Classification of Mental Health Priority in the Tech Workplace

Olivia Lee Jo Yi

BU MET College Computer Science Department

CS699 - Dr. Jae Young Lee

Term Project Report

Fall 2022

Link to Python code:

Google Colab Python Notebook

INTRODUCTION

Open Sourcing Mental Health (OSMH) is an organization that promotes mental health in the tech workplace. The organization has observed a mental health stigma in the tech community and is dedicated to raising awareness, educating, and providing resources to support mental wellness. This analysis utilizes the data from a survey conducted by OSMH that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace.

The purpose of this project is to predict whether or not an employer prioritizes mental health as much as physical health based on a set of attributes collected from OSMH's survey. This project will implement five binary classification algorithms - K-Nearest Neighbors, Naïve Bayesian, Random Forest, Support Vector Machines, and Artificial Neural Networks - and five attribute selection methods - Chi-Square Test, Lasso Regression, Decision Tree Induction, Forward Selection, and Backwards Selection. By implementing various classification algorithms and attribute selection methods, we will be able to understand which attributes are the strongest predictors of mental health priority in the tech workplace.

MATERIALS & METHODS

The Dataset

The dataset used for this analysis is a survey conducted by OSMH that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. The dataset includes 27 attributes, representing the questions asked in the survey, and 1,259 tuples, representing the survey participants. We have identified the variable, *mental_vs_physical*, as the class attribute which represents the survey question "Do you feel that your employer takes mental health as seriously as physical health?" The class attribute includes three unique responses - "Yes," "Don't Know," and "No." From the class attribute, this project aims to determine whether an employer does (*mental_vs_physical* = "Yes") or does not (*mental_vs_physical* = "No") prioritizes mental health as seriously as physical health based on a set of attributes collected from the survey. The list below outlines the remaining 26 attributes and their respective survey questions.

- *Timestamp*: date/time the survey was conducted
- *Age*: age of the survey participant in years
- *Gender*: gender of the survey participant
- *Country*: country of origin of the survey participant
- *State*: state/territory of the survey participant
- *self_employed*: Are you self-employed?

- *family_history*: Do you have a family history of mental illness?
- *treatment*: Have you sought treatment for a mental health condition?
- *work_interfere*: If you have a mental health condition, do you feel that it interferes with your work?
- *no_employees*: How many employees does your company or organization have?
- *remote_work*: Do you work remotely (outside of an office) at least 50 of the time?
- *tech_company*: Is your employer primarily a tech company/organization?
- *benefits*: Does your employer provide mental health benefits?
- *care_options*: Do you know the options for mental health care your employer provides?
- *wellness_program*: Has your employer ever discussed mental health as part of an employee wellness program?
- *seek_help*: Does your employer provide resources to learn more about mental health issues and how to seek help?
- *anonymity*: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
- *leave*: How easy is it for you to take medical leave for a mental health condition?
- *mental_health_consequence*: Do you think that discussing a mental health issue with your employer would have negative consequences?
- *phys_health_consequence*: Do you think that discussing a physical health issue with your employer would have negative consequences?
- *coworkers*: Would you be willing to discuss a mental health issue with your coworkers?
- *supervisor*: Would you be willing to discuss a mental health issue with your direct supervisor(s)?
- *mental_health_interview*: Would you bring up a mental health issue with a potential employer in an interview?
- *phys_health_interview*: Would you bring up a physical health issue with a potential employer in an interview?
- *obs_consequence*: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
- *comments*: Any additional notes or comments

Preparation: Data Cleaning

As part of the data preparation process, the first step is to clean the data. The four steps below outline the process of data cleaning to ensure the data is sound when used for analysis.

1. Remove Duplicate or Irrelevant Data: There are a few attributes that are not relevant to the aims of the analysis. The attributes *Timestamp, Country, State,* and *Comments* were removed as they are irrelevant or highly complex in predicting the priority of a company's mental health. Additionally, we removed tuples in which the response to the class attribute,

mental_vs_physical, is “Don’t Know.” We believe this observation is irrelevant to the analysis due to the ambiguity of the response. Finally, this analysis assumes that all responses in the dataset are unique survey participants.

2. Fix Structural Errors: There are two attributes, *Age* and *Gender*, that have structural errors. For the *Age* attribute, entries less than 18 or greater than 100 were marked as NA to address these assumed user entry errors. For the *Gender* attribute, entries that are misspelled or have incongruent naming conventions were resolved to uniform responses.
3. Handle Missing Data: There are a few attributes, *Age, Gender, Self_Employed,* and *Work_Interfence,* with missing entries. For the continuous variables, missing values were replaced with the attribute mean. For categorical variables, missing values were replaced with the most frequent value of the attribute.
4. Filter Unwanted Outliers: As there is only one continuous variable, *Age*, we examined the distribution of the attribute and identified 15 observations that are considered outliers using the IQR method. We believe these outliers belong to the distribution and decided to keep the outlier values. As part of the preprocessing steps, the variables will be standardized, adjusting the extreme values in this attribute.

Preparation: Data Visualization

The next step in the data preparation processes is diving deeper into the dataset to understand the attributes in relation to the objective of the project. Figure 1 below visualizes the distribution of the values of the class attribute, *mental_vs_physical*. The pie chart illustrates the balanced responses of the class attribute’s labels with 50.2% labeled as “Yes” and 49.8% labeled as "No.” This is important as a balanced dataset is critical when implementing classification models. See Appendix B for additional visualizations of the remaining attributes stratified by the class attribute with histograms for numeric variables and mosaic plots for categorical variables.

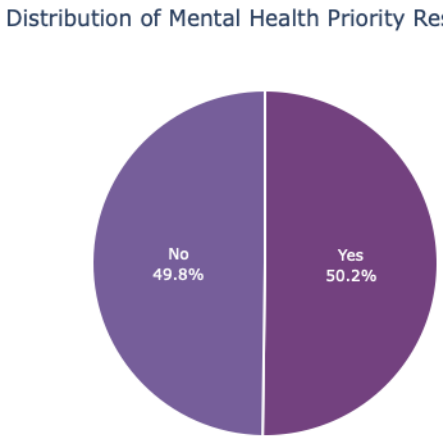


Figure 1: Distribution of Mental Health Priority Responses

Preparation: Data Preprocessing

The last step of the preparation processes is preprocessing the data so that it can be fed into the classification algorithms. The first step is encoding the categorical attributes of the dataset into a numeric form. We employed Label Encoding and One-Hot Encoding to transform the categorical attributes. Label Encoding converts labels to a value between 0 and the number of distinct labels minus 1. Label Encoding was utilized for 1) binary class variables or 2) ordinal variables where the categories of the attribute have a clear order. One-Hot Encoding creates additional features based on the number of unique labels in the categorical feature, where each unique label in the category is added as a feature. One-Hot Encoding was utilized for multi-class variables with no order. The list below summarizes the encoding of the categorical variables' labels into numeric representations. Finally, the attributes are standardized (i.e. scaling the data so that the mean of the data becomes zero and the standard deviation becomes one) to account for the large differences between the attributes' ranges. The data is now preprocessed and ready for implementation.

- *Gender: Male = [1, 0, 0], Non-Binary = [0, 1, 0], Female = [0, 0, 1]*
- *self_employed: No = 0, Yes = 1*
- *family_history: No = 0, Yes = 1*
- *treatment: No = 0, Yes = 1*
- *work_interfere: Never = 0, Rarely = 1, Sometimes = 2, Often = 3*
- *no_employees: 1-5 = 0, 6-25 = 1, 26-100 = 2, 100-500 = 3, 500-100 = 4, More than 1000 = 5*
- *remote_work: No = 0, Yes = 1*
- *tech_company: No = 0, Yes = 1*
- *benefits: No = 0, Don't Know = 1, Yes = 2*
- *care_options: No = 0, Not Sure = 1, Yes = 2*
- *wellness_program: No = 0, Don't Know = 1, Yes = 2*
- *seek_help: No = 0, Don't Know = 1, Yes = 2*
- *anonymity: No = 0, Don't Know = 1, Yes = 2*
- *leave: Very difficult = 0, Somewhat difficult = 1, Don't Know = 2, Somewhat easy = 3, Very easy = 4*
- *mental_health_consequence: No = 0, Maybe = 1, Yes = 2*
- *phys_health_consequence: No = 0, Maybe = 1, Yes = 2*
- *coworkers: No = 0, Some of them = 1, Yes = 2*
- *supervisor: No = 0, Some of them = 1, Yes = 2*
- *mental_health_interview: No = 0, Maybe = 1, Yes = 2*
- *phys_health_interview: No = 0, Maybe = 1, Yes = 2*

- *obs_consequence: No = 0, Yes = 1*

Classification Algorithms

This project implements five classification algorithms to predict whether or not an employer takes mental health as seriously as physical health. To classify the attribute, *mental_vs_physical*, we have selected the following five models: K-Nearest Neighbor, Naive Bayesian, Random Forest, Support Vector Machines, and Artificial Neural Networks. The list below briefly describes the theory of each classification model.

- K-Nearest Neighbors: The K-Nearest Neighbors classification model is a "lazy learner" that works by calculating the distance between the test instance and all the training points. The algorithm selects the k number of points (called nearest neighbors) which is closest to the test instance. The kNN algorithm calculates the probability of the test instance belonging to the classes of the k-training data and the class with the highest probability is selected. The elbow method is employed to determine the optimal number of neighbors that minimizes the error rate of the classifier. See Appendix C for visualization of the elbow method.
- Naïve Bayesian: The Naïve Bayesian classification model is a statistical classifier, meaning it uses probability to predict labels. The algorithm implements Bayes' Theorem of conditional probability to calculate the probability of each class label, given the data point belongs to a particular class. The predicted class is the class with the highest probability.
- Random Forest: The Random Forest classification model consists of a large number of decision trees that operate as an ensemble. The algorithm works by 1) selecting random samples from the dataset, 2) constructing a decision tree for each sample and predicting the classification value for each decision tree, 3) performing a vote for each predicted result, and 4) selecting the prediction result with the most votes as the final prediction.
- Support Vector Machines: The Support Vector Machines (SVM) classification model works by constructing a hyperplane in a multidimensional space to separate the classes. SVM generates the optimal hyperplane in an iterative manner, minimizing the error and maximizing the margin between the support vectors and the dataset. The goal of the algorithm is to find a maximum marginal hyperplane that best divides the classes.
- Artificial Neural Network: The Artificial Neural Network (ANN) classification model is a deep learning method composed of multi-layer fully-connected neural nets. The algorithm works by feeding the inputs into hidden layer(s) composed of nodes. A given node takes the weighted sum of the inputs and passes it through a non-linear activation function. The output of a node becomes the input of another node in the next layer. Finally, the output layer of the model uses the softmax function to classify the observation. Backpropagation is employed to update the weights and biases of the model and minimize the error.

Attribute Selection Methods

This project implements five attribute selection methods to determine the best combination of attributes in predicting whether or not an employer takes mental health as seriously as physical health. Attribute selection methods can be categorized into three techniques - filter methods, wrapper methods, and embedded methods. This project sought to implement a diverse group of attribute selection techniques by executing the following methods: Chi-Square Test, Lasso Regression, Decision Tree Induction, Forward Selection, and Backwards Selection. The list below briefly describes the theory of each attribute selection method.

- Chi-Square Test: The Chi-Square attribute selection method uses the Chi-Square test to determine the independence of two events. The Chi-Square test is performed between each attribute (predictor variable) and the class attribute variable (response variable). If the Chi-Square test concludes that the predictor variable is independent of the class variable, the predictor variable is discarded. Otherwise, the predictor is considered to be important in predicting the class attribute and is selected for the classification model.
- Lasso Regression: The Lasso Regression attribute selection method fits Lasso regression on a scaled version of the dataset. Lasso regression uses a cost function to compute coefficients of each attribute. If the predictor variable and class variable are linearly correlated, the cost function will increase and Lasso Regression will push the coefficient of the less important attributes to zero. Attributes with coefficients of zero are removed during attribute selection and the remaining attributes are selected for the classification model.
- Decision Tree Induction: The Decision Tree Induction attribution selection method uses the Decision Tree Regressor to predict the importance of each attribute. The Decision Tree Regressor is fit on the dataset and uses entropy to calculate the information gain of each attribute. Attributes with an importance of less than 0.5 are removed; attributes with an importance greater than or equal to 0.5 are selected for the classification model.
- Forward Selection: The Forward Selection attribution selection method is a wrapper method where the process is based on a specified machine learning algorithm. Forward Selection is a greedy search approach by evaluating all the possible combinations of features against the performance measure. The method starts with a null model and then fits the model with each individual attribute one at a time and selects the attribute with the minimum p-value. The algorithm repeats this process and iteratively selects additional attributes until the pre-specified stopping criterion is reached or all attributes all included in the model.
- Backwards Selection: The Backwards Selection attribute selection method is a wrapper method where the process is based on a specified machine learning algorithm. Similar to Forward Selection, Backwards Selection is a greedy search approach by evaluating all the possible combinations of features against the performance measure. The method begins with all attributes under consideration (the full model). Next, the algorithm iteratively

removes an attribute with the least significance one at a time. The algorithm continues until a pre-specified stopping criterion is reached or no attribute is left in the model.

Data Mining Tools and Procedure

This project was implemented in Python to prepare and preprocess the data, perform the attribute selection methods, implement the classification algorithms, and evaluate the results. Specifically, we utilized the Python library, Scikit-learn, for preprocessing requirements (one hot encoding, label encoding, standardization, etc.), data validation (cross validation, train/test split, etc.), attribute selection selection (chi-square, lasso regression, decision tree induction, forward selection, and backwards selection), classification algorithms (kNN, Naïve Bayesian, Random Forest, and SVM), and evaluations metrics (accuracy, confusion matrix, precision, recall, f1-measure, etc.). The library, Keras, was utilized to build and train the Artificial Neural Network with specified hyperparameter tuning. Lastly, the library, Plotly, was used to visualize the dataset and results with various charts and graphs. We chose to implement this project in Python as it allows for more flexibility and autonomy in the knowledge discovery process.

This project is housed in Google Colab to organize the Python code with text cells for documentation and comments. The Colab is divided into three main sections: Preparation, Analysis, and Evaluation. To begin executing code, the Preparation section 1) loads relevant libraries and packages, 2) reads the initial dataset into a Pandas dataframe, 3) cleans the data with a four step procedure, 4) explores the data with visualization methods, and 5) implements the necessary preprocessing for the following sections. The Analysis section of the Colab starts by creating definitions for the classification algorithms: K-Nearest Neighbors, Naïve Bayesian, Random Forest, Support Vector Machines, and Artificial Neural Networks. Next, 10-fold cross-validation was implemented on the entire dataset for the five classification algorithms. The average Accuracy, True Positive Rate, False Positive Rate, Precision, Recall, F1-Measure, MCC, and ROC AUC of the 10-folds was collected to examine the effectiveness of classification algorithms. Next, the dataset was split into a training set (66%) and a test set (34%) to implement the five attribute selection methods: Chi-Square Test, Lasso Regression, Decision Tree Induction, Forward Selection, and Backwards Selection. The attributes selected from each method were used as the predictor variables for each of the five classification algorithms. The Accuracy, Confusion Matrix, True Positive Rate, False Positive Rate, Precision, Recall, F1-Measure, MCC, and ROC AUC of the 25 models was collected to examine the effectiveness of the attribute selection methods and the classification algorithms. Finally, the Evaluation section of the Colab summarizes and visualizes the findings from the attributes selected and performance results, determining the overall best model. The ipynb file of the Python code is attached or click the link to visit the Colab directly (Colab Link: https://colab.research.google.com/drive/1YVfjA8N7mahFWUtoVRuafrSicYaLtEs6?usp=sharing)

RESULTS

Attributes Selected

Attribute selection methods were employed to minimize the dimensionality of the dataset in hopes of increasing the accuracy and efficiency of the classification algorithms. As discussed, the project implements five attribute selection methods - Chi-Square Test, Lasso, Regression, Decision Tree Induction, Forward Selection, and Backwards Selection. A threshold for selection criteria was set for each attribute selection method. The Chi-Square Test attribute selection method removed attributes with a p-value greater than 0.05, determining that these attributes were independent of the response variable. The Lasso Regression attribute selection method removed attributes in which the variable coefficient was pushed to zero, indicating that the attribute was not a strong predictor of the response variable. The Decision Tree Induction attribute selection method removed attributes with an importance less than 0.05, indicating that these attributes were not valuable to information gain. Lastly, the Forward Selection and Backwards Selection attribute selection methods added/removed attributes that contributed to a statistically significant performance improvement of the classifier. See Appendix D for a visualization of the criteria and results from the attribute selection methods.

We are interested in which attributes were selected by which attribute selection methods. Figure 2 below summarizes the results from the attribute selection methods as well as the total number of attributes selected for each method. While the Chi-Square Test method selected 13 attributes, the Lasso Regression method only selected 3 attributes. It will be interesting to understand if the classification algorithm's performance increases or decreases with more or less predictor variables. Additionally, notice the attributes that were selected multiple times or not at all. Figure 3 visualizes the number of times an attribute was selected (out of five selection opportunities). Theoretically, the attributes that were selected often are likely strong predictors in determining whether or not an employer takes mental health as seriously as physical health. Notice that the attribute, *leave*, was selected during all five methods, demonstrating that workplaces where it is easy to take medical leave could be associated with an environment of prioritizing mental health. Other attributes that were selected often (3 - 4 times) include *wellness_program, no_employees, supervisor, self_employed, mental_health_consequence,* and *anonymity*. One attribute, *remote_work*, was selected zero times, demonstrating that it was not substantial in predicting workplace mental health priority.

	Chi-Square	Lasso Reg	Decision Tree	Forward Selection	Backwards Selection
Age	0	0	1	1	0
Gender_Male	0	0	0	1	0
Gender_Non-binary	0	0	0	1	1
self_employed	1	0	0	1	1
family_history	0	0	1	0	1
treatment	0	0	0	0	1
no_employees	1	0	1	1	1
remote_work	0	0	0	0	0
tech_company	0	0	0	0	1
benefits	1	0	0	0	0
care_options	0	0	0	1	1
wellness_program	1	1	1	1	0
seek_help	1	0	0	0	0
anonymity	1	0	0	1	1
leave	1	1	1	1	1
mental_health_consequence	1	1	1	0	0
phys_health_consequence	1	0	0	0	0
coworkers	1	0	0	0	0
supervisor	1	0	0	1	1
mental_health_interview	1	0	0	0	0
phys_health_interview	0	0	1	0	0
obs_consequence	1	0	0	0	0
# Attributes	13	3	7	10	10

Figure 2: Table of the Attributes Selected by each Attribute Selection Method

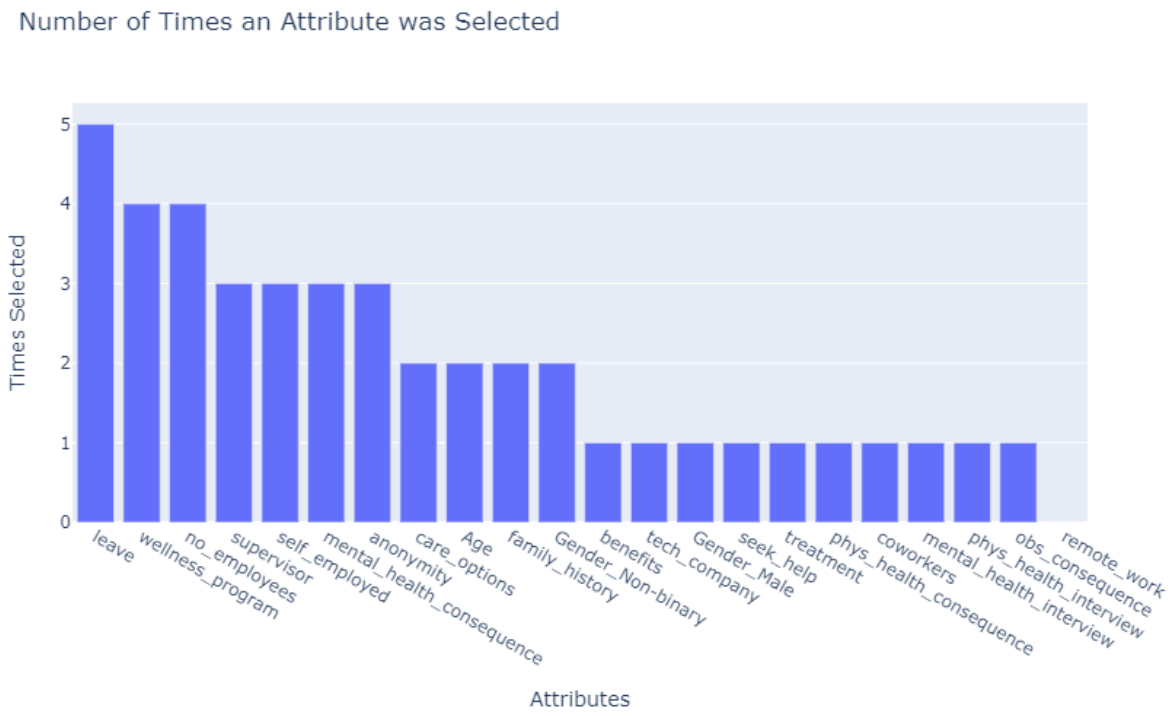


Figure 3: Bar Chart of the Number of Times an Attribute was Selected

Detailed Performance Results

10-Fold Cross Validation Classification Algorithms Results

The following section includes the detailed performance results from the 5 classification algorithms using 10-fold cross validation. Note that all attributes were used for these models as a baseline for evaluating performance. For each classification model, the performance results include the True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F1 Measure, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristics (ROC AUC).

1. K-Nearest Neighbors

	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.821220	0.206363	0.821220	0.782353	0.799331	0.610924	0.803529
Yes	0.793637	0.178780	0.793637	0.824706	0.806679	0.610924	0.803529
Weighted Avg	0.807428	0.192572	0.807428	0.803529	0.803005	0.610924	0.803529

2. Naïve Bayesian

	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.822489	0.191215	0.822489	0.800000	0.808983	0.627955	0.812353
Yes	0.808785	0.177511	0.808785	0.824706	0.814826	0.627955	0.812353
Weighted Avg	0.815637	0.184363	0.815637	0.812353	0.811904	0.627955	0.812353

3. Random Forest

	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.810395	0.175625	0.810395	0.823529	0.814071	0.62985	0.812521
Yes	0.824375	0.189605	0.824375	0.801513	0.809243	0.62985	0.812521
Weighted Avg	0.817385	0.182615	0.817385	0.812521	0.811657	0.62985	0.812521

4. Support Vector Machines

	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.829550	0.174607	0.829550	0.820588	0.823044	0.651602	0.82416
Yes	0.825393	0.170450	0.825393	0.827731	0.824309	0.651602	0.82416
Weighted Avg	0.827472	0.172528	0.827472	0.824160	0.823677	0.651602	0.82416

5. Artificial Neural Networks

	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.819684	0.206141	0.819684	0.786317	0.799572	0.609714	0.802971
Yes	0.793859	0.180316	0.793859	0.819624	0.803777	0.609714	0.802971
Weighted Avg	0.806771	0.193229	0.806771	0.802971	0.801674	0.609714	0.802971

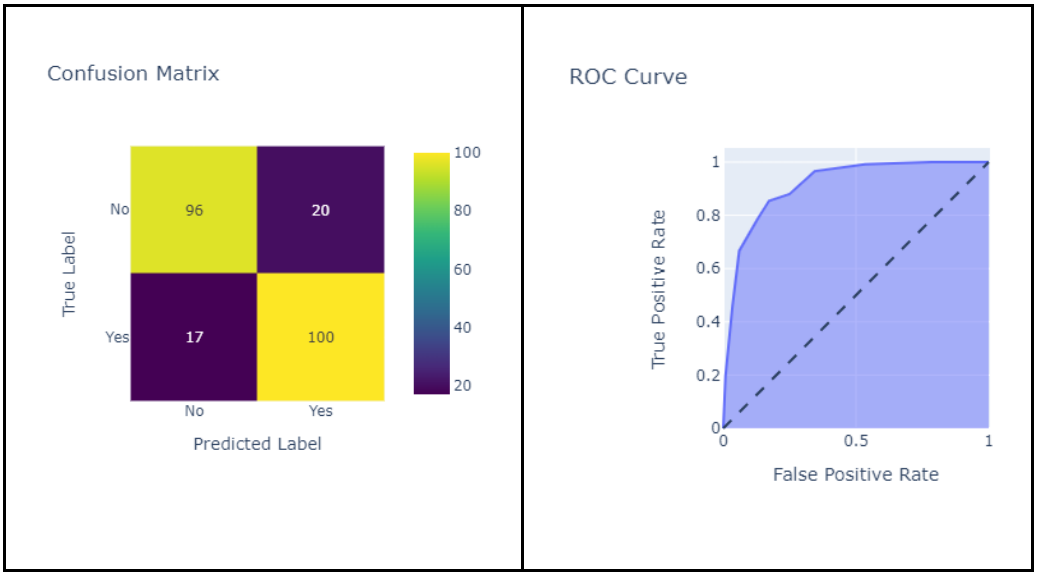
Attribute Selection Methods & Classification Algorithms Results

The following section includes the detailed performance results from the 25 models (5 attribute selection methods x 5 classification algorithms). The list of attributes selected from each attribute selection method is provided. For each classification model, the performance results include the confusion matrix, True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F1 Measure, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristics (ROC AUC).

1. Chi-Square Test

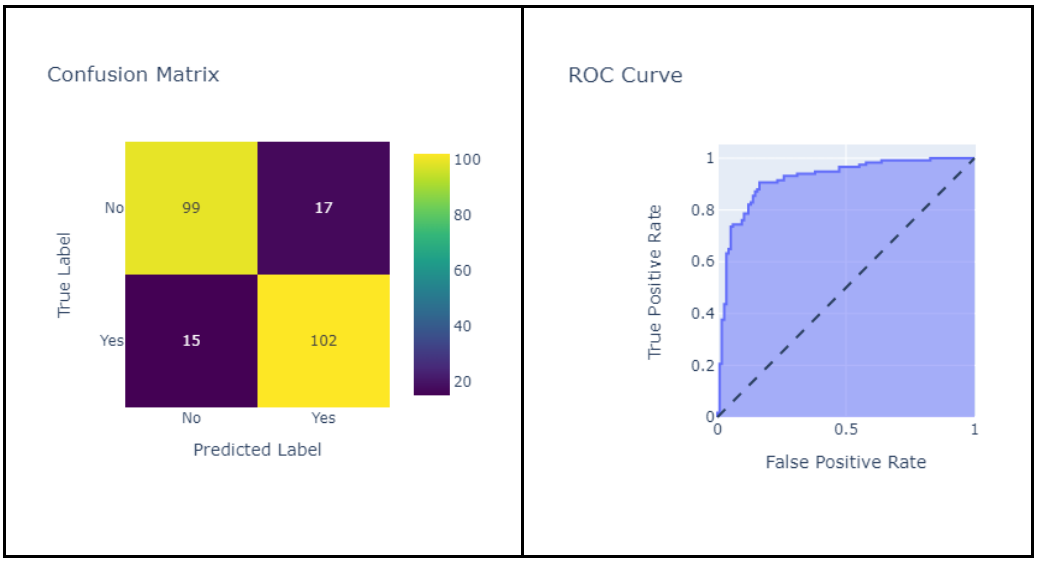
- *Attributes:* 'self_employed', 'no_employees', 'benefits', 'wellness_program', 'seek_help', 'anonymity', 'leave', 'mental_health_consequence', 'phys_health_consequence', 'coworkers', 'supervisor', 'mental_health_interview', 'obs_consequence'

1.1 K-Nearest Neighbors



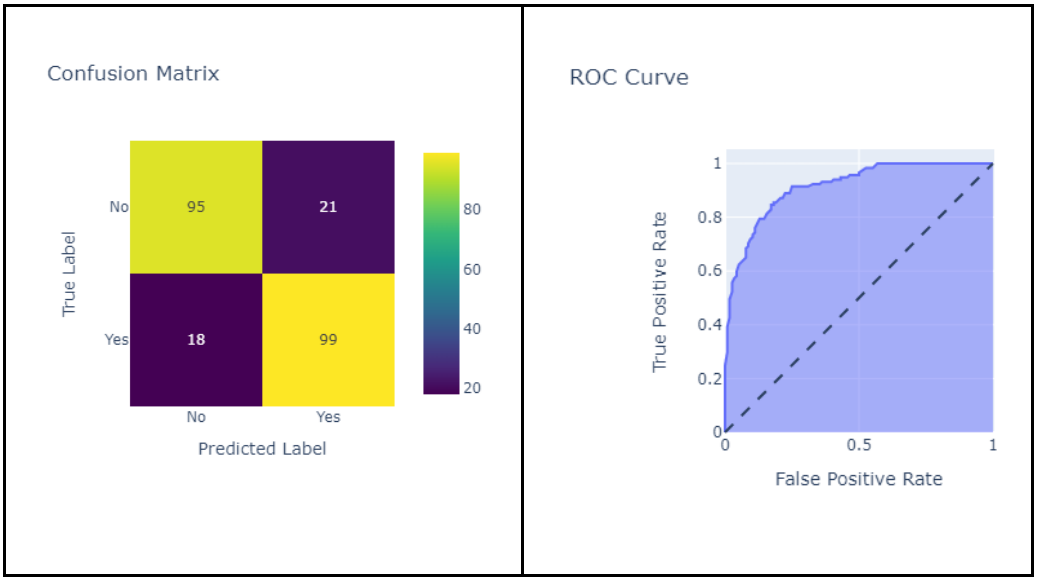
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.827586	0.145299	0.849558	0.827586	0.838428	0.682589	0.914309
Yes	0.854701	0.172414	0.833333	0.854701	0.843882	0.682589	0.914309
Weighted Avg	0.841202	0.158915	0.841411	0.841202	0.841167	0.682589	0.914309

1.2. Naïve Bayesian



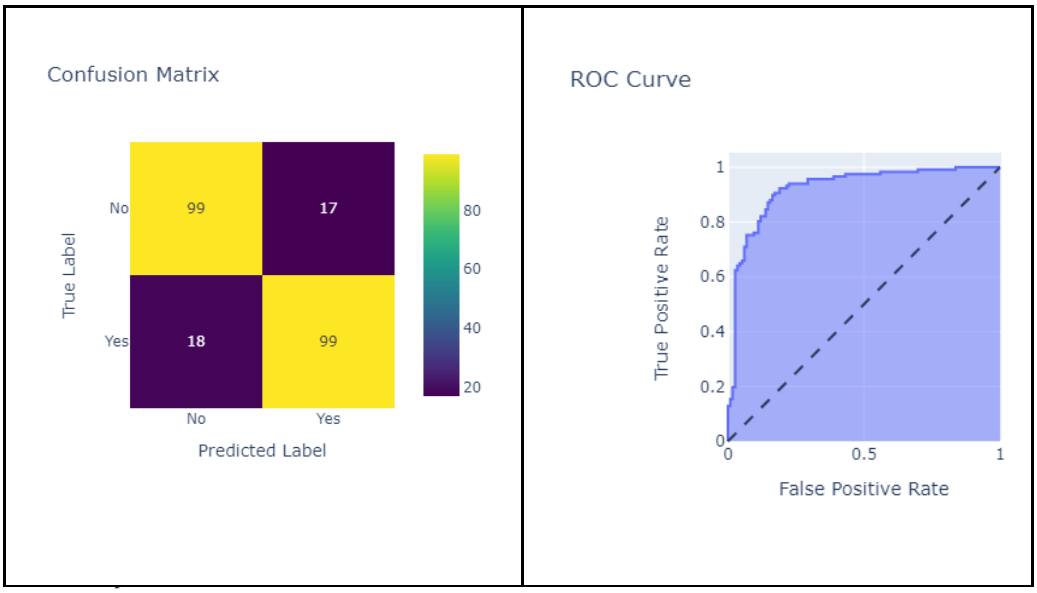
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.853448	0.128205	0.868421	0.853448	0.860870	0.725404	0.917772
Yes	0.871795	0.146552	0.857143	0.871795	0.864407	0.725404	0.917772
Weighted Avg	0.862661	0.137418	0.862758	0.862661	0.862646	0.725404	0.917772

1.3. Random Forest



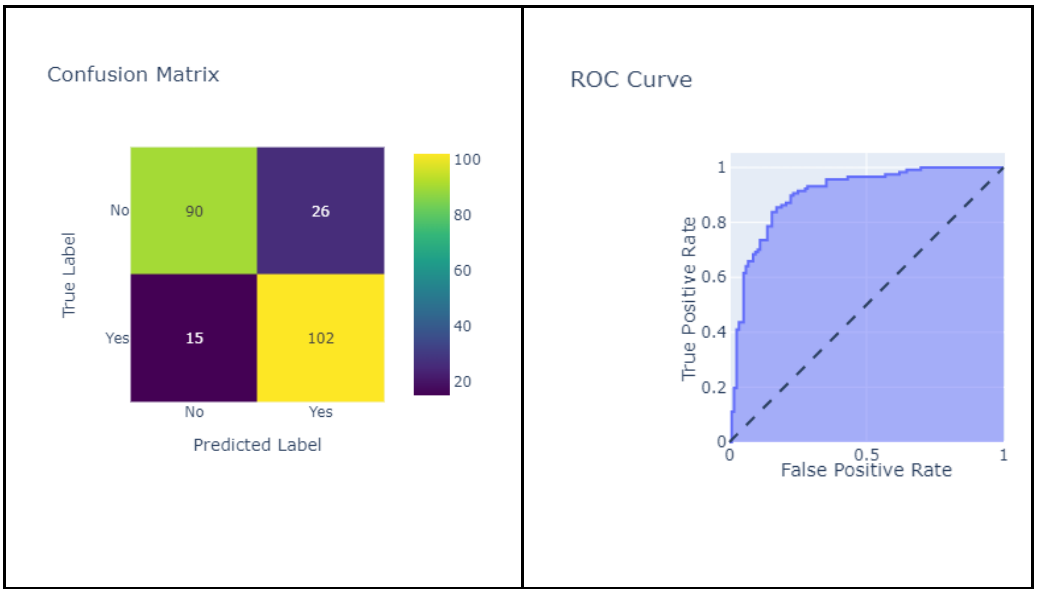
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.818966	0.153846	0.840708	0.818966	0.829694	0.665414	0.914346
Yes	0.846154	0.181034	0.825000	0.846154	0.835443	0.665414	0.914346
Weighted Avg	0.832618	0.167499	0.832820	0.832618	0.832581	0.665414	0.914346

1.4. Support Vector Machines



	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.853448	0.153846	0.846154	0.853448	0.849785	0.699602	0.923077
Yes	0.846154	0.146552	0.853448	0.846154	0.849785	0.699602	0.923077
Weighted Avg	0.849785	0.150183	0.849817	0.849785	0.849785	0.699602	0.923077

1.5. Artificial Neural Network

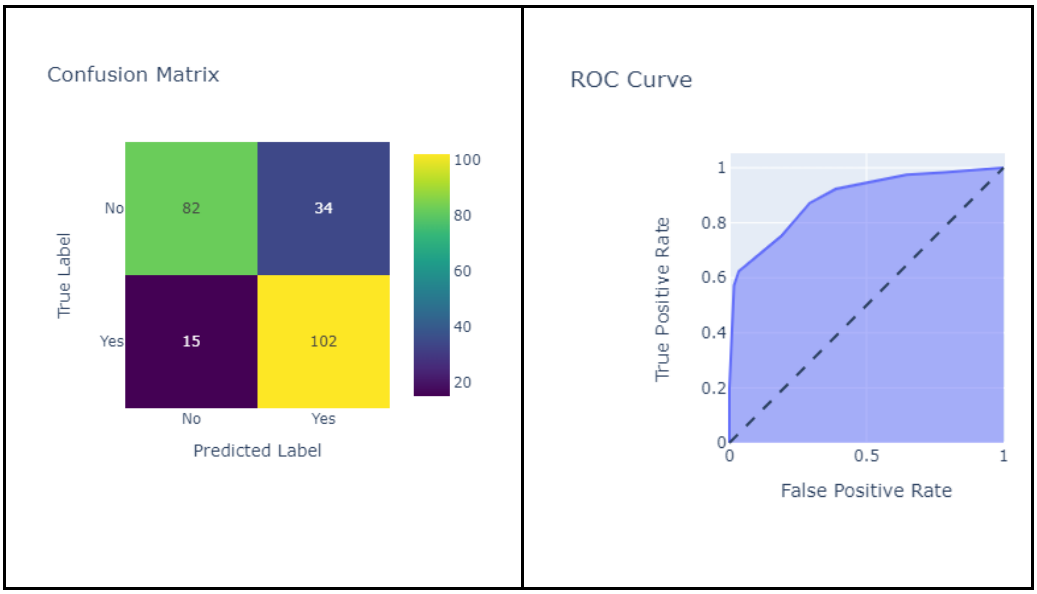


	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.775862	0.128205	0.857143	0.775862	0.814480	0.65083	0.900973
Yes	0.871795	0.224138	0.796875	0.871795	0.832653	0.65083	0.900973
Weighted Avg	0.824034	0.176377	0.826880	0.824034	0.823605	0.65083	0.900973

2. Lasso Regression

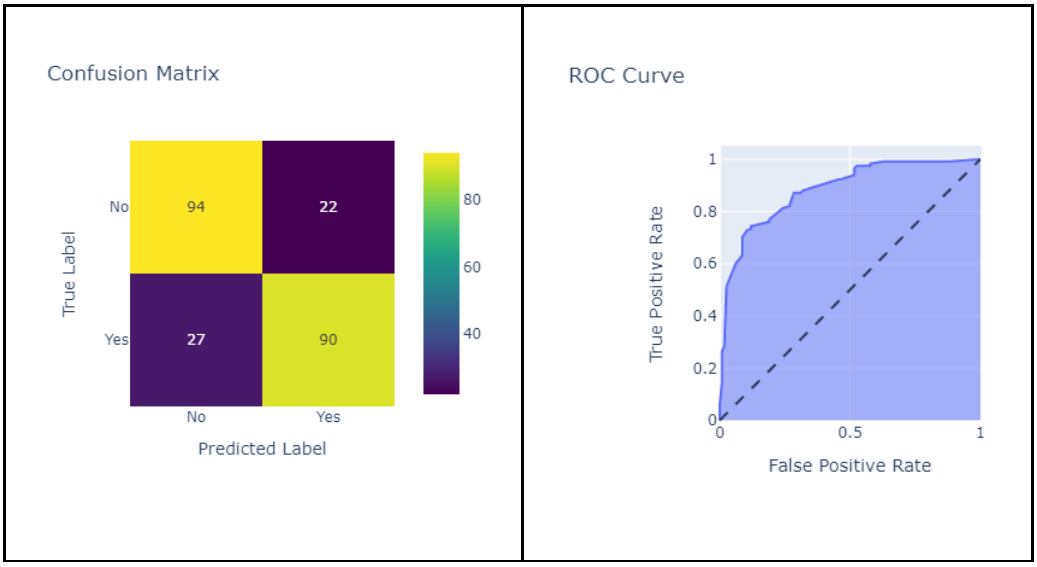
- *Attributes:* 'wellness_program', 'leave', 'mental_health_consequence'

2.1 K-Nearest Neighbors



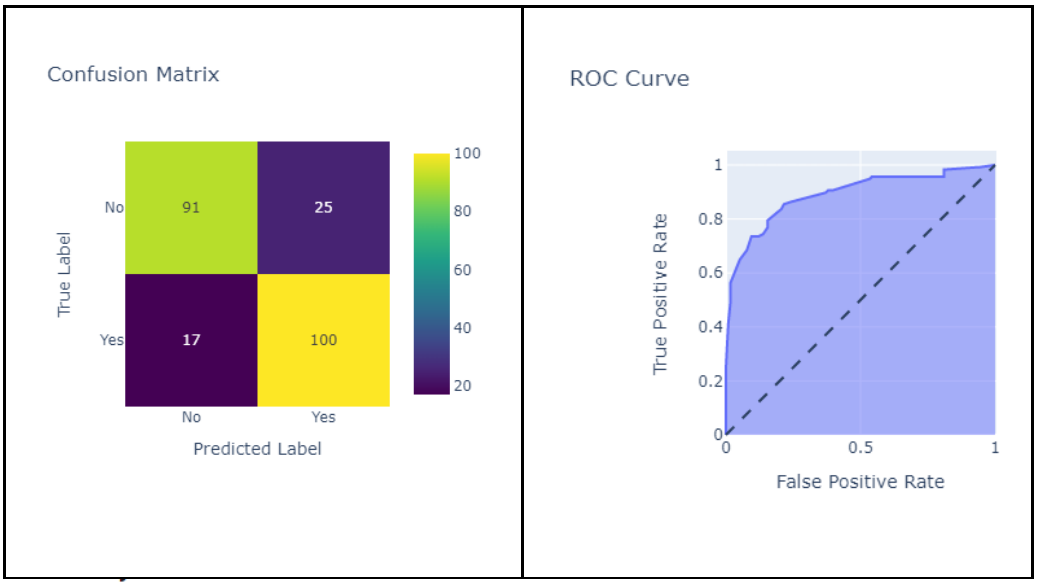
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.706897	0.128205	0.845361	0.706897	0.769953	0.586967	0.886826
Yes	0.871795	0.293103	0.750000	0.871795	0.806324	0.586967	0.886826
Weighted Avg	0.789700	0.211008	0.797476	0.789700	0.788217	0.586967	0.886826

2.2. Naïve Bayesian



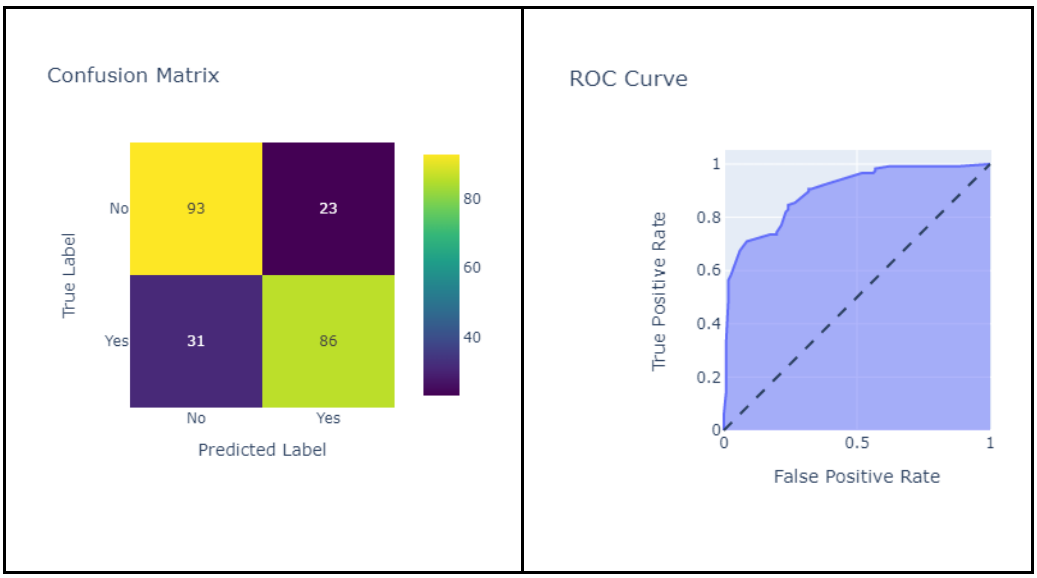
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.810345	0.230769	0.776860	0.810345	0.793249	0.580003	0.883363
Yes	0.769231	0.189655	0.803571	0.769231	0.786026	0.580003	0.883363
Weighted Avg	0.789700	0.210124	0.790273	0.789700	0.789622	0.580003	0.883363

2.3. Random Forest



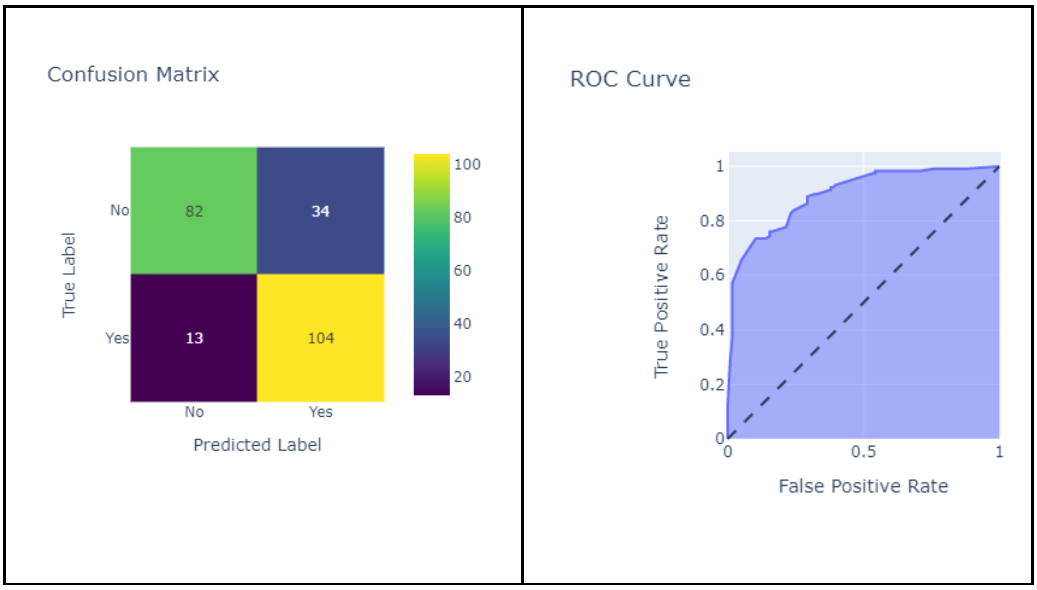
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.784483	0.145299	0.842593	0.784483	0.812500	0.640886	0.890694
Yes	0.854701	0.215517	0.800000	0.854701	0.826446	0.640886	0.890694
Weighted Avg	0.819742	0.180559	0.821205	0.819742	0.819503	0.640886	0.890694

2.4. Support Vector Machines



	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.801724	0.264957	0.750000	0.801724	0.775000	0.537878	0.894268
Yes	0.735043	0.198276	0.788991	0.735043	0.761062	0.537878	0.894268
Weighted Avg	0.768240	0.231473	0.769579	0.768240	0.768001	0.537878	0.894268

2.5. Artificial Neural Networks

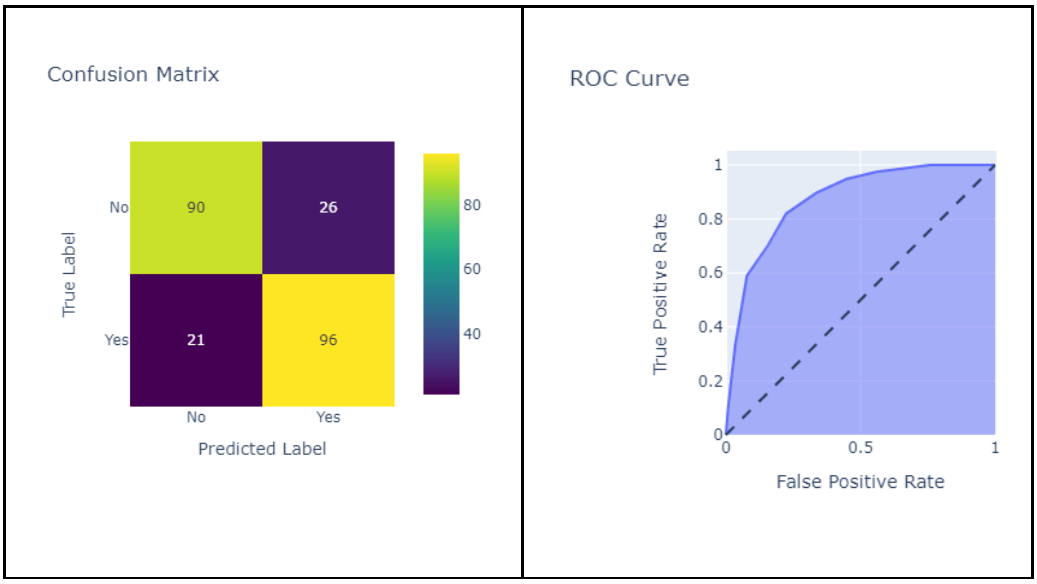


	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.706897	0.111111	0.863158	0.706897	0.777251	0.606192	0.896552
Yes	0.888889	0.293103	0.753623	0.888889	0.815686	0.606192	0.896552
Weighted Avg	0.798283	0.202498	0.808155	0.798283	0.796551	0.606192	0.896552

3. Decision Tree

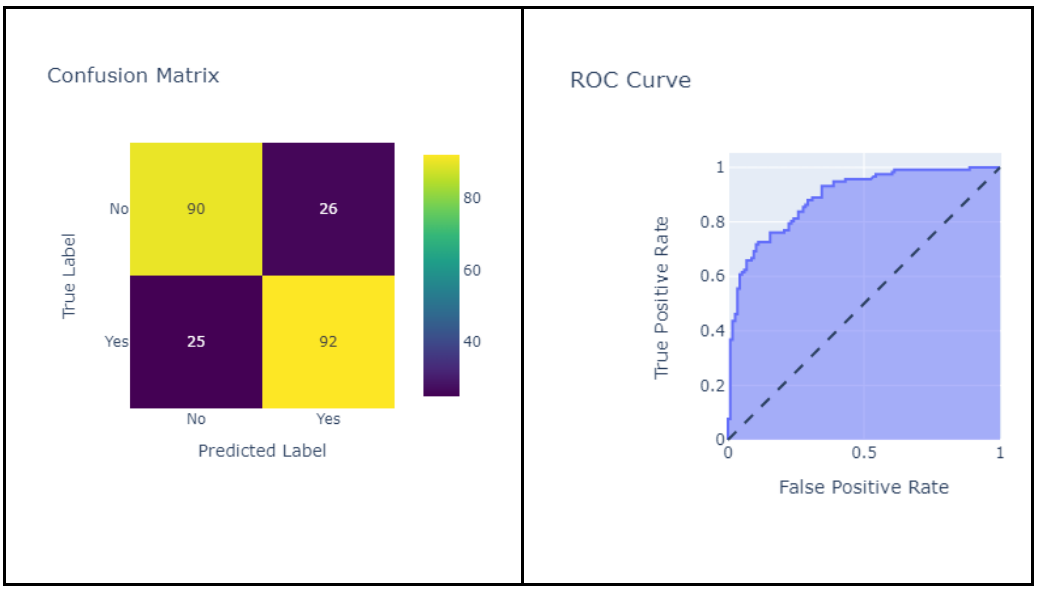
- *Attributes:* 'Age', 'family_history', 'no_employees', 'wellness_program', 'leave', 'mental_health_consequence', 'phys_health_interview'

3.1 K-Nearest Neighbors



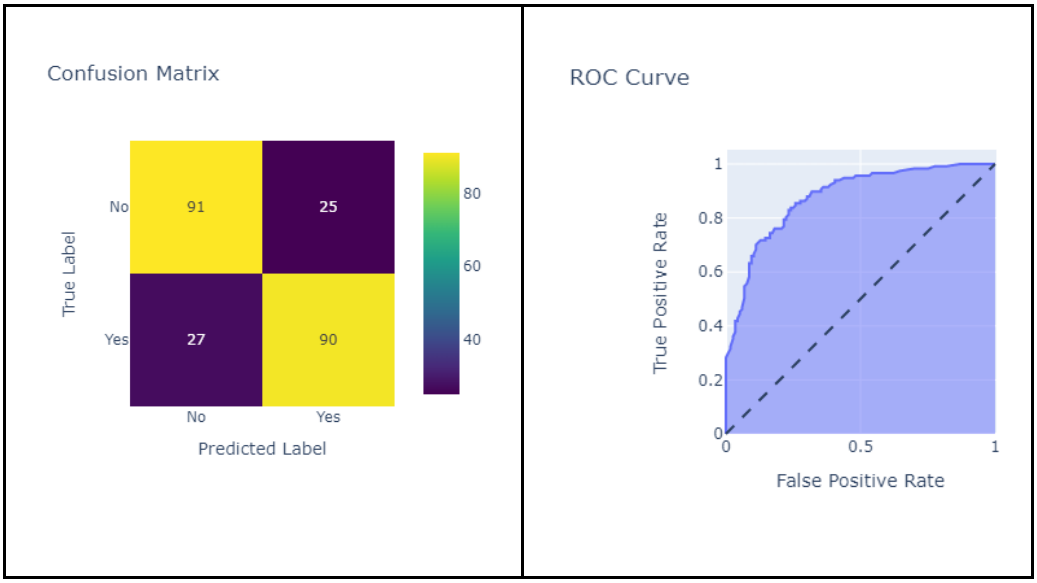
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.775862	0.179487	0.810811	0.775862	0.792952	0.597035	0.873084
Yes	0.820513	0.224138	0.786885	0.820513	0.803347	0.597035	0.873084
Weighted Avg	0.798283	0.201908	0.798797	0.798283	0.798172	0.597035	0.873084

3.2. Naïve Bayesian



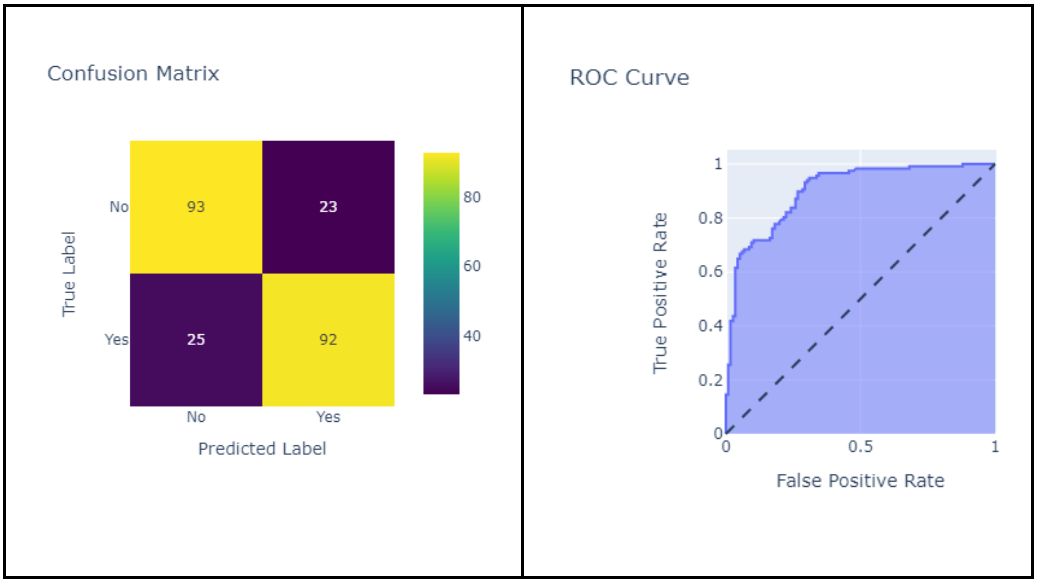
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.775862	0.213675	0.782609	0.775862	0.779221	0.562228	0.889589
Yes	0.786325	0.224138	0.779661	0.786325	0.782979	0.562228	0.889589
Weighted Avg	0.781116	0.218929	0.781129	0.781116	0.781108	0.562228	0.889589

3.3. Random Forest



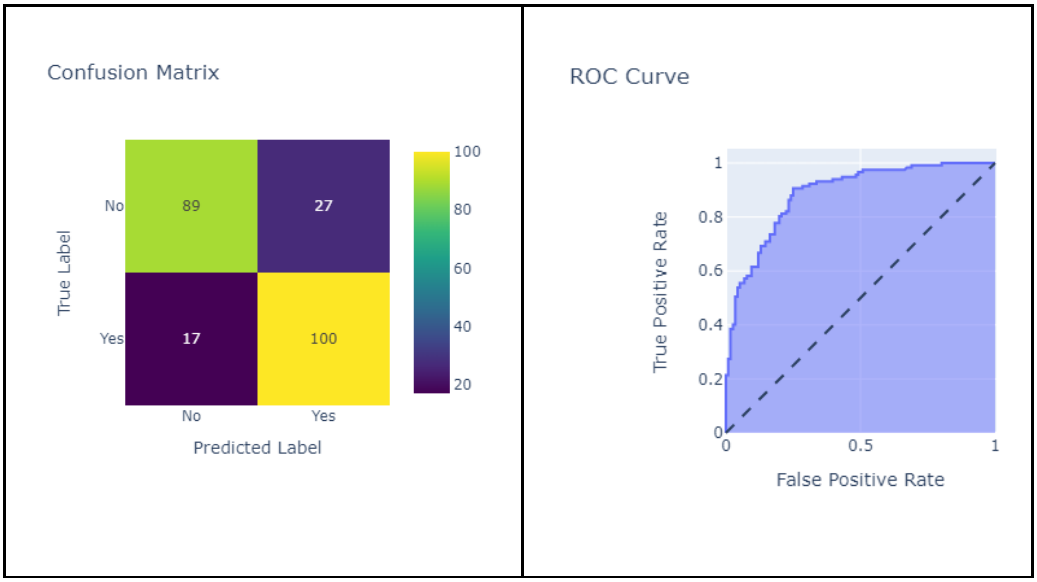
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.784483	0.230769	0.771186	0.784483	0.777778	0.553754	0.877726
Yes	0.769231	0.215517	0.782609	0.769231	0.775862	0.553754	0.877726
Weighted Avg	0.776824	0.223111	0.776922	0.776824	0.776816	0.553754	0.877726

3.4. Support Vector Machines



	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.801724	0.213675	0.788136	0.801724	0.794872	0.588092	0.901267
Yes	0.786325	0.198276	0.800000	0.786325	0.793103	0.588092	0.901267
Weighted Avg	0.793991	0.205942	0.794093	0.793991	0.793984	0.588092	0.901267

3.5. Artificial Neural Networks

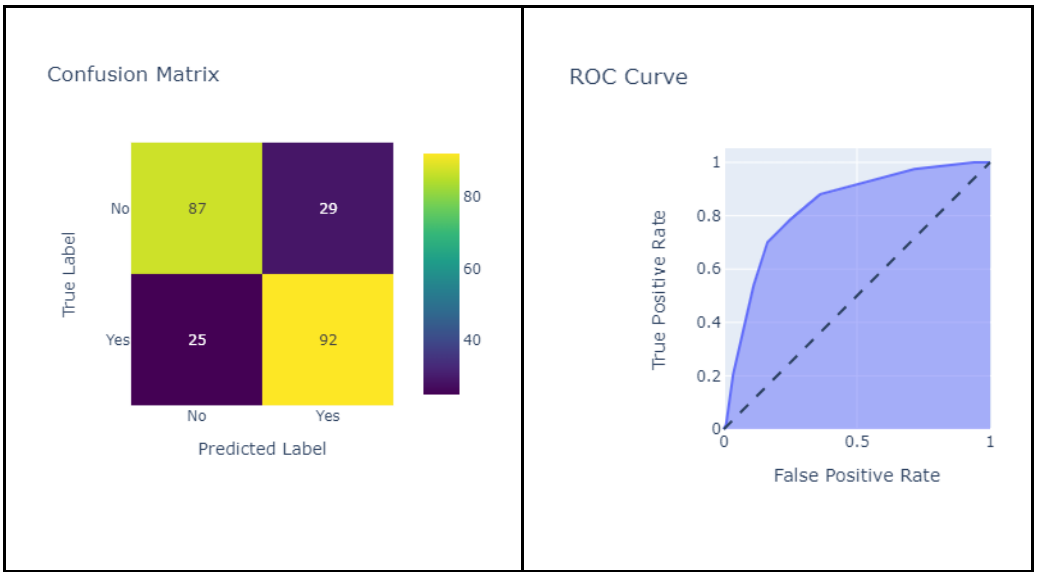


	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.767241	0.145299	0.839623	0.767241	0.801802	0.624478	0.886789
Yes	0.854701	0.232759	0.787402	0.854701	0.819672	0.624478	0.886789
Weighted Avg	0.811159	0.189217	0.813400	0.811159	0.810775	0.624478	0.886789

4. Forward Selection

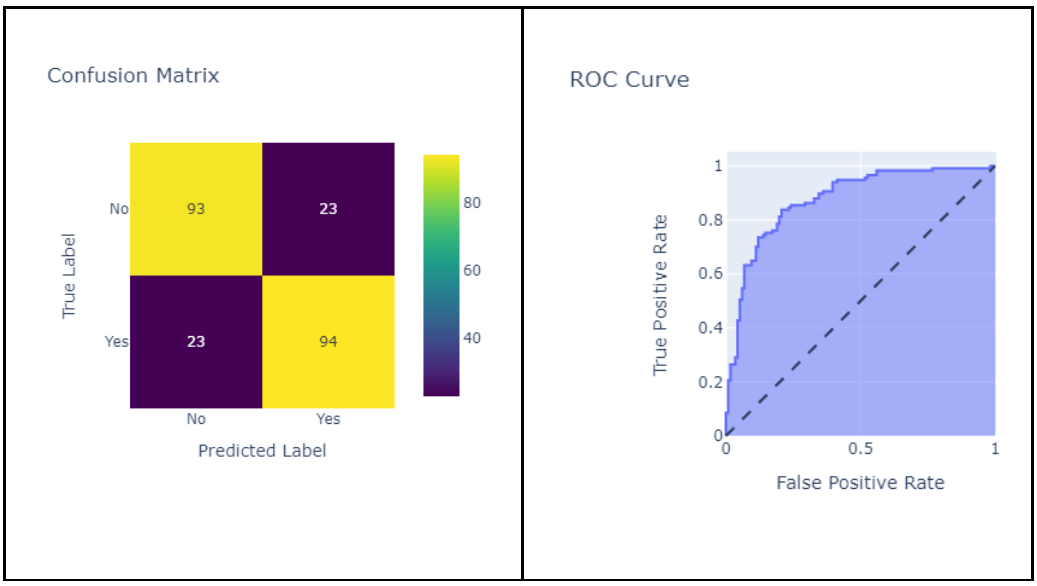
- *Attributes:* 'Age', 'Gender_Male', 'Gender_Non-binary', 'self_employed', 'no_employees', 'care_options', 'wellness_program', 'anonymity', 'leave', 'supervisor'

4.1 K-Nearest Neighbors



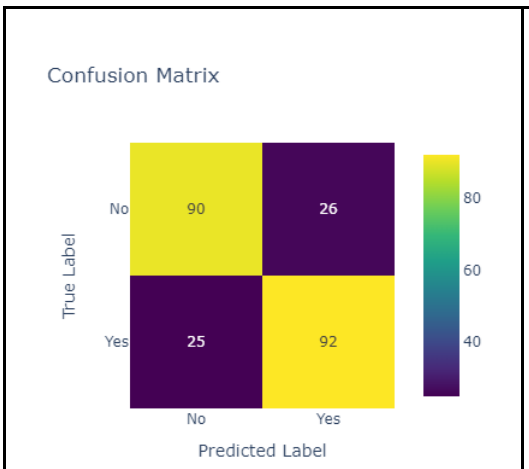
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.750000	0.213675	0.776786	0.750000	0.763158	0.53672	0.83057
Yes	0.786325	0.250000	0.760331	0.786325	0.773109	0.53672	0.83057
Weighted Avg	0.768240	0.231916	0.768523	0.768240	0.768155	0.53672	0.83057

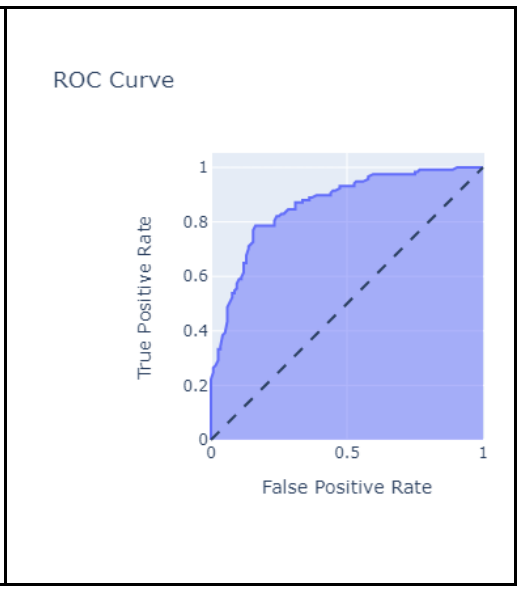
4.2. Naïve Bayesian



	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.801724	0.196581	0.801724	0.801724	0.801724	0.605143	0.876363
Yes	0.803419	0.198276	0.803419	0.803419	0.803419	0.605143	0.876363
Weighted Avg	0.802575	0.197432	0.802575	0.802575	0.802575	0.605143	0.876363

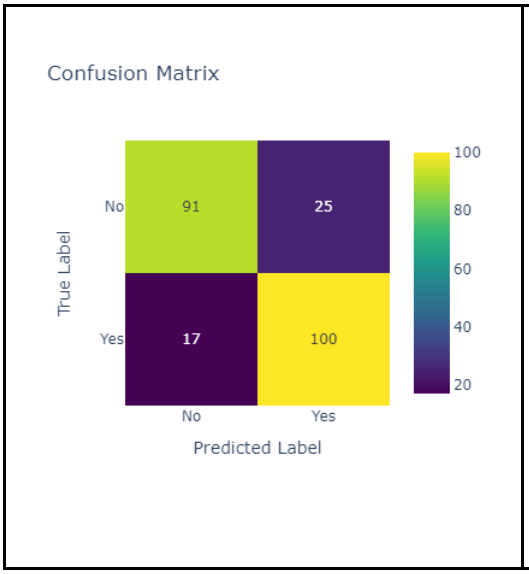
4.3. Random Forest

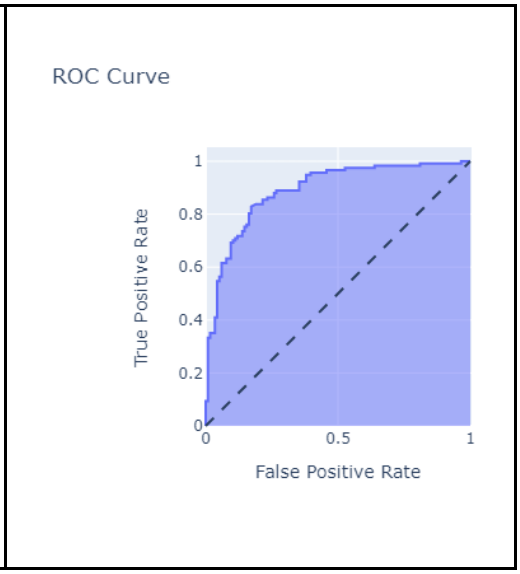




	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.775862	0.213675	0.782609	0.775862	0.779221	0.562228	0.862843
Yes	0.786325	0.224138	0.779661	0.786325	0.782979	0.562228	0.862843
Weighted Avg	0.781116	0.218929	0.781129	0.781116	0.781108	0.562228	0.862843

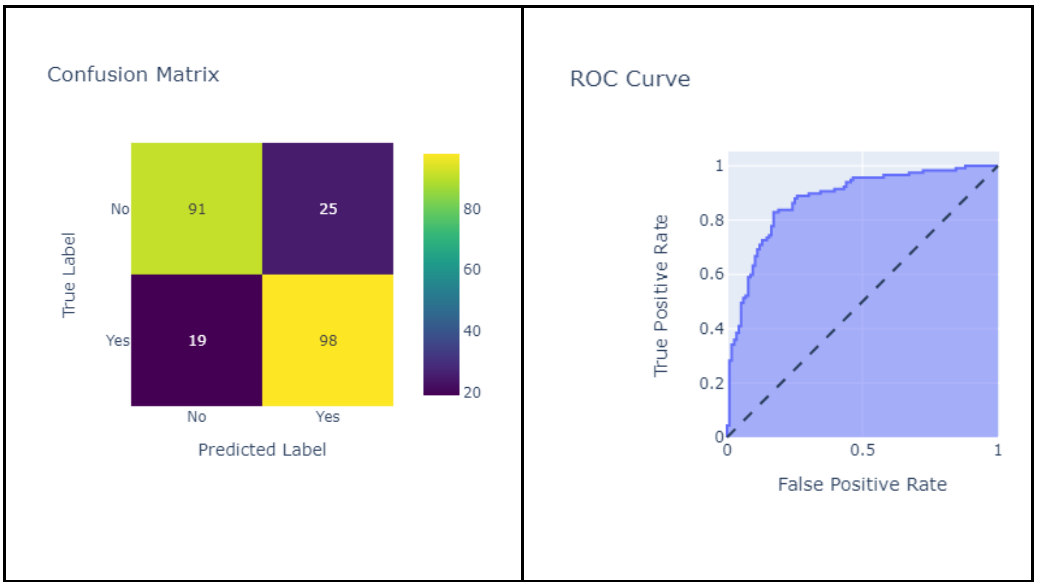
4.4. Support Vector Machines





	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.784483	0.145299	0.842593	0.784483	0.812500	0.640886	0.890068
Yes	0.854701	0.215517	0.800000	0.854701	0.826446	0.640886	0.890068
Weighted Avg	0.819742	0.180559	0.821205	0.819742	0.819503	0.640886	0.890068

4.5. Artificial Neural Networks

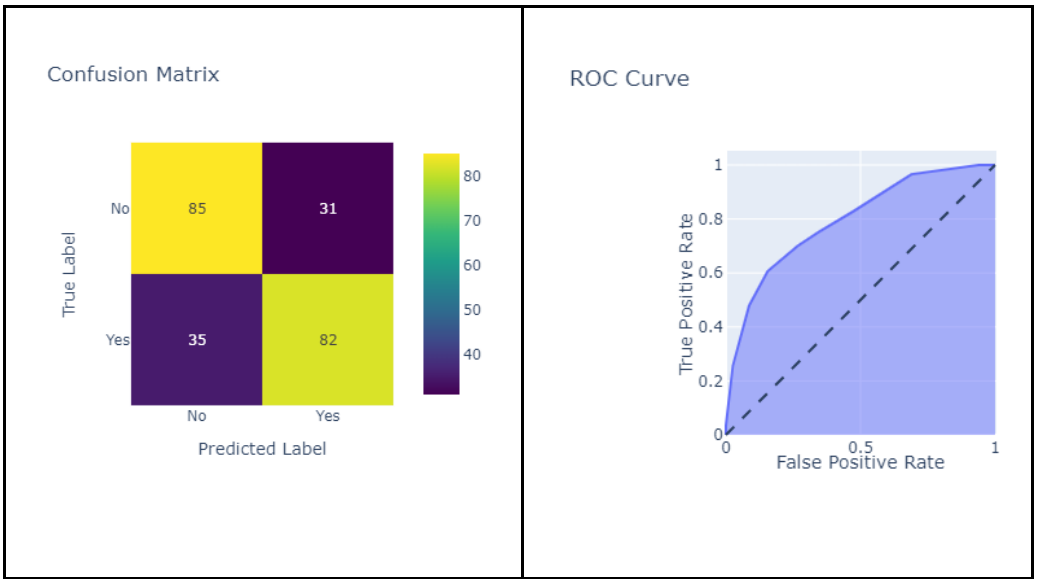


	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.784483	0.162393	0.827273	0.784483	0.805310	0.623054	0.877542
Yes	0.837607	0.215517	0.796748	0.837607	0.816667	0.623054	0.877542
Weighted Avg	0.811159	0.189069	0.811945	0.811159	0.811013	0.623054	0.877542

5. Backwards Selection

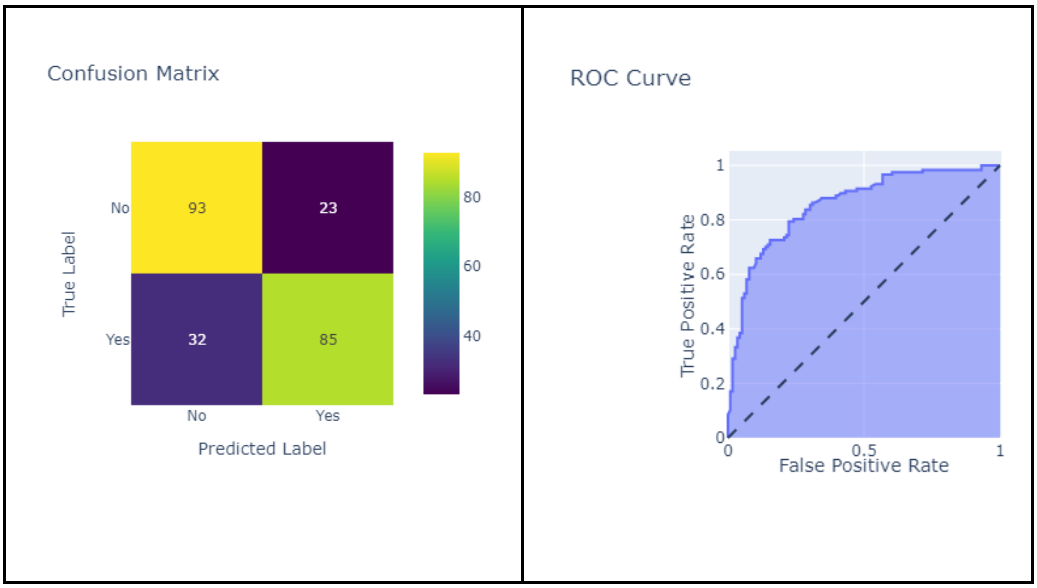
- *Attributes:* 'Gender_Non-binary', 'self_employed', 'family_history', 'treatment', 'no_employees', 'tech_company', 'care_options', 'anonymity', 'leave', 'supervisor'

5.1 K-Nearest Neighbors



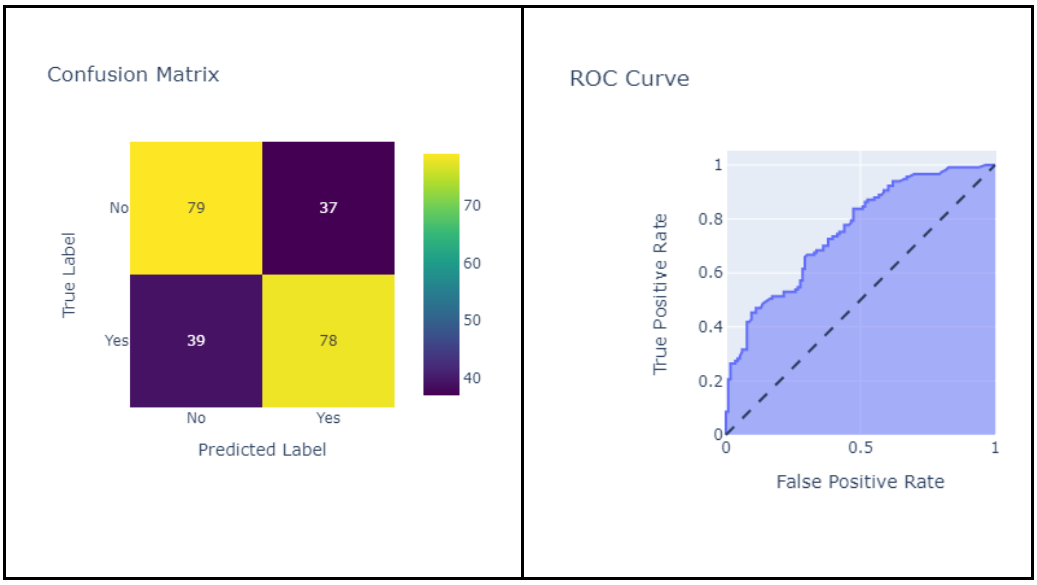
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.732759	0.299145	0.708333	0.732759	0.720339	0.433805	0.794724
Yes	0.700855	0.267241	0.725664	0.700855	0.713043	0.433805	0.794724
Weighted Avg	0.716738	0.283125	0.717036	0.716738	0.716676	0.433805	0.794724

5.2. Naïve Bayesian



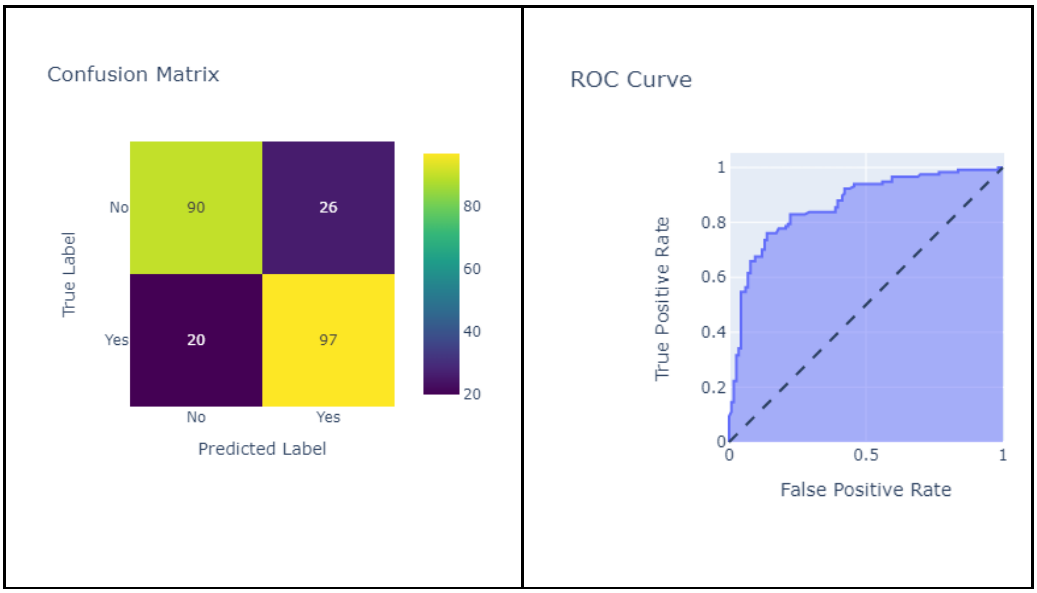
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.801724	0.273504	0.744000	0.801724	0.771784	0.529627	0.856322
Yes	0.726496	0.198276	0.787037	0.726496	0.755556	0.529627	0.856322
Weighted Avg	0.763948	0.235729	0.765611	0.763948	0.763635	0.529627	0.856322

5.3. Random Forest



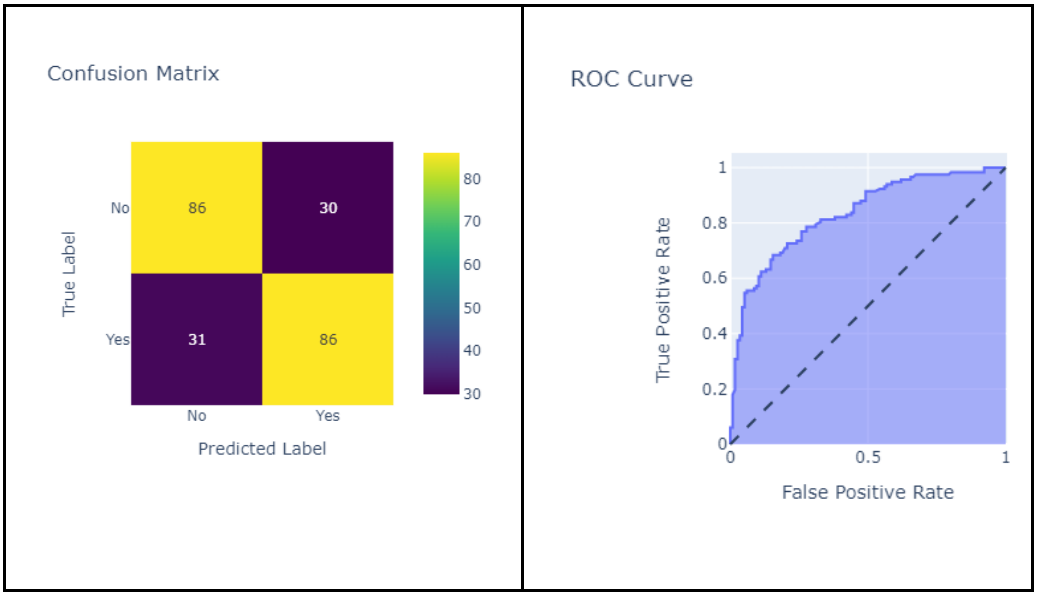
	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.681034	0.333333	0.669492	0.681034	0.675214	0.347727	0.757552
Yes	0.666667	0.318966	0.678261	0.666667	0.672414	0.347727	0.757552
Weighted Avg	0.673820	0.326119	0.673895	0.673820	0.673808	0.347727	0.757552

5.4. Support Vector Machines



	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.775862	0.170940	0.818182	0.775862	0.796460	0.60586	0.863469
Yes	0.829060	0.224138	0.788618	0.829060	0.808333	0.60586	0.863469
Weighted Avg	0.802575	0.197653	0.803336	0.802575	0.802422	0.60586	0.863469

5.5. Artificial Neural Networks



	TPR	FPR	Precision	Recall	F1 Measure	MCC	ROC AUC
No	0.741379	0.264957	0.735043	0.741379	0.738197	0.476422	0.833702
Yes	0.735043	0.258621	0.741379	0.735043	0.738197	0.476422	0.833702
Weighted Avg	0.738197	0.261775	0.738225	0.738197	0.738197	0.476422	0.833702

Summary Performance Results

As seen from the Detailed Performance Results, there are several performance metrics to consider when evaluating a classification model. The goal of this project is to accurately predict whether or not an employer prioritizes mental health as much as physical health. Thus, distinguishing between the responses of the class attribute is most important for the project's aims to understand the factors that contribute to a workplace prioritizing or not prioritizing mental health. For this reason, we have selected the ROC AUC as the performance metric when evaluating the 25 classification models. The ROC Curve is a plot of the True Positive Rate/Sensitivity/Recall (y-axis) versus 1 - True Negative Rate/Specificity (x-axis). Recall that Sensitivity is the ratio of true positives to total positives in the data while Specificity is the ratio of true negatives to total negatives in the data. The ROC Curve contains information such as smaller values on the x-axis indicate lower false positives and higher true negatives while larger values on the y-axis indicate higher true positives and lower false negatives. The area under the ROC curve (AUC) is a measure of separability, distinguishing between classes of the class attribute. The higher the AUC, the better the model is at predicting class 'Yes' as 'Yes' and class 'No' as 'No.' Figure 5 and Figure 6 summarize and compare the results of the 25 classification models based on ROC AUC as the performance metric.

	Chi-Square	Lasso Reg	Decision Tree	Forward Selection	Backwards Selection
kNN	0.914309	0.886826	0.873084	0.830570	0.794724
Naive Bayes	0.917772	0.883363	0.889589	0.876363	0.856322
Random Forest	0.914346	0.890694	0.877726	0.862843	0.757552
SVM	0.923077	0.894268	0.901267	0.890068	0.863469
ANN	0.900973	0.896552	0.886789	0.877542	0.833702

Figure 5: Table of the ROC AUC Performance Results of the 25 Classification Models

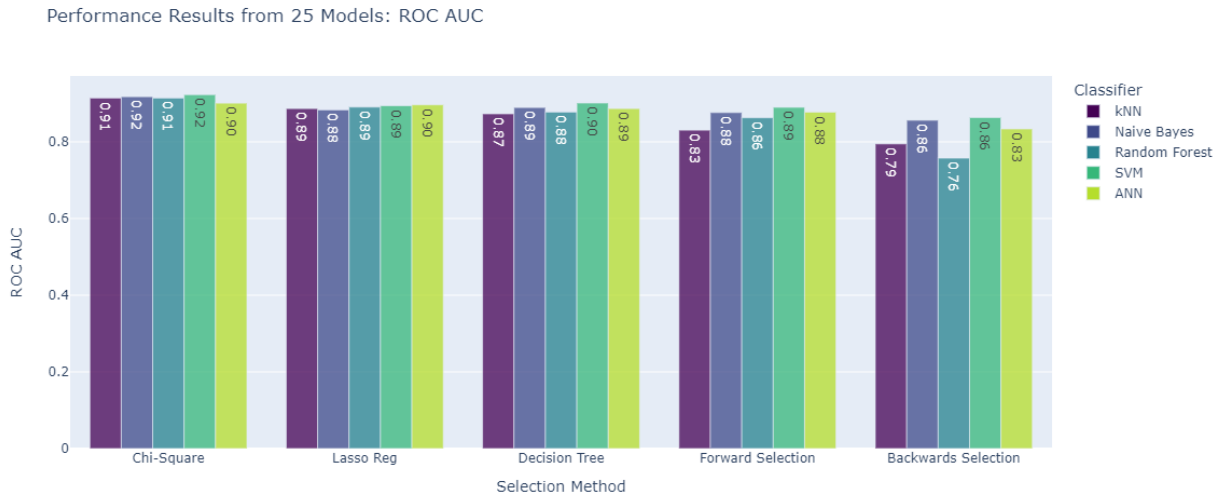


Figure 6: Bar Chart of the ROC AUC Performance Results of the 25 Classification Models

Notice the moderately small variability of the ROC AUC of the 25 classification models; the range of the performance results is 0.75 to 0.92. Before we determine the best model, let's first explore which selection method and classification algorithm performed best based on 1) majority ruling and 2) averaging.

1. Majority Ruling: Figure 7 summarizes the number of times an attribution selection method had the highest ROC AUC for each classification algorithm (five total); notice that the Chi-Square Test for attribute selection was the best method for five out of five of the classification algorithms. Figure 8 summarizes the number of times a classification algorithm had the highest ROC AUC for each attribute selection method (five total); notice that the SVM model was the best algorithm for four out of five of the attribute selection methods.
2. Averaging: Figure 9 visualizes the average ROC AUC by attribute selection method; notice the Chi-Square Test for attribute selection results in the highest average ROC AUC, 0.91, while Backwards Selection for attribute selection results in the lowest average ROC AUC, 0.82. Figure 10 visualizes the average ROC AUC by classification algorithm; notice the SVM model results in the highest average ROC AUC, 0.89, while kNN results in the lowest average ROC AUC, 0.86.

From both the perspectives - majority ruling and averaging - of analyzing the best attribute selection method and classification algorithm, the Chi-Square Test and SVM resulted in the highest ROC AUC.

Method	No. Times Best Method
Chi-Square	5
Lasso Reg	0
Decision Tree	0
Forward Selection	0
Backwards Selection	0

Figure 7: Table of the Number of Times an Attribute Selection Method had the Best ROC AUC

Classifier	No. Times Best Classifier
kNN	0
Naive Bayes	0
Random Forest	0
SVM	4
ANN	1

Figure 8: Table of the Number of Times a Classification Model had the Best ROC AUC

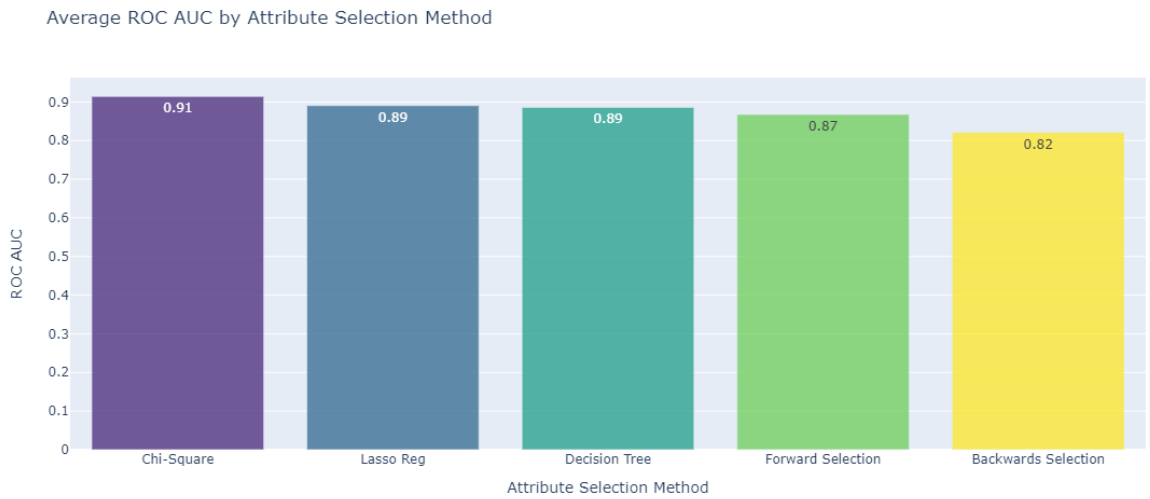


Figure 9: Bar Chart of the Average ROC AUC by Attribute Selection Method

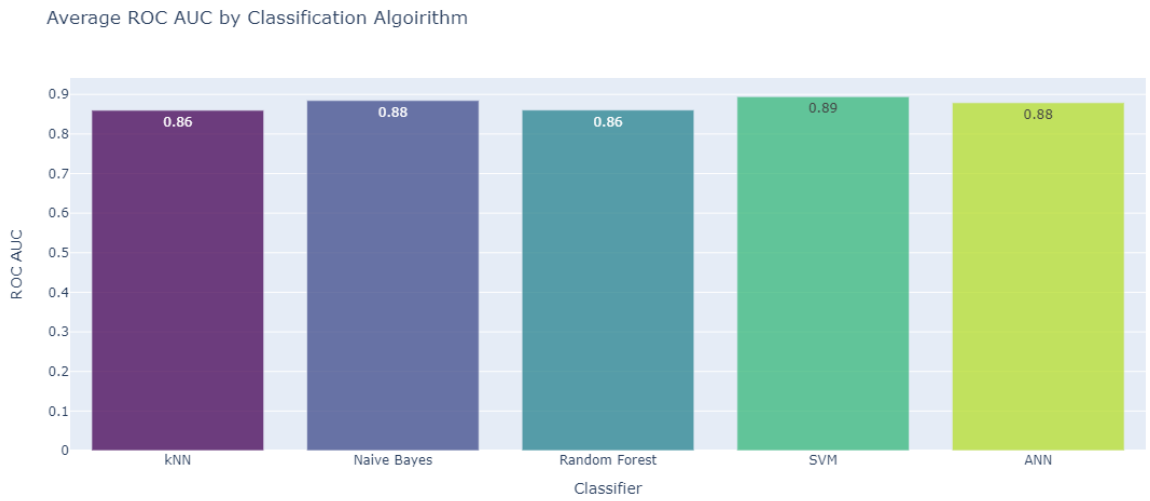


Figure 10: Bar Chart of the Average ROC AUC by Classification Algorithm

Best Model

Overall, the best model is the Chi-Square Test attribute selection method and the Support Vector Machines classification algorithm with the highest ROC AUC of 0.923. The decision to select this model as the best is supported by the stratified analysis above where we determined that the Chi-Square Test is the best attribute selection method and SVM is the best classification algorithm. Recall that the Chi-Square Test selects attributes which are highly dependent on the response variable; this method is favorable as we found some attributes to be independent of the response variable. Additionally, this attribute selection method resulted in the highest number of selected attributes, which performed significantly better than attribute selection methods that resulted in only a small number of selected attributes (i.e. Lasso Regression). Figure 11 outlines the attributes and respective survey questions removed/selected from the Chi-Square Test attribute selection method. This helps us understand which combination of attributes are the strongest predictors for mental health priority in the tech workplace. Lastly, we are interested in how the performance of the best model compares to the performance of the classification algorithm from the best model with all attributes from the dataset. Figure 12 visualizes the ROC AUC results from the SVM classification algorithm on all attributes versus the SVM classification algorithm with the Chi-Square Test attribute selection method. Notice the 0.1 increase (12% increase from 0.82 to 0.92) in the ROC AUC when applying the Chi-Square Test attribute selection method. This effectively demonstrates the importance of dimensionality reduction to increase performance when working with high dimensional datasets.

Attribute	Survey Question
~~*Age*~~	Age of the survey participant in years
~~*Gender*~~	Gender of the survey participant
self_employed	Are you self-employed?
~~*family_history*~~	Do you have a family history of mental illness?
~~*treatment*~~	Have you sought treatment for a mental health condition?
~~*work_interfere*~~	If you have a mental health condition, do you feel that it interferes with your work?
no_employees	How many employees does your company or organization have?
~~*remote_work*~~	Do you work remotely (outside of an office) at least 50 of the time?
~~*tech_company*~~	Is your employer primarily a tech company/organization?
benefits	Does your employer provide mental health benefits?
~~*care_options*~~	Do you know the options for mental health care your employer provides?
wellness_program	Has your employer ever discussed mental health as part of an employee wellness program?
seek_help	Does your employer provide resources to learn more about mental health issues and how to seek help?
anonymity	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?

leave	How easy is it for you to take medical leave for a mental health condition?
mental_health_consequence	Do you think that discussing a mental health issue with your employer would have negative consequences?
phys_health_consequence	Do you think that discussing a physical health issue with your employer would have negative consequences?
coworkers	Would you be willing to discuss a mental health issue with your coworkers?
supervisor	Would you be willing to discuss a mental health issue with your direct supervisor(s)?
mental_health_interview	Would you bring up a mental health issue with a potential employer in an interview?
~~*phys_health_interview*~~	Would you bring up a physical health issue with a potential employer in an interview?
obs_consequence	Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?

Figure 11: Table of the Attributes Removed/Selected from the Chi-Square Attribute Selection Method

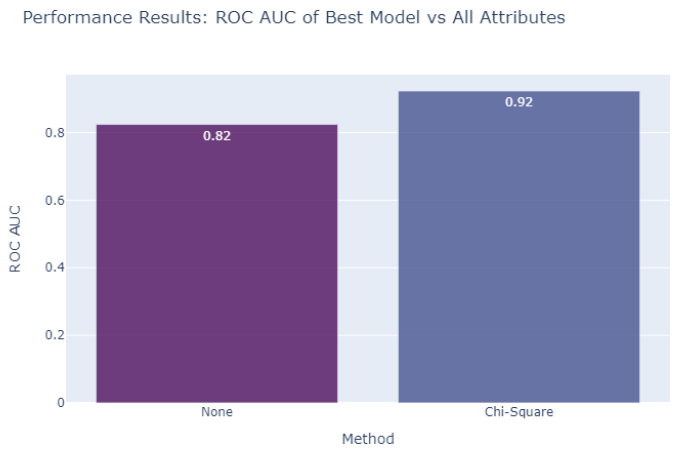


Figure 12: Bar Chart of the ROC AUC of the Best Model with and without the Attribute Selection Method

CONCLUSION

The purpose of this project is to predict whether or not an employer prioritizes mental health as much as physical health based on a set of attributes collected from a survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. Five binary classification algorithms - K-Nearest Neighbors, Naïve Bayesian, Random Forest, Support Vector Machines, and Artificial Neural Networks - were selected to predict the value of the class attributes, *mental_vs_physical*, as 'Yes' (i.e. the survey participant felt that their employer does take mental health as seriously as physical health) or 'No' (i.e. the survey participant felt that their employer does not take mental health as seriously as physical health). Due to the high dimensionality of that dataset (21 attributes after preprocessing), five attribute selection methods - Chi-Square Test, Lasso Regression, Decision Tree Induction, Forward Selection, and Backwards Selection - were employed to determine which attributes are the strongest predictors for mental health priority in the tech workplace. While several performance metrics were collected to evaluate and compare the 25 models, we selected ROC AUC as the performance metric to determine the best model. ROC AUC is most applicable to the project's aims as it is a measure of separability, distinguishing between the classes of the class attribute, allowing for us to understand why a survey participant might respond 'Yes' or 'No' to their feelings about their company's mental health priorities. Based on the ROC AUC performance metric, the Chi-Square Test attribute selection method and the Support Vector Machines classification algorithm is the best model for classifying the priority of mental health in the tech workplace.

This project is a practical implementation of data mining by discovering and extracting interesting patterns from a high dimensional dataset. From this, we learned the benefits of

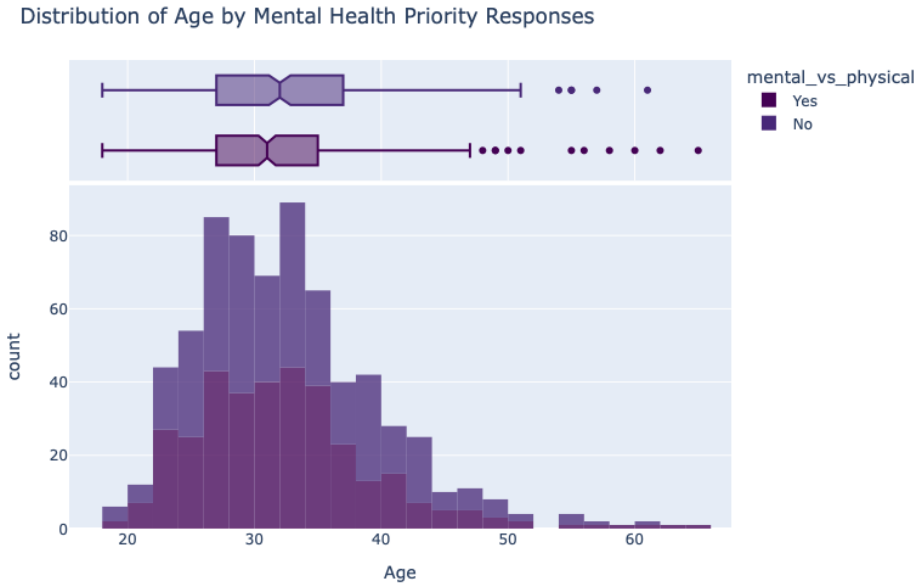
attribute selection methods in reducing the dimensionality of a dataset and increasing the effectiveness of a classification model. Additionally, this project was a good experience in implementing the full life cycle of a data science project by 1) understanding the problem, 2) ingesting and cleaning the data, 3) visualizing the data, 4) preprocessing the data, 5) implementing the models, and 6) evaluating the models. This cycle is an iterative process of understanding and reevaluating the project. We identified a number of classification performance metrics to evaluate/compare models and determined what type of performance metrics are most suitable for our particular scenario. Overall, it was interesting to implement a range of attribute selection methods and classification algorithms to comprehend the disparities among the models.

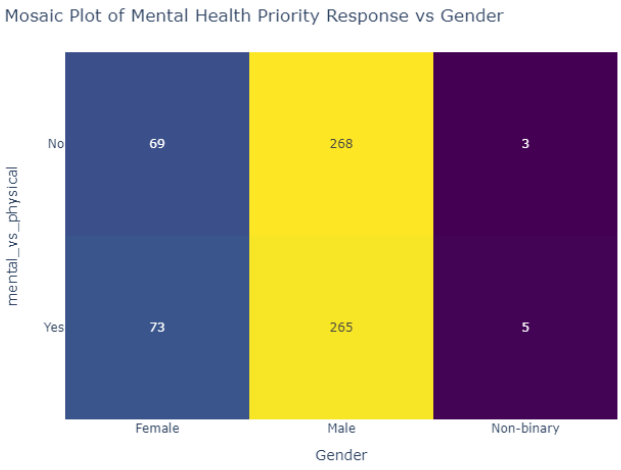
APPENDIX

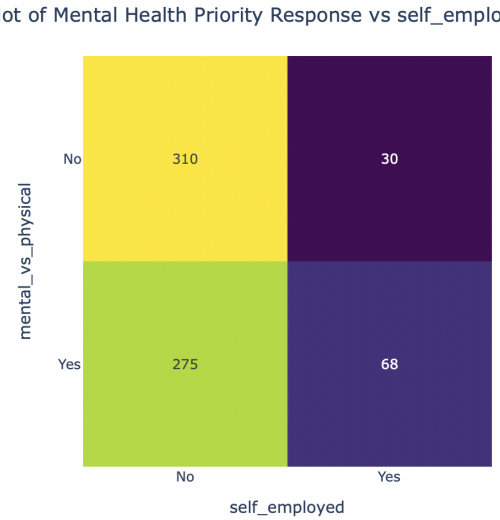
A. Dataset Source

Link: https://www.openml.org/search?type=data&status=active&id=43664&sort=runs

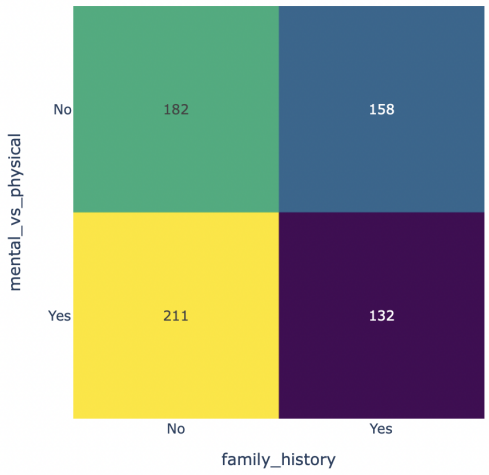
B. Data Visualization: Distribution of Attributes Stratified by Class Attribute



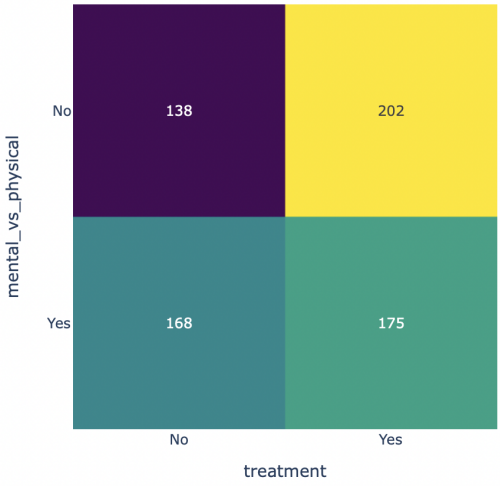




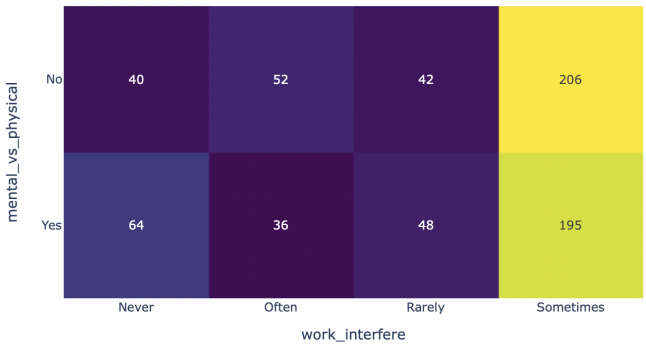
Mosaic Plot of Mental Health Priority Response vs family_history



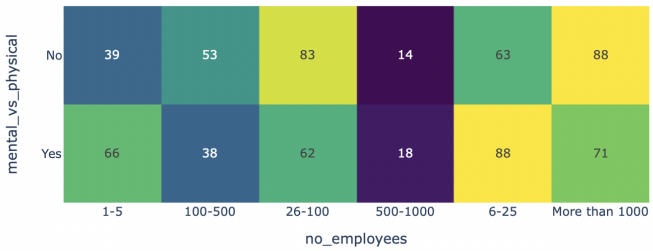
Mosaic Plot of Mental Health Priority Response vs treatment



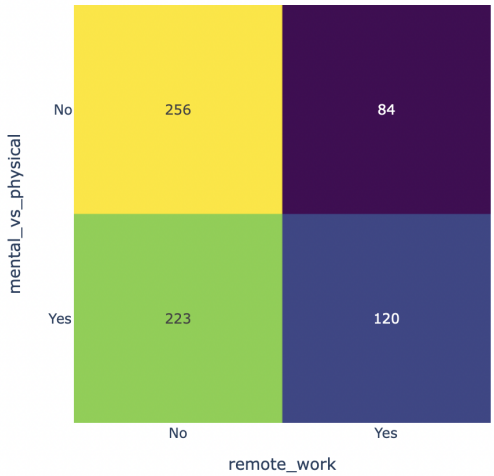
Mosaic Plot of Mental Health Priority Response vs work_interfere



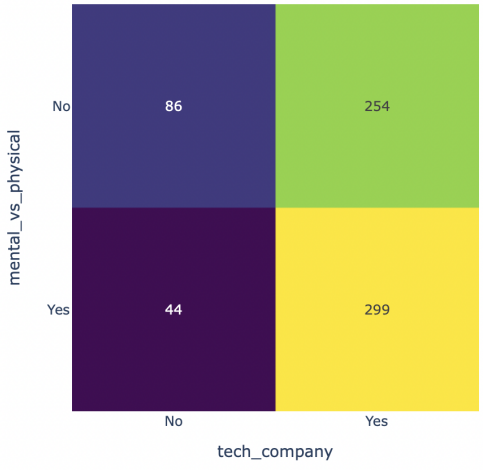
Mosaic Plot of Mental Health Priority Response vs no_employees

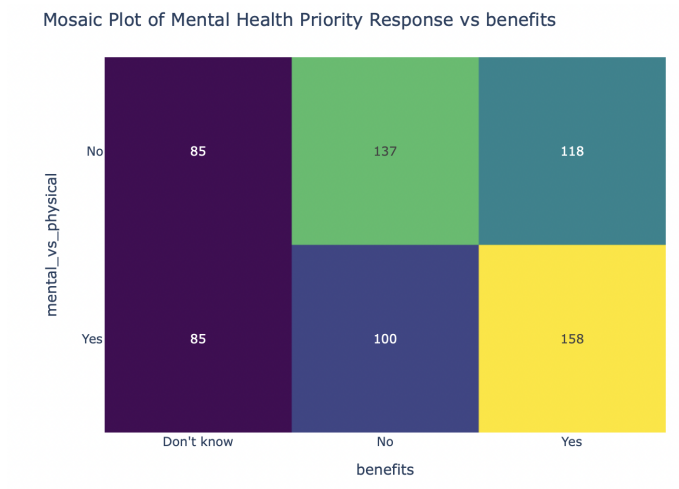


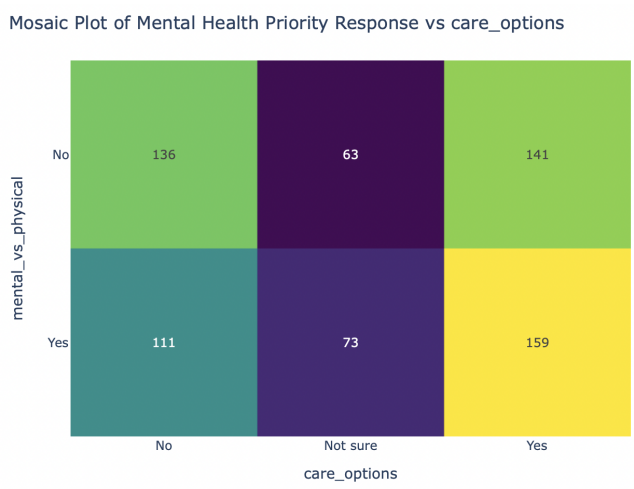
Mosaic Plot of Mental Health Priority Response vs remote_work

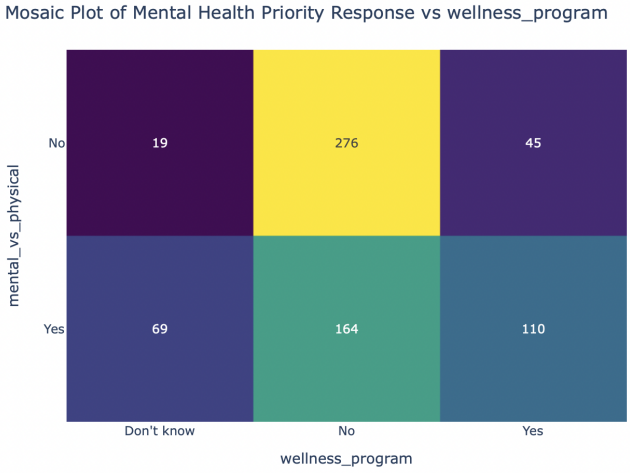


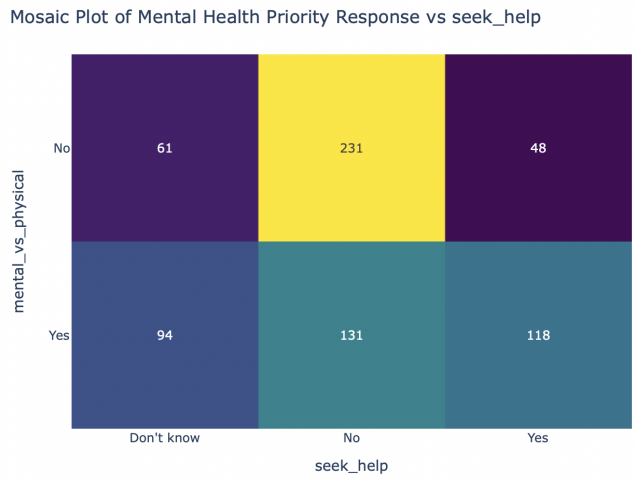
Mosaic Plot of Mental Health Priority Response vs tech_company

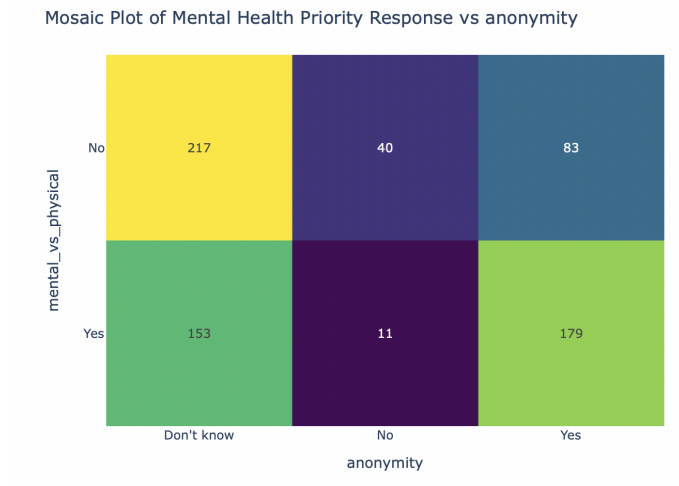


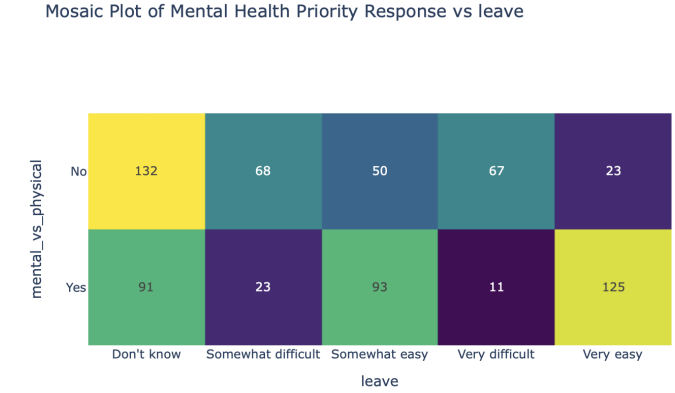

Mosaic Plot of Mental Health Priority Response vs benefits
mental_vs_physical
No
Yes
85
137
118
85
100
158
Don't know
No
Yes
benefits

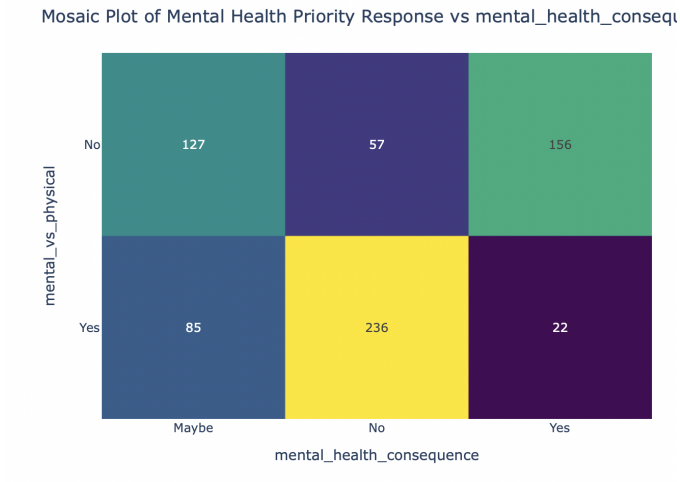

Mosaic Plot of Mental Health Priority Response vs care_options
mental_vs_physical
No
Yes
136
63
141
111
73
159
No
Not sure
Yes
care_options

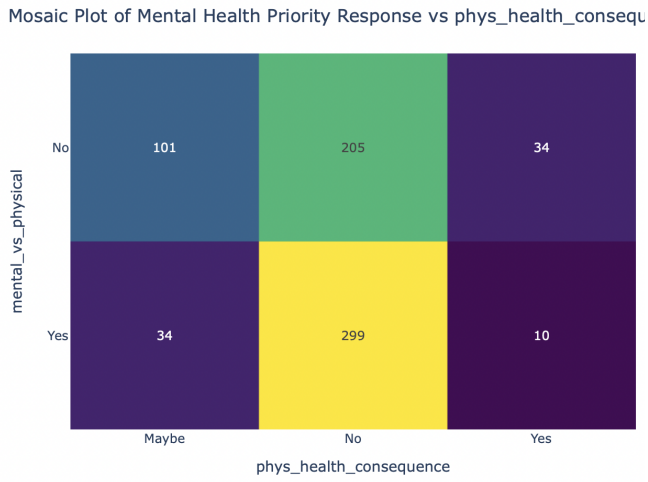

Mosaic Plot of Mental Health Priority Response vs wellness_program
mental_vs_physical
No
Yes
19
276
45
69
164
110
Don't know
No
Yes
wellness_program

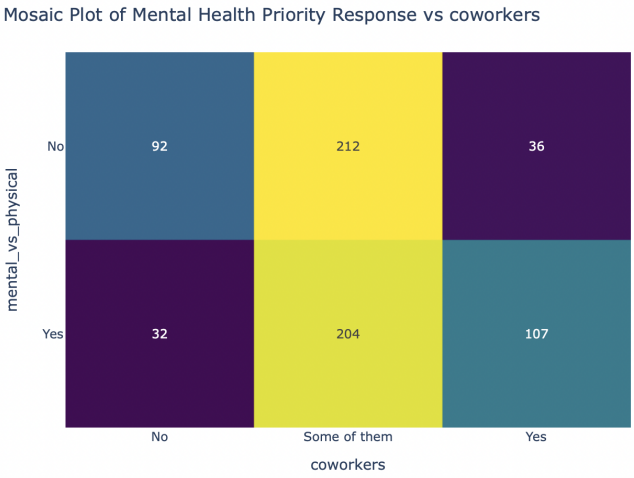

Mosaic Plot of Mental Health Priority Response vs seek_help
mental_vs_physical
No
Yes
61
231
48
94
131
118
Don't know
No
Yes
seek_help

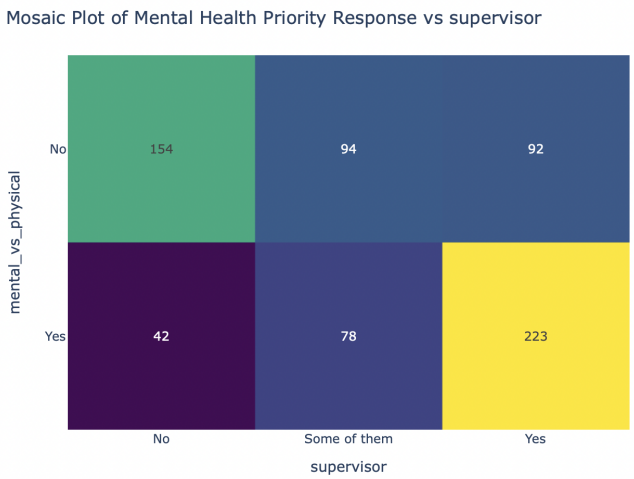

Mosaic Plot of Mental Health Priority Response vs anonymity
mental_vs_physical
No
Yes
217
40
83
153
11
179
Don't know
No
Yes
anonymity

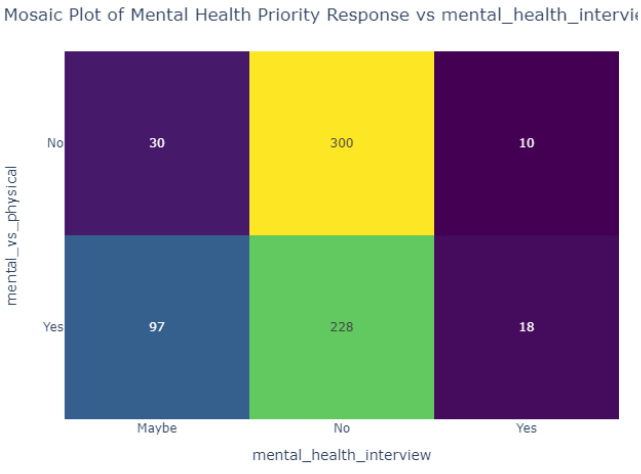

Mosaic Plot of Mental Health Priority Response vs leave
mental_vs_physical
No
Yes
132
68
50
67
23
91
23
93
11
125
Don't know
Somewhat difficult
Somewhat easy
Very difficult
Very easy
leave

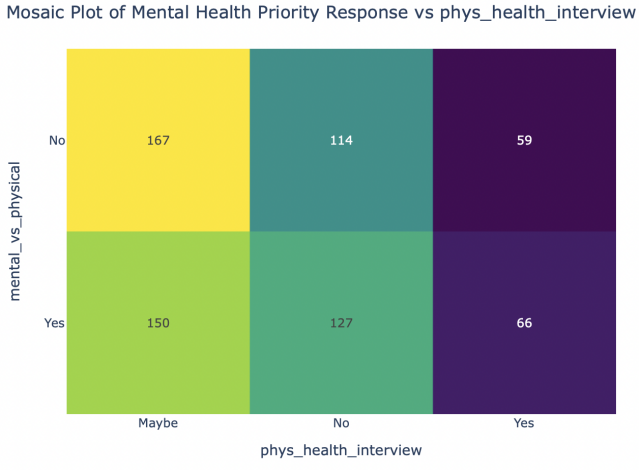

Mosaic Plot of Mental Health Priority Response vs mental_health_conseq
No
Yes
127
57
156
85
236
22
mental_vs_physical
Maybe
No
Yes
mental_health_consequence

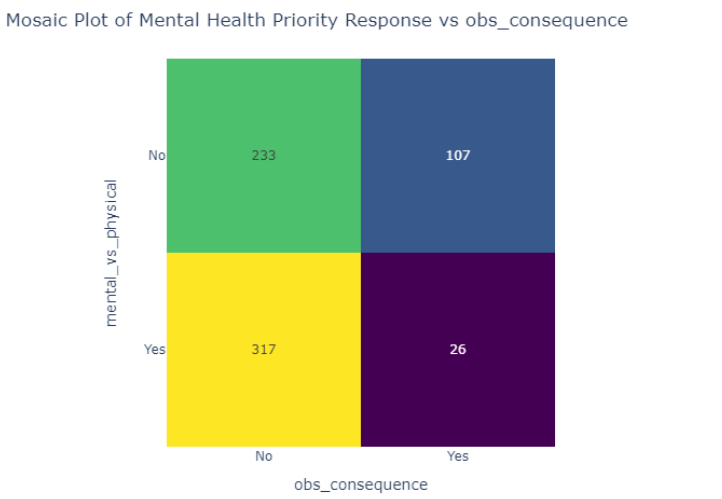

Mosaic Plot of Mental Health Priority Response vs phys_health_conseq
No
Yes
101
205
34
34
299
10
mental_vs_physical
Maybe
No
Yes
phys_health_consequence


Mosaic Plot of Mental Health Priority Response vs coworkers
No
Yes
92
212
36
32
204
107
mental_vs_physical
No
Some of them
Yes
coworkers

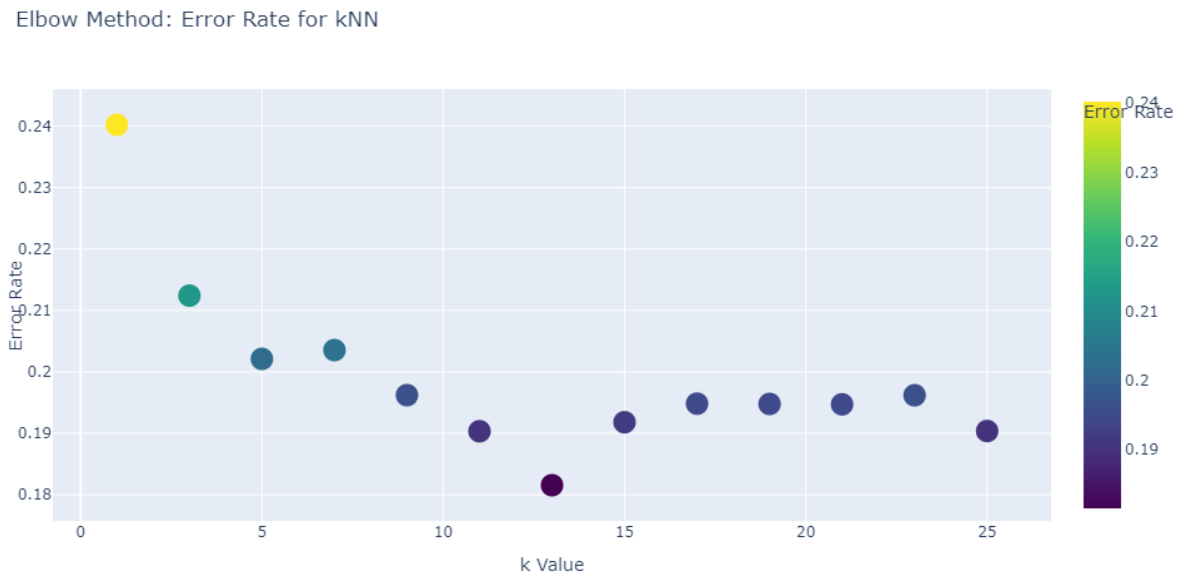

Mosaic Plot of Mental Health Priority Response vs supervisor
No
Yes
154
94
92
42
78
223
mental_vs_physical
No
Some of them
Yes
supervisor


Mosaic Plot of Mental Health Priority Response vs mental_health_intervi
No
Yes
30
300
10
97
228
18
mental_vs_physical
Maybe
No
Yes
mental_health_interview

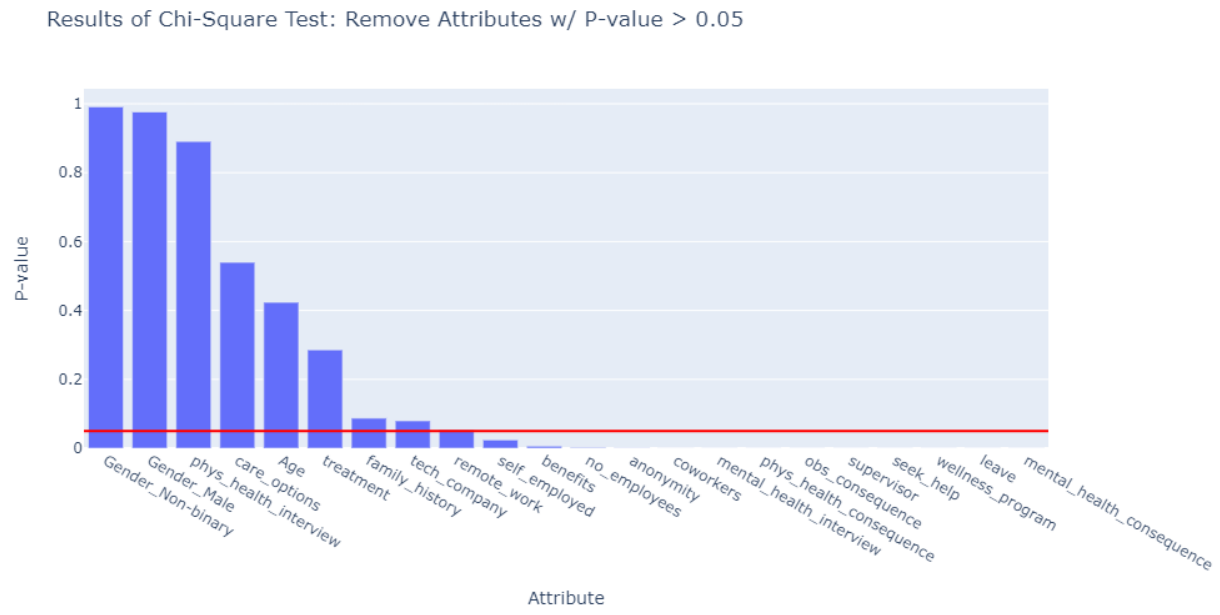

Mosaic Plot of Mental Health Priority Response vs phys_health_interview
No
Yes
167
114
59
150
127
66
mental_vs_physical
Maybe
No
Yes
phys_health_interview

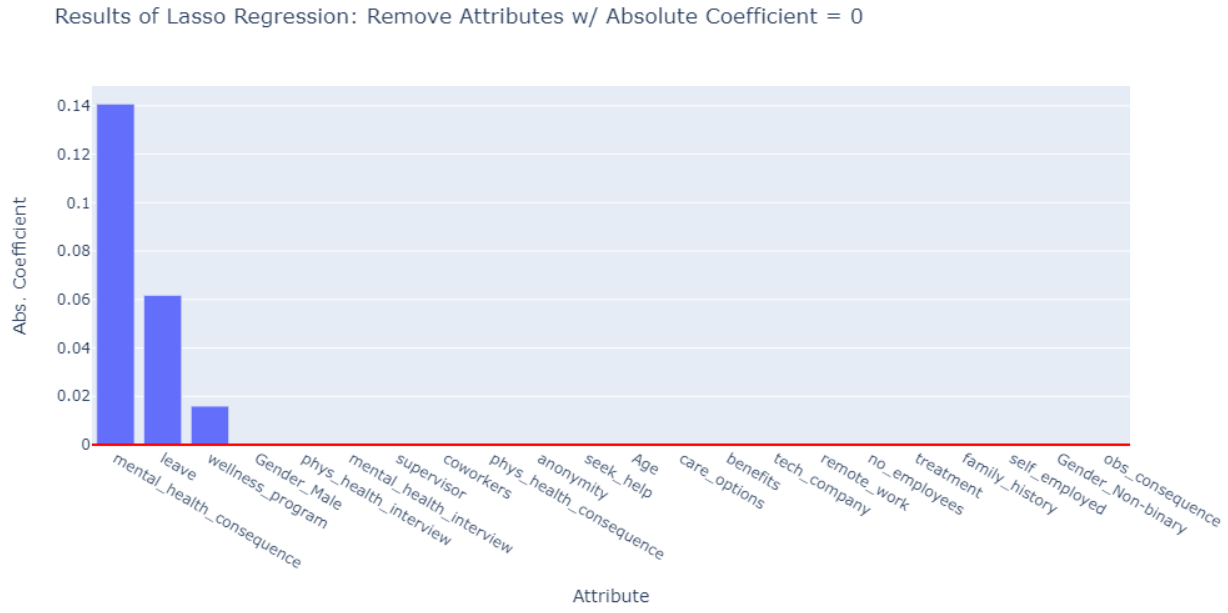


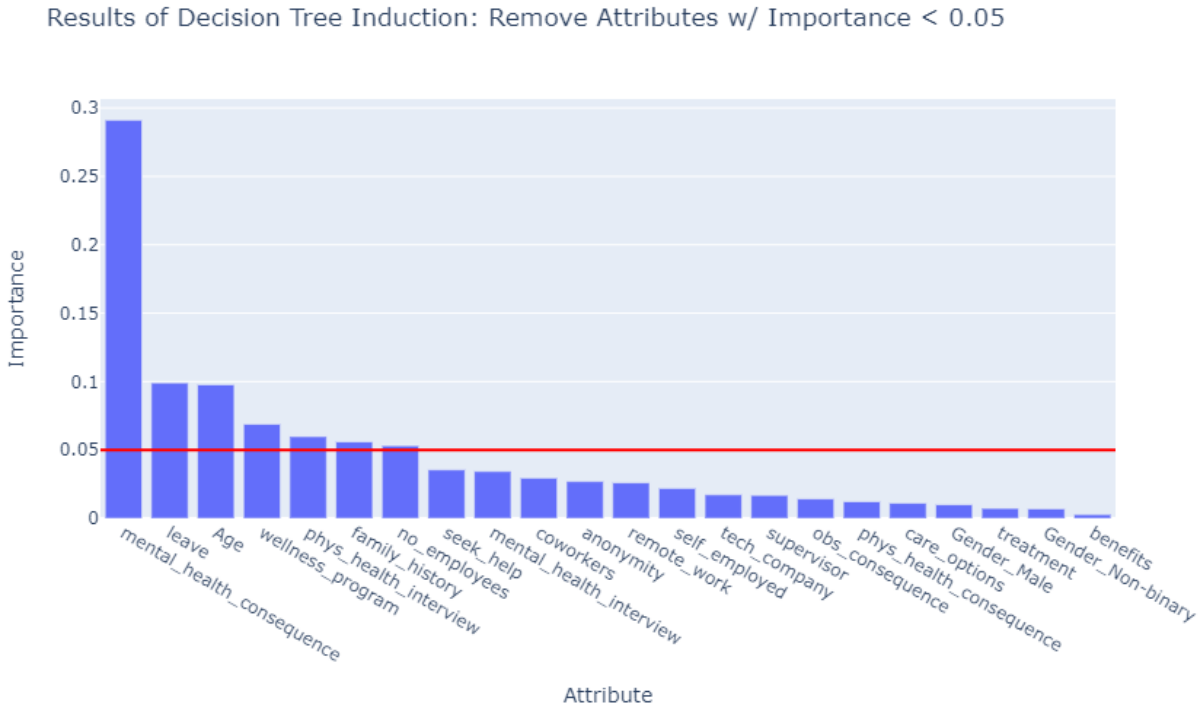
C. K-Nearest Neighbors Classifier: Elbow Method



D. Attribute Selection Methods Criteria






Results of Decision Tree Induction: Remove Attributes w/ Importance < 0.05
Importance
0.3
0.25
0.2
0.15
0.1
0.05
0
mental_health_consequence
leave
Age
wellness_program
phys_health_interview
family_history
no_employees
seek_help
mental_health_interview
coworkers
anonymity
remote_work
self_employed
tech_company
supervisor
obs_consequence
phys_health_consequence
care_options
Gender_Male
treatment
Gender_Non-binary
benefits
Attribute