Vidisha Thakkar[1], Madhuri Patel[1], Shivam Thakkar[1] Kushdip Singh[2]

An Enhanced ML based Approach for Stream Selection of Higher Secondary Education in India: CareerX

Journal of Electrical Systems

Abstract: In the Indian education system, choosing right stream for higher secondary is critical for shaping the future careers and job prospects of the students. It can be overwhelming, as many young students have not yet developed clear interests. Using machine learning approaches, we have created a model to increase the efficiency of this selection. Data, such as academic scores, IQ test results, and personality tests labelled with the students' selected courses, were gathered from multiple sites in Gujarat, India. IQ test results, SSC and SSC grade standard scores were initially used to train the model. Subsequently, we incorporated personality traits into the model to examine their impact on stream selection. Lastly, we only used personality factors to train the models. Accuracy, precision, recall, and F1-score were among the criteria used to assess these model's performance. Our findings demonstrate that the models effectively predict appropriate streams for students, providing a standardized and data-driven approach to streamline the decision-making process. Interestingly, the findings demonstrate how personality traits have a big impact on stream choice. Based on their academic achievement, IQ, and personality traits, this framework has the potential to help students make more informed and individualized educational decisions.

Keywords: Machine Leaning, SVM, Random Forest, Logistic Regression, Decision Tree, XGBoost

1. INTRODUCTION

Early childhood education, primary education, secondary school, higher education, and vocational training are just a few of the levels and types of learning that are included in Indian education. The two two-year cycles of secondary school are Upper/Senior Secondary School and General/Lower Secondary School (Standard X). Students can select a concentration or "stream" in science, business, or the arts and humanities in upper secondary school. After class 10, selecting a stream is an important choice that has a big impact on a student's future professional achievement. This decision defines the appropriate courses to take following class 12 and lays the groundwork for specialized subjects. As a result, when choosing their stream, students need to carefully evaluate their interests, strengths, and future employment prospects. A relationship between personality types and profession choices has been demonstrated by research [1]. Even though these studies were carried out in different nations, Soo's 2013 study in Kenya highlights how crucial it is for counsellors to identify kids' personality types early on[2].

By taking all of these things into consideration, students' innate motivations and skills can be matched with appropriate fields, improving their performance in courses they find enjoyable. Additionally, choosing a stream wisely after class 10 gives them stability and reaffirms their inclinations prior to college specialization. Therefore, choosing the appropriate stream should be the main priority at these crucial points in order to open doors to fulfilling long-term professions.

However, it can be difficult for students in India to choose the correct route after the tenth grade. It's simple to feel overwhelmed and unsure of which course to take when there are so many options accessible [3]. Furthermore, students' interests and decisions are susceptible to influence from an early age, making it challenging for them to make an informed choice. Because of this, a lot of students choose their streams via conventional methods, based on their performance in the first two years of secondary school and their Standard 10th scores.

With the goal of providing students with data-driven stream options following their completion of the Standard 10th Examination, we provide 'CareerX,' a system that identifies the key characteristics needed to select the appropriate stream. Along with their stated hobbies and Standard 10th scores, this approach also takes into account their general cognitive capacity and personality, which are supported by a machine learning model. The suggestions produced by a number of models, such as Decision Tree, Random Forest, Logistic Regression,

SVM, Gradient Boost, and XGBoost, have been put into practice and assessed. To compare the influence of personality on stream selection, the models were evaluated both with and without personality parameter data.

2. RELATED WORK

Akshay Nagpal and Supriya P. Panda developed a system to suggest job paths, helping graduates align their careers with their educational background. The system uses a decision tree algorithm and string-matching techniques. The Career Path Suggestion Algorithm takes the user's career goal (G) and current education level (E) as input, producing a list of suggested career paths. It processes dataset rows by performing Simple Ratio matching between the user's goal and the "Work Position" attribute. This approach relies solely on the most recent educational qualification to predict suitable careers. However, in today's job market, most roles require a combination of qualifications, skills, and experience.To address these complexities, Hmood Al-Dossari and colleagues proposed CareerRec, a recommendation system leveraging machine learning to assist IT graduates in career selection based on their skills[4]. Using a dataset of 2,255 IT employees in Saudi Arabia, the system was tested with five algorithms—KNN, Decision Trees (DT), Bagging meta-estimator, Gradient Boosting, and XGBoost. XGBoost achieved the highest accuracy of 70.47%, effectively predicting careers among developer, analyst, and engineer roles[5].

Min Nie and collaborators introduced the ACCBOX model, which employs XGBoost with regularization to predict career choices by clustering students based on factors such as their reading interests, professional skill mastery, behavioral patterns, and family economic status.Armaan et al., in their study presented at ICCIDS-2019, developed a tool to assist students in selecting universities for master's programs by evaluating profiles against admission criteria[6]. The study compared machine learning models—Linear Regression, SVM, Decision Trees, and Random Forest—using parameters like GRE, TOEFL, GPA, and research experience, benchmarking them against various performance metrics.

K. G. Kaushalya[7] and colleagues proposed a Subject Stream Prediction system, utilizing General Certificate of Education Ordinary Level (G.C.E. O/L) exam results, skills, and career preferences to recommend subject streams. The system also offers ten alternative streams with relevant jobs and qualifications if the user is dissatisfied with the initial recommendation. Among the tested algorithms—Decision Tree, Random Forest, KNN, and SVM—the Random Forest Classifier demonstrated the highest accuracy at 72%.

3. CAREERX STREAM CLASSIFIER

We have developed a predictor named 'CarrerX. This stream predictor is developed under four phases.

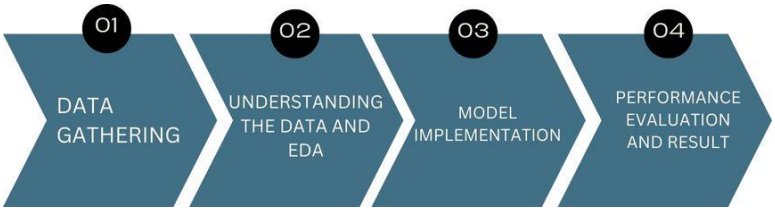

Figure 1. stream classification process

Phase 1: Data Gathering

This dataset includes a wide range of factors, such as Stanford-Binet IQ test scores, results from the Big Five personality assessment, and academic performance records.The data was collected from college students excelling in their chosen streams across various locations in Gujarat, India, including Ahmedabad, Baroda, and Surendranagar. Students provided their responses through Google Forms, which were distributed electronically to enable convenient access via any internet-enabled device. The responses were automatically collated into a centralized database for analysis.

To enhance the dataset and address any gaps, additional data was generated using Gretel AI, an advanced data synthesis platform. Gretel AI employs sophisticated algorithms to create synthetic data closely resembling real-world data while ensuring privacy and confidentiality. This augmented data increased the diversity and robustness of the dataset.

Tests Conducted to Generate Features for the Stream Selection Model are:

Stanford-Binet Test

The Stanford-Binet test [8] is a widely recognized tool for measuring intelligence, especially in the United States. The fifth edition of the test has undergone rigorous reliability assessments, including standard measurement error tests, plotting of information curves, and retest stability evaluations. Its internal consistency is considered highly accurate compared to other IQ assessments.

Big Five Personality Test (OCEAN Test)

The Big Five personality test [9] evaluates five core traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN). This test provides insights into the interplay of these parameters, offering a comprehensive understanding of personality traits.

Phase II: Understanding Data and EDA

Once the data has been successfully gathered and the necessary supplementary data has been generated with Gretel AI, it's crucial to comprehend the characteristics of the collected data. The data undergoes pre-processing through different statistical techniques and visualization tools. The gathered information is a labelled dataset made up of 5000 entries, where the data is categorized as PCB, PCM, COMMERCE, and ARTS. All the attributes have a data range of 0-100, inherently normalized.

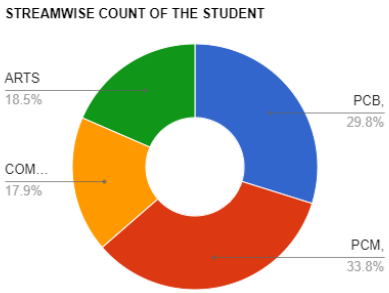

Figure 2. Stream wise Count of the Students

As shown in the Figure 3, the data is plotted and Gaussian curve is plotted above it to represent the underlying distribution of the variables, IQ, O, C, E, A and N. We can identify the value of each parameter mentioned along with their distribution of the probability.

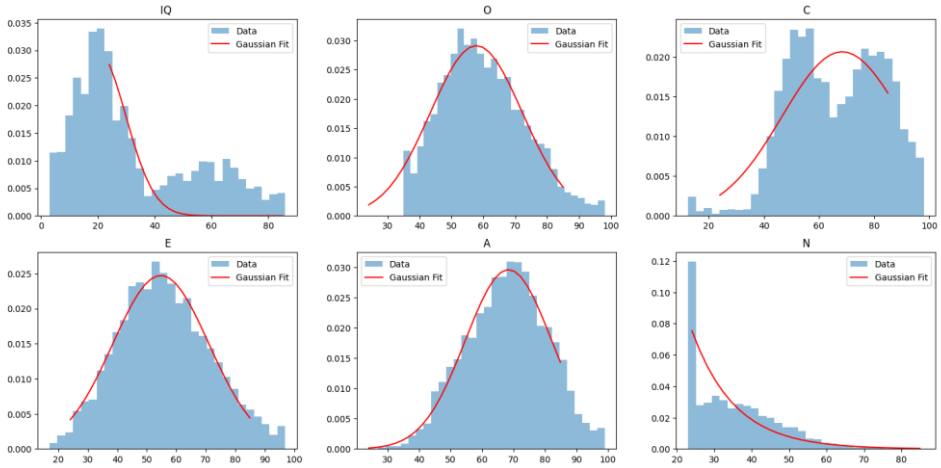

Figure 3. Gaussian Curve over the features of IQ and OCEAN

The parameters **E** and **A** closely align with a normal distribution based on their Gaussian fits. In contrast, parameters such as **IQ**, **C**, and **N** exhibit significant deviations from normality, suggesting that a Gaussian model may not be suitable for these features. The parameter **O** shows some skewness but is reasonably well represented by the Gaussian fit. While certain parameters appear to follow a normal distribution, the non-normality of several features necessitates the use of models that do not assume data normality, such as Decision Tree, Random Forest, and Gradient Boosting.

Correlation Analysis

Understanding the correlation between features is crucial for identifying relationships, simplifying data, and ensuring the robustness of statistical models. A correlation matrix provides a comprehensive view of these interrelationships and aids in data-driven decision-making. The heatmap reveals that the 10th and 12th-grade results are moderately correlated, reflecting students' consistent cognitive abilities—students who perform well in the 10th grade tend to perform well in the 12th grade. However, most other correlations are weak, indicating largely independent relationships among these variables. This independence enhances the dataset's ability to provide diverse insights for modelling.

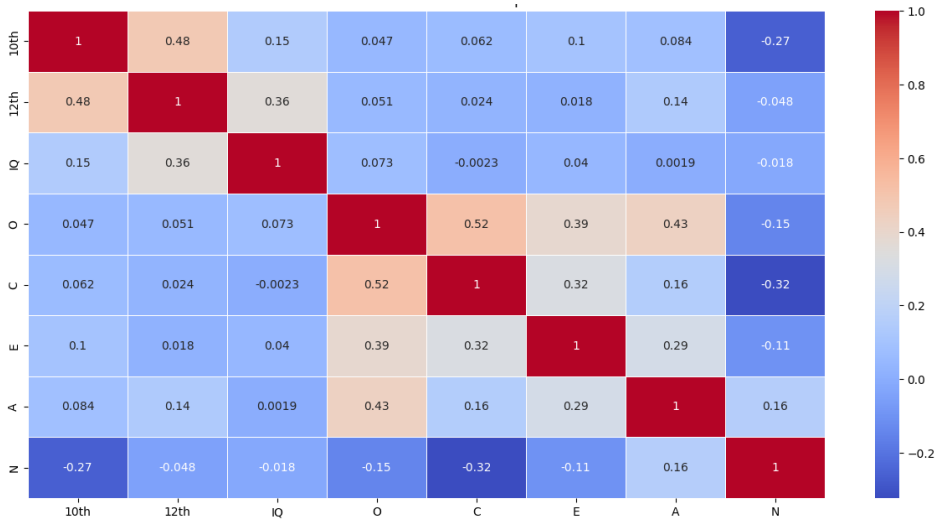

Figure 4. Correlation Matrix of the Features

The influence of personality: Figure 5 illustrates how each personality trait varies among students from various streams, making it a crucial component of our analysis.

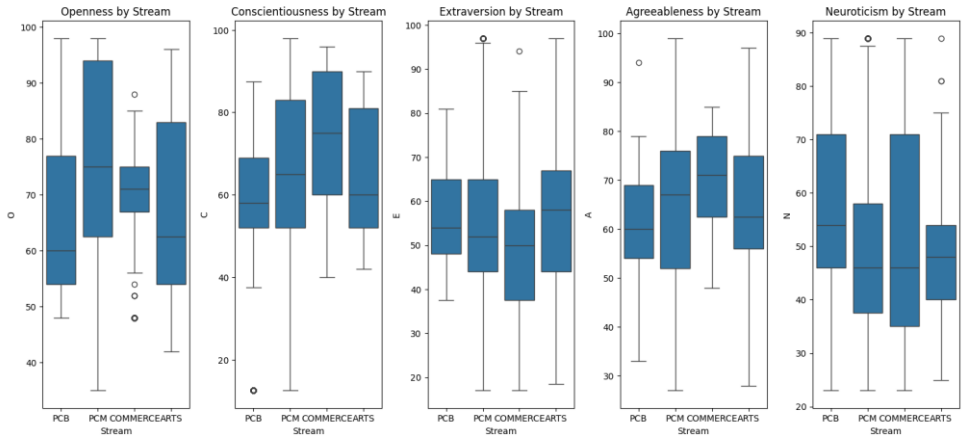

Figure 5. Box Plot OCEAN v/s Stream

Therefore, it can be concluded from a thorough examination of the dataset's various properties that the student's academic record, cognitive capacity, and general personality can all be crucial in determining their stream. Three distinct combinations of these parameters are thus used to train the model.

1) The streams were classified as Science Mathematics (PCM), Science-Biology (PCB), Commerce, and Arts based on the usage of standard 10th and 12th grade results and IQ scores.

2) To carry out the classification, personality parameters-OCEAN are added to the usual 10th, 12th, and IQ scores.

3) Only OCEAN parameters are used to perform the classification of the given dataset.

Table 1 shows Input features set.

Input Features			**Output**
Feature Set-1(FS-10	***Feature Set -2(FS-2)***	***Feature Set -3(FS-3)***	
		Personality parameters(numeric-continuous):	
10^{th}-grade percentage (numerical-continuous)	10^{th}-grade percentage (numerical- continuous)	O-Openness C-Conscientiousness E- Extraversion A- Agreeableness N- Neuroticism	
12^{th}-grade percentage (numerical-continuous)	12^{th}-grade percentage (numerical- continuous)		The output is the stream wise classification of the students.
IQ score (numerical-continuous)	IQ score (numerical-continuous)		
	Personality parameters(numeric-continuous): O-Openness C-Conscientiousness E- Extraversion A-Agreeableness N- Neuroticism		

Table 1. Input Feature Set

Phase III: Model Implementation

Choosing the right models is crucial, depending on the problem's nature and the data's properties. Therefore, the models listed below are used for choosing the student's stream:

Model	Significance of the model
Logistic Regression	It's a good starting point for classification tasks and provides a benchmark for comparing more complex models.
SVM	Effective for classification tasks with high-dimensional data. It can also capture non linear relationship using kernel function.
Decision Tree	A Decision Tree is a non-parametric supervised learning method used for classification. This method provides a clear and interpretable model, where each decision path can be easily understood.
Random Forest	This ensemble method is robust to overfitting and can handle a mix of continuous and categorical variables. It's also good for feature importance analysis, and identify which features (grades, IQ scores, personality scores) are most influential in stream classification.
Gradient Boosting	Gradient Boosting is a powerful machine learning technique used for both classification and regression tasks. It builds an ensemble of weak learners (typically decision trees) in a sequential manner, where each new tree corrects the errors of the previous ones.
XGBoost	Known for its performance and speed, XGBoost handles complex data well and provides high accuracy. It's particularly effective in classification tasks and can capture non-linear relationships in the data.

Each Machine learning models shown in above table works as discussed below.

Logistic Regression: It is a simple and interpretable model that is a good starting point for classification tasks and provides a benchmark for comparing more complex models.

Consider the feature vector **x** represent each student's data, where x=[x1,x2,…,xn] includes grades from standard 10th and 12th, IQ score (and personality scores) or personality score.The target variable **Y** represents the stream.

The softmax function here generalizes logistic regression to multiple classes (Here PCB, PCM, Commerce, Arts) as,

$$P(y=i|\mathbf{x}) = \frac{e^{\mathbf{w}_i^T\mathbf{x}+b_i}}{\sum_{j=1}^{K} e^{\mathbf{w}_j^T\mathbf{x}+b_j}} \quad (1)$$

Where, *K* is the number of classes (streams: PCB, PCM, Commerce, Arts), wi and *bi* are the weight vector and bias for class i.The cost function here would be cross entropy loss depicted as:

$$J(\mathbf{w},\mathbf{b}) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log(P(y=k|\mathbf{x}_i)) \quad (2)$$

where *m* is the number of training examples, *yik* is an indicator function that is 1 if the class label of the *i*-th example is *k*, and 0 otherwise.

A. Support Vector Machine (SVM): Support Vector Machine is a supervised machine learning algorithm effective for classification tasks with high-dimensional data, also if the data is not linearly separable, SVM performs well using the kernel functions.

Consider the feature vector **x** represent each student's data, where x=[x1,x2,…,xn] includes grades from standard 10th and 12th, IQ score (and personality scores) or personality score. The target variable **Y** represents the stream. SVM attempts to find a hyperplane that separates the classes (streams) with the maximum margin, with the decision function as f(x),

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{3}$$

Where w is the weight vector and b being the bias.

The goal is to find *w* and *b* such that the margin between the hyperplane and the nearest data points (support vectors) is maximized.

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \tag{4}$$

Subject to the constraints

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \forall i \tag{5}$$

B. Decision Tree and Random Forest: A powerful version of decision trees, Random Forest is an ensemble method is robust to overfitting and can handle a mix of continuous and categorical variables. It's also good for feature importance analysis, and can help to understand which features (grades, IQ scores, personality scores) are most influential in stream classification.

Random Forest generates multiple decision trees using bootstrap sampling. Where, each decision tree is constructed using a different bootstrap sample.

Consider the feature vector **x** represent each student's data, where x=[x1,x2,…,xn] includes grades from standard 10th and 12th, IQ score (and personality scores) or personality score.

The target variable **Y** represents the stream.

At each node in the tree, a random subset of features is selected to find the best split.

The best split is determined based on criteria like Gini impurity or entropy (information gain).

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2 \tag{6}$$

$$Entropy(D) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{7}$$

Once all trees are built, Random Forest combines their predictions. For classification, each tree casts a vote for the class label. The final prediction is made based on majority voting.

$$\hat{y} = \text{mode}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T) \tag{8}$$

where y^i is the prediction from the i-th tree and *T* is the total number of trees.

Feature Set : 10th and 12th Academic Score and IQ					
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC AUC**
XGBoost	0.87	0.870692	0.87	0.869895	0.969315
Random Forest	0.86	0.861351	0.86	0.859669	0.970991
Gradient Boosting	0.866	0.867362	0.866	0.865436	0.964688
Decision Tree	0.852	0.85208	0.852	0.851569	0.916603
Logistic Regression	0.353	0.339934	0.353	0.304698	0.576299
SVM	0.602	0.5954	0.602	0.595817	0.820074

Table 3. Performance of the model against FS-1

C.Gradient Boosting and XGBoost : By leveraging the power of gradient boosting and regularization, XGBoost provides robust and accurate classification of students into streams based on their academic performance, IQ, and personality traits.

Consider the feature vector **x** represent each student's data, where x=[x1,x2,…,xn] includes grades from standard 10th and 12th, IQ score (and personality scores) or personality score.

The target variable **Y** represents the stream.

4. IMPLEMENTATION AND RESULTS

The training dataset have been trained with Random Forest, Decision tree, Gradient Boosting, Logical Regression, SVM and XGBoost. The model is later tested on testing dataset against various performance parameters such as Accuracy, Precision, Recall, F1-score and ROC AUC.

With the feature set consisting of 10th and 12th Academic Score and IQ parameter, the performance of the said models is as mentioned in Table 3.

With the feature set consisting of 10th and 12th Academic Score, IQ and Personality parameters OCEAN, the performance of the said models is as mentioned in Table 4.

Feature Set : 10th and 12th Academic Score, IQ and OCEAN					
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC AUC**
XGBoost	0.959	0.959608	0.959	0.958569	0.997232
Random Forest	0.965	0.965588	0.965	0.964659	0.997327
Gradient Boosting	0.957	0.958718	0.957	0.956428	0.997028
Decision Tree	0.942	0.942652	0.942	0.94184	0.974775
Logistic Regression	0.557	0.581346	0.557	0.540216	0.769818
SVM	0.914	0.915013	0.914	0.912842	0.98475

Table 4. Performance of the model against FS-2

Taking into the consideration with only personality parameters of OCEAN, without, the models performance is shown in Table 5.

Feature Set : OCEAN					
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC AUC**
XGBoost	0.961	0.962669	0.961	0.960355	0.996902
Random Forest	0.962	0.96385	0.962	0.961552	0.996917
Gradient Boosting	0.962	0.963653	0.962	0.96148	0.996223
Decision Tree	0.951	0.952313	0.951	0.950632	0.982524
Logistic Regression	0.543	0.54784	0.543	0.51057	0.755793
SVM	0.923	0.922743	0.923	0.92238	0.985082

Table 5. Performance of the model against FS-3

It is significant to highlight that, overall, XGBoost is the most reliable model for this dataset, especially when "OCEAN" features are taken into account. Furthermore, it appears that most models perform better when personality traits (OCEAN) are included.

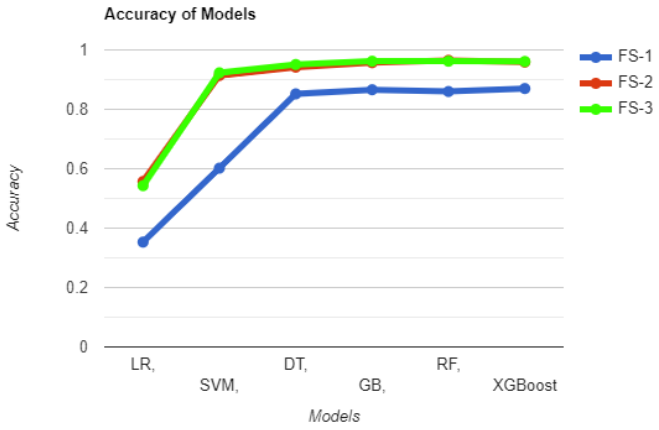

Figure 6. Accuracy Comparative of the Models against various feature-sets

Figure 6 shows that only OCEAN parameters as the feature set work as well for all implemented models.

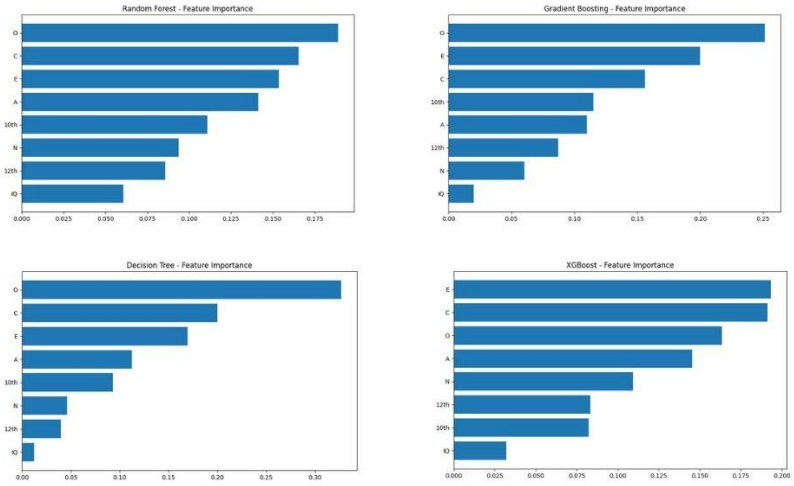

Figure 7. Importance of features in various models

As shown in figure 7, the feature importance chart for all the well performing models also shows the impact of personality parameters' impact in the overall performance of the model.

5. CONCLUSION

This paper sheds light on how several machine learning models are used to determine a student's stream after 10th grade. The combination of three feature sets is compared to the best models for multiclass classification problems and to uncover underlying data patterns. It is noteworthy that the feature set with personality parameters outperforms those without, suggesting that a student's personality has a greater influence on branch selection than both academic and IQ scores. Second, Random Forest, Gradient Boosting, Decision Tree, and XGBoost have consistently outperformed all feature set combinations among all the models taken into consideration. However, when it comes to speed and performance for the specified problem statement, XGBoost is the best model available.

REFERENCES

[1] Akshay Nagpal and Supriya P Panda. Career path suggestion using string matching and decision trees. arXiv preprint arXiv:1505.06306, 2015.

[2] Hmood Al-Dossari, FA Nughaymish, Z Al-Qahtani, Mohammed Alkahlifah, and Asma Alqahtani. A machine learning approach to career path choice for information technology graduates. Engineering, Technology & Applied Science Research, 10(6):6589–6596, 2020.

[3] Min Nie, Zhaohui Xiong, Ruiyang Zhong, Wei Deng, and Guowu Yang. Career choice prediction based on campus big data—mining the potential behavior of college students. Applied Sciences, 10(8):2841, 2020.

[4] M.S. Acharya, A. Armaan, and A.S. Antony,A comparison of regression models for prediction of graduate admissions, In 2019 international conference on computational intelligence in data science (ICCIDS) (pp. 1-5). IEEE, 2019.

[5] K. G. Kaushalya Abeywardhane , Anjalie Gamage, " Subject Stream Prediction: A Machine learning Approach to Select the Suitable Subject Stream for Senior Secondary Students in Sri Lanka",International Journal of Innovative Science and Research Technology, Volume 7, Issue 12, December, 2022

[6] El Guabassi, I., Bousalem, Z., Marah, R., & Qazdar A. A Recommender System for Predicting Students' Admission to a Graduate Program using Machine Learning Algorithms. International Journal of Online and Biomedical Engineering (iJOE), 17(02), pp. 135–147,2021.

[7] https://stanfordbinettest.com/history-stanford-binet-test

[8] https://www.truity.com/test/big-five-personality-test

[9] Momberg, Christine. The relationship between personality traits and vocational interests in a South African context. University of Pretoria (South Africa), 2006.

[10] McPherson, B., & Mensch, S. (2007). STUDENTS'PERSONALITY TYPE AND CHOICE OF MAJOR. Journal of Management Information and Decision Sciences, 10(2), 1.

[11] Onoyase, D., & Onoyase, A. (2009). The relationship between personality types and career choice of secondary school students in Federal Government Colleges in Nigeria. The Anthropologist, 11(2), 109-115.