

Classification and Analysis of Mushrooms with Machine Learning and Data Mining

Pravesh Kumar¹, Dr. Rohitashwa Pandey²

¹Research Scholar, ²Associate Professor

praveshka2000@gmail.com, rohitashwapandey1982@gmail.com

Abstract:

It is advised that we consume mushrooms as part of our diet because they are rich in essential nutrients, proteins, minerals, and vitamins. Due to their comparable qualities and abundance, it can be difficult to distinguish between edible and dangerous mushrooms. Data mining methods, specifically classification, can be used to identify the type of mushrooms by helping to uncover crucial patterns from millions or even billions of data records. Existing research on mushroom classification focuses on applying ML techniques separately, and certain algorithms perform better in terms of accuracy than others. There is a paucity of research on mushroom classification. Based on this research, an integrated model was suggested, which would merge the decisions produced by the most accurate techniques into a single decision rather than treating them independently.

Introduction: Mushrooms are important organisms that play a significant role in the ecosystem, agriculture, food production, and pharmaceutical research. They belong to the fungi kingdom and exhibit a wide variety of physical characteristics, including cap shape, cap color, gill structure, odor, stalk shape, and habitat. Based on these characteristics, mushrooms can be classified into different categories, including edible and poisonous varieties. However, identifying mushroom species accurately can be challenging because some species have similar physical appearances but differ significantly in their properties.

k-Nearest Neighbors (kNN) Algorithm

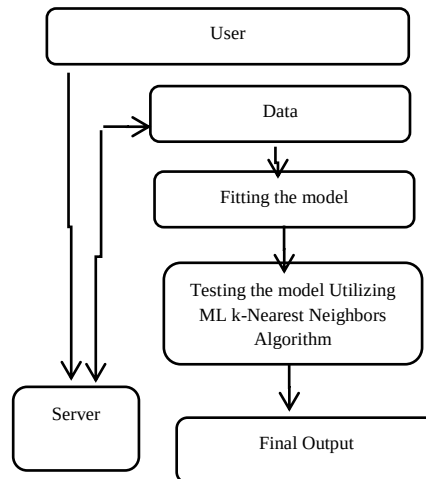


Figure 1: Flow graph of the proposed method

KNN has been modified by many authors to improve its efficiency. A class-wise KNN (C-KNN) algorithm has been designed and tested against mushroom dataset. Here, the testing data is assigned a class label based on the lowest class-wise distance. Accuracy achieved for C-KNN is 78.16%. K means and KNN algorithms have been combined as amalgam KNN model to classify instances of the mushroom database. Here the quality of data is improved by removing the noise and thereby increasing the efficiency. K-means removes the incorrectly classified instances and classification is done using KNN algorithm.

The choice of K in KNN depends upon the data. Larger values of k can reduce noise in classification. A good k value can be selected by cross validation technique. By estimating the k value with ten-fold cross validation, a classification accuracy of 97.4% has been obtained (Xiong et al 2007). The concept of KNN algorithm is illustrated in Figure 2.

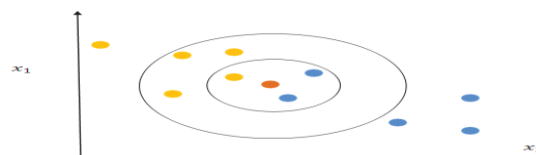


Figure 2:K nearest neighbour algorithm

DATA SET DESCRIPTION

The mushroom dataset might be found at UCI mushroom data, where it was downloaded [12]. The data collection has a total of 8124 rows of data records and 22 different qualities to consider. Each type of mushroom is classified according to whether or not it is edible or harmful. These rows are broken up into 4208 edible mushrooms and 3916 toxic mushrooms.

Before continuing with the process, the initial step is to organize the data. In order to prepare the data, it will be necessary to get rid of any null values and features that are repeated. Python packages that were utilized in the creation of datasets from raw data. The 22 characteristics are necessary for determining if a mushroom is edible or inedible. It is necessary for us to investigate these characteristics in order to determine which ones make a significant contribution to the classification procedure. We have come to the conclusion that "stalk-root" and "stalk-root" should both be removed. The reason for this is that the "stalk-root" feature is lacking 2480 data, and the "veil-type" feature only contains a single value, meaning that it cannot assist us in classifying anything. Furthermore, it is anticipated that "gill-color" will have the greatest contribution to the classification, so this factor needs to be taken into consideration.

Table 1: Mushroom dataset attribute information

No	Attribute	values
1	cap-shape	bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2	cap-surface	fibrous=f,grooves=g,scaly=y,smooth=s
3	cap-color	brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
4	bruises	bruises=t,no=f
5	odor	almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
6	gill-attachment	attached=a,descending=d,free=f,notched=n
7	gill-spacing	close=c,crowded=w,distant=d
8	gill-size	broad=b,narrow=n
9	gill-color	black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
10	stalk-shape	enlarging=e,tapering=t
11	stalk-root	Bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
12	stalk-surface-above-ring	fibrous=f,scaly=y,silky=k,smooth=s
13	stalk-surface-below-ring	fibrous=f,scaly=y,silky=k,smooth=s
14	stalk-color-above-ring	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
15	stalk-color-below-ring	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
16	veil-type	partial=p,universal=u
17	veil-color	brown=n,orange=o,white=w,yellow=y
18	ring-number	none=n,one=o,two=t
19	ring-type	cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
20	spore-print-color	black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
21	population	abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
22	habitat	grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

This dataset contains descriptions of hypothetical samples belonging to 23 different species of gilled mushrooms in the Agaricus and Lepiota Family Mushrooms. These descriptions were taken from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is classified as either absolutely safe to consume, absolutely poisonous, or of unknown edibility and is not advised for consumption. This more recent class was brought together with the toxic one [4]. This dataset has 22 attributes, and there are 8124 different mushrooms included in it. The information regarding the dataset's attributes may be found in table 1.

Data Preprocessing

The dataset is divided into two categories: those that are edible and those that are poisonous. A bar graph is plotted in order to determine whether or not each is in balance. Because the data is categorical, the ordinal representation requires being converted via Label Encoder. Label Encoder makes a numerical representation of each value contained in a column [18]. Figure 3 depicts the total count for each class, while Figure 4 displays the dataset following the application of labels.

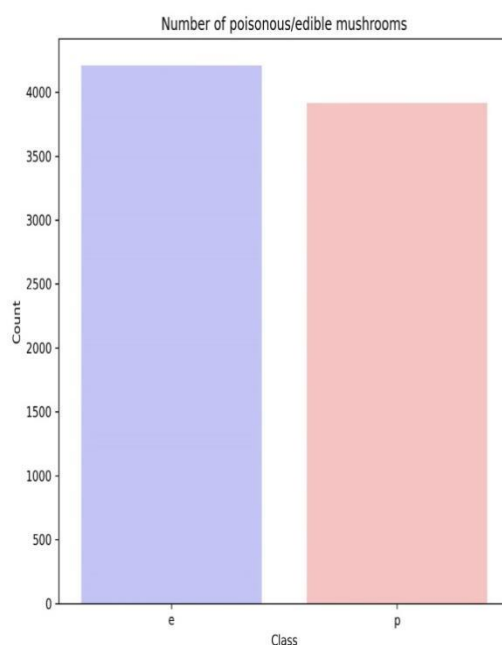


Figure 3: Count of edible and poisonous mushrooms

	cap- shape	cap- surface	cap- color	bruises	odor	gill- attachment	gill- spacing	gill- size	gill- color	stalk- shape	stalk- surface- below- ring	stalk- color- above- ring	stalk- color- below- ring	veil- type	veil- color	ring- number	ring- type	spore- print- color	population	habitat	
0	5	2	4	1	6	1	0	1	4	0	...	2	7	7	0	2	1	4	2	3	5
1	5	2	9	1	0	1	0	0	4	0	...	2	7	7	0	2	1	4	3	2	1
2	0	2	8	1	3	1	0	0	5	0	...	2	7	7	0	2	1	4	3	2	3
3	5	3	8	1	6	1	0	1	5	0	...	2	7	7	0	2	1	4	2	3	5
4	5	2	3	0	5	1	1	0	4	1	...	2	7	7	0	2	1	0	3	0	1

Figure 5: Label Encoding

In exploratory data analysis, correlation matrices are an essential tool that must be used. It is helpful to have an understanding of the connection between the variables and the columns. In order to visually show the degree of correlation between the variables, a heatmap is displayed.

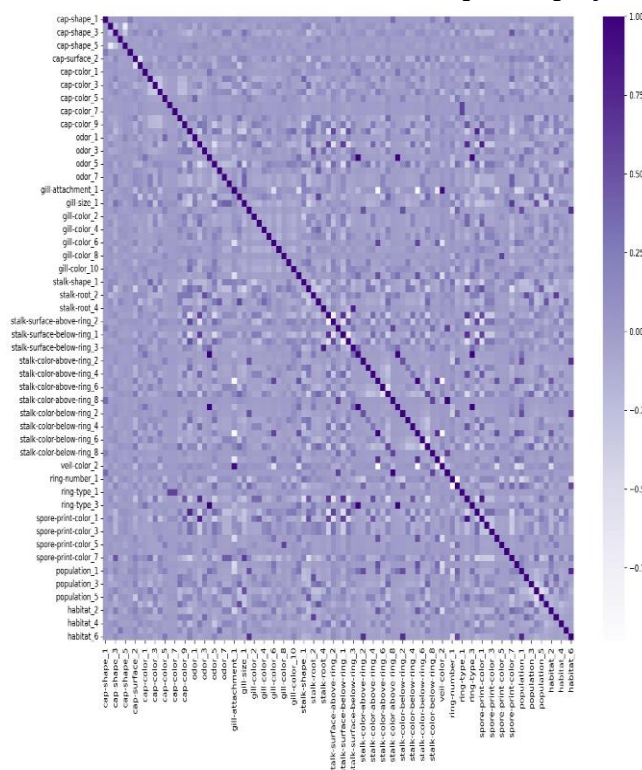


Figure 4: Correlation between the dummy/indicator variables

Classification Modelling

The data that was gathered is then subjected to supervised machine learning methods following the feature extraction and selection processes. The following methods of machine learning are going to be utilized, as was covered in the prior section:

- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Random Forest (RF)

Result and Analysis

In this experimental work, we utilized a total of three different machine learning methods. KNN, SVM, and RF are the names of these algorithms. The UCI Mushroom Classification Dataset was subjected to analysis using each of these algorithms. The data were split up into two parts: the training data and the testing data. The training data and the testing data each had 70% of the total data and the testing data had 30%. Scaling of the features was accomplished with the help of the StandardScaler() function. In order to lessen the number of dimensions and choose the most useful characteristics from the dataset, the Principal Component Analysis (PCA) algorithm was applied with n_components set to 2. The same

dataset was used for testing all three algorithms, and results were gathered from that testing. Accuracy of prediction is the primary criterion for evaluation that is applied to this body of work. The algorithm's overall success rate is referred to as its accuracy.

Table 1 displays the True Positives (TP), True Negatives (TN), False Negatives (FN), and False Positives (FP) that were predicted by each of the algorithms. For our purposes, the term "TP" refers to edible mushrooms. TN, these are the true dangerous mushrooms. FP, which is known to be poisonous yet is thought to be edible. FN, believed to be poisonous despite the fact that it may be eaten.

The visualizations of training and test set are given here:

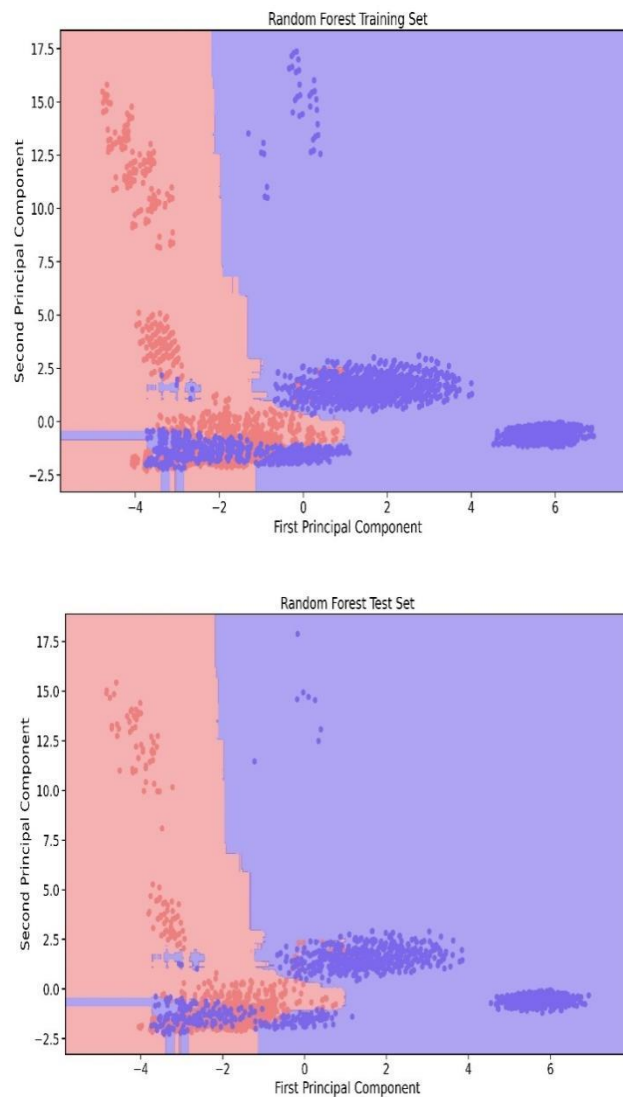


Figure 6:Random Forest (RF) Training and Test Set

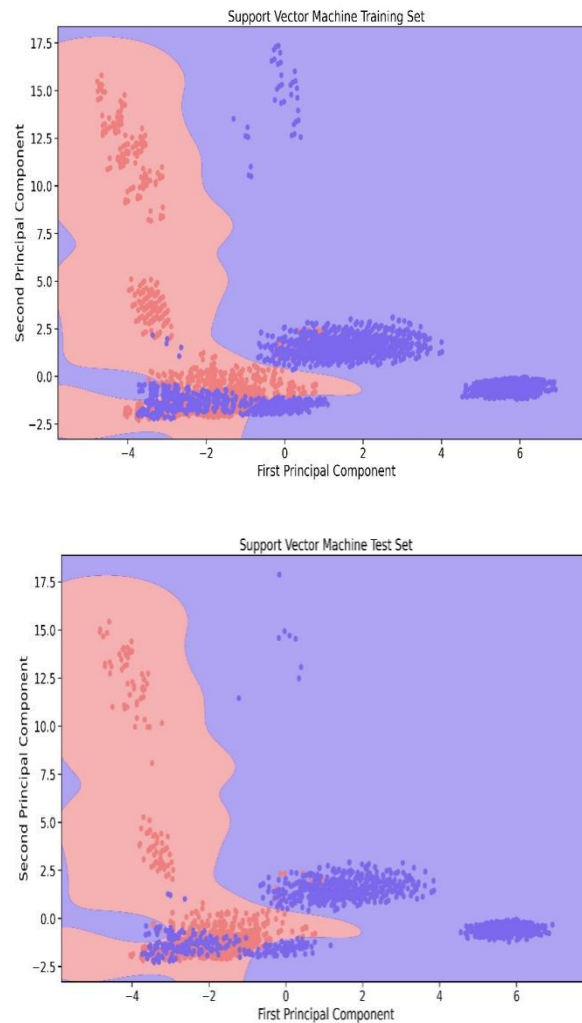
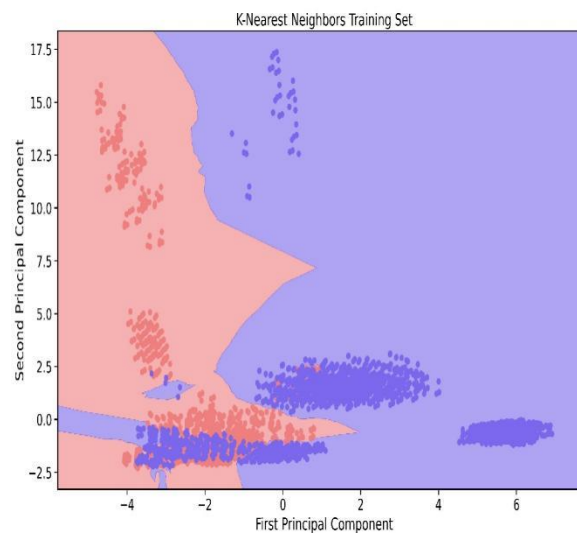


Figure 7:Support Vector Machine (SVM) Training and Test Set



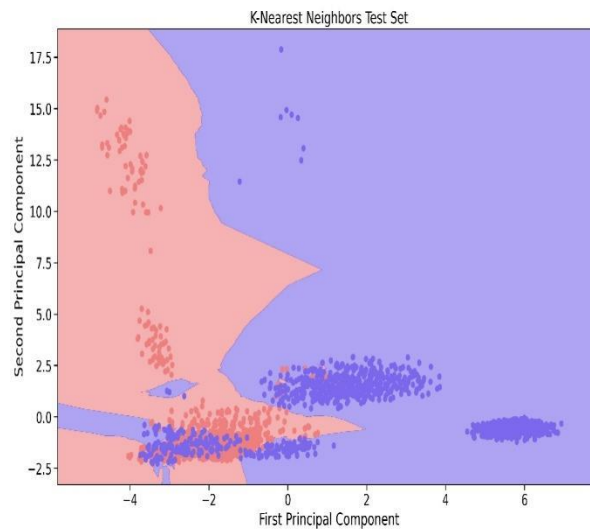


Figure 8: KNN Training and Test Set

The Training and Test Sets have been plotted in figure 6 to figure 8 for Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors algorithms.

A Receiver Operator Characteristic (ROC) curve, which is an evaluation metric for binary classification problems, specifically mushroom classification, was plotted by ourselves. It is a probability curve that plots the TPR against the FPR at a variety of threshold values and, in essence, differentiates the "signal" from the "noise." The area under the curve, often known as AUC, is a metric that is used to evaluate the capacity of a classifier to differentiate between different categories. A graph that displays the performance of a classification model at all classification thresholds is referred to as a ROC curve. A higher performance can be inferred from classifiers that produce curves that are located closer to the upper left corner. Hence, the ROC curve can be summarized using AUC and shown in figure 9. The highest ROC value is 0.9502 for KNN classifier that means KNN has best performance among all three classifiers.

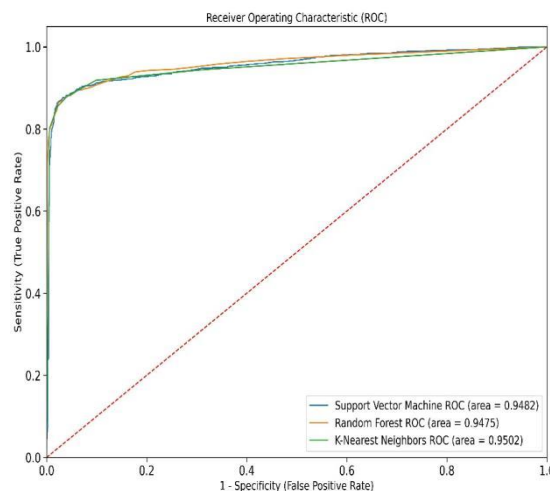


Figure 9: Receiver Operator Characteristic (ROC) Curve

6.1 Conclusion

The mushroom is a significant contributor of both necessary proteins and vitamin content. On the other hand, the vast majority of identified mushroom types are toxic. ML learning was utilized to determine whether or not certain types of mushrooms were edible based on their properties. Previous research utilized ML techniques alone, and it was found that certain strategies performed significantly better in terms of accuracy than others. Consequently, it is difficult to decide which approaches should be taken because there are many viable options. Within the scope of this investigation, we have utilized a number of different machine learning classifiers on the "mushroom data" dataset. The dataset obtained from the repository on the UCI campus. Following an examination of the data, we have discovered that one feature known as "stalk-root" is missing a significant number of values, while another feature known as "veil-type" has the same values for each row. These two characteristics are removed so as to eliminate any potential impact they may have had on the classifications. In this case, the "odor_n" attribute is the most significant and influential aspect in the decision-making process. On the basis of the findings, we are able to draw the conclusion that an inedible mushroom is more likely to have an odor.

REFERENCES:

1. Karun, N. C and Sridhar, K. R. 2013. Occurrence and distribution of Termitomyces (Basidiomycota, Agaricales) in the Western Ghats and on the West Coast of India. Czech Mycology, 65 (2), 233-254.
2. Jonathan, G., Adetolu, A., Ikpebie, O and Donbebe, W. 2006. Nutritive value of common wild edible mushrooms from Southern Nigeria. Global Journal of Biotechnology and Biochemistry, 1 (1), 16-21.
3. Khourdifi, Youness, and Mohamed Bahaj. "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification." In 2018 International Conference on Electronics, Control, Optimization and Computer Science, ICECOCS 2018, 1–5. IEEE, 2019. <https://doi.org/10.1109/ICECOCS.2018.8610632>.
4. Khourdifi, Youness, and Mohamed Bahaj. "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification." In 2018 International Conference on Electronics, Control, Optimization and Computer Science, ICECOCS 2018, 1–5. IEEE, 2019. <https://doi.org/10.1109/ICECOCS.2018.8610632>.
5. Ren, S, Lian, Y & Zou, X 2014, ' Incremental naïve Bayesian learning algorithm based on classification contribution degree', Journal of Computers, vol. 9, no. 8, pp. 1967-1974.
6. Srinivas, K. Rani, BK & Govrdhan, A 2010, 'Applications of data mining techniques in healthcare and prediction of heart attacks', International Journal on Computer Science and Engineering, vol. 2, no. 2, pp. 250-255.
7. Wu, X., Kumar, V, Quinlan, JR, Ghosh, J, Yang, Q, Motoda, H & Steinberg, D 2008, 'Top 10 algorithms in data mining', Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37.
8. Anburajan, M, Sivanandam, S, Bidyarsmi, S, Venkatraman, B, Menaka, M & Raj, B 2011, Changes of skin temperature of parts of the body and serum asymmetric dimethylarginine

- (ADMA) in type-2 diabetes mellitus Indian patients, Proceedings of the annual international conference of the engineering in medicine and biology society, pp. 6254-6259.
9. Koklu, M & Unal, Y 2013, 'Analysis of a population of diabetic patients databases with classifiers', World Academy of Science, vol. 7, no.8, pp. 167-169.
10. KR, Anandakumar & Ashwinkumar UM 2010, 'Ethical and legal issues for medical data mining', International Journal of Computer Applications, vol. 1, no. 28, pp. 7-11.
11. Fayyad, UM 1996, 'Data mining and knowledge discovery in databases: applications in astronomy and planetary science', Proceedings of the thirteenth national conference on Artificial intelligence, vol. 2, pp. 1590-1592.
12. Manjusha, KK, Sankaranarayanan, K. & Seena, P 2014, 'Prediction of different dermatological conditions using naïve Bayesian classification', International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 1, pp. 864-868.
13. C.kalaiselvi, G.m.Nasira, 2014. "A New Approach of Diagnosis of Diabetes and Prediction of Cancer using ANFIS", IEEE Computing and Communicating Technologies, pp 188-190
14. Velu C.M, K.R.Kashwan, 2013. "Visual Data Mining Techniques for Classification of Diabetic Patients", IEEE International Advance Computing Conference (IACC), pp-1070-1075.