Precision in Progress: Leveraging Data Mining Technique to Empower Career Path Selection for Incoming Senior High School Students

Romeo, Jr. E. Bejar

Teacher III, Department of Education, Region XII, Cotabato Division

DOI: https://dx.doi.org/10.47772/IJRISS.2024.801014

Received: 18 December 2023; Revised: 23 December 2023; Accepted: 27 December 2023; Published: 24 January 2024

ABSTRACT

This research investigated the application of data mining, specifically the Random Forest classification model, to optimize career path selection for incoming Senior High School students in the Philippines. Given the diverse tracks and strands within the SHS program, the traditional decision-making process relies on anecdotal advice, limited exposure, and personal perceptions, often resulting in sub optimal choices. Focused on addressing the complexities introduced by the K-12 educational reform, the study analyzed the data of 1,020 students from three public schools including the Sibsib National High School. The Random Forest model achieved high accuracy (91.2%) and precision (72.6%), with critical attributes identified as Career Prospects, Personal Interests or Skills, and the Monthly Salary Bracket of Parents. While the model excelled overall, there is room for improvement in predicting certain academic tracks, particularly Humanities and Social Sciences (HUMSS). The study recommends refining the model, emphasizing enhancements for specific tracks and continual updates to accommodate evolving student data patterns.

Keywords: Data Mining, Career Path Selection, Random Forest Model, Senior High School, K-12 Educational Reform.

INTRODUCTION

In recent years, the field of education has witnessed a transformation driven by the integration of technology and the systematic analysis of data. Within this evolving landscape, data mining, a specialized branch of artificial intelligence and machine learning, has emerged as a powerful tool. Its role is to reveal hidden patterns effectively (Gargano & Raggad, 1999), predict trends and user behaviors (Puscasiu, et al., 2020), and seek to extract useful information (Mackinnon & Glick, 1999) from extensive datasets allowing its users to make proactive, knowledge-driven decisions (Tamaskar & Raut, 2015). This approach has been effectively applied in diverse sectors, ranging from financial analysis to healthcare management, and it is now knocking on the doors of education.

The Senior High School (SHS) program in the Philippines had been implemented as part of the K-12 educational reform. The program was designed to provide an additional two years of specialized education to high school students. It offers various tracks and strands designed to cater to students' diverse interests and career goals. These tracks and strands serve as specialized pathways that students can choose based on their aspirations and aptitudes. The available tracks include Academic, Arts and Design, Sports, and Technical-Vocational-Livelihood (TVL). Within these tracks, there are different strands, each focusing on specific subjects and skill sets. This program aims to provide ample time to acquire sufficient knowledge and mastery of skills that will better prepare Senior High School graduates for various paths after graduation (Gestiada, et al., 2017), including employment, entrepreneurship, college education, and middle skills development. SHS graduates who choose the employment path are expected to have acquired skills and knowledge relevant to their chosen tracks, making them more employable in their respective fields.

Graduates with an entrepreneurial focus are better equipped to start and manage their businesses or contribute effectively to existing enterprises. Moreover, some graduates are also expected to meet the entrance requirements of colleges and universities, both in the Philippines and abroad. Lastly, graduates with middle skills development training can pursue immediate employment opportunities or choose to further their education.

This new educational paradigm, consequently, adds an extra layer of complexity to the career selection process of the Junior High School completers. It has amplified the need for precise guidance for students to choose the right SHS track among the academic, arts and design, sports, and technical-vocational-livelihood tracks, and the right strand within those tracks. Each track and each strand lead to a multitude of career opportunities, and the choice made by a student at this stage profoundly influences his future (Dangoy & Madrigal, 2020). In certain situations, changing one's chosen academic path could be a cause of having the wrong decision in choosing a career, potentially resulting in a waste of time and resources, and career frustration (Dangoy & Madrigal, 2020). Therefore, the need for a solution that helps students navigate this maze of choices becomes evident.

Sibsib National High School (SNHS), located at Sibsib, Tulunan, Cotabato, is a DepEd school offering both Junior High School (Grades 7 to 10) and Senior High School (Grades 11 and 12) programs. Its Senior High School offerings encompass the Humanities and Social Sciences (HumSS) Strand within the Academic Track, as well as specialized programs in Computer Systems Servicing (CSS) and Shielded Metal Arc Welding (SMAW) within the Technical-Vocational-Livelihood Track. For the current school year, it houses 185 Grade 7 students, 203 Grade 8, 176 Grade 9, 154 Grade 10, 170 Grade 11, and 128 Grade 12 students.

Currently, Grade 10 students often rely on anecdotal advice, limited exposure, and personal perceptions when choosing their SHS track. This process lacks the rigor and objectivity needed to accurately align their strengths and interests with their chosen path. As a result, students may end up on paths that do not best suit their abilities and passions, potentially leading to challenges, dissatisfaction, and suboptimal career prospects in the future. To address this issue, there is a pressing need for a more data-driven and precise method that leverages students' unique characteristics to guide them toward the most suitable SHS tracks. However, in the context of the said school, the role of data mining techniques in optimizing this crucial decision-making process remains largely unexplored.

With this, the researcher investigated the potential of using data mining technique, specifically the Random Forest classification technique, to improve the accuracy and dependability of the process through which incoming Senior High School students select their career paths. This stands as the core challenge that this research endeavors to tackle comprehensively. By addressing this issue, the research provided not only a deeper understanding of how data-driven approaches can benefit career guidance but also practical solutions to enhance the decision-making process for these students. Ultimately, the goal is to equip incoming Senior High School students with the tools and insights they need to make more informed and confident choices about their future educational and career pathways.

Objectives of the Study

The main objective of the study is to leverage the Random Forest classification technique to improve the accuracy and dependability of the process through which incoming Senior High School students select their career paths.

Specifically, this study aimed to answer the following research questions:

1. What attributes are considered for the prediction of career path selection?
2. How effective random forest classification model in the prediction of career path selection?
3. What is the output of the random forest classification model and the prediction of career path

selection?

REVIEW OF RELATED LITERATURE

The transition from basic education to senior high school marks a critical phase in a student's academic journey and is regarded as one of the highly challenging events in the life of a student (Zeedyk, et al., 2003). This stage of a student's life may have a negative impact on psychological well-being and academic achievement (Evans, Borriello, & Field, 2018). Moreover, at this juncture, students are required to make pivotal decisions regarding their chosen SHS tracks and strands, which will significantly influence their future career paths (Dangoy & Madrigal, 2020). In recent years, this process has garnered increasing attention from educators and school administrations due to its lasting impact on students' lives. To address this, researchers have turned to data mining techniques, a branch of artificial intelligence, to enhance the precision of career path selection for incoming SHS students. This review explores the existing literature and studies related to this area of inquiry.

Data mining is a field of intersection of computer science and statistics (Agarwal, 2013) to find unexpected, valuable, or interesting structures (Hand, 2000), to discover patterns (Agarwal, 2013), and to extract useful data and trends (Gera & Goel, 2015) from large data sets. This can be done by using techniques like clustering, classification, association, and regression (Gera & Goel, 2015). Data mining has revolutionized various fields. Journals and Smita (2014) provided a survey of data mining techniques and their applications in sectors such as marketing, fraud detection, manufacturing, and telecommunication. Keleş (2017) discussed the impact of data mining in sectors including banking and finance, telecommunication, health, public, construction, engineering, and science. Chen et al. (1996) provided an overview of data mining techniques from a database perspective, emphasizing their relevance in information-providing services and online platforms. Chinchuluun et al. (2010) focused specifically on the importance of data mining techniques in agriculture and environmental sciences, highlighting their role in knowledge discovery and prediction.

In education, data mining has emerged as a potent tool to analyze vast datasets, allowing educators and policymakers to gain insights into student performance, learning behaviors, and career choices. Sachin and Vijay (2012) discussed the history and applications of data mining techniques in education, emphasizing their potential in areas such as prediction, classification, relationship mining, clustering, and social networking. Cheng (2017) focused on the capabilities of data mining approaches in education, particularly in improving learning experiences and institutional effectiveness. Kaur (2015) provided a review of the applications of data mining in education, highlighting its usefulness in extracting useful patterns and rules from educational data. Algarni (2016) emphasized the importance of educational data mining in various areas, including identifying at-risk students, assessing institutional performance, and optimizing curriculum renewal.

Data mining techniques have been employed to identify patterns in students' academic records and behaviors, aiding in the development of interventions and career guidance in the Philippine educational context. Rosado et al. (2019) predicted the performance improvement of Grade 7 Junior High School students for A.Y. 2016-2017 and 2017-2018 from the Basic Education Department of the University of Perpetual Help System Laguna. The data mining technique applied was classification using Naïve Bayes. They determined the Grade 7 Junior High School students' overall average scores based on gender, compared the academic performance between males and females, pinpointed the subjects in which students excel or struggle, assessed performance variations among students from different previous schools, analyzed the students' academic achievements concerning their parents' occupations, and offered predictive insights to assist decision-makers in formulating targeted marketing strategies for schools with low enrollment. Go et al. (2023) analyzed the factors of e-learning in a higher education institution in the Philippines. A data

mining approach was used to predict the satisfaction of higher education students given certain features of e-learning. The findings revealed that those features can be used to accurately predict the student satisfaction towards e-learning of higher education students in the Philippines. Sarte and Palaoag (2019) designed a data warehouse and data mining architecture for the analysis of drop-out, retention, and migration patterns of students. The data warehouse architecture would help school decision-makers in performing student academic status and record analysis. It would also help them determine the critical patterns and trends of students who might drop out, change track, grade repeater, or transfer out.

The importance of precise career path selection during the SHS years cannot be overstated. Ill-informed decisions can result in wasted time, resources, and academic frustration (Dangoy & Madrigal, 2020). Students often rely on anecdotal advice, limited exposure, and personal perceptions when choosing their SHS track. This process lacks the rigor and objectivity needed to accurately align students' strengths and interests with their chosen path. As a result, students may end up on paths that do not best suit their abilities and passions, potentially leading to challenges, dissatisfaction, and suboptimal career prospects in the future. Recent research endeavors have demonstrated the potential of data mining techniques to address these challenges. Nazareno et al. (2019) applied artificial neural networks to predict career strands based on students' grades, achieving an accuracy of 74.1%. Atienza et al. (2022) developed a web-based career track recommender system using a deep neural network, which predicted academic strands with an accuracy of 83.11%. Gestiada et al. (2017) used social cognitive career theory and analytic hierarchy process intending to guide students in career track selection. Hernandez and Atienza (2021) proposed a career track recommender system using the Deep Neural Network (DNN) model. The result of their study showed that the DNN algorithm performs reasonably well in predicting the academic strand of students with a prediction accuracy of 83.11%. They concluded that the recommender system served as a decision tool for counselors in guiding their students to determine which Senior High School track was suitable for students with the utilization of the DNN model. Despite the increasing use of data mining techniques in educational contexts, there is a lack of research focused on leveraging the Random Forest model to enhance the precision and dependability of career path selection for incoming SHS students. Consequently, exploring the untapped potential of Random Forest Classification in the context of SHS track prediction stands as a critical research gap that necessitates comprehensive investigation and analysis.

Nonetheless, it is noteworthy that prevailing studies predominantly emphasize academic performance and socioeconomic factors as the primary determinants in predicting career path selection for incoming Senior High School students. This provides a knowledge gap that the researcher aims to address. Various studies found several factors that might be of great importance in this research. The selection of SHS tracks or strands was evidently influenced by the student's intended college courses and parental guidance (Nazareno et al., 2021), job opportunities (Dublin et al., 2020), and personality factors (Vallejo, 2019). These factors can be deemed valuable additions to the analysis. Consequently, data mining emerges as a potent tool to bridge this gap by offering personalized recommendations that will consider those factors. This inclusive approach aligns to empower all students to make informed decisions.

As the education landscape continues to evolve, leveraging these techniques to empower students in their career decisions is a promising avenue. This research study, "Precision in Progress: Leveraging Data Mining Technique to Empower Career Path Selection for Incoming Senior High School Students," seeks to contribute to this important endeavor by addressing knowledge gaps and exploring innovative solutions.

Conceptual Framework

The following diagram shows the conceptual framework of this research.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS |Volume VIII Issue I January 2024

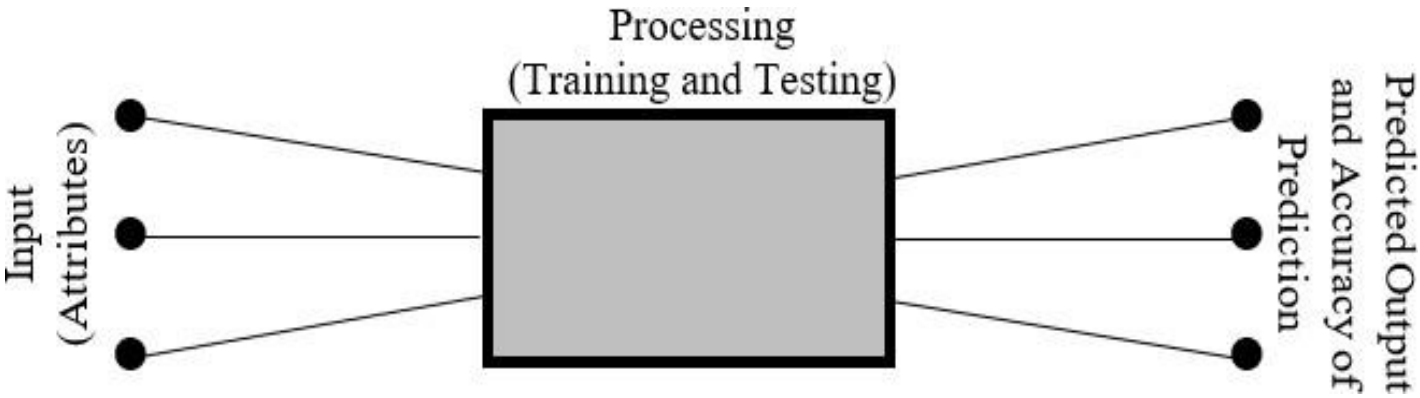

Figure 1. Conceptual Framework

The process involves three primary stages. The first stage, input design, encompasses attributes used for predicting career path selection. The second stage involves (1) Data Preprocessing, (2) Normalization, and (3) Training, Validating, and Testing. In the final stage, the model provides two key outputs: the predicted career path selection and the accuracy of the prediction.

METHODOLOGIES

Data Collection

The students' subject grades and socio-demographic profiles were collected and used to create the random forest classification model. The researcher thoroughly followed ethical procedures before, during, and after the actual data collection. Grade 10 subject grades and socio-demographic data were obtained through a structured online questionnaire, focusing on the necessary information. To uphold confidentiality, the names of the respondents were removed to ensure that no student could be identified in this study.

Data Cleaning

The collected datasets were downloaded and organized in an Excel sheet. In the data pre-processing phase, the researcher meticulously cleaned and replaced any inappropriate data related to students. This data cleaning was carefully implemented to ensure its suitability for the model. All null values were systematically removed, facilitating a smoother learning process for the predictive model.

Normalization

Normalization is helpful as it standardizes data to have similar distributions. Attributes related to students, especially text data, were converted into numerical values. Subsequently, normalization was implemented by scaling students' subject grades between 0 and 1. The study opted for the min-max type of normalization, scaling every feature value between its minimum (0) and maximum (1).

Data Preprocessing

All subject grades were normalized ensuring that each grade varied within the same range from 0 to 1. Additionally, categorical data, such as for sex, were converted into indexes. Male was assigned the value of 1, while Female was assigned the value of 0. Furthermore, the rest of the attributes were also converted to indexes.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS |Volume VIII Issue I January 2024

Tools and Techniques

JASP was used in analyzing the dataset, creating, and testing the random forest classification model, and visualizing the analysis of the dataset. JASP stands for Jeffrey's Amazing Statistics Program in recognition of the pioneer of Bayesian inference Sir Harold Jeffreys. This is a free multi-platform open-source statistics package, developed and continually updated by a group of researchers at the University of Amsterdam. JASP provides a user-friendly interface and aims to make statistical analysis accessible to a broad audience. It covers a wide range of statistical methods and techniques, including descriptive statistics, t-tests, ANOVA, regression, Bayesian statistics, and more. JASP 0.11 adds the Machine Learning module with 13 brand new analyses that can be used for supervised (regression and classification) and unsupervised (clustering) learning.

RESULTS AND DISCUSSION

Attributes Considered for the Prediction

Table I presents the attributes considered for the prediction.

Table I Student-Related Attributes

No.	Attributes	Category	Frequency
1	Track and Strand	Academic – ABM	170
		Academic – HUMSS	170
		Academic – STEM	170
		TVL – HE	170
		TVL – IA	170
		TVL – ICT	170
2	Grade 10 Subject Grades	Filipino; English; Mathematics; Science; Araling Panlipunan; Edukasyon sa Pagpapakatao; Music, Arts, Physical Education, and Health; and Technology and Livelihood Education	1,020
3	Sex	Male	430
		Female	590
4	Number of Siblings	1	166
		2	194
		3	245
		4	142
		5	72
		6	64
		7	48
		8	36
		9	29
		10	24
5	Monthly Salary Bracket of Parents	1 – Less than 9,250	588
		2 – Between 9,250 – 19,040	232
		3 – Between 19,040 – 38,080	137
		4 – Between 38,040 – 66,640	43

		5 – Between 66,640 – 114,240	17
		6 – Greater than 114,240	3
6	Personal Interests or Skills		1,020
7	Career Prospects		1,020

Table I outlines various student-related attributes encompassing tracks and strands, grade 10 subject grades, gender distribution, the number of siblings, parents' monthly salary brackets, and students' personal interests or skills, alongside their perceived or desired career prospects. The dataset includes a total of 1,020 student respondents, with 16.7% distribution for each strand.

Figure 2 shows the sample of input variables for the first 10 students.

1	Track and S	Filipino	English	Math	Science	AP	EsP	MAPEH	TLE	Sex	Number of	Monthly Sa	Personal In	Career Pros
2	1	92	94	92	90	89	94	89	96	1	1	1	4	4
3	1	91	91	92	90	89	92	93	99	2	2	2	1	1
4	1	95	94	95	91	96	94	96	94	1	2	3	1	1
5	1	89	91	94	90	89	97	98	94	2	4	4	1	1
6	1	95	90	90	90	95	92	98	90	2	6	2	1	2
7	1	95	97	93	95	93	93	97	95	2	1	4	1	1
8	1	90	86	91	92	84	91	90	82	2	4	1	9	4
9	1	85	84	88	80	89	86	90	88	1	3	3	10	10
10	1	94	89	90	90	95	93	90	93	1	3	3	10	5

Figure 2. Sample of Input Variables

Figure 2 shows the format of the final dataset used in the training, validation, and testing of the model. From the dataset collected, 63.92 percent of the data was utilized for training, 16.08 percent for validation, and 20 percent of the data was reserved for testing.

Table II below presents the feature importance.

Table II Feature Importance

Feature	**Mean Decrease in Accuracy**	**Total Increase in Node Purity**
Career Prospects	0.289	0.028
Personal Interests / Skills	0.183	0.010
Monthly Salary Bracket of Parents	0.014	0.003
Filipino	0.122	4.574×10^{-4}
Sex	0.011	-4.695×10^{-4}
MAPEH	0.069	-0.001
English	0.091	-0.003
EsP	0.089	-0.008
Science	0.045	-0.009
TLE	0.088	-0.010
Math	0.100	-0.013
Number of Siblings	0.045	-0.013
AP	0.110	-0.017

Table II outlines the feature importance of the random forest classification model. Two metrics were utilized: "Mean Decrease in Accuracy" and "Total Increase in Node Purity." Mean Decrease in Accuracy

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS |Volume VIII Issue I January 2024

measures the average decrease in model accuracy when a particular feature is excluded from the model. Features with higher mean decrease in accuracy are considered more important in maintaining the overall predictive performance of the model. Career Prospects, with mean decrease accuracy of 0.289, contributes the most to decreasing the model's accuracy when excluded. This suggests that this feature plays a crucial role in making accurate predictions. The Monthly Salary Bracket of Parents and Sex, with mean decrease of accuracy of 0.014 and 0.011 respectively, have the lower impact. On the other hand, the Total Increase in Node Purity refers to how well a node separates instances of different classes. Features with a higher total increase in node purity are considered more important in maintaining the homogeneity of decision tree nodes. Filipino has 4.574×10^{-4} total increase in node purity, while Sex has -4.695×10^{-4}. Some subjects (MAPEH, English, EsP, Science, TLE, Math, AP) negatively impact accuracy when excluded, emphasizing their importance in the model.

Figures 3 and 4 illustrate the Mean Decrease in Accuracy and Total Increase in Node Purity plots.

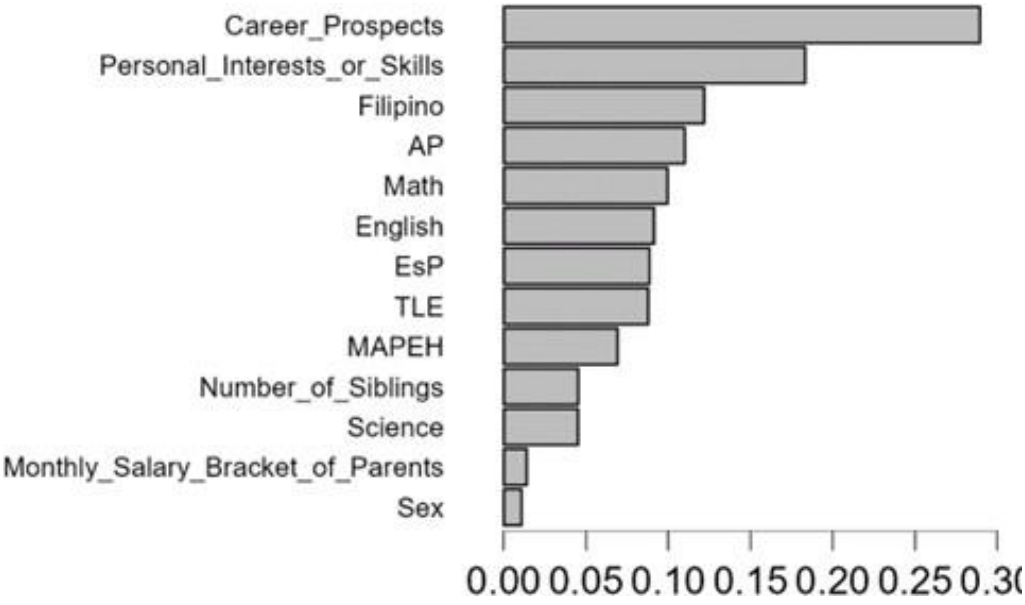

Figure 3. Mean Decrease in Accuracy

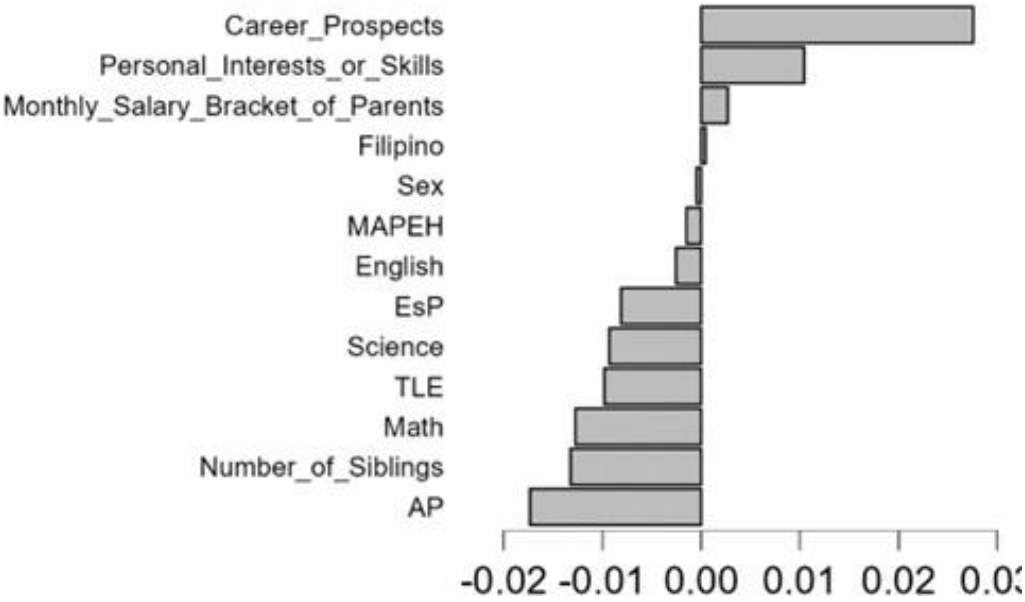

Figure 4. Total Increase in Node Purity

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS |Volume VIII Issue I January 2024

Performance and Evaluation Metrics

Table III presents the random forest classification model's validation, test, and OOB accuracy.

Table III Random Forest Classification Model Accuracy

Trees	Features per Split	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy	OOB Accuracy
34	3	652	164	204	0.799	0.735	0.929

Note. The model is optimized with respect to the *out-of-bag accuracy*.

Table III shows the results of a Random Forest classification model along with its key parameters and performance metrics. The model is configured with 34 trees and splits based on 3 features at each node. The training dataset (n(Train)) consists of 652 instances, while the validation dataset (n(Validation)) comprises 164 instances, and the test dataset (n(Test)) includes 204 instances. The validation accuracy is reported as 0.799, indicating the model's ability to correctly classify instances in the validation dataset. The test accuracy is slightly lower at 0.735, representing the model's performance on an independent set of data. Notably, the out-of-bag (OOB) accuracy is reported as 0.929, and it is important to note that the model is optimized based on this metric. OOB accuracy is a valuable measure as it reflects the model's performance on unseen data during the training process. The result indicates a high level of accuracy on the out-of-bag samples, suggesting that the model performs well on data not used during the training of individual trees within the ensemble.

Table IV presents the confusion matrix in determining the accuracy of track prediction. This summarizes the number of correct and incorrect predictions made by the model in tabular format. The observed value and predicted value are indicated in the table below. The performance of the model is also evaluated using the accuracy, precision, and recall performance metrics. Accuracy is a proportion of the total number of correct predictions of a strand, while Precision indicates the proportion of correct positive observations. The recall is a proportion of positives correctly predicted as positive.

Table IV Confusion Matrix

		Predicted						Classification Overall	Recall
		Acad. ABM	Acad. HUMSS	Acad. STEM	TVL HE	TVL IA	TVL ICT		
Observed	Acad. ABM	27	0	2	1	0	1	31	0.8710
	Acad. HUMSS	3	16	12	1	3	2	37	0.4324
	Acad. STEM	5	10	21	1	0	3	40	0.5250
	TVL HE	0	2	0	28	0	0	30	0.9333
	TVL IA	3	2	0	0	32	2	39	0.8205
	TVL ICT	0	0	0	0	1	26	27	0.9630
Truth Overall		38	30	35	31	36	34	204	0.7353
Precision		0.7105	0.5333	0.6000	0.9032	0.8889	0.7647		

Table IV shows the confusion matrix that evaluates the performance of the random forest classification model across different classes. The model correctly predicted 27 instances of ABM, with a Recall of 87.10%, indicating that it effectively captures 87.10% of the actual ABM cases. The Precision for ABM is

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS |Volume VIII Issue I January 2024

71.05%. Moreover, it correctly predicted 16 instances of HUMSS, with a Recall of 43.24% and a Precision of 53.33%. Additionally, the model correctly predicted 21 instances of STEM, with a Recall of 52.50% and a Precision of 60.00%. For TVL strands, the model correctly predicted 28 instances of TVL HE, with a high Recall of 93.33%. The Precision for TVL HE is 90.32%. Furthermore, it correctly predicted 32 instances of TVL IA, with a Recall of 82.05% and a Precision of 88.89%. Lastly, the model correctly predicted 26 instances of TVL ICT, with a Recall of 96.30% and a Precision of 76.47%. The result reveals that the classification model performs reasonably well across different academic tracks, with varying degrees of accuracy for each category. The model excels in predicting certain tracks and strands, such as TVL HE and TVL ICT, as evidenced by high Recall values (93.33% and 96.30%, respectively) and commendable Precision scores (90.32% and 76.47%). These results imply that the model effectively identifies and minimizes false positives in these tracks. However, challenges are observed in accurately predicting instances of HUMSS, with a lower Recall of 43.24%. This suggests a limitation in capturing a substantial portion of actual HUMSS cases. The Precision for HUMSS is moderate at 53.33%, indicating that while positive predictions are reasonably reliable, there is room for improvement in minimizing false positives.

Table IV also indicates that the overall test accuracy of the model is 73.53%. This implies that the model demonstrates a moderate level of overall test accuracy, suggesting a reasonable ability to make correct predictions across different academic tracks.

Table V shows the summary of the evaluation metrics for the random forest classification model.

Table V Evaluation Metrics

Evaluation Metrics	**Average/Total**
Support	204
Accuracy	0.912
Precision (Positive Predictive Value)	0.726
Recall (True Positive Rate)	0.735
False Positive Rate	0.053
False Discovery Rate	0.267
F1 Score	0.726
Matthews Correlation Coefficient	0.691
Area Under Curve (AUC)	0.928
Negative Predictive Value	0.947
True Negative Rate	0.947
False Negative Rate	0.242
False Omission Rate	0.053
Threat Score	1.496
Statistical Parity	1.000

Note: All metrics are calculated for every class against all other classes.

Table V presents various evaluation metrics for the random forest classification model, offering a comprehensive assessment of its performance across multiple dimensions. The dataset consists of 204 instances. The overall accuracy of the model is 91.2%, representing the proportion of correctly predicted instances. Precision is 72.6%, indicating the proportion of true positive predictions among all positive predictions. The recall is 73.5%, representing the proportion of actual positives that the model correctly identifies. The false positive rate is 5.3%, indicating the proportion of actual negatives that are incorrectly

classified as positives. The false discovery rate is 26.7%, representing the proportion of positive predictions that are incorrect. The F1 Score is 72.6%, which considers both Precision and Recall, providing a balanced measure that considers both false positives and false negatives. The Matthews Correlation Coefficient is 0.691, offering a correlation between predicted and observed classifications. The Area Under Curve (AUC) is 92.8%, indicating the area under the receiver operating characteristic (ROC) curve. The negative predictive value is 94.7%, representing the proportion of true negatives among all negative predictions. The true negative rate is 94.7%, providing the proportion of actual negatives correctly identified by the model. The false negative rate is 24.2%, indicating the proportion of actual positives incorrectly classified as negatives. The false omission rate is 5.3%, representing the proportion of incorrect negative predictions. The threat score is 1.496, providing an assessment of the model's accuracy in predicting positive instances while penalizing false positives. The statistical parity has a value of 1.000 indicating equal predictive performance across all classes.

The results indicate that the model demonstrates strong overall accuracy (91.2%) and balanced performance based on the F1 Score (72.6%) suggesting that the model is effective, but it is crucial to consider individual metrics for specific insights. Its AUC reflects the model's effectiveness in distinguishing between classes indicating good discriminatory power between classes. The Matthews Correlation Coefficient offers a comprehensive assessment of classification performance. The Statistical Parity indicates fairness in predictions across all classes.

Sample Output of the ANN Model

Figure 5 presents the sample output of the trained random forest classification model for student respondents 171 to 190.

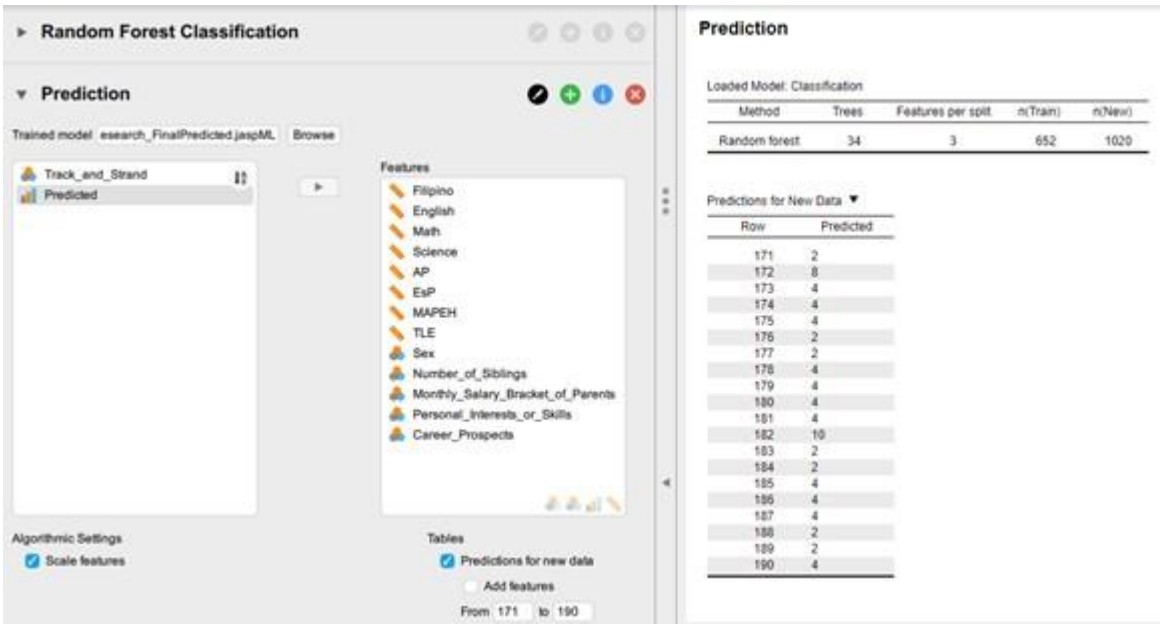

Figure 5. Sample Output of the Random Forest Classification Model

Figure 5 shows the sample output of the trained random forest classification model. It is notably clear that the model predicts the track and strand of the respondents given the above mentioned attributes or features.

CONCLUSION AND RECOMMENDATION

Conclusion

The study successfully utilized a Random Forest classification model to predict academic tracks and strands

based on various student-related attributes. The dataset included 1,020 student respondents, with attributes ranging from subject grades to personal interests and career prospects. The model achieved a validation accuracy of 79.9%, a test accuracy of 73.5%, and an out-of-bag accuracy of 92.9%. Feature importance analysis highlighted the significance of attributes like Career Prospects, Personal Interests or Skills, and Monthly Salary Bracket of Parents in predicting SHS tracks and strands. The confusion matrix revealed varying accuracies across tracks, with notable precision and recall for TVL HE and TVL ICT but challenges in predicting HUMSS. Overall model evaluation metrics demonstrated strong performance, with an accuracy of 91.2%, a balanced F1 Score of 72.6%, and good discriminatory power (AUC of 92.8%). Despite these positive outcomes, the model exhibited room for improvement in certain areas, emphasizing the need for continuous refinement and consideration of individual track performance metrics.

Recommendation

To enhance the model's performance, it is recommended to explore specific improvements for predicting the HUMSS track, where the model showed lower recall. This might involve additional feature engineering, data augmentation, or considering alternative modeling techniques. Continuous monitoring and refinement of the model, particularly in addressing class-specific challenges, will contribute to a more robust and accurate prediction system. Additionally, it would be beneficial to gather more data and explore potential correlations between attributes and SHS tracks and strands, ensuring a more comprehensive understanding of the factors influencing predictions. Regular model updates and refinement cycles should be implemented to adapt to evolving patterns in student data.

REFERENCES

1. Agarwal, S. (2013). Data Mining: Data Mining Concepts and Techniques. 2013 International Conference on Machine Intelligence and Research Advancement, Katra, India, 203-207. doi:10.1109/ICMIRA. 2013.45
2. Algarni, A. (2016). Data Mining in Education. International Journal of Advanced Computer Science and Applications(ijacsa), 7(6). doi:10.14569/IJACSA.2016.070659
3. Atienza, J. D., Hernandez, R. M., Castillo, R. L., De Jesus, N. M., & Buenas, L. E. (2022). A Deep Neural Network in a Web-based Career Track Recommender System for Lower Secondary Education. 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 1-6. doi:10.1109/ASIANCON55314.2022.9908965
4. Chen, M.-S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6), 866-883. doi:10.1109/69.553155
5. Cheng, J. (2017). Data-Mining Research in Education. ArXiv, abs/1703.10117. Retrieved from https://api.semanticscholar.org/CorpusID:16027966
6. Chinchuluun, A., Xanthopoulos, P., Tomaino, V., & Pardalos, P. M. (2010). Data Mining Techniques in Agricultural and Environmental Sciences. International Journal of Agricultural and Environmental Information Systems, 1, 26-40.
7. Dangoy, J. E., & Madrigal, D. V. (2020). Career Preferences and Factors Influencing the Career Choice of Senior High Students of a Catholic School. Philippine Social Science Journal, 3(2), 95-96. doi:10.52006/main.v3i2.235
8. Dublin, B. C., Logrosa, A. A., Sosing, M. D., & Cornillez, E. C. (2020). Factors influencing career preference of junior high school students for senior high school study. TARAN-AWAN Journal of Educational Research and Technology Management, 1(1), 29-38. Retrieved from https://journal.evsu.edu.ph/index.php/tjertm/article/view/210
9. Evans, D., Borriello, G. A., & Field, A. P. (2018). A Review of the Academic and Psychological Impact of the Transition to Secondary Education. Frontiers in Psychology, 9:1482. doi:10.3389/fpsyg.2018.01482

10. Gargano, M. L., & Raggad, B. G. (1999). Data mining - a powerful information creating tool. OCLC Systems & Services: International digital library perspectives, 15(2), 81-90. doi:10.1108/10650759910276381
11. Gera, M., & Goel, S. (2015). Data Mining – Techniques, Methods and Algorithms: A Review on Tools and their Validity. International Journal of Computer Applications, 113(18), 22-29.
12. Gestiada, G. A., Nazareno, A. L., & Roxas-Villanueva, R. L. (2017). Development of a Senior High School Career Decision Tool Based on Social Cognitive Career Theory. Retrieved from https://api.semanticscholar.org/CorpusID:52242395
13. Gestiada, G. A., Nazareno, A. L., & Roxas-Villanueva, R. M. (2017). Development of a Senior High School Career Decision Tool Based on Social Cognitive Career Theory. Retrieved from https://api.semanticscholar.org/CorpusID:52242395
14. Go, M. B., Golbin, R. A., Velos, S. P., Dayupay, J. P., Cababat, F. G., Baird, J. C., & Quiñanola, H. (2023). A data mining approach to classifying e-learning satisfaction of higher education students: a Philippine case. International Journal of Innovation and Learning, Inderscience Enterprises Ltd, 33(3), 314-329.
15. Hand, D. J. (2000). Data Mining: New Challenges for Statisticians. Social Science Computer Review, 18(4), 442–449. doi:10.1177/089443930001800407
16. Hernandez, R., & Atienza, R. (2021). Career Track Prediction Using Deep Learning Model Based on Discrete Series of Quantitative Classification. Applied Computer Science, 17(4), 55-74. doi:10.23743/acs-2021-29
17. Journals, I., & Smita, P. S. (2014). Use of Data Mining in Various Field: A Survey Paper. IOSR Journal of Computer Engineering, 16, 18-21.
18. Kaur, H. (2015). A Review of Applications of Data Mining in the Field of Education. Retrieved from https://api.semanticscholar.org/CorpusID:110301793
19. Keleş, M. K. (2017). An overview: the impact of data mining applications on various sectors.
20. Mackinnon, M. J., & Glick, N. (1999). Applications: Data Mining and Knowledge Discovery in Databases – An Overview. Australian & New Zealand Journal of Statistics, 41(3), 255-275. doi:doi.org/10.1111/1467-842X.00081
21. Nazareno, A. L., Lopez, M. F., Gestiada, G. A., Martinez, M. P., & Roxas-Villanueva, R. M. (2019). An artificial neural network approach in predicting career strand of incoming senior high school students. Journal of Physics: Conference Series, 1245. doi:10.1088/1742-6596/1245/1/012005
22. Nazareno, A. L., Lopez-Relente, M. F., Gestiada, G. A., Martinez, M. P., De Lara, M. D., & Roxas-Villanueva, R. (2021). Factors Associated with Career Track Choice of Senior High School Students. Philippine Journal of Science, 150(5), 1043-1060.
23. Puscasiu, A., Fanca, A., -I, D., & Valean, H. (2020). Data mining for identifying trends in markets. 2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania, 1-6. doi:10.1109/AQTR49680.2020.9130020
24. Rosado, J. T., Payne, A. P., & Rebong, C. B. (2019). eMineProve: Educational Data Mining for Predicting Performance Improvement Using Classification Method. 2019 World Symposium on Smart Materials and Applications (WSSMA 2019). doi::10.1088/1757-899X/649/1/012018
25. Sachin, R. B., & Vijay, M. S. (2012). A Survey and Future Vision of Data Mining in Educational Field. 2012 Second International Conference on Advanced Computing & Communication Technologies, Rohtak, India, 96-100. doi:10.1109/ACCT.2012.14
26. Sarte, E. T., & Palaoag, T. D. (2019). K-12 Students‟ Academic Status: A Data Warehouse Architecture Framework. International Journal of Recent Technology and Engineering (IJRTE), 8(2S11), 2749-2752. doi:10.35940/ijrte.B1337.0982S1119
27. Tamaskar, S. D., & Raut, A. B. (2015). DATA MINING: A TECHNIQUE WITH WIDENING APPLICATIONS AND HAVING SOME ISSUES. Retrieved from https://api.semanticscholar.org/CorpusID:64413718
28. Vallejo, O. T. (2019). Personality and Socio-Economic Factors Influencing the Choice of Academic Track among Senior High Schoolers. International Journal of Academic Research in Education and Review, 7(3), 28-33. doi:10.14662/IJARER2019.020

29. Zeedyk, M. S., Gallacher, J., Henderson, M., Hope, G., Husband, B., & Lindsay, K. (2003). Negotiating the Transition from Primary to Secondary School: Perceptions of Pupils, Parents and Teachers. School Psychology International, 24(1), 67–79. doi:10.1177/0143034303024001010