Abstract: As students progress through their academics and pursue their desired courses, it is critical that they assess their capabilities and interests in order to determine which professional path their interests and capabilities will take them to. There is a tendency among students to choose career paths based on the choices of their peers or the highest-paying roles. They are unable to recognize their own abilities and select a vocation at random, resulting in job dissatisfaction and demoralization. Furthermore, while hiring prospects, recruiters must evaluate them on a variety of levels. As a result, there is a need for a system that can assist students in determining the ideal career role for them based on their ability and other evaluation indicators. As a result of advances in machine learning, this is now achievable. This paper proposes the development of a hybrid student's career path recommender system using Ensemble technique. The system takes into account individual's personal interests and academic records to recommend correct Computer Science career path that would be best suited for them.

Keywords: Ensemble method, Machine learning, Career path, career recommender, computer science.
Development Of A Hybrid Students’ Career Path Recommender System Using Machine Learning Techniques

Developer, Business Intelligence Analyst, Security Administrator and Technical Support).

Related Work

In a study conducted by Vidyashree Ram and Muthukumaravel (2021) the research work which deals with the career prediction of students whether they will continue their education beyond their current graduation level using machine learning concepts such as Support Vector Machine, AdaBoost, Random Forest and Decision Tree. In their study, RF performed better, with a level of accuracy of 89.3 percent.

Bhumi Chitr, et al. (2017) proposed “a recommendation system that recommends University elective courses based on similarities between the courses and the course taken by the student”. On the dataset of academic records of university students, Pearson Correlation Coefficient (PCC) and Alternating Least Square (ALS) were subjected to analysis. The Alternating Least Square (ALS) performs greatly in the recommended system. The researcher also suggests using data other than student enrollment data to incorporate the student’s behavior for further guidance.

Shankarmani et al. (2020) They offered a recommender system in their work that maps students to courses based on their understanding of several job domains as the target student, in order to alleviate the problem of students having a large number of courses to pick from. The next stage was to choose a small number of clusters by removing clusters where courses were taken by different students to guarantee no points had less commonalities. Calculating the weighted median values yields the courses taken by the majority of students identical to the target student. However, the solution recommends only one programme to the student and the courses that belong to the same domain were wrongly predicted. The researcher noted the importance of recommending more than one courses to the students.

Mondal et al. (2020) they suggested “a recommendation system based on past learning details and performance that employs a machine learning approach to suggest acceptable courses to learners.” A K-Means clustering algorithm was used to classify students based on their performance ratings and collaborative filtering techniques were applied to the clusters to find appropriate courses for the students. Following that, the user (student) will be examined in the prescribed courses. In the future, the researcher proposes adding a knowledge base to uncover commonalities so that more students with comparable areas of interest and target needs can be identified.

In a research authored by Kurniadi et al. (2019) proposes “Intelligent Recommender System (IRS) architecture” for higher education. Their research framework addresses issues such as forecasting student’s performance, graduating on time and recommending subjects based on their career interests and performance. All of which are beneficial for educational interventions for student future growth. The use of machine learning and data mining techniques in anticipating and offering suggestions was inextricably linked to the success of the planned IRS framework’s development and implementation. Naïve Bayes, Support Vector Machine (SVM) and k-Nearest Neighbour (k-NN). These methods were used to solve student-related difficulties and make relevant suggestions. Decision Tree outperform other algorithms used.

Madhan et al. (2021) proposed a career prediction method that can assist students in their undergraduate or postgraduate studies in choosing the right career path for them. The model would suggest a career path for the learner based on his abilities in various subjects and locations. The algorithm used focused mainly on applicants in the computer science and engineering domain. The proposed model will assist in determining which career field the candidate should be considered for. The algorithms employed in their research were XGBoost and Decision Tree, which were implemented using the R program. The researcher suggested that the scope be broadened to include additional fields.

In a study conducted by Min Nie et al. (2020) in their paper, they proposed a model called “the Approach Cluster Centers Based On XGBOOST (ACCCBOX)’ to predict students' career choices. The experimental result of predicting students’ career choices clearly demonstrates the superiority of their research method compared to the existing state-of-the-art techniques by evaluating the behavioral data of over four thousand students.

In another research by Prasanna and Haritha (2019) proposed “feasible forecasts for student’s field selection based on their marks and choice of interest”. Choosing the correct field in IT stream is very important for his or her future. The researcher noted that if the decision went wrong it will be a mismatch between student capability, personal interest and aptitude. Their main objective was to develop a “Smart Career Guidance Recommendation System” for recommending courses and certification in the IT domain. The dataset used to build their model was skill tests and questionnaire to extract the information regarding their abilities and interests. The algorithm used ware Support Vector Machine, Decision Tree Classifier, Random Forest Classifier, Multinomial naïve bayes, Gaussian naïve bayes, Passive Aggressive Classifier, K-Nearest Neighbors, Logistic Regression, Linear Discriminant Analysis and Ada Boost Classifier where used, though Linear Discriminant Analysis outperform the rest algorithms.

Natividad et al. (2019) offers a career recommender model to assist senior high school students in deciding which career path to take. Dataset were collected from 716 senior high school students in the Philippines. To deliver appropriate recommendation, the proposed recommender system was developed using a fuzzy-based engine. They introduced 72 fuzzy model rules, and their model generates reasonable decision-making results. In order to assist prospective students in selecting relevant IT businesses in Nigeria, Ogunde and Idialu (2019) proposed a recommender model. The data was collected using an online survey that received 200 responses. The C4.5 method was used to classify the dataset and generate a decision tree model from the training dataset in this collaborative filtering recommendation technique (with 78.84 percent accuracy). Students may input their preferences and view firm suggestions by utilizing the constructed model as a
Development Of A Hybrid Students’ Career Path Recommender System Using Machine Learning Techniques

knowledge base for an extremely useful front-end web application
In a research presented by Upendran et al. (2016) they
proposed a “course recommendation system to find out the
courses which are apt for a student pursuing admission to the
college”. Typically, their predictions were based on the
present job trend or the career goal. According to the
proposed system, the prediction formulated was based on the
grades of previous academic performance and cognitive
ability of the student. A model was developed from the
legacy dataset or data from the students who have completed
the course successfully. The developed model was used for
predicting courses for new students. The idea behind this
research approach was that, when a student with specific set
of skills is successful in a course then another student with
likely set of skills may have a higher success probability in
the said course, Apriori principle was used.
In Kiran et al. (2018) research, they developed a model that
will provide recommendations for job seekers by matching
their profiles with persons with similar profile (e.g.,
professional skills and educational background). The data
used was gotten through a Google survey distributed on
social media in Pakistan. The researcher used the Apriori
algorithm to mine and extract association rules from the
collected data. The algorithms were implemented using R
Studio and 62 association rules were generated to support the
recommendations.
Grewal and Kaur (2015) study, they developed a
recommender model for educational counseling to assist
students selecting courses. The proposed viable prediction
for student course selection was based on their grades and
choice of job interests. Students interested in disciplines
such as medicine, engineering, the arts and business were the
study’s target group. Data was collected from 1500 students
in India. Clustering methods such as K-Means Clustering
algorithm were used to find structures and relationships
within the data. The Association rule was employed to
examine the associations linking the subgroups. This
procedure was used effectively to identify student traits that
correspond to individual characteristics. Lastly,
classification using fuzzy set theory and rough sets were
used. The model recommended appropriate information
depending on courses, jobs and activities to aid a student’s
decision-making process. According to the study, the
students were able to make decisions related to their studies.
The students completed a feedback form and their
satisfaction was expressed in 95% of the cases.
In a related research authored by Alhassan et al. (2020)
studied the “impact of assessment grades and online activity
data in the Learning Management Model on students’
academic performance”. Their data set included 241 records
of undergraduate students from six different courses taught
between 2017 to 2019. Their data was gotten from the
Deanship of E-Learning and Distance Education at King
Abdulaziz University. The data gathered comprises the
students’ evaluation grades and blackboard activity of the
students. Random forest, Decision tree, Multilayer
perceptron, Sequential minimum optimization and Logistic
regression were the algorithms employed in this
investigation. In terms of forecasting student academic
success, the random forest algorithm surpasses all other
algorithms, followed by the decision tree.

Methods
In our research, we used four machine learning classifiers
namely, Naïve Bayes, Decision Tree, K-Nearest Neighbors
and Support Vector Machine. Below are brief description of
the classifiers used in this research.

Decision Tree (DT)
A Decision tree is a form of predictive modeling method that
uses supervised learning. A decision tree is a graphical
depiction of all possible solutions for a given set of
circumstances. A decision tree is built from the root using a
top-down technique that incorporates data segmentation and
the calculation of data homogeneity using entropy. Both
categorical and numerical data may be used with this
approach.

Naïve Bayesian (NB)
Naïve Bayes is a well-known data classification approach. It
is based on the premise of probability theory idea and
assumes that predictors are independent of one another. In
other words, the presence of one feature in a class is assumed
to be unconnected to the presence of other features in the
class.

K-Nearest Neighbors (KNN)
K-nearest neighbor algorithm is a classifier that develops
multiple categories of cases based on similarity measures.
It's a supervised machine learning method that's used to
solve problems like classification and regression. It is non-
parametric because it makes no assumptions about how data
is distributed. Learning in a classification system is based on
'how similar' one data is to another.

Support vector machines (SVMs)
Support vector machines (SVMs) are supervised machine
learning techniques that may be used for both regression and
classification. However, they are frequently employed in
categorization problems. SVMs have a unique
implementation strategy when compared to other machine
learning algorithms. They are widely used because of their
capacity to modify a variety of continuous and categorical
variables. Support Vector Machine model can be described
as a reflection of several groups in a hyper plane
multidimensional space. The hyper plane will be produced
iteratively by Support Vector Machine in order to reduce the
error. In order to discover a maximal marginal hyper plane,
SVM divides the datasets into groups (MMH).

Methodological approach
This section of the research discusses the approach
employed to accomplish the defined objectives of the
proposed System. To achieve these objectives, the following
steps were adopted. This involves a broad process of finding
knowledge in data and highlights the use of specific machine
learning methods to a high degree. The Java programming
language was used to create this system.
Data Collection
In this phase, the first step was data collection, this was achieved through a predefined structure since the target output is a multiclass classification, the career data was collected via Electronic Google forms. The dataset consists of 700 entries from students of the Federal University Lokoja and 12 attributes.

Preprocessing
During this phase, the data collected was thoroughly checked for missing values, removal of irrelevant data that would affect our model, some irrelevant attribute that would not help to predicting the optimum career path was removed. Features such as nationality, gender and email, were removed from the dataset. Finally in this phase, dataset collected is pre-processed, imported, transformed, structured and make ready for further process in the next phase.

Modeling
The model was built with the following supervised machine learning algorithms: Naïve Bayes, Decision Tree, k-Nearest Neighbour (K-NN), Support Vector Machine and Ensemble learning to train the model and as well make recommendation. Using resampling techniques, the dataset was divided into 90% for training and 10% for testing, and the modell's performance was evaluated using testing data.

Model Evaluation
The confusion matrix was used to assess the model's performance. The model developed was evaluated and validated at this point, such that the outcomes of the metrics determine whether to do some adjustment for a further improvement or not. If the desirable performance is achieved, the model would be deployed.

Ensembles techniques
When compared to a single classifier on the dataset, ensemble approaches boost prediction accuracy. We used one ensemble strategy to increase the performance of classification algorithms in this study. Bagging classifier, one of the most prominent ensemble approaches, was used to integrate the findings of the four machine learning classifiers.

Bagging Classifier: The bagging technique was employed to decrease the calculated variability of the classifier. The bagging ensemble method splits the dataset across numerous training subsets that are chosen at random with substitution. After that, the classifier was utilized to train these data subsets. The average of the results obtained by each data subset is now employed, producing better results than a single classifier.

Figure 1 shows the structure of methodology approached used in this research work.
Results
Before applying machine learning classifiers, the dataset was visualized using a pie chart. The analysis of the dataset and implementation of classification in this study was done using the Weka tool. The student dataset was divided into 90% as training set and 10% as a testing dataset using 10-fold cross validation. Below pie charts shows the rate of performance accuracy of various classifier in percentage.
The analysis of dataset and implementation of classification in this paper has been done using Weka tool. The figures above shows the performance evaluation of each of the classifier visualization in percentage as will be tabulated in the next table. The student dataset is divided into 90% as training set and 10% as test dataset using 10-fold cross validation.

Mathematically, the evaluation metrics formulas applied on all the above respective classifier algorithms are given below:

i. Accuracy = \( \frac{TP+TN}{TP+FP+TN+FN} \)

ii. Error Rate = \( \frac{FP+FN}{TP+FP+TN+FN} \)

iii. Precision = \( \frac{TP}{TP+FP} \)

iv. Recall = \( \frac{TP}{TP+FN} \)

v. F1 Score = \( \frac{2 \times Precision \times Recall}{(Precision+Recall)} \)

Where:
- TP = True Positive
- FP = False Positive
- FN = False Negative
- TN = True Negative

In tabular form, the respective parameters measured and their values are shown below:

Table 1: Summary of the evaluation results on the classifier algorithms used

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Error Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.503</td>
<td>0.117</td>
<td>0.516</td>
<td>0.504</td>
<td>0.506</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.618</td>
<td>0.786</td>
<td>0.635</td>
<td>0.619</td>
<td>0.620</td>
</tr>
<tr>
<td>DT</td>
<td>0.820</td>
<td>0.806</td>
<td>0.831</td>
<td>0.820</td>
<td>0.821</td>
</tr>
<tr>
<td>SVM</td>
<td>0.661</td>
<td>0.766</td>
<td>0.070</td>
<td>0.680</td>
<td>0.662</td>
</tr>
</tbody>
</table>

The table above shows the Summary of the Output of Evaluation from the Algorithms used. When we compare the results from the table above, we see that the Decision Tree classifier scored the highest accuracy of 82.01 percent. According to the table above, virtually all of the classifiers predicted their accuracy between 50% and 82 percent, demonstrating that the selection of these classifiers is optimal for recommending the student's career path.

We presented the bagging ensemble approach to increase the performance of machine learning classifiers. To enhance the accuracy of machine learning classifiers, ensemble approaches are used. Table 3 shows the method's output after employing the Bagging ensemble methodology.

Table 2: Output of Ensemble Technique

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Error Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagging</td>
<td>0.906</td>
<td>0.068</td>
<td>0.907</td>
<td>0.906</td>
<td>0.906</td>
</tr>
</tbody>
</table>

According to the table above, ensemble approach enhances classifier performance when compared to simple machine learning technique, since the accuracy of the bagging classifier is 90.65 percent, which is also around 9 percent greater than the single machine learning classifier, Decision Tree.

Conclusion

The main aim of this research is to develop a hybrid machine learning system that is capable of recommending student’s career path with a clear road map. In order to achieve this aim, we introduced ensemble technique to improve the prediction accuracy of student’s performance. Machine learning techniques and ensemble methods are widely used in student performance prediction lately. Ensemble method is used to improve the result of single machine learning classifiers. In this research, four machine learning classifiers, namely Decision Tree, Naïve Bayesian, K-nearest Neighbor and Support Vector Machine, are used as base learning algorithms and then an ensemble technique is applied. Bagging was used to enhance the results of single-base learners. The best accuracy among these different machine learning classifiers is 82.01% from Decision Tree and 90.65% in bagging ensemble technique.

The results obtained by this research can be used as a clear road map in Students’ career recommendation in order to encourage the non-performing students and to pay more attention to these students to improve their performance. This can improve the quality of higher education and may be beneficial for higher education institutions and the society at large.

Review of Contribution and Achievements

- We have been able to design a system that is user friendly such that users can interact with the system through a friendly user interface.
We were able to include machine learning algorithms in the hybrid recommendation approach on two levels: the recommender algorithm itself; and the hybridization management.

**Recommendations**

The effectiveness and efficiency of this system have made it possible to solve real-world problems in the form of decision making in higher institutions. Based on the results from this work, it is highly recommended that higher institutions of learning should start adopting similar systems in their educational network to vastly improve educational experiences and render timely and effective solutions to its students. Moreover, future works should include experiments using more advanced algorithms. The comparative analysis should also be done using various techniques. Finally, data from computer science discipline was used for the analysis in this work; future works should consider data from other disciplines covering the entire academic domain in the universities.

**Suggested Areas for Future Research**

The experiment conducted in this project to validate the proposed system as a usable tool for learning and experimenting within the educational environment, resulted in several important learned lessons and future possibilities for developing the concept and building upon it. Among the most important future prospective for this research are:

- Expanding the system functionality to provide descriptions and explanations of the career topics.
- Enabling recommendation explanations for students, and thus increasing the student’s trust in the recommendation.

Conflict of Interest

The authors declare no conflict of interest, financial or otherwise.

**References**


