Proper Drug Prediction

1st Jyothi Sai Moganti
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States of America
jmoganti@stevens.edu

2nd Anirudh Nagendra Srivatsa
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States of America
asrivats@stevens.edu

3rd Lakshmi Narasimha Reddy Guda
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States of America
lguda@stevens.edu

***Abstract*—In this modern world, people are easily affected by diseases. A specific disease must be treated with the appropriate drug, and not everyone with that disease responds to the same treatment. So how do we discover the ideal medication for a disease? Based on a person's many traits, the right medication can be predicted. This can be accomplished by utilizing data collection that includes comprehensive details about a patient suffering from a sickness. One of the project's primary goals is to identify these traits. The features can be represented visually to assist us to decide which are important. For this, we are using three machine learning algorithms: Logistic Regression, Decision Tree, and Random Forest.**

I. Introduction

Millions of people are impacted each year by improper pharmaceutical prescriptions, making drug advice essential. Negative side effects might result from drug interactions and different drugs used to treat disorders of the same kind. It is essential to find the best treatment for a condition depending on a person's qualities, and this process requires many different sorts of attributes and a vast amount of data. Females who are over 40 typically have elevated cholesterol levels. For each sex, high cholesterol manifests itself differently. In the United States, there are more than 59 million men and 71 million women who have high or very high (240 mg/dL or greater) cholesterol, so it's crucial to know how your sex influences your risk and what you can do to lower it. Women in their twenties and thirties have the lowest normal diastolic readings (115.5-70.5) and the lowest normal systolic readings (110.5/72.5), respectively. Women between the ages of 56 and 60 have the highest normal blood pressure readings (132.5/78.5), while men between the ages of 31 and 35 have the lowest readings (114.5/75.5). Men between the ages of 61 and 65 have the highest normal blood pressure measurement (143.5/76.5). While prescribing a medication, all of these aspects should be taken into account. We therefore seek to accomplish this by combining Python programming with machine learning methods. Algorithms for machine learning can be used to decide which medicine is most effective for a specific ailment. The goal of this project is to provide various classification-based insights. Three algorithms—Random Forest, Decision Trees, and Logistic Regression—will be used for this project.

II. Related Work

Despite the regulatory difficulties that medical device manufacturers now face, the application of machine learning in digital health has become more widespread. Based on a dataset of 90 patients, Karanasiou et al.15 [1] used eleven classification algorithms, such as support vector machines (SVM) and Bayesian networks, to predict the adherence of patients with heart failure, and Franklin et al.16 used ten different machine learning models to predict patient adherence over the course of the next 30, 60, and 90 days using data from Medicare enrolment files and medical and pharmacy claims. In this study, we set out to demonstrate the potential and utility of such technologies in comprehending and perhaps even resolving adherence at a considerably more appealing data scale.

The forecasting of a binary assay result, which shows whether a certain drug, for instance, binds to a particular receptor, inhibits a particular pathway, or causes harmful effects. More specifically, despite the fact that the assays under consideration all have the same biomolecular target, each ChEMBL [2] experiment is treated as a separate classification problem. Support vector machines28 (SVMs) and K-nearest-neighbors (KNNs) as examples of similarity-based classification methods, and random forests29 (RFs) as an example of a feature-based classification method, are just a few of the methods that Andreas Mayr used to compare the prediction performances of several deep learning architectures. He also included the target prediction techniques naive bayes (NB) and SEA30-32 in the comparison because he saw them as examples of approaches created particularly for the aim of drug discovery.

III. Our Solution

In order to forecast the correct outcomes, machine learning techniques can be used. Our objective is to anticipate which medication is best for a patient with a given ailment based on the patient's age, gender, blood pressure, cholesterol, and sodium to potassium levels.

A. Description of Dataset

The Dataset is not too large. The data contains characteristics of the person with a particular disease. Consider yourself a medical researcher gathering information for a study. You have gathered information on a group of patients who were all afflicted with the same ailment. Each patient responded to one of five drugs—Drug A, Drug B, Drug C, Drug X, and Drug Y—during the course of their treatment. Building a model to determine which medication would be suitable for a future patient with the same ailment is part of your work. The target is the treatment that each patient responded to, and the features of this dataset include the patient's age, sex, blood pressure, cholesterol, and sodium

to potassium levels(The ratio of sodium to potassium levels in the person's blood). It is a sample of a multiclass classifier, and you may use the dataset's training portion to create a decision tree, which you can then use to determine the patient's class or to recommend a medication to a new patient.

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23.0	F	HIGH	HIGH	25.355	drugY
1	47.0	M	LOW	HIGH	13.093	drugC
2	47.0	M	LOW	HIGH	10.114	drugC
3	NaN	F	NORMAL	HIGH	NaN	drugX
4	61.0	F	LOW	HIGH	18.043	drugY
5	22.0	F	NORMAL	HIGH	8.607	drugX
6	49.0	F	NORMAL	HIGH	16.275	drugY
7	NaN	M	LOW	HIGH	11.037	drugC
8	60.0	M	NORMAL	HIGH	15.171	drugY
9	43.0	M	LOW	NORMAL	19.368	drugY

Fig. 1. Sample of the Dataset

B. Machine Learning Algorithms

We are using the following three algorithms:

1. Logistic Regression

2. Decision Trees

3. Random Forest

Multinomial logistic regression is used when you have a categorical dependent variable with two or more unordered levels (i.e. two or more discrete outcomes). It is practically identical to logistic regression, except that you have numerous possible outcomes instead of just one. The dependent variable's first level is picked as the reference level. This is often the most common or the most frequent category. The likelihood of being in the reference category is contrasted with the likelihood of being in any of the other categories. These relative probabilities are the expected log odds (the logarithmic of the odds). Software is typically used to accomplish this kind of regression. Essentially, a succession of individual binomial logistic regressions for M - 1 categories will be performed by the software (one calculation for each category, minus the reference category). Ordered logistic regression, logistic regression, and multinomial logistic regression all have the same results when M = 2. You would have performed each of these separate regressions and then compared the results before the invention of computer software. The software takes away that effort, and calculates parameters simultaneously, resulting in better efficiency.Assumptions include: There are no unnecessary variables and the model is adequately described, Cases are independent, The independent variables do not exhibit multicollinearity. Similar to other types of regression, multinomial logistic regression searches for a correlation between the independent and dependent variables. You will receive sets of coefficients for each variable in the output. Each software package will produce a different result. When it comes to interpreting results, UCLA provides several great resources.

The supervised learning algorithms family includes the decision tree algorithm. The decision tree technique, in contrast to other supervised learning methods, is capable of handling both classification and regression issues. To decide whether to divide a node into two or more sub-nodes, decision trees employ various algorithms like ID3, CART, etc.

Random Forest is a part of the supervised learning methodology. It can be applied to Machine Learning issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating various classifiers to address difficult issues and enhance model performance. Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Instead of depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions. More trees in the forest result in increased accuracy and mitigate the overfitting issue.

C. Implementation Details

To gather information and create predictions, the project will utilize techniques including Logistic Regression, Decision Trees, and Random Forest. The first stage is preprocessing, which involved filling in all the missing values with the column's average value because missing values were leading to the erroneous feature presentation. We have encoded the two classes F and M in the Age column with 1 and 0, respectively. The three grades Low, Normal, and High in the BP column are encoded with 0, 1, and 2 accordingly. We have encoded the Normal and High classes of the cholesterol column with 0, and 1, respectively. Additionally, the target column, the Drug column, is encoded as well. Given one or more independent variables, Multinomial Logistic Regression is a classification technique that extends the logistic regression algorithm to tackle multiclass probable outcome issues. The probabilities of the categorically dependent variable, which has two or more possible result classes, are predicted using this model. The model predicts that the instance belongs to that class if the estimated probability is higher for that class. It is a multi-class classifier as a result. It makes use of the softmax function, which accepts a vector as input and produces another vector, each of whose components is a probability that the input vector belongs to the specified class.

For a better understanding of the data, we have created various visualizations which help us in improving our comprehension of the data and assist us in selecting the most crucial aspects. We have shown which medicine, DrugY, with the highest rank, is more suited for females and the medicine that DrugY maintains the highest rank in and is, therefore, more compatible for males is depicted in Fig 3. In Fig. 4, the cholesterol levels for different age groups are displayed visually, and it is clear that the greater than 55 age group has the highest cholesterol levels.

Figures 5 and 6 show the ratios of sodium to potassium for various age groups and genders, respectively. We can see

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23.000000	1.0	2.0	1.0	25.355000	4.0
1	47.000000	0.0	0.0	1.0	13.093000	2.0
2	47.000000	0.0	0.0	1.0	10.114000	2.0
3	45.121693	1.0	1.0	1.0	16.564952	3.0
4	61.000000	1.0	0.0	1.0	18.043000	4.0
5	22.000000	1.0	1.0	1.0	8.607000	3.0
6	49.000000	1.0	1.0	1.0	16.275000	4.0
7	45.121693	0.0	0.0	1.0	11.037000	2.0
8	60.000000	0.0	1.0	1.0	15.171000	4.0
9	43.000000	0.0	0.0	0.0	19.368000	4.0

Fig. 2. Sample of the Dataset after Pre-processing

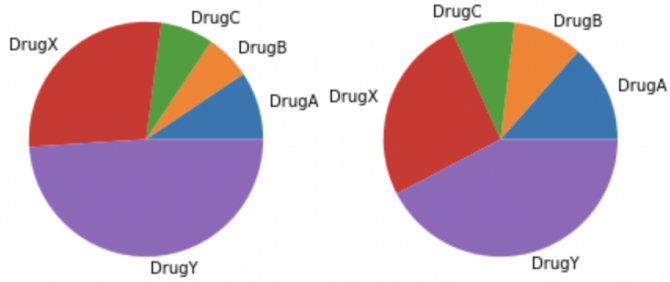


Fig. 3. Drug Compatibility for females and males

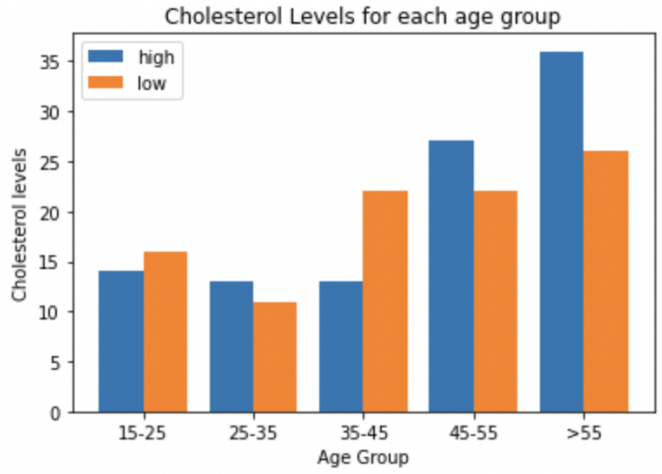


Fig. 4. Cholesterol Levels for each age group

that the sodium-to-potassium ratio is high for people between the ages of 15 and 25 in Fig. 5, and we can see that it is high for females in Fig. 6.

The blood pressure is shown in Figs. 7 and 8 for each age group, as well as for males and females, respectively. Figure 7 shows that the age group of 55 has a high proportion of people with high blood pressure, as well as a high proportion of people with normal blood pressure due to the age group's high population density, while the age group of 45 to 55 has

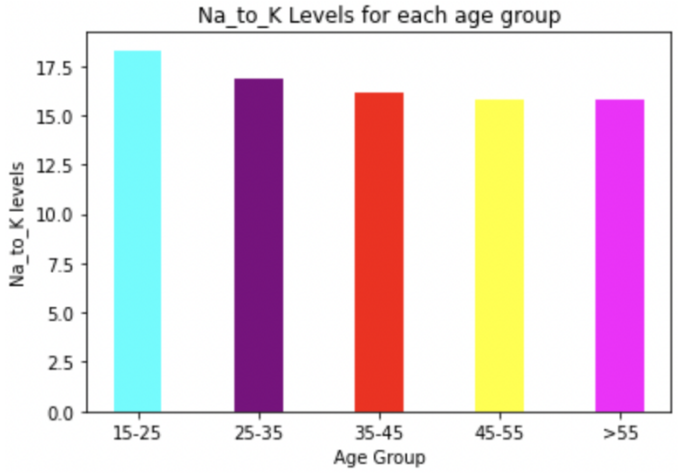


Fig. 5. Na to K Levels for each age group

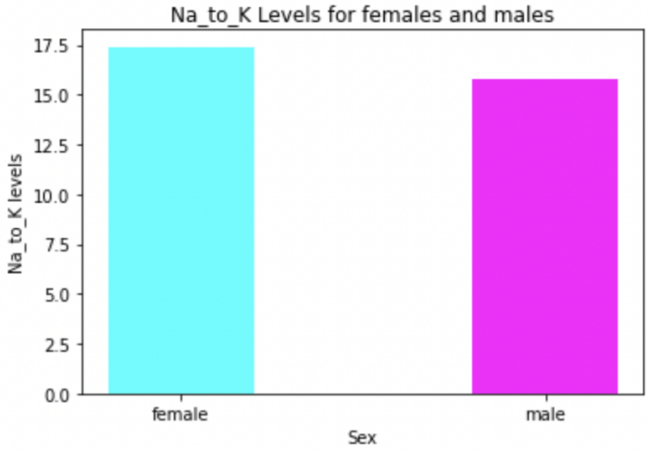


Fig. 6. Na to K Levels for females and males

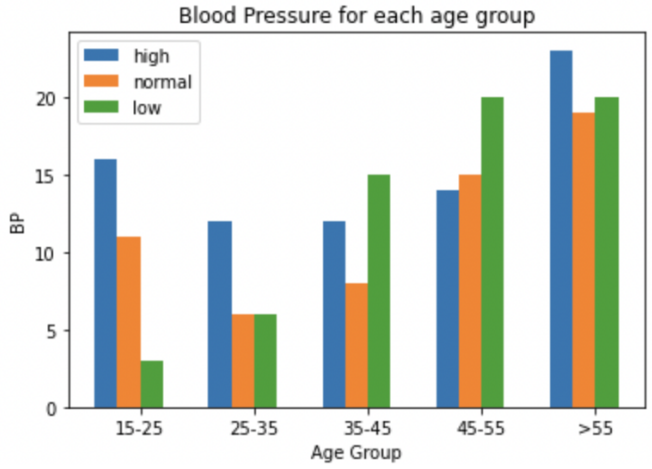


Fig. 7. Blood Pressure for each age group

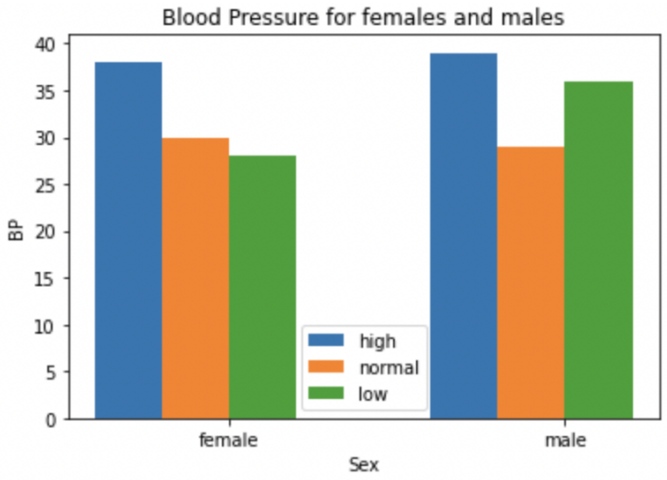


Fig. 8. Blood Pressure for females and males

a high proportion of people with low blood pressure. Figure 8 shows that men have a higher proportion of persons with high and low blood pressure, while women have a higher proportion of people with normal blood pressure. The algorithms that will be used in the project are listed below.

Since the Multinomial Logistic Regression algorithm performs best when there are three or more potential outcomes, this method was chosen. The five target classes in our dataset—drugA, drugB, drugC, drugX, and drugY—have been encoded with the numbers 0, 1, 2, 3, and 4 in the Drug column. The accuracy from Logistic Regression we got is 82.5 percent and Fig 9 represents the confusion matrix. Fig 10 represents a classification report i.e the precision, recall, and F1-score for different classes of drugs.

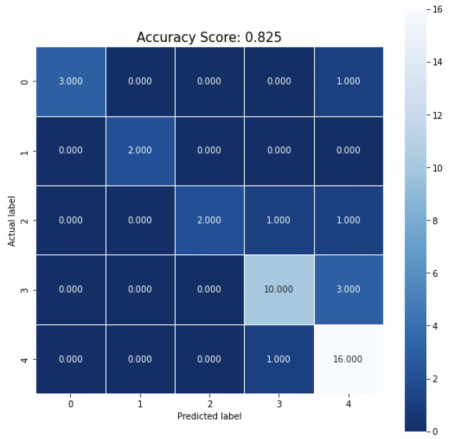


Fig. 9. Confusion Matrix

A softmax function is used in Multinomial Logistic Regression [3] to model the relationship between the predictors and probabilities of each class. Finally, it predicts the class with the highest probability out of all possible classes. The softmax function is demonstrated below.

$$P(y=j|z^{(i)}) = \phi_{softmax}(z^{(i)}) = \frac{e^{z^{(i)}}}{\sum_{j=0}^{k} e^{z_k^{(i)}}} \quad (1)$$

Accuracy Score: 0.825

	precision	recall	f1-score	support
0.0	1.00	0.75	0.86	4
1.0	1.00	1.00	1.00	2
2.0	1.00	0.50	0.67	4
3.0	0.83	0.77	0.80	13
4.0	0.76	0.94	0.84	17
accuracy			0.82	40
macro avg	0.92	0.79	0.83	40
weighted avg	0.84	0.82	0.82	40

Fig. 10. Classification Report

Here, j is the class of the input observation I and can range from 0 to k, where k is the number of classes that can be assigned to the input observation. The term $\sum_{j=0}^{k} e^{z_k^{(i)}}$ normalizes the distribution by making the sum of the probabilities for each class equal to one. z is the net input vector and is given as

$$z = w_1x_1 + w_2x_2 + ... + w_mx_m + b = \sum_{l=1}^{m} w_lx_l + b = w^Tx + b \quad (2)$$

.

w is the weight vector, x is the feature vector of 1 training sample, and b is the bias unit.

For a training sample, we already know the feature vector X. The goal is to find the weight vectors w and b that bring the actual and predicted classes as close together as possible. The Maximum Likelihood criterion is used to determine this. It is given below for a Multinomial Logistic Regression.

$$P(Y|X) = \prod_{i=1}^{n} P(y^{(i)}|x^{(i)}) - \log P(Y|X) = \prod_{i=1}^{n} -\log P(y^{(i)}|x^{(i)}) \quad (3)$$

Maximizing likelihood function $P(Y|X)$ is the same as minimizing the function $-\log P(Y|X)$ .As a result, we can define a loss function as follows:

$$l = -\log P(Y|X) = -\sum_{j} y_j \log y_j \quad (4)$$

Typically, y is a one-hot encoded vector with j=1 for one class and j=0 for the other. Furthermore, because y_j is a probability ranging from 0 to 1, their logarithms are never greater than 0. When y_j, the loss function is 0. As a result, if the classifier correctly classifies the class, the loss function is minimized. This loss function is also known as cross-entropy loss. We can further simplify the loss function by substituting the value of the softmax function for y_j, as shown below.

$$l = \log \sum_{k} exp(z_k) - \sum_{j} y_j z_j \quad (5)$$

Decision Trees are a type of supervised machine learning in which the training data is continuously segmented based on a particular parameter, with you describing the input and the corresponding output. We can predict the most appropriate medication for a disease by dividing the training data into specific parameters. The decision tree algorithm is represented by a tree in which nodes represent attributes, branches represent rules, and leaf nodes represent outcomes (discrete and

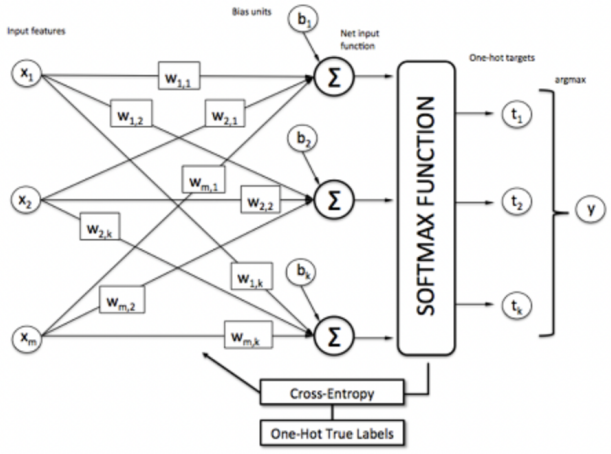


Fig. 11. Multinomial logistic regression

continuous). Attribute selection is commonly performed using two techniques: 1) Information Gain and 2) Gini Index.

Information gain is the opposite of entropy, which measures entropy decrease. Information gain computes the difference between the dataset's entropy before and after splitting based on given attribute values.

$$Gain(T, X) = Entropy(T) - Entropy(T, X) \quad (6)$$

where T is the target variable and X is the feature to be split on.

The Gini index measures inequality in a sample. Its value ranges from 0 to 1. A Gini index of 0 indicates that the sample is perfectly homogeneous and that all elements are similar, whereas a Gini index of 1 indicates that there is maximal inequality among elements. It is the sum of the squares of each class probabilities. It is depicted as follows:

$$GiniIndex = 1 - \sum_{(i=1)}^{n} P_i^2 \quad (7)$$

Classification Report:

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	4
1.0	1.00	1.00	1.00	2
2.0	1.00	0.75	0.86	4
3.0	0.93	1.00	0.96	13
4.0	1.00	1.00	1.00	17
accuracy			0.97	40
macro avg	0.99	0.95	0.96	40
weighted avg	0.98	0.97	0.97	40

Fig. 12. Classification Report before pruning

In Decision Tree pruning, the branches of the decision tree are removed to overcome the overfitting condition of the decision tree. This pruning can be done by hyperparameter tuning. Hyperparameter tuning can be done by using GridSearchCV. Grid search is a technique for tuning hyperparameters that can help you build and test models for any combination of algorithm parameters per grid. To find the best value for that tuning hyperparameter, we could use 10 fold cross-validation. Decision criterion parameters such as max depth, min sample split, and so on. These are known as hyperparameters. The Grid Search method will be used to find the simplest set of hyperparameters. In the Grid Search, all hyperparameter mixtures will be passed through the model one by one, and the score for each model will be calculated. It returns the set of hyperparameters with the highest score. It requires the model or objects to be trained as well as various hyperparameter values. the error for various hyperparameter values, allowing you to select the best ones. It then computes the error for various hyperparameter values, allowing you to select the best ones.

Classification report:

	precision	recall	f1-score	support
0.0	1.00	0.75	0.86	4
1.0	1.00	1.00	1.00	2
2.0	1.00	0.75	0.86	4
3.0	0.87	1.00	0.93	13
4.0	1.00	1.00	1.00	17
accuracy			0.95	40
macro avg	0.97	0.90	0.93	40
weighted avg	0.96	0.95	0.95	40

Fig. 13. Classification Report after pruning

As its name suggests, a random forest is made up of a lot of different decision trees working together as an ensemble. The class with the most votes becomes the prediction made by our model. The random forest's individual trees each spit out a class prediction. Consequently, given that there are five target classes (drugA, drugB, drugC, drugX, and drugY), the drug class with the most votes will be the most accurate in predicting a person's disease. The decrease in node impurity weighted by the probability of reaching that node is used to calculate feature importance. The number of samples that reach the node divided by the total number of samples yields the node probability. The more important the feature, the higher the value.

$$ni_j = w_j C_j - w_{left(j)} * C_{left(j)} - w_{right(j)} * C_{right(j)} \quad (8)$$

ni sub(j)= the importance of node j,w sub(j) = weighted number of samples reaching node j, C sub(j)= the impurity value of node j, left(j) = child node from left split on node j, right(j) = child node from right split on node j.

The significance of each feature on a decision tree is then calculated as follows:

$$fi_i = \frac{\sum_{j:nodejsplitsonfeaturei} ni_j}{\sum_{k \in allnodes} ni_k} \quad (9)$$

fi sub(i)= the importance of feature i, ni sub(j)= the importance of node j.
These can then be normalized to a range of 0 to 1 by dividing

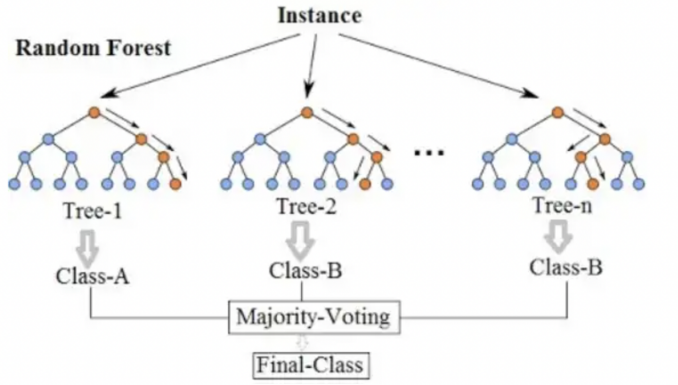
by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\sum_{j \in allfeatures} fi_j} \tag{10}$$

At the Random Forest level, the final feature importance is its average across all trees. The total number of trees is divided by the sum of the feature importance values on each tree:

$$RFfi_i = \frac{\sum_{j \in alltrees} normfi_{ij}}{T} \tag{11}$$

RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model, normfi sub(ij)= the normalized feature importance for i in tree j, T = total number of trees



Various decision trees are used in a random forest system. Decision nodes, leaf nodes, and root node make up every decision tree. The leaf node of each tree represents the decision tree's final output. The final output is chosen using a majority-voting system. In this case, the output chosen by the majority of decision trees becomes the random forest system's final output.

Classification Report:

	precision	recall	f1-score	support
0.0	1.00	0.75	0.86	4
1.0	1.00	1.00	1.00	2
2.0	1.00	0.50	0.67	4
3.0	0.92	0.85	0.88	13
4.0	0.81	1.00	0.89	17
accuracy			0.88	40
macro avg	0.95	0.82	0.86	40
weighted avg	0.89	0.88	0.87	40

Fig. 14. Classification Report of Random Forest

IV. Comparison

After implementing all 3 algorithms it has been found that the decision tree algorithm has the least RMSE with a value of 0.15 and the highest accuracy of 97.5 percent. The Multinomial Logistic regression manages to achieve a value of 0.79 of RMSE and an accuracy of 82.5 percent. The Decision Tree manages to achieve a value of 0.15 of RMSE and an accuracy of 97.5 percent. The Random Forest manages to achieve a value of 0.75 of RMSE and an accuracy of 87.5 percent.

V. Future Directions

In the future, we hope to find commonalities or features across individuals and forecast which medication would work best for a given patient. We would find the most accurate algorithm for prediction. Provided some more time, we could increase the accuracy of correctly predicting a proper drug for the patient.

VI. Conclusion

After implementation of all three machine learning algorithms, we noticed that the RMSE value of the Decision Tree algorithm is the least which gives us a better prediction on our problem statement.

References

[1] Y. Gu, M. Liu, and A. Hall, "Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data," 2021.

[2] A. Mayr, G. Klambauer, and M. Steijaert, "Large-scale comparison of machine learning methods for drug target prediction on chembl," *Chemical Science*, 2018.

[3] J. Zhang and A. Enam, "Multinomial logistic regression," *Mapping the Travel Behavior Genome*, 2020.