	0.000	0.000	0.000	0.000	0.000	0.971	0.083	Expertise Data and Quantitative Analysis Decision Ana
	0.000	0.000	0.000	0.000	0.000	0.000	0.980	0.183	SKILLS Bitcoin Ethereum Solidity Hyperledger Beginner
	1.000	0.063	0.143	1.000	0.250	0.366	0.968	0.143	Good logical and analytical skills Positive attitude
	0.000	0.000	0.000	0.000	0.000	0.000	0.968	0.143	COMPUTER PROFICIENCY Basic MS Office PowerPoint word (
	0.000	0.000	0.000	0.000	0.000	0.000	0.968	0.143	Computer Skills Proficient in MS office Word Basic Ex
	0.000	0.000	0.000	0.000	0.000	0.000	0.968	0.143	Willingness to a ept the challenges Positive thinkin
	0.000	0.000	0.000	0.000	0.000	0.000	0.968	0.143	PERSONAL SKILLS Quick learner Eagerness to learn new
	0.000	0.000	0.000	0.000	0.000	0.000	0.968	0.143	COMPUTER SKILLS SOFTWARE KNOWLEDGE MS Power Point MS (
	0.000	0.000	0.000	0.000	0.000	0.000	0.968	0.143	Skill Set OS Windows XP 7 8 8 1 10 Database MYSQL sql
Weighted Avg.	0.107	0.007	0.038	0.107	0.052	0.057	0.972	0.155	

Fig 9: Accuracy class (PART)

Confusion matrix:

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	ab	ac	ad	ae	af	ag	ah	ai	aj	ak	al	am	an	ao	ap	aq	ar	as	at	au
0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig 10: Confusion matrix (PART)

True class	Predicted class	
	a-Yes	b+c-No
a-Yes	TP-0	FN-3
b+c-No	FP-5	TN-1

Table 1 : Prediction of 0,3,5,1

True class	Predicted class	
	b-Yes	a+c-No
b-Yes	TP-0	FN-3
a+c-No	FP-2	TN-4

Table 2 : Prediction of 0,3,2,4

True class	Predicted class	
	c-Yes	a+b-No
c-Yes	TP-0	FN-3
a+b-No	FP-2	TN-4

Table 3 : Prediction of 0,3,2,4

Rule OneR:

Correctly Classified Instances are 12 which is 1.3857 % of the dataset and Incorrectly Classified Instances are 854 which is 98.6143 % of the dataset. Kappa statistics is 0, Mean absolute error is 0.0119, Root mean squared error is 0.109, Relative absolute error is 99.2349%, Root relative squared error is 100%, Relative absolute error is 140.7714% and Total Number of Instances 866.Time taken to build model: 0 seconds and Time taken to test model on training split: 0.02 seconds.

Detailed Accuracy By Class:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.004	Skills Programming Languages Python pandas numpy scip
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.004	Education Details May 2013 to May 2017 B E UIT RGPV D
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.004	Areas of Interest Deep Learning Control System Design
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.004	Skills R Python SAP HANA Tableau SAP HANA SQL SAP HAN
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.004	Education Details MCA YMCAUST Faridabad Haryana Data
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.004	SKILLS C Basics IOT Python MATLAB Data Science Machin
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.004	Skills Python Tableau Data Visualization R Studio Mac
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.004	Education Details B Tech Rayat and Bahra Institute of
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.004	Personal Skills Ability to quickly grasp technical as
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.008	SOFTWARE SKILLS Languages C C java Operating System
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.008	SKILLS Bitcoin Ethereum Solidity Hyperledger Beginn
	1.000	0.063	0.143	1.000	0.250	0.366	0.968	0.143	Good logical and analytical skills Positive attitu
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.010	COMPUTER PROFICIENCY Basic MS Office PowerPoint wor
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.010	Computer Skills Proficient in MS office Word Basic
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.010	Willingness to a ept the challenges Positive think
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.010	PERSONAL SKILLS Quick learner Eagerness to learn ne
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.010	COMPUTER SKILLS SOFTWARE KNOWLEDGE MS Power Point M
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.010	Skill Set OS Windows XP 7 8 8 1 10 Database MYSQL s
Weighted Avg.	0.106	0.010	0.028	0.106	0.043	0.048	0.548	0.028	

Fig 11: Accuracy class (OneR)

Confusion matrix:

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	ab	ac	ad	ae	af	ag	ah	ai	aj	ak	al	am	an	ao	ap	aq	ar	as	at	au	a
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Fig 12: Confusion matrix (OneR)

True Positive Rate (TP Rate):

- Also known as Sensitivity or Recall.
- It measures the proportion of actual positive cases that were correctly identified by the model as positive.
- TP Rate = TP / (TP + FN), where TP is True Positives and FN is False Negatives.
- For ZeroR: TP Rate is 0.019
- For PART: TP Rate is 0.107
- For OneR: TP Rate is 0.106

False Positive Rate (FP Rate):

- It measures the proportion of actual negative cases that were incorrectly classified as positive by the model.
- FP Rate = FP / (FP + TN), where FP is False Positives and TN is True Negatives.
- For ZeroR: FP Rate is 0.019
- For PART: FP Rate is 0.007
- For OneR: FP Rate is 0.010

Precision:

- Precision quantifies the accuracy of positive predictions made by the model.
- It is the ratio of true positive predictions to all positive predictions made by the model.
- Precision = TP / (TP + FP).
- For ZeroR: Precision is 0.000
- For PART: Precision is 0.038
- For OneR: Precision is 0.028

Recall:

- Recall, also known as Sensitivity or True Positive Rate, measures the ability of the model to identify all relevant instances.
- It is the ratio of true positive predictions to all actual positive instances.
- Recall = TP / (TP + FN).
- For ZeroR: Recall is 0.019
- For PART: Recall is 0.107
- For OneR: Recall is 0.106

F-Measure:

- F-Measure provides a single score that balances both precision and recall.
- It is the harmonic mean of precision and recall.
- F-Measure = 2 * (Precision * Recall) / (Precision + Recall).
- For ZeroR: F-Measure is 0.001
- For PART: F-Measure is 0.052
- For OneR: F-Measure is 0.043

Matthews Correlation Coefficient (MCC):

- MCC is a correlation coefficient used to evaluate the quality of binary classifications, especially when dealing with imbalanced datasets.
- It ranges between -1 and +1, where +1 indicates perfect prediction, 0 indicates no better than random prediction, and -1 indicates total disagreement between prediction and observation.
- For ZeroR: MCC is 0.000
- For PART: MCC is 0.057
- For OneR: MCC is 0.048

ROC Area:

- ROC (Receiver Operating Characteristic) curve is a graphical plot that illustrates the diagnostic ability of a binary classification model.
- ROC Area represents the area under the ROC curve, which quantifies the model's ability to discriminate between positive and negative classes.
- For ZeroR: ROC Area is 0.336
- For PART: ROC Area is 0.572
- For OneR: ROC Area is 0.548

PRC Area (Precision-Recall Curve Area):

- PRC Area represents the area under the Precision-Recall curve.
- It provides a comprehensive measure of a model's performance across different thresholds, particularly useful when dealing with imbalanced datasets or when the focus is on positive instances.

- For ZeroR: PRC Area is 0.007
- For PART: PRC Area is 0.155
- For OneR: PRC Area is 0.028

```
Plot the margins
plt.figure(figsize=(10, 6))
plt.plot(sorted(margins), marker='o')
plt.xlabel('Sample index')
plt.ylabel('Margin')
plt.title('Margin Curve for ZeroR Classifier')
plt.show()
```

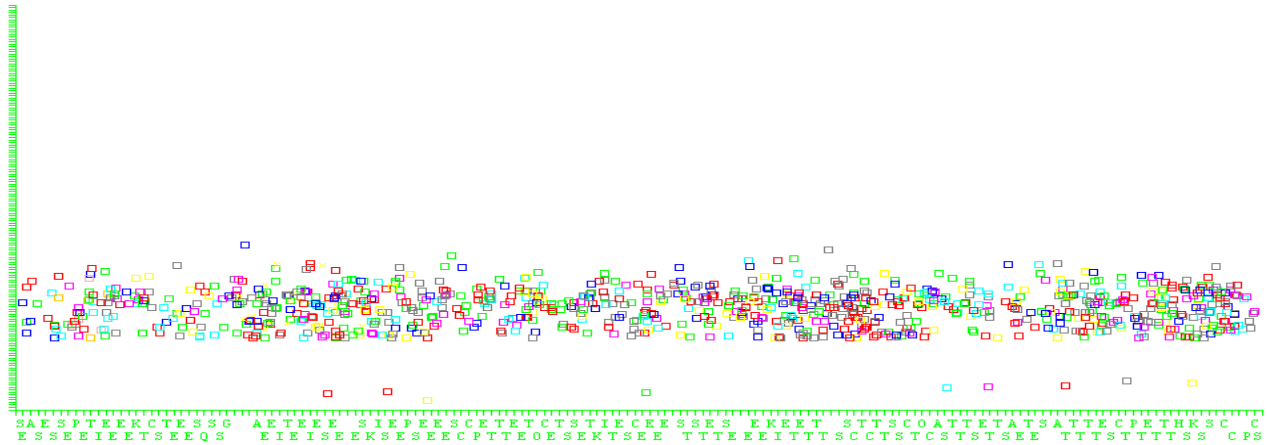

Fig 13: Visualize classifier errors(ZeroR)

Interpretation since the ZeroR classifier predicts the most frequent class for all instances , the margin will be relatively low as it doesn't actually "learn" from the features. The margin curve for ZeroR will help us understand the baseline confidence of predictions, which should be compared against more sophisticated models to evaluate their performance improvements.

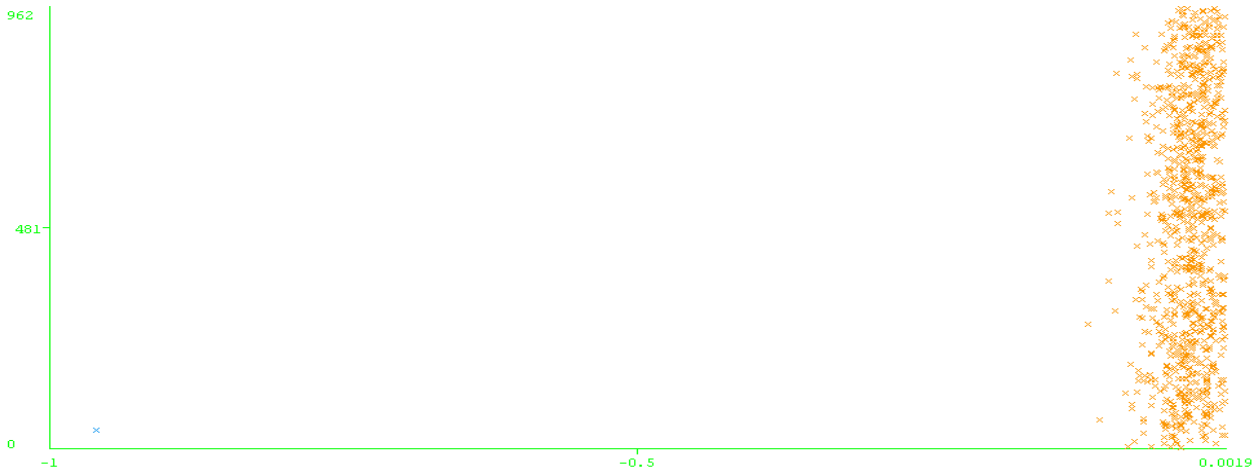

Fig : Visualize margin curve (ZeroR)

The ZeroR classifier is a simple baseline classifier that ignores all input features and predicts the majority class (the most frequent class in the training set). It's often used as a benchmark to compare the performance of more complex classifiers.

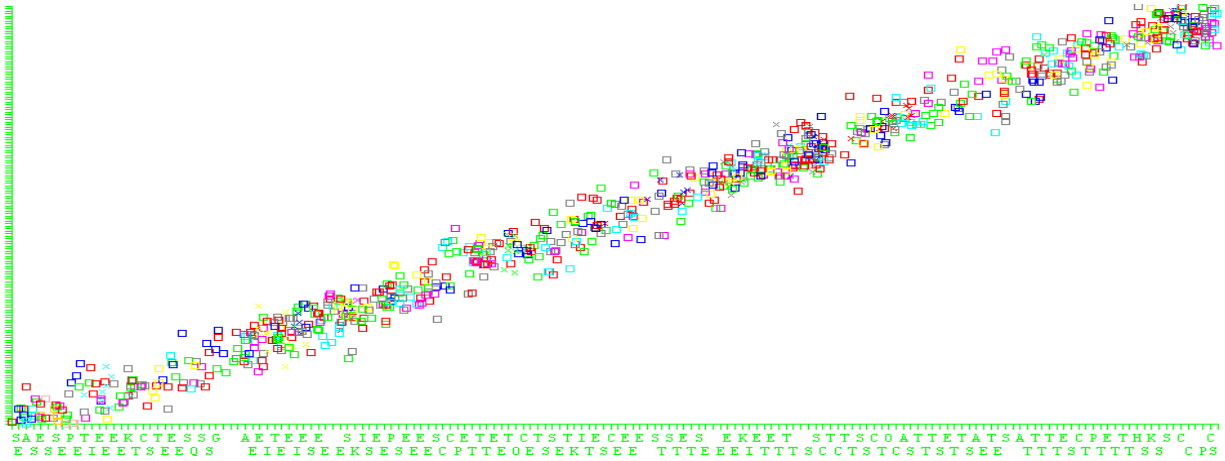

Fig 14: Visualize Classifer (PART)

The margin curve for the PART classifier (approximated with a decision tree) will help us understand how confident the classifier is in its predictions. A higher margin indicates more confident predictions, while lower margins indicate less confident predictions. By comparing this to the ZeroR classifier's margin curve, we can evaluate the improvement in confidence and accuracy.

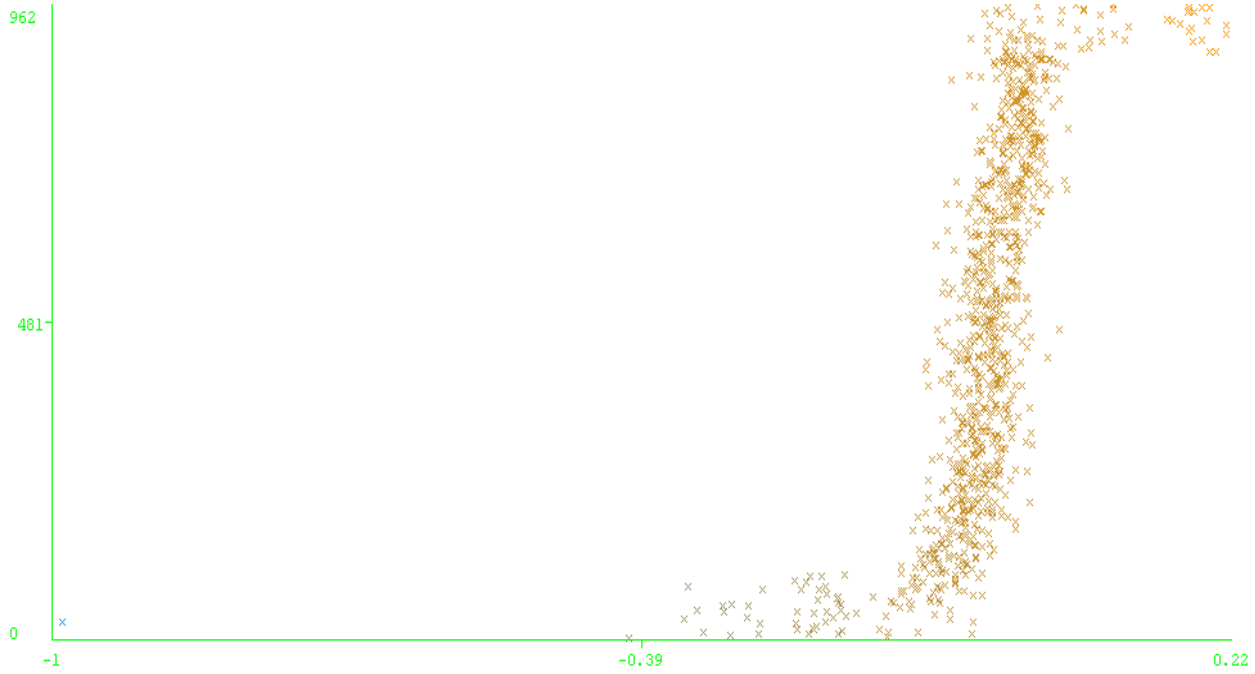

Fig : Visualize margin curve (PART)

```
plt.figure(figsize=(10, 6))

plt.plot(sorted(margins), marker='o')
```

```
plt.xlabel('Sample index')

plt.ylabel('Margin')

plt.title('Margin Curve for Decision Tree Classifier (PART Approximation)')

plt.show()
```

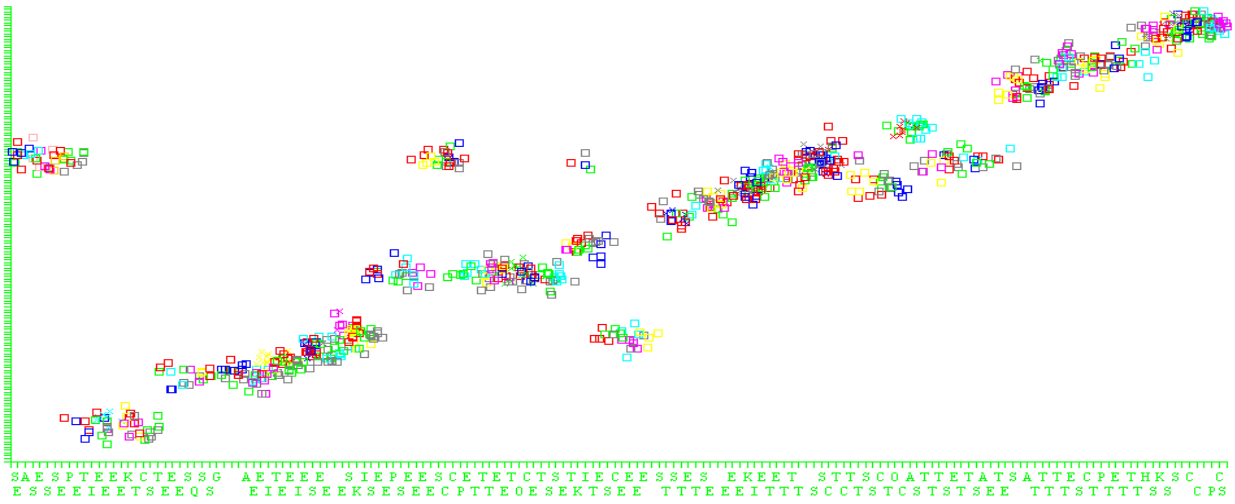

Fig 17: Visualize classifier errors(OneR)

The OneR (One Rule) classifier is a simple, interpretable classifier that makes predictions based on a single feature. It selects the feature that provides the best performance (typically measured by the lowest error rate) and then creates rules based on the values of that feature.

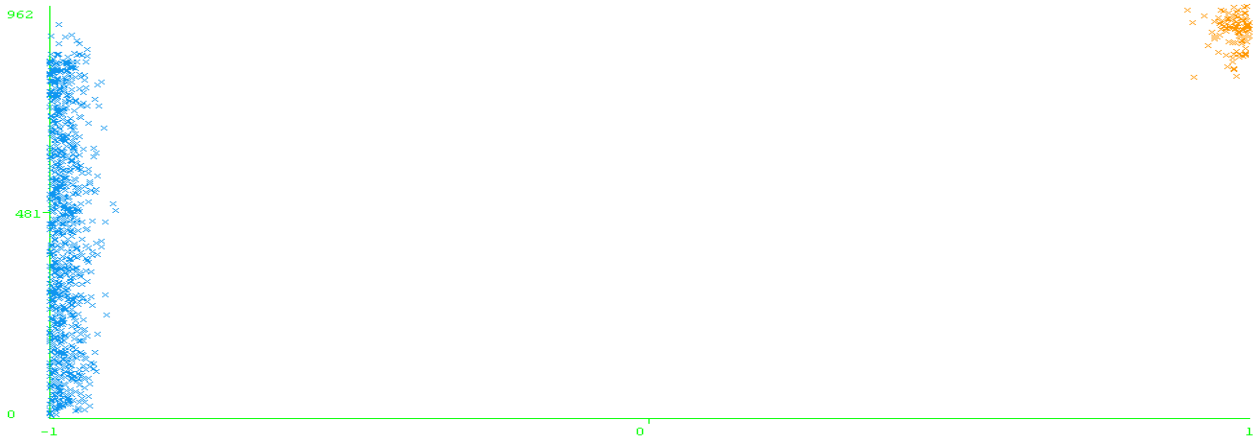

Fig 18: Visualize margin curve(OneR)

Conclusion:

Based on the provided evaluation metrics and the analysis of the confusion matrix, it can be concluded that the PART classifier outperforms both the ZeroR and OneR classifiers. The evaluation metrics, including the weighted averages of TP rate, indicate that the PART classifier achieves higher performance in terms of overall accuracy and predictive power. Additionally, the analysis of the

confusion matrix provides further evidence supporting the superiority of the PART classifier. Therefore, it can be confidently stated that among the models evaluated, PART is the best-performing classifier for the given task.

Reference :

1. Velusamy, P., Rajendran, S., Mahendran, R.K., Naseer, S., Shafiq, M., Choi, J.-G.:Unmanned Aerial Vehicles (UAV) in Precision Agriculture: Applications and Challenges. Energies. 15, (2022). https://doi.org/10.3390/en15010217
2. Beckman, J., Countryman, A.M.: The importance of agriculture in the economy:impacts from COVID-19. Am. J. Agric. Econ. 103, 1595–1611 (2021)
3. Maes, W.H., Steppe, K.: Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture. Trends Plant Sci. 24, 152–164 (2019)
4. R. Naveen Kumar* and M. Anand Kumar Enhanced Fuzzy K-NN Approach for Handling Missing Values in Medical Data Mining Indian Journal of Science and Technology, Vol 9(S1), DOI: 10.17485/ijst/2016/v9iS1/94094, December 2016
5. Hussain, M., Wang, Z., Huang, G., Mo, Y., Kaousar, R., Duan, L., Tan, W.: Comparison of Droplet Deposition, 28-Homobrassinolide Dosage Efficacy and Working Efficiency of the Unmanned Aerial Vehicle and Knapsack Manual Sprayer in the Maize Field. Agronomy 12, 385 (2022)
6. Petkovics, I., Simon, J., Petkovics, Á., Čović, Z.: Selection of unmanned aerial vehicle for precision agriculture with multi-criteria decision making algorithm. In: 2017 IEEE 15th international symposium on intelligent systems and informatics (SISY). pp. 151–156. IEEE (2017)
7. Hamurcu, M., Eren, T.: Selection of unmanned aerial vehicles by using multicriteria decision-making for defense. J. Math. 2020, 4308756 (2020). https://doi.org/10.1155/2020/4308756
8. Atanassov, K.T.: Intuitionistic fuzzy sets. In: Intuitionistic fuzzy sets. pp. 1–137.Springer (1999)
9. Vlachos, I.K., Sergiadis, G.D.: Intuitionistic fuzzy information–applications to pattern recognition. Pattern Recognit. Lett. 28, 197–206 (2007) Google Scholar
10. Xu, Z., Chen, J., Wu, J.: Clustering algorithm for intuitionistic fuzzy sets. Inf. Sci. (Ny) 178, 3775–3790 (2008) MathSciNet MATH Google Scholar
11. Hung, W.-L., Yang, M.-S.: On the J-divergence of intuitionistic fuzzy sets with its application to pattern recognition. Inf. Sci. (Ny) 178, 1641–1650 (2008) MathSciNet MATH Google Scholar
12. R. Naveen Kumar1 and Dr. M. Anand Kumar2 “A Novel Feature Selection Algorithm with Dempster Shafer Fusion Information for Medical Datasets” IJAE Research ISSN 0973-4562 Volume 12, Number 14 (2017) pp. 4205-4212 Scopus.
13. Dr R.Naveenkumar “An Empirical Research Approach on Confusion Matrix Using Existing Musical Industry Dataset” International Journal of Scientific Research in Engineering and Management (IJSREM) Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930.