# Predicting Internship Participation across Machine Learning Algorithms

**Kristen Moy, University of Florida**

Kristen Moy is third year undergraduate researcher pursuing a bachelor's degree in Computer Science and a minor in Statistics at the University of Florida

**Divya Verma, University of Florida**

Divya Verma is a third year Computer Science student at the University of Florida minoring in Business Administration with a certificate in AI.

**Wyatt Harris, University of Florida**

Wyatt Harris is an Undergraduate Researcher at the University of Florida, pursuing a Bachelor of Science in Computer Engineering in the Department of Electrical and Computer Engineering within the Herbert Wertheim College of Engineering.

**Dr. Christina Gardner-McCune, University of Florida**

Christina Gardner-McCune is an Associate Professor in the Department of Computer & Information Science & Engineering at the University of Florida.

**Dr. Amanpreet Kapoor, University of Florida**

Amanpreet Kapoor is an Instructional Associate Professor in the Department of Engineering Education, and he teaches computing undergraduate courses in the Department of Computer and Information Science and Engineering (CISE).

# Predicting Internship Participation across Machine Learning Algorithms


## Abstract

Securing an internship during undergraduate studies is a valuable milestone for computing students, as it provides students with hands-on experience, enhances their technical skills, and increases their employability for full-time job opportunities. However, we have limited knowledge about our ability to identify which students in computing programs are likely or unlikely to participate in internships. To address this limitation, we conducted a survey-based study to examine predictors of internship participation among 512 undergraduate computing students enrolled at a large public university in the United States. We previously modeled the data, but in this study we evaluated the efficacy of four common predictive models - K-Nearest Neighbors, Random Forest, Logistic Regression, and Neural Networks - to understand student participation in internships. Our findings indicate that the Random Forest and Logistic Regression algorithm achieved the highest evaluation metrics in accuracy, precision, recall, specificity, and area under the curve in predicting internship participation, while the K-Nearest Neighbors and Neural Networks model performed the worst in these metrics. Our results provide baselines for future models and contribute to a better understanding of factors predicting internship participation. Additionally, they can inform academic institutions and career services in supporting students' professional development.


## 1 Introduction

Internships play a crucial role in shaping the career trajectories of undergraduate computing students by providing them with valuable practical experience and essential industry exposure [1, 2]. These experiential learning opportunities not only enhance technical proficiency - such as coding, software development, and data analysis—but also contribute significantly to professional skill development, including teamwork, communication, and problem-solving abilities [3]. Collectively, these skills improve students' overall employability and support their professional growth, better preparing them for the competitive demands of the workforce [3]. Despite these clear benefits, prior work has shown that only 60% of the graduating computing students participate in one or more internship(s) during their academic undergraduate programs [4]. The relatively low participation rate suggests that there are barriers to internship access that go beyond academic preparedness alone [1]. While prior studies have identified factors associated with internship participation through both qualitative and quantitative analysis, such as identifying barriers and modeling participation, they primarily focused on descriptive insights rather than

predictive efficacy. This work will test the efficacy of predictive models and motivate why and how these models can increase internship participation among undergraduate computing students.

Addressing this research gap, our study seeks to understand the efficacy of four machine learning (ML) algorithms - K-Nearest Neighbors (KNN), Random Forest Classifier, Logistic Regression, and Neural Networks - for predicting internship participation among undergraduate computing students. We investigate which model offers the highest accuracy and reliability in forecasting internship participation. The central research question guiding this study is: *Which machine learning algorithm is the most effective in predicting an undergraduate computing student's likelihood of participating in an internship?* Through this investigation, we aim to advance the understanding of internship participation and offer actionable insights that can inform students, academic advisors, and career services. Our work contributes to practice and literature strategies for enhancing computing students' career development and increasing internship accessibility, thereby supporting their successful transition from academia to industry.

## 2 Background

### 2.1 Impact of Internships on Career Outcomes

Galbraith and Mondal [5] explored the relationship between internships and career outcomes using a survey of recent and past graduates. The researchers identified that by participating in an internship, students were more likely to receive a full-time offer [5]. Furthermore, through the survey responses, it was identified that a majority of the respondents felt that their internships helped give them real-world experience [5]. Over 80% of the respondents noted they made a career transition within 2 years of receiving the full-time offer, which indicates the internship was effective in preparing them for the real world, including job transitions. The researchers concluded that internships do bring value not just to the student, but also to the employer [5]. Additionally, recruiters commonly evaluate internship participation as a part of the recruitment process, viewing it as a reflection of a candidate's skills and work readiness. Individuals who have participated in internships were more successful in obtaining more full-time positions and a higher initial salary [4]. This mutually beneficial effect achieved from internships supports the importance of students pursuing internships instead of just focusing on their university education.

### 2.2 Modeling & Predicting Factors that Influence Internship Participation

A study conducted by Hoekstra modeled factors influencing internship participation among undergraduate majors, identifying predictors such as race, gender, age, first-generation status, educational plans, and high-impact involvement [1]. Her findings showed that Asian American, male, older, first-generation students, or students with less involvement were less likely to intern.

Building on this study, our prior work identified key factors associated with computing undergraduate students' participation in internships [4]. We found that the year in school, socioeconomic status, involvement in activities outside the curriculum, and lower identity

diffusion scores were significantly associated with student participation in internships. Additionally, Gillespie et al. [6] identified the key psychosocial factors associated with participation in internships among college students. This study found that interns experience greater job satisfaction when they are more emotionally open and have closer relationships with their supervisors [6]. Although these studies modeled student participation in internships, our work investigates the efficacy of the predictability of models to understand internship participation.

Other studies have used machine learning models to predict students' participation in jobs [7, 8]. Recent research used machine learning algorithms to find internship effectiveness, specifically regarding the employability of a student [9]. For instance, one study focused on predicting job prospects for computer science students utilizing 25 different machine learning models, comparing the accuracy, precision, and recall scores of each model to determine which is best for prediction [2]. They concluded that the Random Forest Classifier was the best-performing model. Our work additionally studies the permutation importance of each model in order to better analyze the strengths and weaknesses of individual models.

## 3 Methods

### 3.1 Study Design and Research Question

Our study is based on Concurrent Triangulation Design, which involves data collection through multiple methods, specifically surveys and interviews. In this design, data is analyzed later after collection to allow for data triangulation [10]. Concurrent triangulation design is useful as it increases the validity of the study by using data from multiple sources. After a pilot study which was performed in 2016, a cross-sectional survey was conducted in 2019 with several participants at a large public university [4]. Interviews were held with selected participants but for this paper, we use only survey data [4]. The sample size of this study was five times larger than the initial pilot study and the findings from the pilot study were used to inform the survey and interviews. This paper utilizes data from our previous study that models this survey data to identify students' internship participation [4]. This paper utilizes various machine learning algorithms to predict the likelihood of a student receiving an internship based on the factors received in the survey. The difference between our previous and current analysis is that in this paper we create models to predict internship participation and evaluate them on a test set, while the previous study aimed to model only existing data. We aim to answer the following question: *Which machine learning algorithm is the most effective in predicting an undergraduate computing student's likelihood of participating in an internship?*

### 3.2 Research Site and Participants

The population of our study is traditional undergraduate college students enrolled in a CS-related major, such as CS, CE (Computer Engineering), and Digital Arts & Sciences (DAS), at a large US public university where acceptance is selective but flexibility to switch one's major at any time is provided. Internship participation is not mandatory for students prior to graduation, but is encouraged. The Institutional Review Board approved our study at the research site.

| Year | | | | | Gender | | Race/Ethnicity | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5-6 | M | F | White | Asian | Hispanic /Latinx | African American | Other |
| 27% | 18% | 32% | 17% | 5% | 73% | 27% | 47% | 25% | 20% | 6% | 3% |

Figure 1: Demographics of students in our corpus (N = 518)

Students were surveyed from CS1, CS2, software engineering, HCI, and OS courses. If they participated in the study, students were given the option of 1% extra credit or a chance to win a gift card. Students who opted out of the survey were given a substitute assignment of equal effort. After eliminating 41 duplicates, our survey was answered by 698 students. The survey had a response rate of 43% (N=698, Total course enrollments=1620). Data was discarded from students who were not pursuing CS-related majors or were CS minors (n=78), non-traditional students over age 24 (n=45), students who completed under 80% of the survey (n=20), students with a high proportion of pertinent missing data (n=20), students not in our undergraduate program (n=15), and non-gender-conforming students (n=2). This data was discarded for the following reasons: (1) our goal was to assess traditional college students' participation in internships, focusing on students enrolled in CS-related majors, (2) there was a lack of metric data essential for accurate analysis, and (3) the sample's certain population was an inadequate representation of the larger group.

After discarding this data, our final corpus consists of 518 students who were enrolled in CS (66%), CE (26%), DAS (4%), or double (4%) majors. The average age of the participants was 20.2 (Min=18, Max=24, SD=1.4), and Figure 1 displays demographic data representative of the CS program's student population at our university.

### 3.3 Data Collection

Data was acquired through a 74-question survey, which averaged 41.5 minutes to complete. It included 11 sections, covering demographics, professional goals and identity, degree experiences, social support, and external activity involvement, with a total of 74 questions – 49 multiple choice, 15 open-ended, and 10 short response. Participants had the option to skip any question. The questions were crafted from the findings of our prior work [4], NCWIT Student Experience of the Major Survey [11], CRA Data Buddies Survey [12], and the revised version of Bennion and Adams' Extended Objective Measure of Ego Identity Status (EOM-EIS) validated instrument [13, 14] to measure Marcia's identity statuses [15]. This study used 16 quantitative questions that were relevant for analysis and replication context.

### 3.4 Data Analysis

We conducted our data cleaning, preprocessing, and analysis in Microsoft Excel and Python, utilizing Python libraries scikit-learn, pandas, numpy, matplotlib, seaborn, and researchpy in our prior and current research [4]. In our prior research, discarding missing data and multicollinearity was addressed. Multicollinearity exists when two or more predictors are moderately or highly

correlated in a regression model. This can lead to a higher variance, overfitting, and instability in model interpretation [4]. Therefore, *age* was eliminated as it was correlated with year in school.

### 3.4.1 Dependent and Independent variables.

Our study's response (dependent) variable is a discrete binary variable denoting whether a student participated or was participating in an internship the following summer (coded as "Yes") or did not participate in an internship ("No"). Our study also utilized 26 explanatory (independent) variables: High School Courses, Enrollment Status, GPA, Employment Status, Income, Gender, Race, Moratorium Score, Diffusion Score, Achievement Score, Foreclosure Score, Year in School, Leading a Team Project, Leading a Student Organization, Practicing Interview Questions, Participation in Independent Projects, Entrepreneurial/Consulting Projects, Computing Competitions, Career Fairs, Tech Conferences, Hackathons, Online Courses, Receiving Mentoring, Providing Mentoring, Participation in Research, Participation in Other Activities outside of Classroom.

We used James Marcia's identity status theory [15] to examine students' professional identity formation and used four independent variables for identity statuses that were aggregated scores in our survey using questions from the validated identity survey instrument [14]. The four identity statuses reflect two key dimensions: exploration and commitment. The four statuses are identity diffusion (low exploration and low commitment), foreclosure (high commitment but low exploration, often influenced by authority figures), moratorium (high active exploration but low commitment), and identity achievement (high commitment following high meaningful exploration).

### 3.4.2 Models

We used four Machine Learning models to evaluate the data and predict student participation in an internship: K-Nearest Neighbors, Random Forest Classifier, Binary Logistic Regression, and a Feed-Forward Neural Network. These models were chosen for their popularity, robustness, and to test a variety of strengths and weaknesses. The Python scikit-learn library consisted of machine learning algorithms to fit, transform, and predict data.

**K-Nearest Neighbors.** This algorithm follows the Euclidean distance formula to calculate the distance between test and train samples to make predictions [16]. The KNN algorithm is used with the following equation:

$$\mathrm{D(a,\ b)} = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2}$$

This model helps to calculate the closeness of the numerical values encoded with the internship variable based on other defined points. The five nearest neighbors in this study represent the most similar candidates based on 26 features as described in 3.4.1. This algorithm predicts internship participation based on features of similar students and was selected for its widespread use in prior studies modeling student employment outcomes [17, 16]. While past work predicted job attainment following internships, our focus is on predicting internship participation itself.

**Random Forest Classifier.** Another model we used is a Random Forest Classifier (RFC) algorithm. This algorithm builds on decision trees by using multiple trees and merging them to improve accuracy, performance, stability, and prevent overfitting,compared to a single decision tree [18]. The RFC algorithm uses a different subset, each randomly sampled, from the data for each tree used and at each split in the tree the algorithm randomly selects a subset of the features in order to prevent every tree from being the same. The trees are allowed to grow as much as possible in order to capture as much detail as possible and finally the results from each individual tree are aggregated in the end through a majority vote [18]. The first equation used is the Gini Impurity which is used to decide the feature and value the tree splits based on which split minimizes the weighted average of the impurity across the child nodes. The equation is as follows [19]:

$$\text{Gini} = 1 - \sum_{i=1}^{n} p_i^2$$

where $p_i$ is the proportion of observations belonging to class $i$ in the node and $n$ is the number of classes. Once the various gini impurity values are found, the node with the lowest impurity value will be selected for the split to ensure the subsequent subgroups will be as homogenous as possible [19]. The next equation used is for the majority voting process to combine all the results [18].

$$y = \text{mode}(y_1, y_2, \ldots, y_T)$$

where $y_1, y_2, \ldots, y_T$ are the predictions made by the individual decision trees and $y$ is the final predicted class. RF emerged as the top contender in predicting job placement success from a number of similar variables in Mazid-Ul-Haque et al.'s study [2]. This algorithm was chosen because of its ability to handle overfitting, its robustness, and accuracy [2].

**Logistic Regression.** A binary logistic regression model was also used in this study. This model accepts categorical and continuous explanatory variables and presents its relationship with a dichotomous response variable [20]. Logistic regression models utilize logit, the natural logarithm of an odds ratio that can determine the successes and failures of the model. An odds ratio is a numerical representation of the likelihood that an outcome occurs given another variable. Logistic regression can be represented by the equation:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{26} X_{26}$$

where: $P$ is the probability of event $(Y = 1)$, $\beta_0$ is the constant coefficient, $X_1, X_2, \ldots, X_{26}$ are explanatory variables, and $\beta_1, \beta_2, \ldots, \beta_{26}$ are the coefficients of the explanatory variables. Positive coefficients display positive relationships between the dichotomous response variable and an explanatory variable, whereas negative coefficients represent negative relationships between these same variables. Logistic Regression was chosen for its simplicity, effectiveness, lack of baselines, and use in our prior study [4]. Binary logistic regression models utilize a random state, which ensures that the permutations are reproducible [21]. This binary logistic regression model utilized a random state of 16, selected arbitrarily.

**Neural Networks.** We also used a Neural Network model for predictions. This model mimics decision making in the human brain from learning patterns in data to make predictions [22]. Neural Networks contain many processing units, referred to as neurons, which perform a series of computations from a given input to produce an output. We use a feed-forward network, also known as multi-layer perceptrons [23]. We chose this network because it works well with basic tasks, such as regression and classification. Deng et al. [24] conducted a study on job matching predictions. The neural network model was held superior to a random forest model in effectiveness and accuracy. The feed-forward network is comprised of three layers - an input, hidden, and output layer. The input layer only receives the raw data in which the neurons in the layer correspond to the features. The hidden layers perform the computations, applying a set of weights and biases to the received data [25, 26, 27]. This also requires an activation function to determine how neurons respond to the input. We chose the Rectified Linear Unit (ReLU) because this overcomes the vanishing gradient problem, which allows the model to perform better as it backpropagates [27]. In addition, we applied the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) solver for weight optimization. LBFGS is more optimal for smaller datasets, less than 100,000 points, since this approach converges faster than the typical Adam optimizer [28]. The output layer produces the predictions and outputs them. Here, the neurons correspond to the number of classes. Our neural network model utilized a random state of 42, selected arbitrarily.

### 3.4.3 Model Evaluation: Train-Test Split.

We used an 80:20 train (n=408) and test (n=104) split to train and evaluate our models with data that was selected at random. This split was used to avoid overfitting in our machine learning models. Overfitting is when an algorithm fits too closely to the training data [5]. In our study, a train file was used to train our models on existing data, then a test file was used to predict unseen data; in this case, to predict student participation in an internship.

### 3.4.4 Model Evaluation: Metrics.

We also used the Python scikit-learn library to calculate metrics for evaluating the accuracy of prediction models such as the accuracy score, precision score, recall score, confusion matrix, and a receiving operating characteristic (ROC) curve using the metrics module. To calculate these metrics, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were computed. TP represents when the model correctly predicts that a student will have participated in an internship and TN is when it is correctly predicted that a student will not have participated in an internship. FP occurs when the model incorrectly predicts that a student will have participated in an internship, and FN indicates when the model incorrectly predicts that a student will not have participated in an internship. The accuracy score measures the overall percentage of correct predictions using the following formula:

$$\text{Accuracy Score} = = \frac{TP + TN}{TP + TN + FP + FN}$$

The precision score calculates how often positive predictions (participation in internships) are correct and is represented as:

$$\text{Precision Score} = = \frac{TP}{TP + FP}$$

The recall score computes the proportion of correctly predicted positives and is regarded as:

$$\text{Recall Score} = = \frac{TP}{TP + FN}$$

The specificity score determines how well the model predicts the negative case. The score is depicted as:

$$\text{Specificity Score} = = \frac{TN}{TN + FP}$$

The confusion matrix is a table comprising TP, TN, FP, and FN which is used to evaluate the performance of the classification model by comparing predicted outcomes with actual values. This metric is used to assess the accuracy of the predictions. A high FP or FN indicates that a low accuracy of the model. We also compute the ROC curve to show the trade-off between recall and specificity, also referred to as the true positive rate and false positive rate respectively. The ROC curve displays the performance of each model with different decision thresholds. With this curve, the performance of a model can be obtained by calculating the area under the curve (AUC) to measure diagnostic accuracy. A higher AUC closer to 1.0 reflects a stronger model performance and indicates a better ability to differenciate between classes [29].

Given our test set (n=104), it was shown that 62 of the individuals did not participate in an internship (59.6%) and 42 participated in at least one or more (40.3%). The majority class rate was 59.6%, so we used this baseline to determine which models were acceptable in comparing their overall accuracy to this rate. Additionally, models exceeding 59.6% in specificity and 40.3% in recall were considered to demonstrate meaningful predictive value in correctly identifying their respective classes.

Lastly, to identify the most important feature predictors, we used a permutation importance method. *Permutation importance* evaluates the change in model accuracy on the test data when the relationship between a feature and the target variable is broken (unbiased). This occurs when the data points of a single attribute are randomly shuffled, holding the target and all other attributes in place. The randomly shuffled attribute should cause less accurate predictions, since the shuffled data no longer corresponds with the other observed attributes. The performance deterioration of the model measures the importance of the attribute that was shuffled. This approach provides a holistic view of how significant each feature contributes to the overall model performance [30, 31]. However, it only displays the influence of a specific feature and not whether it positively or negatively affects the dependent variable. Furthermore, some features may have negative importance values, which suggests that the predictions on the shuffled data happen to be more accurate than the actual values. In such cases, these features likely have little to no impact on the model's predictions, but random chance caused the shuffled data to be more accurate [31].

| Model | Accuracy | Precision | Recall | Specificity | AUC |
|---|---|---|---|---|---|
| **K-Nearest Neighbors** | 64.42% | 57.14% | 47.62% | 75.81% | 0.67 |
| **Random Forest** | 71.15% | 64.29% | 64.29% | 75.81% | 0.77 |
| **Logistic Regression** | 71.15% | 62.50% | 71.43% | 70.97% | 0.77 |
| **Neural Network** | 58.65% | 48.94% | 54.76% | 61.29% | 0.65 |

Table 1: Performance of Machine Learning Algorithms on Internship Participation Prediction

## 4 Results

### 4.1 Individual Model Findings

#### 4.1.1 K-Nearest Neighbors

The KNN algorithm predicted internship participation with an accuracy of 64.42%, precision of 57.14%, specificity of 75.81%, and recall of 47.62%. The confusion matrix showed 20 true positives (TP), 22 false negatives (FN), 15 false positives (FP), and 47 true negatives (TN). Based on the ROC curve, the algorithm achieved an AUC of 0.67. Permutation importance revealed household income as the most influential feature, followed by identity foreclosure, career fair participation, high school CS coursework, and team-based project experience.

#### 4.1.2 Random Forest Classifier

The results from the RF Classifier yielded an accuracy of 71.15%, which outperforms the baseline classifer by around 11%. Furthermore, the model produced a precision of 64.29%, a recall score of 64.29%, and a specificity score of 75.81%. The precision and recall scores are balanced, while the specificity score indicates the model is strong at identifying students at risk of not securing an internship. The algorithm successfully predicted 27 instances as TP, while it misclassified 15 FN, 15 FP, and correctly identified 47 TN. The ROC curve generated for the RF algorithm had an AUC of 0.77, which indicates a model with good discriminatory power. The permutation importance determined career fairs, year in school, participation in hackathons, household income, and GPA as the most telling in internship participation.

#### 4.1.3 Logistic Regression

The binary logistic regression model achieved an accuracy score of 71.15% for internship participation prediction, a precision score of 62.50%, a recall score of 71.43%, and a specificity score of 70.97%. The confusion matrix for this algorithm consisted of 30 TP, 12 FN, 18 FP, and 44 TN. The ROC curve generated for the logistic regression algorithm yielded an AUC of 0.77. Permutation importance found that participation in career fairs, household income, identity diffusion scores, leading team projects, and year in school were the main determinants of internship participation.

#### 4.1.4 Neural Network

Our neural network model received an accuracy score of 58.65% for internship participation prediction, a precision score of 48.94%, a recall score of 54.76%, and a specificity score of

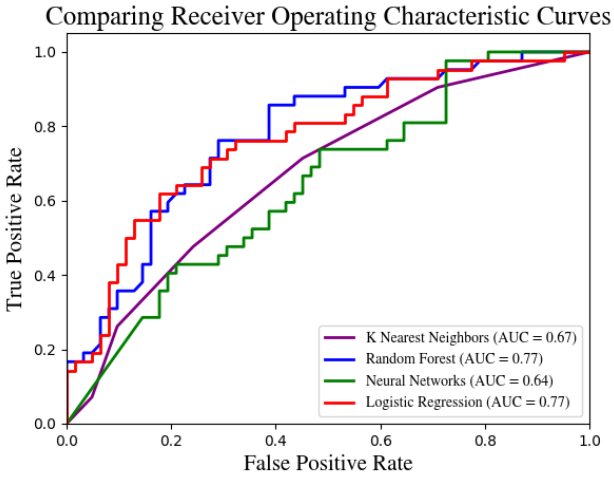


Figure 2: Receiver Operating Characteristic (ROC) Curves of Algorithms for Internship Participation Prediction

61.29%. The generated ROC curve resulted in an AUC of 0.65. The confusion matrix for this algorithm consisted of 27 TP, 15 FN, 28 FP, and 34 TN. Permutation importance was performed on this model to determine the key factors of internship participation. This analysis determined that participation in career fairs, year, employment status (full-time or part-time), provided mentoring, and computing competitions were the most impactful factors for predicting internship participation.

### 4.2 Comparing Models

A graph displaying the ROC curves of four models tested is shown in Figure 2. Our study found that Logistic Regression and Random Forest models both performed the best with an AUC of 0.77, whereas the Neural Network model performed the worst with an AUC of 0.65. Thus, our Logistic Regression and Random Forest models were best at distinguishing between positive and negative classes.

## 5 Discussion

### 5.1 Individual Model Discussions

#### 5.1.1 K-Nearest Neighbors.

The KNN model underperformed relative to other models in our analysis. As shown in Table 1, KNN achieved an accuracy of 64.42% with an AUC of 0.67, which exceeds our overall accuracy baseline. This suggests that while the model has some capacity to distinguish between students who did and did not secure internships, it lacks robustness in generalization, particularly when compared to logistic regression and random forest models, both of which achieved AUC scores of 0.77.

The KNN model had a recall score of 47.62% and a specificity of 75.81%, which is exceeds both of our baselines in these aspects. of categorizing The ROC curve reinforces the predictive ability

of KNN with an AUC of 0.67, meaning the curve is only modestly better than the diagonal reference line. We can deem this model to be reliable in predicting internships.

The permutation importance analysis identifies key factors driving predictions in the KNN model. Among the top associated features are household income, identity foreclosure, career fair participation, and experience with CS courses in high school. This indicates that students with a higher socioeconomic status and those more involved with attending career fairs or taking CS courses in high school were more likely to participate in an internship.

### 5.1.2 Random Forest Classifier

The Random Forest model performed with an accuracy rate of 71.15% and produced an AUC of 0.77, identical to the Logistic Regression model. The overall accuracy surpasses the overall baseline accuracy of 59.6%. Since random forest models are able to capture complex patterns and nonlinear relationships, this model performed generally better than the other models we used. Additionally, the built-in averaging and feature sampling helped the model reduce overfitting, while still having high predictive strength. The Random Forest confusion matrix showed a high true positive and true negative rate, which shows the reliance of recall and specificity in comparison to our baseline. Random Forest obtained a 64.29% recall, exceeding our 40.9% baseline by around 24%. Similarly, the specificity excelled by surpassing the baseline by 16.21%. These show that the model accurately placed the students in the correct category of obtaining internships. The false positive and false negative rates were equal and lower than the true positive/negative rates. Additionally, the ROC curve supports the effective ranking capability of the model, where positive cases scored higher than the negative ones. The curve's shape and high AUC confirm the model's ability to categorize accurately with minimal error.

### 5.1.3 Logistic Regression

Our Logistic Regression model had a 71.15% accuracy rate with an AUC of 0.77, making it one of the best-performing models and well exceeding the overall accuracy baseline of 59.6%. These results indicate that the variables have a clear linear decision boundary and a relatively simple pattern. The limited parameters of this model, compared to other models in the study, may also have reduced the overfitting, providing more accurate predictions. This makes our logistic regression model one of the most reliable. The model accurately predicted a significant number of true positives and negatives while limiting the number of false positives and negatives. Specifically, this model procured a recall of 71.43% and specificity of 70.97%. Both of these metrics transcend the baseline metrics by 31.13% and 11.37%, respectively. This model accurately predicted students that obtained internships correctly the most out of all the models and well surpassed the baseline, which we may regard as an acceptable model.

### 5.1.4 Neural Network

Our feed-forward Neural Network model performed the worst, with 58.65% accuracy and an AUC of 0.64. The overall accuracy did not meet our baseline of 59.6%, missing the mark by 0.95%. It struggled to generalize, misclassifying many cases due to high false positives. The

model likely underfit due to the small dataset and poorly defined decision boundaries, limiting its ability to capture complex patterns.

Though missing the benchmark for overall accuracy, the model meets the baseline in recall and specificity, exceeding it by 14.46% and 1.69%, respectively. We can state that the model is sufficient in producing well-enough results for obtaining and not obtaining an internship correctly, but generates many false negatives and positive, resulting in a insufficent overall accuracy. The precision score particularly presents the lowest scores across all model with 48.94%, representing a low rate of correct positive predictions. The model struggled to generalize, misclassifying students' internship outcomes and reducing reliability in identifying participation.

The ROC curve displayed a low AUC (0.64), meaning the model had only a limited ability to rank predictions by likelihood. Additionally, this model likely struggled with underfitting due to a smaller dataset, which prevented the model from finding more complex patterns. Despite using ReLU and LBFGS, which are suited for small datasets, the model underperformed, likely due to limited data and simple parameters leading to poor convergence.

### 5.2 Determinants of Internship Participation across Models

To visualize our model findings, we used a permutation importance graph, enabling consistent, unbiased rankings across all models. We utilized each model's built-in permutation importance function to find the key determinants of internship participation. Based on intersecting the top five most important features among the models, the features that were shown as an influential predictive feature in two or more models were career fair participation, household income, year in school, and leading a team project.

All models exhibited career fairs as influencing internship participation. Household income and year in school was deemed an important predictive feature among three different models and leading a team project was shown in two models. Establishing household income as an important predictive feature demonstrates the potential for structural barriers being present that limit participation. With these features being identified, institutions can push students to attend more career fairs, provide more resources for lower income and younger students, and open opportunities for those to lead team projects among computing students. While several of these factors have been reported by our previous work as well [4], this study also identified GPA, experience with CS courses in high school, identity foreclosure and employment status using these new models as additional factors that are associated with computing students' internship participation extending work by Hoekstra [1]. Our study builds on a prior study that modeled internship participation using a Logistic Regression model [4]. We compared different ML models, and determined the effectiveness of each model in predicting internship participation, providing a deeper understanding of the strengths and weaknesses of each model. This knowledge can be used to identify significant factors of internship participation across models and can be used by higher education stakeholders for identifying at-risk students and developing interventions that improve students' employment outcomes.

## 6 Limitations

While interpreting the results of this study, there are several limitations that should be considered. One limitation is the dataset size, which was fairly small and centralized to one institution. Therefore, the findings may have limited generalization, since internship participation can vary across different geographical regions and student cultures. This smaller dataset may have posed as a disadvantage to the Neural Network model, as they typically require large datasets to accurately identify complex patterns. Additionally, the feature set was based on primarily self-reported data collected via survey. This creates as an additional limitation, for self-reported data may contain inaccuracies and can introduce response bias. It is also important to note the limitations of permutation importance. While permutation importance allows for the consistent reporting of feature relevance across all models, it does not necessarily imply causality or a positive effect on the dependent variable. Thus, the identified determinants should be interpreted as correlational. Furthermore, while models were evaluated utilizing the standard metrics (accuracy, precision, recall, AUC), these metrics may not fully represent the practical impact of misclassifications.

Despite these limitations, this study provides a comparative baseline analysis of Logistic Regression, Neural Network, Random Forest, and KNN models. The evaluations of these machine learning models for predicting internship participation offers key insights into associated factors that can inform future research.

## 7 Conclusion

Our study evaluated the predictive performance of four ML algorithms - KNN, Random Forest, Logistic Regression, and Neural Networks - to predict internship participation. Among these, Logistic Regression and Random Forest performed the best, demonstrating the highest accuracy, precision, recall, and area under the ROC curve. Logistic Regression performed particularly well due to its simplicity and effectiveness in handling linearly separable data for binary classification. This model surpassed the other models in correctly predicting students who had no internship participation. The Random Forest model performed well at capturing non-linear relationships for a more robust performance. This model's specificity score tied with the KNN model and was the best at correctly predicting students who had internship experience. The KNN model struggled with its other metrics, likely due to its sensitivity to high-dimensional data and insufficient scaling and tuning. The KNNs performance did not perform as well as the prior two, perhaps due to its complex classification. However, we still deem this model to be acceptable from exceeding our baseline. The Neural Network model underperformed in almost all evaluation metrics likely due to the limited dataset size. While neural networks are effective at modeling complex and non-linear patterns, they typically require large datasets to perform well and often outperform traditional models when sufficient data is available. We do not consider this model to be reliable since it generated high false positives, which caused the overall accuracy to drop below our baseline.

To conclude, Logistic Regression and Random Forest models performed the best for our dataset. Despite similar performance, the models have unique strengths that can help determine when to use each model. Logistic Regression performs best when focusing on interpretability and

efficiency, while Random Forest performs best when focusing on flexibility and non-linearity. Our work contributes baseline performance of these models on a small dataset of students enrolled at a single institution. Future work could investigate our findings using new datasets from diverse geographical contexts and institution types, as well as explore alternative ML algorithms and additional features that may improve the prediction of student participation in internships.

With an equity mindset, these models should be applied cautiously and with clear intention. Given that the dataset represents students from a single institution and therefore embeds the historical patterns, the resulting models may inadvertently reflect or reinforce existing inequities. Consequently, the final model should not be used for decision-making that restricts opportunities. Rather, it should serve as a diagnostic tool to help identify structural barriers, highlight which student groups may be underserved, and inform the development of targeted support mechanisms. Interpretability plays a critical role in this process; for example, the coefficients produced by a Logistic Regression model enable practitioners to examine which factors are most strongly associated with internship participation and to assess whether those relationships reflect authentic differences in access or underlying systemic biases.

Ultimately, the model should guide interventions that broaden participation, such as proactive outreach, advising initiatives, and changes in resource allocation, while simultaneously being routinely evaluated for disparate impacts across demographic groups. When used in this capacity, the model can contribute to more equitable internship participation.

[1]

## References


[1] Caitlin E. Hoekstra. *Examining Demographic and Environmental Factors that Predict Undergraduate Student Participation in Internships*. PhD thesis, State University of New York, 2021.

[2] Md. Mazid-Ul-Haque, Md Saef Ullah Miah, Abhijit Bhowmik, S M Abdullah Shafi, and Noorhuzaimi Mohd Noor. Predictive analysis of internship and job placement success in computer science education. *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, pages 1–8, 2024. doi: 10.1109/COMPAS60761.2024.10796641.

[3] Mia Minnes, Sheena Ghanbari Serslev, and Omar Padilla. What do cs students value in industry internships? *ACM Trans. Comput. Educ.*, 21(1), March 2021. doi: 10.1145/3427595. URL `https://doi.org/10.1145/3427595`.

[4] Megan Wolf, Amanpreet Kapoor, Charlie Hobson, and Christina Gardner-McCune. Modeling determinants of undergraduate computing students' participation in internships. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2023, page 200–206, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394314. doi: 10.1145/3545945.3569754. URL `https://doi.org/10.1145/3545945.3569754`.

[5] Diane Galbraith and Sunita Mondal. The potential power of internships and the impact on career preparation. *Journal of STEM Education: Innovations and Research*, 21(2):35–39, 2020. URL `https://files.eric.ed.gov/fulltext/EJ1263677.pdf`.


[1] Note that we used ChatGPT and Writefull to verify grammar and reword small portions of the text in this paper.

[6] Iseult Gillespie, Jiahong Zhang, and Matthew Wolfgram. Psychosocial factors and outcomes of college internships: An integrative review. *Center for Research on College-Workforce Transitions, University of Wisconsin-Madison (Literature Review No. 3). http://ccwt. wceruw. org/documents/CCWT_report_LR% 20Psychosocial*, 20, 2020.

[7] Hyun Soo Kim, Sunho Kim, and Do Hyun Kim. Development of the machine learning-based employment prediction model for internship applicants. *Journal of the Semiconductor & Display Technology*, 2022.

[8] Maglioni Arana, Maria Camborda, Severo Calderon, Fidel Castro, Maribel Arana, and Bryan Huaricapcha. Prediction of labor insertion in graduates based on the pre-professional internship process through a supervised machine learning model. *2024 International Symposium on Accreditation of Engineering and Computing Education (ICACIT)*, 2024. doi: 10.1109/ICACIT62963.2024.10788618.

[9] Md. Sakibul Hassan Omi, Tanvirul Islam, Ayon Mazumder, M Saif Sartaz, Arpit Christian, and Md. Hasan Imam Bijoy. Using machine learning to predict internship effectiveness in employability. In *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–7, 2025. doi: 10.1109/ECCE64574.2025.11013797.

[10] Layne D. Bennion and Gerald R. Adams. A revision of the extended version of the objective measure of ego identity status: An identity instrument for use with late adolescents. *Journal of Adolescent Research*, 1986.

[11] Student experience of the major (sem). URL `https://www.ncwit.org/sites/default/files/resources/sem_survey_in_a_box_0.pdf`.

[12] Data buddies - center for evaluating the research pipeline. URL `https://cra.org/cerp/data-buddies/`.

[13] Gerald R. Adams. Objective measure of ego identity status: A reference manual. *Utah State University*, 1989.

[14] Layne D Bennion and Gerald R Adams. A revision of the extended version of the objective measure of ego identity status: An identity instrument for use with late adolescents. *Journal of Adolescent research*, 1(2): 183–197, 1986.

[15] James E. Marcia. Development and validation of ego-identity status. *Journal of Personality and Social Psychology. 3, 5 (1966), 551–558*, 1966. doi: https://doi.org/10.1037/h0023281.

[16] Animesh Giri, M Vignesh V Bhagavath, Bysani Pruthvi, and Naini Dubey. A placement prediction system using k-nearest neighbors classifier. In *2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)*, pages 1–4, 2016. doi: 10.1109/CCIP.2016.7802883.

[17] Jyoti Khandelwal, Gaurav Pareek, Ratul Dey, and Shilpa Pareek. The study of machine learning classification algorithm for student placement prediction. In *2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)*, pages 1–7, 2023. doi: 10.1109/IEMECON56962.2023.10092294.

[18] Sruthi. Random forest algorithm in machine learning. *Analytics Vidhya*, May 2025. URL `https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/`.

[19] Himanshi. Singh. Splitting decision trees with gini impurity. *Analytics Vidhya*, November 2024. URL `https://www.analyticsvidhya.com/articles/gini-impurity/`.

[20] Debbie Hahs-Vaughn and Richard Lomax. An introduction to statistical concepts. *Routledge*, 2013.

[21] What is scikit-learn random state in splitting dataset? *GeeksForGeeks*, 2024. URL `https://www.geeksforgeeks.org/what-is-scikit-learn-random-state-in-splitting-dataset/`.

[22] Nikolaus Kriegeskorte and Tal Golan. Neural network models and deep learning. *Current Biology*, 29(7): R231–R236, 2019. ISSN 0960-9822. doi: 10.1016/j.cub.2019.02.034. URL `https://doi.org/10.1016/j.cub.2019.02.034`.

[23] George Bebis and Michael Georgiopoulos. Feed-forward neural networks. *IEEE Potentials*, 13(4):27–31, 1994. doi: 10.1109/45.329294.

[24] Yu Deng, Hang Lei, Xiaoyu Li, and Yiou Lin. An improved deep neural network model for job matching. In *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 106–112, 2018. doi: 10.1109/ICAIBD.2018.8396176.

[25] Magdi Zakaria, AS Mabrouka, and Shahenda Sarhan. Artificial neural network: a brief overview. *neural networks*, 1:2, 2014.

[26] N. Karunanithi, Darrell. Whitley, and Yashwant K. Malaiya. Using neural networks in reliability prediction. *IEEE Software*, 9(4):53–59, 1992. doi: 10.1109/52.143107.

[27] Mohammed Aljame. Prediction of student dropout and academic performance. Master's thesis, Lund University, Lund, Sweden, 2016. URL `https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8929048&fileOId=8929050`.

[28] Nazri Mohd Nawi, Meghana R. Ransing, and Rajesh S. Ransing. An improved learning algorithm based on the broyden-fletcher-goldfarb-shanno (bfgs) method for back propagation neural networks. In *Sixth International Conference on Intelligent Systems Design and Applications*, volume 1, pages 152–157, 2006. doi: 10.1109/ISDA.2006.95.

[29] Auc roc curve in machine learning. *GeeksForGeeks*, 2025. URL `https://www.geeksforgeeks.org/machine-learning/auc-roc-curve/`.

[30] scikit-learn developers. Permutation importance vs random forest feature importance (mdi). `https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html`.

[31] Alexis Cook. Permutation importance. `https://www.kaggle.com/code/dansbecker/permutation-importance/tutorial`.