iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IX **Month of publication:** September 2021

DOI: https://doi.org/10.22214/ijraset.2021.38203

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IX Sep 2021- Available at www.ijraset.com

Drug Classification using Black-box models and Interpretability

Pooja Thakkar[1], Kartik Mane[2]
[1, 2]*Department of Information Technology, Shri Govindram Seksaria Institute of Technology and Science, Indore, India*

Abstract: ***The focus of this study is on drug categorization utilising Machine Learning models, as well as interpretability utilizing LIME and SHAP to get a thorough understanding of the ML models. To do this, the researchers used machine learning models such as random forest, decision tree, and logistic regression to classify drugs. Then, using LIME and SHAP, they determined if these models were interpretable, which allowed them to better understand their results. It may be stated at the conclusion of this paper that LIME and SHAP can be utilised to get insight into a Machine Learning model and determine which attribute is accountable for the divergence in the outcomes. According to the LIME and SHAP results, it is also discovered that Random Forest and Decision Tree ML models are the best models to employ for drug classification, with Na to K and BP being the most significant characteristics for drug classification.***
Keywords: Machine Learning, Back-box models, LIME, SHAP, Decision Tree

I. INTRODUCTION

Drug classification is one of the most significant endeavours in the area of pharmacology, and it is also one of the most difficult. In medicine, drug repurposing alludes to using existing medications to treat diseases that were not initially specified as criterion for the use of the medication when first created. Because clinical trials are streamlined in this scenario [1], [2], it is less costly and faster to utilise this technique than it is to develop a new drug. 279 drugs out of 6733 prescription medications were reprocessed between 1998 and 2016, as per the United States Food and Drug Administration [2]. For example, the administration of an antibiotic including erythromycin for the treatment of gastric motility disorder or the administration of an anti-emetic drug including thalidomide for the treatment of multiple myeloma [1] are both illustrations of such situations. It is necessary to identify new adverse effects from pharmaceuticals, both helpful and harmful, in order for medication repurposing to be effective. As a result of a lack of necessary information and the need of a considerable amount of time [3], it is challenging to examine all of the presently available medications for each disease and doing so would consume an excessive amount of time. Finding new possible adverse drug reactions is also an important task, since unexpected side effects can have a severe influence on the patient's health, and in some cases their lives, if they occur. In order to address this problem, this research article offers a drug categorization technique that would aid in the classification of all medicines according to their intended use. This will make the process of repurposing drugs much simpler.

A. Aim

This research aims to use machine learning models such as random forest, decision tree, and logistic regression to classify drugs. Then, employing LIME and SHAP, they determined if these models were interpretable, which allowed them to better understand their results. Towards the conclusion of the project, a comparison study will be conducted to determine which model achieved the best performance.

II. RELATED WORK

Large quantities of pharmacological information are being produced as a result of the growth of the pharmaceutical business and high-throughput analysis of the genetic code, and numerous clinical and medication databases are now accessible to the public on the internet. Databases such as DrugBank [4], the Therapeutic Target Database (TTD) [5], and the Drug Gene Interaction Database (DGIdb) [6] comprise experimentally demonstrated drug-target and medication details, which allow to obtain thorough multiomics drug information from a single source. Due to the sheer fast development of high-throughput methods, large numbers of pharmacological gene transcription profiles, like the Connectivity Map (CMap) [7], are being amassed. Additionally, other medication information, such as enzymes, adverse effects, and pathways, may be found in a number of databases (e.g., DrugBank, KEGG, and SIDER). The Medication Combination Database (DCDB) [8], for example, collects existing drug combinations through data portals like the Drug Combination Database (DCDB).

©IJRASET: All Rights are Reserved

The DCDB has a total of 1363 pairings of drugs in total which is a result of drug classification. One may tackle data-driven prediction issues, such as medication combinations [8] using these enhanced sources of pharmacological knowledge.

Following the premise that "similar medicines exhibit comparable activity," numerous new approaches and procedures have been developed in order to anticipate the synergistic effects of pharmacological combinations, and these techniques and methods are still being refined. These techniques attempted to anticipate novel drug combinations based on their resemblance to already discovered medication combinations [9]. Cheng et al. [10] used four similarity-based characteristics to apply five different types of algorithms naive Bayes, decision tree, k-nearest neighbour, logistic regression, and support vector machine (SVM)] to four different types of data. When Vilar and colleagues [11] developed the chemical structural similarity-based prediction technique, they predicted a significant number of novel combinations. This problem has been studied by Zang [12], who built prediction models by combining methods such as neighbour recommender technique, random walk method, and matrix perturbation process. They went on to further investigate matrix factorization technique and ensemble approach on this problem. For the detection of unknown medication combinations, Shi et al. [13] devised an iterative matrix factorization approach that included a DDI network and a drug side consequences vector feature. Lee et al. [14] developed a deep learning network using autoencoders in order to correctly identify more medication combinations with greater precision. In order to routinely identify unknown medication mixes with multi-feature drug characteristics, these techniques provide promising and practicable methodologies that are easy to implement.

All the above studies revolve around drug combination but none of them have mentioned anything regarding drug classification. This is what this report will try to address and to even enhance it more, along with drug classification, LIME and SHAP models will be used to find interpretability.

III. METHODOLOGY AND DATASET

A. *Machine Learning and Interpretability*

Machine learning is amongst the best prominent topics at now since it enables computers to comprehend data from data and to generate predictions without needing to be explicitly programmed for that goal. Supervised learning is amongst two important machinery-learning divisions which, once educated on past data, enables a framework to predict future results. One uses input/output pairings or labelled information to form the framework in attempt to produce a feature that is sufficiently approximated to be capable of forecasting outputs for fresh inputs when supplied with the model, the main goal of supervised learning. The two kinds of supervised learning tasks that can be met are challenges with regression and classification. If the outputs are uniform, a regression problem will occur, whereas a classification difficulty occurs when the outcomes are categorical.

And besides, why shouldn't one just accept and ignore the reasoning behind his choice if an algorithm for machine learning works well? There is not enough one single metric, such as classification accuracy, to describe the bulk of actual operations. Five instances of this are offered by Doshi-Velez and Kim [15]. See now more closely why interpretability is so important. When it comes to predictive modelling, two options must be chosen: Do users just want to know what the prediction says? For example, the chance of a customer leaving or the efficacy of a certain patient's medicine. Otherwise, do users desire to see the reason why a prediction is produced and are ready to pay the price for interpretability of reducing predictive accuracy? The rationale behind a decision is not always essential; it is enough to realise that in some cases the projected ability on the test dataset was acceptable. However, knowing the "why" might help people learn more about the problem, the data, and the rationale why a system can strive to do as anticipated.

The perception of a model of machine learning reflects how easy it is for people to understand the methods used by the algorithm to get their findings. Artificial intelligence (AIs) algorithms have long been notorious for being "black boxes," not to understand their internal functions and to prevent the findings obtained in the wake of regulators and other stakeholders from being justified. Interpretability is considered excellent in the case of basic models such as logistic regression. When features are introduced, however, even more complicated methods such as deep learning become more difficult to understand.

B. *Decision Tree*

There are three components of a decision tree: decision nodes, leaf nodes and a root node (or node at the bottom of the tree). A decision tree approach divides a training set into branches, which are split by a decision tree methodology into like branches. This is continued until the first node of the leaf is attained. The node of the leaf can no longer be detached [16]. In order to predict the outcome, the features reflected by the decision tree nodes are used. The decision nodes are used to link the leaves. Figure 1 shows the three distinct vertical types in a decision tree.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IX Sep 2021- Available at www.ijraset.com

Information gains are considered while training decision-making trees. It helps to alleviate the uncertainty of these trees. A substantial gain of information shows that the circumstance has removed a major uncertainty (information entropy). Entropy and information gain are important variables in the separation of branches, which is a key component for the construction of decision trees.

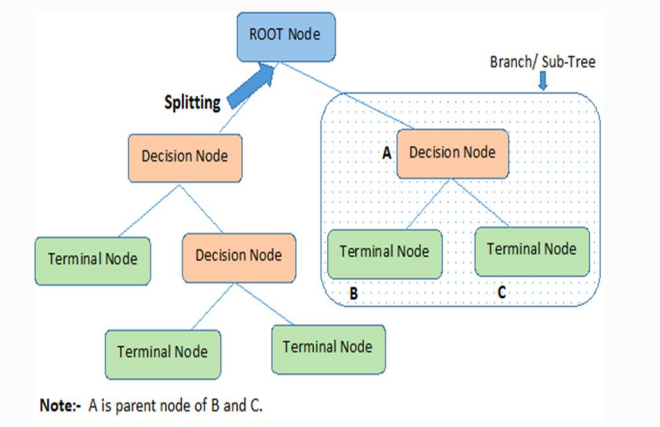


Figure 1: Decision Tree

C. Random Forest

Supervised machine learning techniques such as decision trees are used to construct a random forest. A random forest technique consists of a collection of decision trees that are connected to one another. The bagging or bootstrapping merging techniques are used to teach the random forest approach how to create the resulting "forest." Bagging is a meta-algorithm that aggregates information and improves the effectiveness of machine learning systems. The (random forest) approach generates a result based on the predictions provided by the decision trees in a random manner. In order to predict, it considers the average or median of the results from various trees. The correctness of the output is improved by increasing the number of trees used. A random forest classifier overcomes the limitations of a decision tree classifier. It reduces overfitting of data while increasing accuracy [17].

The primary difference between the decision tree and the random forest approach is that in the latter, the process of generating root nodes and splitting nodes is done randomly rather than sequentially. The bagging method is used by the random forest in order to generate the required prediction. Bagging is the process of using many data samples (training data) rather than a single occurrence of a phenomenon. A training dataset is a collection of events and attributes that are used to generate predictions by a machine learning algorithm. Random forest models use training data to produce a variety of outputs, which are determined by the decision trees generated by the random forest model. Following this evaluation, the most favourable result will be selected for use as the final outcome.

D. Logistic Regression

When the dependent variable (target) is categorical, the Logistic Regression method is employed. Based on an input variable, logistic regression may be used to predict the likelihood of a discrete result occurring in the future. One of the most frequent results in logistic regression models is a binary result; anything that can be expressed as true or false, yes, or no, and so on. Logistic regression is a straightforward and highly efficient technique for categorising issues that are either binary or linear in nature. When dealing with linearly separable classes, this classification model is extremely simple to implement and produces very excellent results in terms of speed. In the industrial sector, it is a widely used method for classification purposes.

E. Dataset Pre-processing

The dataset was obtained from Kaggle.com, and it is a Drug Classification dataset [18] in nature. The drug type is the characteristic that is being sought.

The feature sets are:

1) Age
2) Sex
3) Blood Pressure Levels (BP)
4) Cholesterol Levels
5) Na to Potassium Ration

©IJRASET: All Rights are Reserved

In ability to use NumPy and pandas, the first stage is to import these necessary libraries. The dataset is then presented, with columns for age, gender, blood pressure, cholesterol, sodium, and drug categorization. Figure 2 depicts the specifics of this situation.

```
data.head()
```

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23	F	HIGH	HIGH	25.355	DrugY
1	47	M	LOW	HIGH	13.093	drugC
2	47	M	LOW	HIGH	10.114	drugC
3	28	F	NORMAL	HIGH	7.798	drugX
4	61	F	LOW	HIGH	18.043	DrugY

```
data.columns
```

```
Index(['Age', 'Sex', 'BP', 'Cholesterol', 'Na_to_K', 'Drug'], dtype='object')
```

Figure 2: Dataset details

Figure 3 has the detailed information about the dataset, including its count, average, standard deviation, and more. The mean age is 44.3, and the count of age column has 200 values.

```
data.describe()
```

	Age	Na_to_K
count	200.000000	200.000000
mean	44.315000	16.084485
std	16.544315	7.223956
min	15.000000	6.269000
25%	31.000000	10.445500
50%	45.000000	13.936500
75%	58.000000	19.380000
max	74.000000	38.247000

Figure 3: Data description

Figure 4 shows the dataset has no null values.

```
Age            0
Sex            0
BP             0
Cholesterol    0
Na_to_K        0
Drug           0
dtype: int64
```

Figure 4: Null value

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IX Sep 2021- Available at www.ijraset.com

Figure 5 contains a chart of column age, showing the maximum age in the sample and that there are more individuals aged 40-50 with count of 20-30.

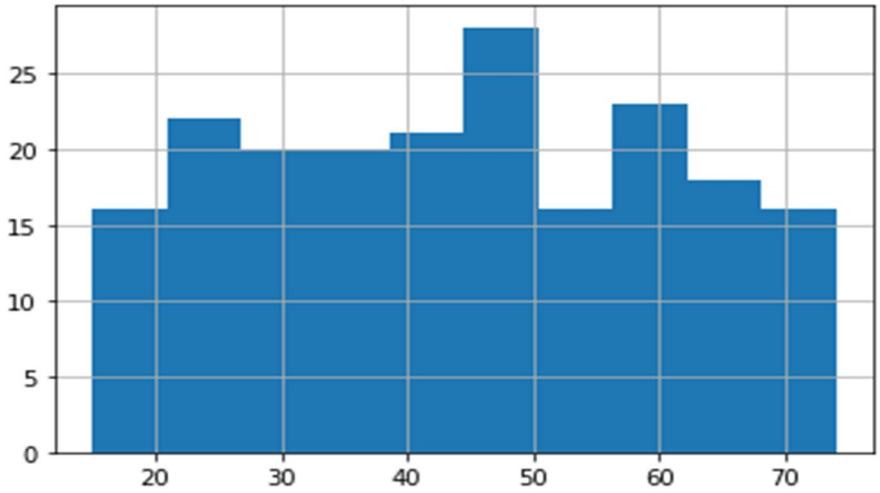


Figure 5: Age graph

The chart shown in figure 6 reveals the gender statistics. In the dataset, it is noted that men tend to have a score closer to 0.5, while females have a lower score, on average.

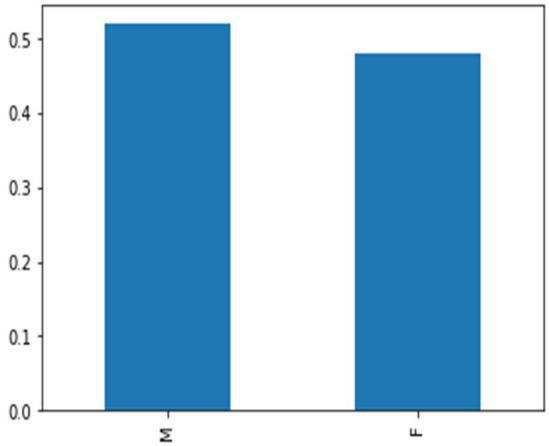


Figure 6: Sex graph

The procedure proceeds by labelling data, as seen in Figure 8. To get a number for each class, all values between 0 and the number of classes, less 1, are included. Python utilises a technique called label encoding. Classes may be referred to as being of "four distinct classes" when the number of classes that comprise the category parameter value is five (0, 1, 2, 3, and 4).

```
[ ]  # Label encoding

     from sklearn.preprocessing import LabelEncoder
     from sklearn.pipeline import Pipeline

     columns_list = ['Sex', 'BP', 'Cholesterol', 'Drug']

     encoding_pipeline = Pipeline([
         ('encoding',MultiColumnLabelEncoder(columns= columns_list))
         # add more pipeline steps as needed
     ])
```

Figure 7: Label encoding

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IX Sep 2021- Available at www.ijraset.com

One can see how each column is encoded in Figure 9. Each column value is 0, 1, 2, 3, or 4 for all the distinct classes of the column.

	Sex	BP	Cholesterol	Drug
0	0	0	0	0
1	1	1	0	3
2	1	1	0	3
3	0	2	0	4
4	0	1	0	0
...	...	...	...	...
195	0	1	0	3
196	1	1	0	3
197	1	2	0	4
198	1	2	1	4
199	0	1	1	4

200 rows × 4 columns

Figure 8: Encoded data

Figure 10 shows the final dataset.

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23	0	0	0	25.355	0
1	47	1	1	0	13.093	3
2	47	1	1	0	10.114	3
3	28	0	2	0	7.798	4
4	61	0	1	0	18.043	0

Figure 9: Final dataset

The last step is to separate the dataset into training and testing sets in an attempt to do forecasts.

IV. APPROACH

In this section, the implementation of decision tree, random forest, and logistic regression is done. LIME and SHAP models will be used to analyse the output and feature importance of these models.

In this part, the researcher will use decision tree, random forest, and logistic regression to execute the dataset for drug classification. For analysis of output and feature significance, the researcher will use LIME and SHAP models.

The technique, LIME, which refers to Local Interpretable Model-Agnostic Explanations, describes classifier and regressor outputs in a local way by utilising an interpretable model. LIME, as contrast to a global interpretable framework, avoids making general statements about the process. It modifies a single data sample's feature value, then looks at the modifications' impact on the final result. It provides explanations for the predictions each test dataset generates. A local interpretability [19] is a result of the LIME feature which influences a prediction on a single sample by organising the many reasons each feature has an effect.

©IJRASET: All Rights are Reserved

SHAP is an acronym that exists for Shapley Additive exPlanations, which is a special algorithm developed by Lloyd Shapley to address the issues faced by cooperative game theory. It was developed in 1951, the same year Lloyd Shapley suggested his Shapley value solution as a method to solve cooperative game theory's issues. With any kind of machine learning or deep cognitive technique, including deep learning models, SHAP is very successful [20].

A. Decision Tree

The Decision Tree Classifier is used to import the decision tree, which has to be fitted with the data. The model scored 0.975 on its accuracy. The number of classifications a model correctly anticipates, divided by the total number of classifications the model predicts, represents accuracy. The model's F1 score was computed in attempt to comprehend its predictions, and the result was 0.974. When this step is complete, the feature significance of the model is computed, and it's unspecific as the following - Feature importance: [0.14028556 0. 0.27325642 0.12002209 0.46643593].

The problem was fixed by using LIME SHAP models.

1) LIME: To make predictions, the report will utilise the rows - 19 and 25.

Figure 11 shows the classifications and prediction rates for drugs a, b, c, x, and y. In a way, LIME can determine how each characteristic relates to a drug's categorization, via the feature's properties and value. The most significant divergence is to drugY, where the variation is 1.00. The characteristics that lead to drugY are shown in blue, whereas those that route to drugB is shown in green. Na to K is best at forecasting which drug it will benefit, and the deviation is pushing it into drugY. BP deviates in this situation towards non-drug B.

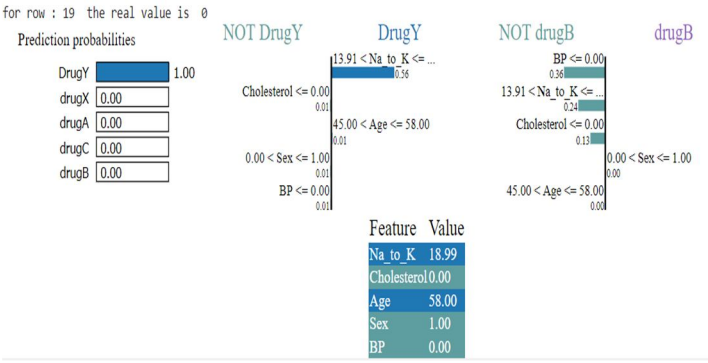


Figure 10: LIME for row 19

Figure 12 shows the classifications and prediction rates for drugs a, b, c, x, and y. In a way, LIME can determine how each characteristic relates to a drug's categorization, via the feature's properties and value. The most significant divergence is to drugC, where the variation is 1.00 in BP. The characteristics that lead to drugC are shown in blue, whereas those that route to drugB is shown in green. BP is best at forecasting which drug it will benefit, and the deviation is pushing it into drugC. Cholesterol deviates in this situation towards non-drug B and Na to K deviates towards drugB.

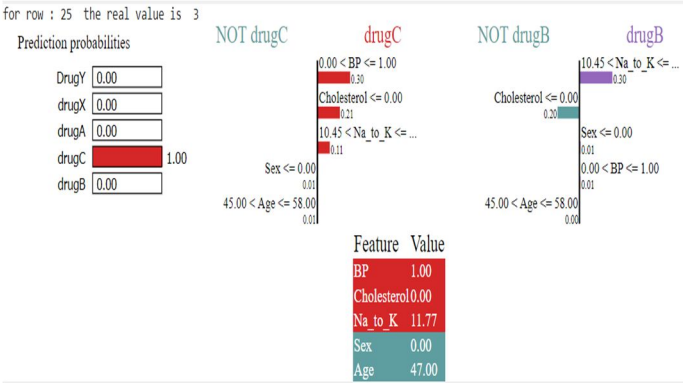


Figure 11: LIME for row 25

©IJRASET: All Rights are Reserved

2) SHAP

Figure 13 shows a summary graph of SHAP. The following notation specifies what class an attribute belongs to which class it's associated with:

DrugY – 4
drugX – 0
drugA – 1
drugC – 3
drugB – 2

For example, the purple line is greater in feature Na to K, indicating that class 0 is dominating this feature and class 0 Equals drugX. Every other feature follows the same logic.

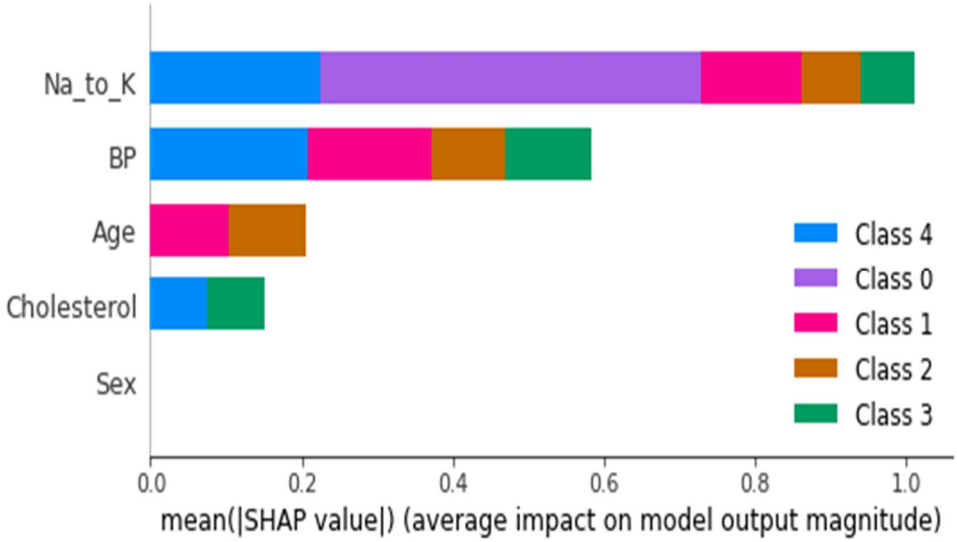


Figure 12: SHAP for Decision Tree

B. Random Forest

The Random Forest Classifier is used to import the random forest model, which has to be fitted with the data. The model scored 0.975 on its accuracy. The number of classifications a model correctly anticipates, divided by the total number of classifications the model predicts, represents accuracy. The model's F1 score was computed in attempt to comprehend its predictions, and the result was 0.974. The confusion matrix array is given below:

```
array([[20,  0,  0,  0,  0],
       [ 0,  3,  0,  0,  0],
       [ 0,  0,  3,  0,  0],
       [ 0,  0,  0,  3,  0],
       [ 1,  0,  0,  0, 10]])
```

When this step is complete, the feature significance of the model is computed, and it's unspecific as seen in figure 14. The problem was fixed by using LIME SHAP models.

```
print("feature importance value : ", result.importances_mean)

feature importance value :  [0.10875 0.      0.32625 0.08625 0.4775 ]
```

Figure 13: Feature importance random forest

©IJRASET: All Rights are Reserved

1) LIME: To make predictions, the report will utilise the rows - 19 and 25.

Figure 15 shows the classifications and prediction rates for drugs a, b, c, x, and y. In a way, LIME can determine how each characteristic relates to a drug's categorization, via the feature's properties and value. The most significant divergence is to drugY, where the variation is 0.94. The characteristics that lead to drugY are shown in blue, whereas those that route to drugA is shown in green. Na to K is best at forecasting which drug it will benefit, and the deviation is pushing it into drugY. BP deviates in this situation towards drugA.

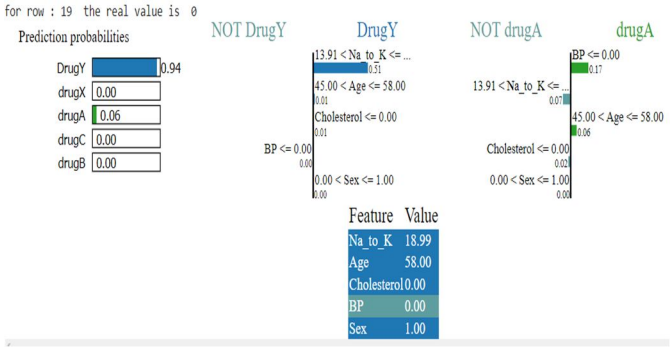


Figure 14: Lime for row 19

Figure 16 shows the classifications and prediction rates for drugs a, b, c, x, and y. In a way, LIME can determine how each characteristic relates to a drug's categorization, via the feature's properties and value. The most significant divergence is to drugC, where the variation is 1.00 in BP. The characteristics that lead to drugC are shown in red, whereas those that route to drugB is shown in green. BP is best at forecasting which drug it will benefit, and the deviation is pushing it into drugC. Cholesterol deviates in this situation towards non-drug B and Na to K deviates towards drugB.

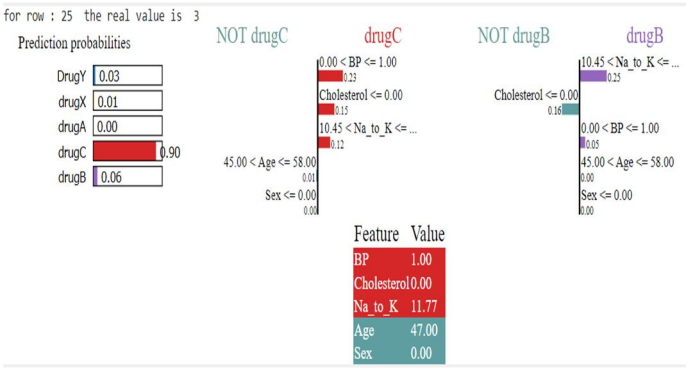


Figure 15: Lime for row 25

2) SHAP: First, SHAP is imported, and the SHAP Tree Explainer is utilised to do so. Figure 17 shows which characteristic has the most influence on drug categorization.

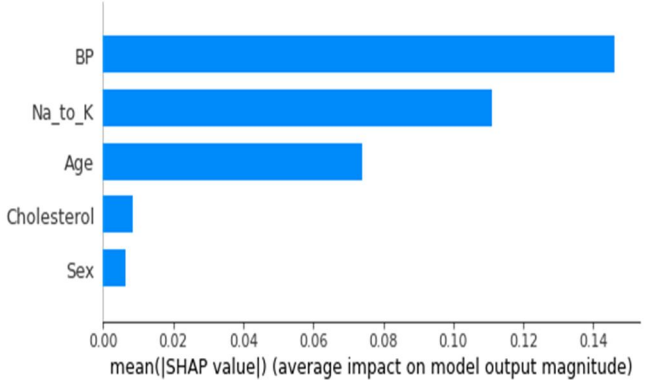


Figure 16: SHAP

©IJRASET: All Rights are Reserved

Figure 18 shows a summary graph of SHAP. The following notation specifies what class an attribute belongs to which class it's associated with. For example, the purple line is greater in feature Na to K, indicating that class 0 is dominating this feature and class 0 Equals drugX. Every other feature follows the same logic.

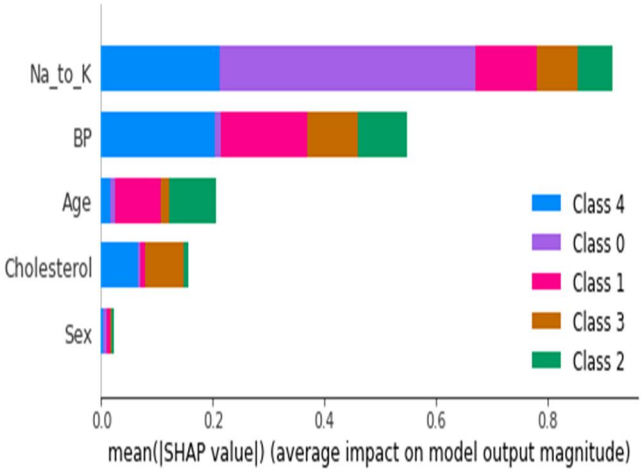


Figure 17: SHAP for Random Forest

C. Logistic Regression

The Gradient Boosting Regressor is used to import the logistic regression model, which has to be fitted with the data. The model scored 0.9 on its accuracy. The number of classifications a model correctly anticipates, divided by the total number of classifications the model predicts, represents accuracy. The model's F1 score was computed in attempt to comprehend its predictions, and the result was 0.86.

1) LIME: To make predictions, the report will utilise the rows - 19 and 25.

Figure 19 shows the classifications and prediction rates for drugs a, b, c, x, and y. In a way, LIME can determine how each characteristic relates to a drug's categorization, via the feature's properties and value. The most significant divergence is to drugY, where the variation is 0.73. The characteristics that lead to drugX are shown in blue, whereas those that route to drugA is shown in green. Na to K is best at forecasting which drug it will benefit, and the deviation is pushing it into drugY. BP deviates in this situation towards drugX. Here, Age is also playing a part in deviating towards not drugY.

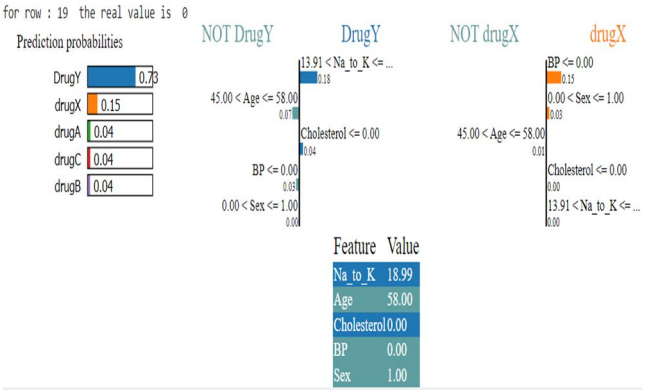


Figure 18: LIME for row 19

Figure 20 shows the classifications and prediction rates for drugs a, b, c, x, and y. In a way, LIME can determine how each characteristic relates to a drug's categorization, via the feature's properties and value. The most significant divergence is to drugB, where the variation is 0.38. The characteristics that lead to drugB are shown in purple, whereas those that route to drugY is shown in green. Na to K is best at forecasting which drug it will benefit, and the deviation is pushing it into drugB. Cholesterol deviates in this situation towards drugY and Na to K deviates highly towards not drugY.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IX Sep 2021- Available at www.ijraset.com

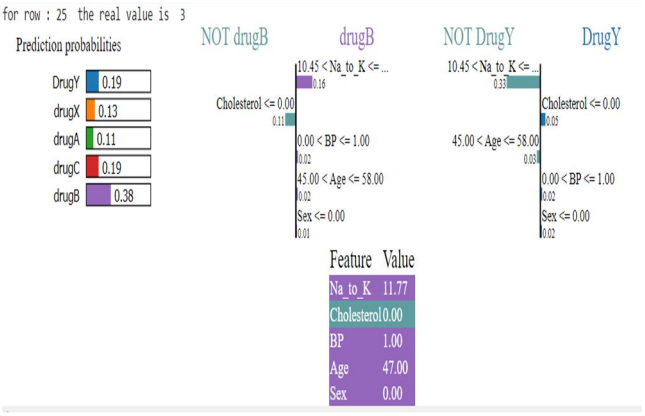


Figure 19: LIME for row 25

2) SHAP: Figure 21 shows a summary graph of SHAP. The following notation specifies what class an attribute belongs to which class it's associated with. For example, the purple line is greater in feature Na to K, indicating that class 0 is dominating this feature and class 0 Equals drugX. Every other feature follows the same logic.

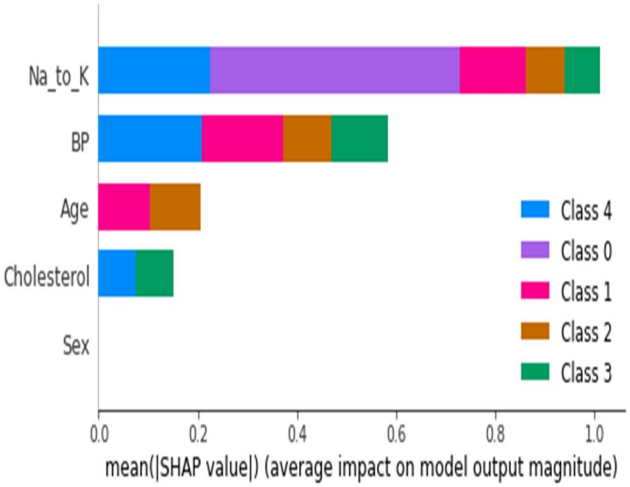


Figure 20: SHAP for Logistic Regression

V. RESULTS AND DISCUSSION

According to the findings of the above mentioned implementation, decision tree and random forest models possess identical accuracies and F1 score, indicating that for this dataset, they both function effectively and can predict accurate outcomes with an accuracy of 0.975. Finally, logistic regression seems to have the least accuracy of 0.9, indicating that it is less accurate than the other two models. In addition, the algorithms LIME and SHAP are used to explain the model's prediction concisely and correctly. LIME provides a simple tabular form that correctly represents the values of the attributes that affect the model's prediction. LIME's interpretation is based on the variables of a linear formula. LIME summarises each prediction made by a model in a simple manner. Shapley then continues on to explain how the model anticipates outcomes based on the individuals' characteristics. SHAP has not been shown to be superior to LIME in this regard, since LIME is better at forecasting and pinpointing the characteristics that influence the results.

VI. CONCLUSION

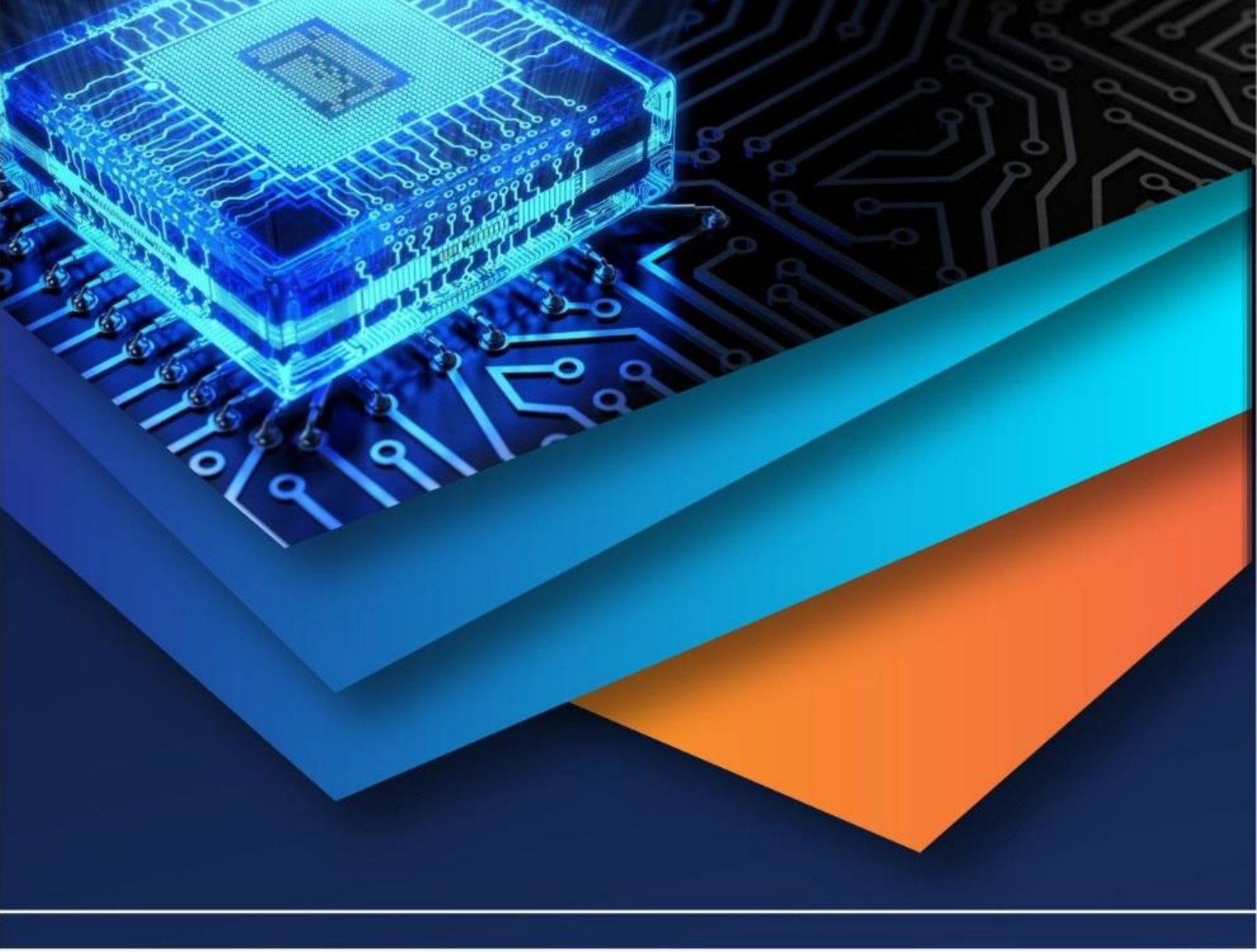
It can be stated at the conclusion of this paper that LIME and SHAP can be utilised to get insight into a Machine Learning model and determine which feature is accountable for the divergence in the outcomes. According to the LIME and SHAP results, it is also discovered that Random Forest and Decision Tree ML algorithms are the best models to employ for drug classification, with Na to K and BP being the most significant characteristics for drug classification.

©IJRASET: All Rights are Reserved

REFERENCES

[1] T. Oprea and J. Mestres, “Drug repurposing: far beyond new targets forold drugs,” The AAPS journal, vol. 14, no. 4, p. 759–763, 2012.
[2] L. Medina-Franco, M. A. Giulianotti, G. S. Welmaker and R. A. Houghten, “Shifting from the single to the multitarget paradigm in drugdiscovery,” Drug discovery today, vol. 18, no. 9, p. 495–501, 2016.
[3] H.-M. Lee and Y. Kim, “Drug repurposing is a new opportunity fordeveloping drugs against neuropsychiatric disorders,”,” Schizophrenia research and treatment, vol. 20, 2016.
[4] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo and Y. Liu, “DrugBank 4.0: shedding new light on drug metabolism,” in Nucleic Acids Resource, 2014.
[5] V. Solovyev and V. Ivanov, “Knowledge-driven event extraction inRussian: corpus-based linguistic resources,” Computational intelligenceand neuroscience, pp. 16-20, 2016.
[6] A. H. Wagner, A. C. Coffman, B. J. Ainscough, N. C. Spies, Z. L. Skidmore and K. M. Campbell, “DGIdb 2.0: mining clinically relevant drug-gene interactions,” in Nucleic Acids Resorce, 2016.
[7] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat and M. J. Wrobel, “The connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease,” Science, vol. 313, p. 1929–1935, 2014.
[8] Y. Liu, Q. Wei, G. Yu, W. Gai, Y. Li and X. Chen, “DCDB 2.0: a major update of the drug combination database,” Database , 2014.
[9] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. E. Leonard and J. H. Holmes, “Identifying potential adverse effects usingthe web: A new approach to medical hypothesis generation,” Journal ofbiomedical informatics, vol. 44, no. 6, p. 989–990.
[10] F. Cheng and Z. Zhao, “Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties,” Journal of the American Medical Informatics Association , vol. 21, p. 278–286, 2014.
[11] S. Vilar, E. Uriarte, L. Santana, N. P. Tatonetti and C. Friedman, “Detection of drug-drug interactions by modeling interaction profile fingerprints,” PLoS One 8, 2013.
[12] W. Zhang, K. Jing, F. Huang, Y. Chen, B. Li and J. Li, “SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions,” Information Science , vol. 497, p. 189–201, 2019.
[13] J. Y. Shi, H. Huang, J. X. Li, P. Lei, Y. N. Zhang and K. Dong, “TMFUF: a triple matrix factorization-based unified framework for predicting comprehensive drug-drug interactions of new drugs,” BMC Bioinformatics, vol. 19, pp. 411-420, 2018. G. Lee, C. Park and J. Ahn, “Novel deep learning model for more accurate prediction of drug-drug interaction effects,” BMC Bioinformatics, vol. 20, no. 415, 2019.
[14] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” arXiv: Machine Learning, 2017.
[15] M. Li, H. Xu and Y. Deng, “Evidential Decision Tree Based on Belief Entropy,” School of Computer Science and Engineering, University of Electronic Science and Technology of China, 2019.
[16] N. Donges, “A complete guide to the random forest algorithm,” 16 June 2019. [Online]. Available: https://builtin.com/data-science/random-forest-algorithm.
[17] P. Tripathi, “Drug Classification,” Kaggle, 2020. [Online]. Available: https://www.kaggle.com/prathamtripathi/drug-classification. [Accessed 1 September 2021].
[18] V. Dibia, “ML Interpretability: LIME and SHAP in prose and code,” 8 May 2020. [Online]. Available: https://blog.cloudera.com/ml-interpretability-lime-and-shap-in-prose-and-code/.
[19] D. Dataman, “Explain Your Model with the SHAP Values,” 14 September 2019. [Online]. Available: https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)