A matching model for major orientation of computer science students based on random forest model

Di Wu

Department of Sociology, Hohai University, Jiangsu, Nanjing, 211100, China

hhshxwd@163.com

Abstract. With the advancement of computer innovation, there are increasingly major computer bearings, such as software development, network security, artificial intelligence, data science, etc. When choosing the direction of computer science, students have to consider their interests, techniques, soft skills and numerous other components to utilize their qualities and accomplish their career objectives. This article centres on how to use machine learning methods to anticipate the direction of computer science students. There's a solid relationship between students' interests, techniques, and soft skills when choosing the direction of computer science majors. By employing a random forest machine learning show to prepare students' ability characteristics, it is conceivable to anticipate which direction of specialization students are more appropriate for. Within the explore, we isolated the training, validation, and test sets, agreeing to the proportion of 6:2:2. We utilized the exactness rate to judge the model's forecast exactness. After preparing, the expectation precision of this show increments, and the precision rate of the test set comes to 96%. This implies that this machine learning show can reasonably foresee computer science students' heading and give them way better career improvement proposals. In the future, we will encourage optimising the demonstration to make strides in the forecast exactness and apply it to a broader run of areas.

Keywords: Correlation analysis, Random Forest, Major Orientation.

1. Introduction

With the fast improvement of computer innovation, increasingly branches of computer majors are covering program advancement, arrange security, counterfeit insights, information science, and numerous other areas [1]. When choosing a computer major, students must comprehensively consider their interests, techniques, soft skills, and other variables to utilize their qualities better and realize their career objectives and life values [2].

Within the current setting, orientation matching of computer science students has become the centre of scholarly investigation [3]. The inquiry about this theme points to supplying students with the most suitable computer science orientation proposals by comprehensively analyzing their capacities, techniques, soft skills, and dominance of programming languages to help them more viably arrange their personal career advancement [4].

Matching and foreseeing different aspects of computer science majors' capacities is indispensable to orientation matching research [5]. Understudy competencies incorporate, but are not constrained to, consistency in considering math and cooperation [6]. By analyzing students' capacities, it is possible to decide the most appropriate computer science orientation for them to use their potential qualities more effectively [7].

Soft skills are basic for matching computer students' directions [8]. Soft skills incorporate communication, leadership, self-management, and advancement [9]. These abilities are similarly crucial for computer experts. Soft skills can be vital in making strides in cooperation effectiveness, extending administration, and viable communication with others. In this manner, by analyzing students' soft skills, we are able to give students more comprehensive career arranging counsel [10].

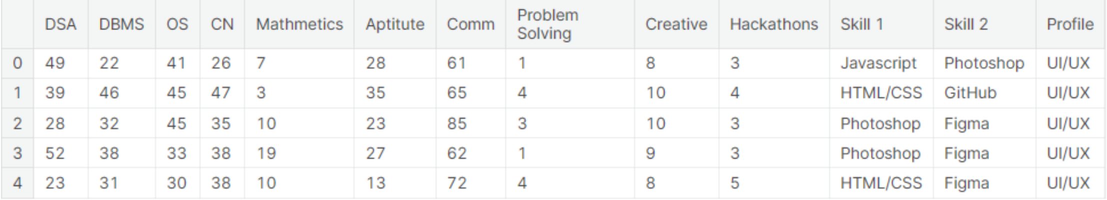
Mastery of programming language is additionally primary in coordinating computer students' courses. Students learn Java, Python, C++, and other programming languages when researching computer science. Diverse computer specialization headings require diverse programming language aptitudes. Subsequently, I can exhort students on the foremost suitable direction for their computer majors by dissecting their dominance of programming language.

In conclusion, matching computer students' bearings may be a complex and efficient issue. It requires cautious thought of components such as students' interests, techniques, soft skills and mastery of programming language. In this way, we have chosen freely available UCL datasets to analyze this issue. By analyzing these factors, we will give students the foremost practical advice when choosing the heading of their computer science major. This makes a difference for them to arrange their career more viably, reach their career objectives and realize the value of their life.

2. Data preprocessing

The dataset utilized in this article is taken from UCL's freely accessible dataset. This dataset gives generalized measurements on students' specialized and soft skills, as well as their support in competitions and capability in programming languages. The dataset has data on DSA, DBMS, working frameworks, maths and different delicate abilities and gives generalized information on each student's capacity. This information can be utilized to analyze the students' interests, techniques, soft skills, and mastery of programming language and offer assistance to decide their reasonableness for different directions of computer science majors such as UI/UX, improvement, information science, etc. Figure 1 shows a few of the data.

The profile is the foremost reasonable direction of a computer program for the students. Other markers outside of this column are the students' different capacities and soft skills. This article builds machine learning for this dataset to foresee the appropriate direction for computer program students. This show can come about by contributing different markers of students' capacities and, in this way, giving courses and recommendations to the students.



	DSA	DBMS	OS	CN	Mathmetics	Aptitute	Comm	Problem Solving	Creative	Hackathons	Skill 1	Skill 2	Profile
0	49	22	41	26	7	28	61	1	8	3	Javascript	Photoshop	UI/UX
1	39	46	45	47	3	35	65	4	10	4	HTML/CSS	GitHub	UI/UX
2	28	32	45	35	10	23	85	3	10	3	Photoshop	Figma	UI/UX
3	52	38	33	38	19	27	62	1	9	3	Photoshop	Figma	UI/UX
4	23	31	30	38	10	13	72	4	8	5	HTML/CSS	Figma	UI/UX

Figure 1. Selected datasets. **(**Photo credit: Original**)**

The indicators were visualised and analysed to observe the distribution of the data, and DSA, DBMS, OS, CN, Mathmetics and Aptitute were selected for presentation, and the results are shown in Figure 2.

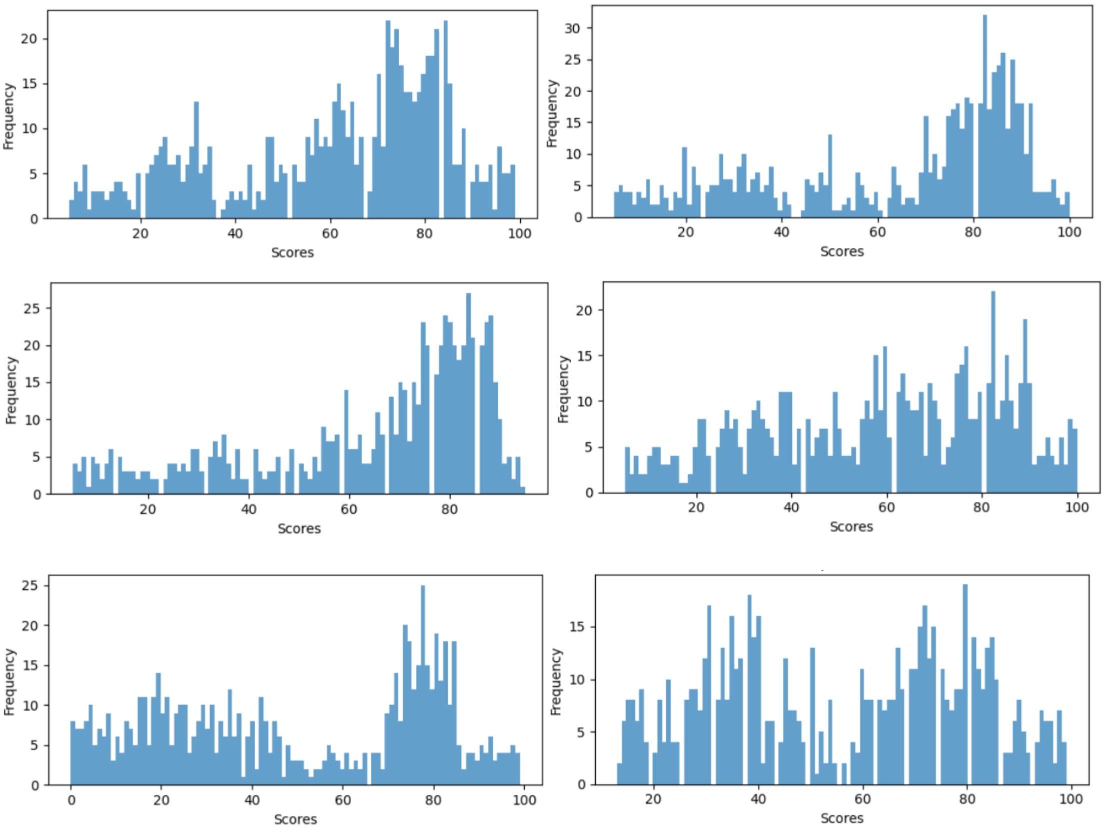


Figure 2. Visualisation and analysis. (Photo credit: Original)

3. Relevance analysis

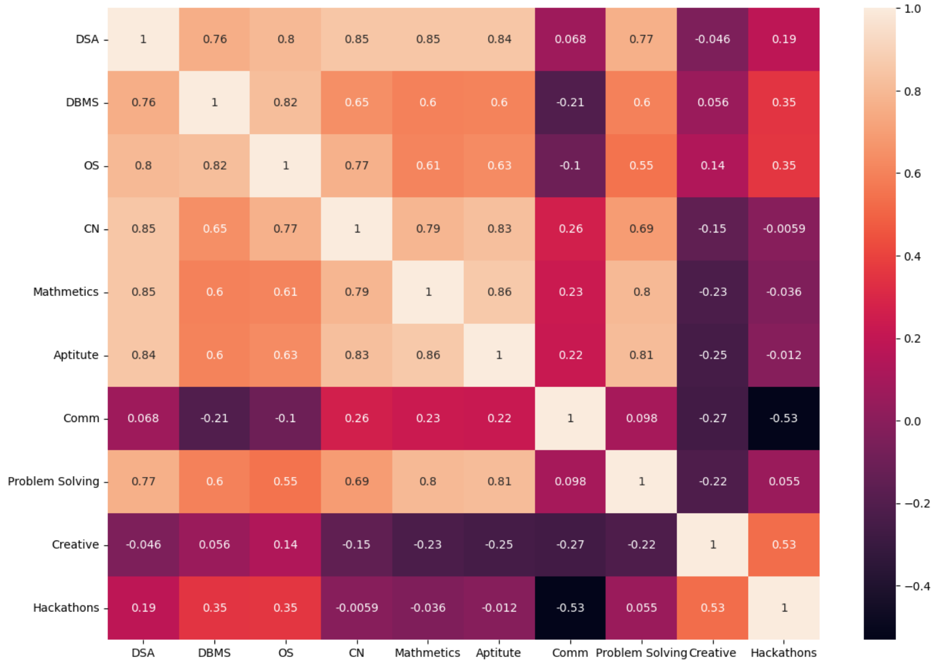


Figure 3. Correlation heat map. (Photo credit: Original)

Pearson correlation analysis is a method used to assess the strength of linear correlation between two variables. The result obtained is the Pearson correlation coefficient, which was obtained by calculating the covariance of two variables and dividing it by the product of the standard deviation of the two variables. The value of this coefficient ranges from -1 to 1. The two variables are positively correlated when the correlation coefficient is 1. When the correlation coefficient is -1, the two variables are entirely negatively correlated. When the correlation coefficient is 0, there is no linear correlation between the two variables. The correlation coefficients between the indicators are calculated. Figure 3 shows the results of the correlation heat map.

The correlation heat map shows that there is a strong correlation between the indicators of students' competence, which can be further explored using machine learning methods.

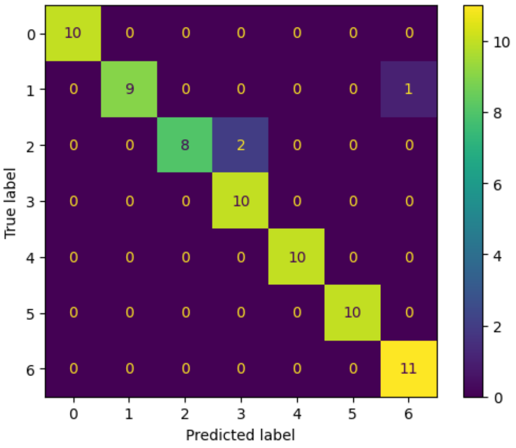
4. Random forest model

Random Forest is an integrated learning algorithm consisting of multiple decision trees. Each decision tree is a decision tree. The advantage of Random Forestd is that it can deal with high-dimensional data and solve the overfitting problem. When constructing decision trees, Random Forests randomly extracts a portion of the features of the data and uses only that portion to construct the decision tree. This avoids using too many data features, which reduces the possibility of the overfitting problem. In addition, feature selection can be performed using Random Forest. When constructing a random forest, the importance of each feature can be calculated to determine which features have a more significant impact on the classification results. This can help us select the most essential features, thus improving classification accuracy.

Constructing a random forest is usually divided into two phases: training and prediction. In the training phase, a portion of data is randomly selected from the original dataset, which is used to construct the corresponding decision tree. This process is repeated many times. Moreover, each time, the selected dataset and the construction process of each decision tree are randomized. This means that each decision tree is different. In the prediction phase, Random Forest inputs the data to be predicted into each decision tree and obtains the prediction results of each decision tree. Subsequently, the prediction results of all the decision trees are integrated through the voting mechanism to arrive at the final comprehensive prediction result.

5. Experiments and Results

The training set, validation set and test set are divided according to the ratio of 6:2:2, and the random forest machine learning model is used for training, the accuracy rate is used to judge the prediction accuracy of the model and to draw the graph of the training process with the confusion matrix, and the results are shown in Fig. 4 and Fig. 5.



	0	1	2	3	4	5	6
0	10	0	0	0	0	0	0
1	0	9	0	0	0	0	1
2	0	0	8	2	0	0	0
3	0	0	0	10	0	0	0
4	0	0	0	0	10	0	0
5	0	0	0	0	0	10	0
6	0	0	0	0	0	0	11

Figure 4. Confusion matrix. (Photo credit: Original)

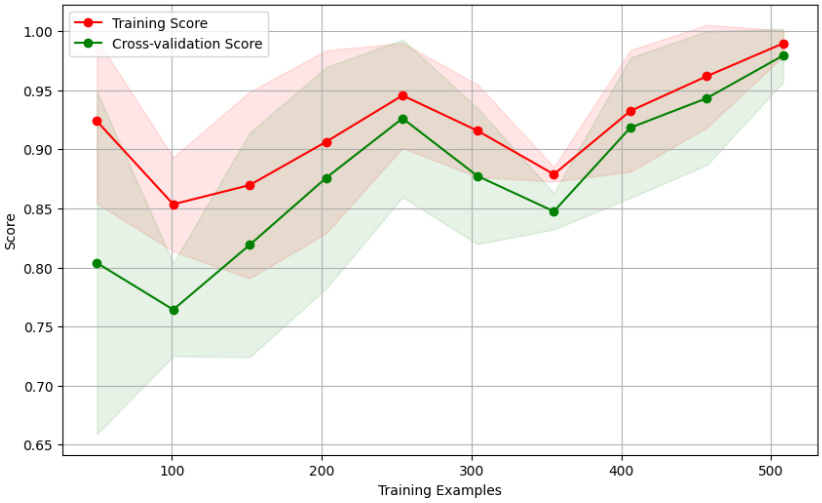


Figure 5. Training process. (Photo credit: Original)

From the results, it can be seen that the prediction accuracy of the training and validation sets is increasing as the training process proceeds. From the confusion matrix, it can be seen that 3 samples out of 71 samples in the test set are incorrectly predicted and 68 samples are correctly predicted, and the accuracy of the test set is 96%, which can be a good prediction of the direction of computer science students, and in the future, we can input the characteristics of the students' various abilities to match them with the appropriate professional direction.

6. Conclusion

This article uses machine learning methods to analyze the professional techniques and soft skills characteristics of computer science majors, which provides a basis for predicting the development direction of computer science majors.

Machine learning methods to analyze computer science majors' ability characteristics can objectively evaluate the student's abilities and strengths to predict the most suitable professional direction for each student to choose. In this article, we use the random forest machine learning model for training, and by dividing the training set, validation set and test set, we derive the results of predicting the direction of computer majors. During the training process, we find that the prediction accuracy of the training sets and validation sets increases with the number of training times. This shows that our model has some generalization ability and can adapt well to new data. Meanwhile, the accuracy of the test set is 96%. This shows that our model can reasonably predict the suitable central direction for students.

By analyzing the confusion matrix, we will see that out of 71 tests in the test set, three were erroneously anticipated, and 68 were accurately anticipated. In spite of the fact that some tests are erroneously anticipated, in common, the prescient capacity of this model show is still solid. This, moreover, outlines the potential of applying machine learning methods to foresee the significant direction of computer science students.

In conclusion, this article gives a reference for students to select their central direction in computer science, as well as a few thoughts for the application of machine learning in instruction. In the future, we will further make strides in the model to extend prediction exactness and apply machine learning strategies to more instructive fields to supply way better suggestions for students.

References

[1] Diamond B ,P. Q S ,D. M S , et al.How attending and resident reactions influence medical student specialty selection[J].Global Surgical Education - Journal of the Association for Surgical Education,2023,3(1):

[2] Shirin E D ,Madalina M ,Dionne G , et al. Medical Students' Medical Studentsâ. Perspectives on Family Planning and Impact on Specialty Choice[J].JAMA Surgery,2023

[3] L. L D ,M. S D ,Yuan H H . STEM Magnet High Schools and Student Intent to Declare a STEM Major[J].The Educational Forum,2024,88(1):79-93.

[4] Marc H ,Romuald M ,Ismaël M .Role models and revealed gender-specific costs of STEM in an extended Roy model of major choice[J].Journal of Econometrics,2024,238(2):105571.

[5] Melinda D ,Janeve D ,Aliya K , et al. The Choice! The challenges of trying to improve medical students' satisfaction with their specialty choices[J]. Canadian medical education journal,2023,14(5):49-55.

[6] Masataka N, Norihiro O, Yasuaki K, et al. Malware detection for IoT devices using hybrid system of whitelist and machine learning based on lightweight flow data[J]. Enterprise Information Systems,2023,17(9):2142854.

[7] Naveen K ,Marjorie G ,Todd O .The perceived impact of curricular and non-curricular factors on specialty interests and choice during medical school at a single center in the United States[J].BMC medical education,2023,23(1):730-730.

[8] Michael B ,M R S , Zsolt Z , et al. Profiling medical specialties and informing aspiring physicians: a data-driven approach[J].Advances in health sciences education : theory and practice,2023:1-12.

[9] Lea J ,Sarah P ,Viktor O , et al. Matching of advanced undergraduate medical students' competence profiles with the required competence profiles of their specialty of choice for postgraduate training[J].BMC medical education,2023,23(1):647-647.

[10] Xu C ,Xiang F ,Duan R , et al. An Analysis of Factors Influencing Chinese University Students' Major Choice from the Perspective of Gender Differences[J]. Sustainability, 2023,15(18):14037.