International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Classification of Gene Expression Data Set using Support Vectors Machine with RBF Kernel

M Ramachandro, Ravi Bhramaramba

Abstract: ***The huge amount of data being generated by different organizations and its underlying advantages in multiple fields like decision making, data security, research purposes have made data classification a very important and mandatory process now-a-days. Data Classification is the process of grouping data of similar characteristics into categories. Classification can be done based on the output we are looking forward to. Hence it is considered very useful. Classifying data allows us to predict the nature of future data-sets and discover useful patterns among them. This project aims at classifying gene data sets. Gene data sets are the information collected from a set of genes put to a specific test. It can be used for medical research purposes; by studying the pattern in the datasets allows us to predict the kind of genes that are more vulnerable to a particular disease there by allowing us to prevent the manifestation of the disease right at its beginning, just as they say, prevention is better than cure. In this paper, such classification is effort using a supervised machine learning algorithm – SVM (Support Vector Machine). There are many algorithms in existence to perform classification but this algorithm has its own lead over the others. It is capable of both classification and regression. It works well with structured, semi-structured and unstructured data too. It contains a kernel function which when used appropriately can solve any complex problem. The summary of this project is, taking gene data sets as input and obtaining classified clusters as output.***

Keywords: Data Classification, Gene expression, Support Vector Machine (SVM), Machine Learning, Supervised Algorithms

I. INTRODUCTION:

SVM is a supervised machine learning algorithm. As discussed earlier, the current plethora of data cannot be classified just by using linear classification techniques. Hence we use SVM techniques which contain both linear and non-linear classification techniques. SVM is very useful for both regression and classification. In case of classification, the algorithm is very simple. It creates a line called the line of choice separating the distinct classes.
The aim of the algorithm is to maximize.The class margin, t he benefit is that after grading it is easier to predict the target classes for new datasets.

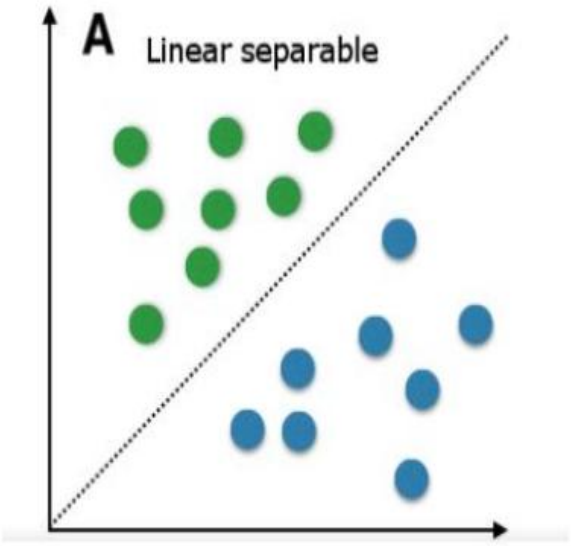

Figure 1: Linear Separable With Two Dimensions

The above picture clearly depicts what linear classification means. However, it is not always the case i.e, there will be cases when a straight line cannot classify the classes. It is called non-linear classification. SVM can handle such kind of data too.

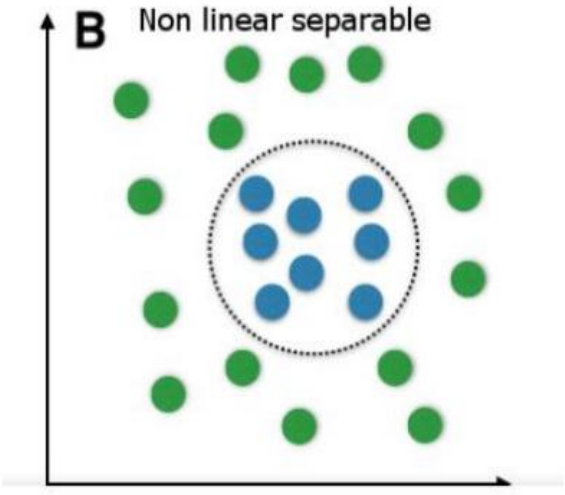

Figure 2: Non linear Separable dimensions

In the above picture we can see that, there is no straight line that can separate the data into different classes. These are the kind of cases where other methods become obsolete but SVM is still applicable. This is where the kernel trick used by SVM comes into picture. Using this kernel function, SVM classifies non- linear data.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
M Ramachandro*, Dept of Computer Science &Engg, GMR Institute of Technology, Andhra Pradesh
Ravi Bhramaramba, Dept of Information Technology, GITAM, Andhra Pradesh

Choosing the appropriate kernel function is the key for SVM because with proper kernel, any complex problem can be solved. Some common applications of SVM are: Prediction of protein structure, detection of intrusion, recogn ition of handwriting, detection of steganography in digital i mages and diagnosis of breast cancer.

II. PROBLEM STATEMENT:

As there is a huge amount of data in existence, people have started using the stored data to un earth useful patterns which enable them to make better business decisions and provide opportunities to predict the nature of future data sets. Classification of data has proven to make any businesses efficient and less loss-prone. However, the major problem is that the kind of data that is generated is very varied and huge i.e., it consists of structured, semi-structured and unstructured data. Studies have shown that more than 70% of the data being generated lately is unstructured. Hence classifying this data requires more sophisticated algorithms. The existing classification techniques encounter problems when handling unstructured and semi structured data.However, machine learning provides a better and efficient alternative. Machine learning is basically building an artificially intelligent system.Machine learning algorithms are widely categorized into – supervised and unsupervised learning is basically like a teacher supervising you. Here, the teacher is the algorithm; the student is the input data. The algorithm trains the data so as to get the desired output. On the other hand, in unsupervised learning, there is no teacher, there is no desired output. Algorithms are designed so as discover interesting structures in the data. Here, we use one such supervised machine learning algorithm called the SVM – Support Vector Machine.

III. LITERATURE SURVEY

Classification of data means [1] separating it into different classes based on some constraints. Different classification techniques exist and each of these techniques follow three approaches – statistical, neural networks and machine learning for classification. Statistical strategies are typically characterised through having a precise fundamental likelihood model which presents a probability of being in each category rather of simply a classification. The field of Neural Networks has arisen from diverse sources ranging from understanding and emulating the human intelligence to broader problems of copying human competencies such as speech and can be used in a number of fields such as banking, legal, medical, news, in classification software to categorise records as intrusive or normal. Machine Learning usually covers computerized computing methods based totally on logical operations that examine a assignment from a series of examples. Some classification algorithms are: k nearest neighbours, naïve bayes, ANN algorithm, SVM algorithm.

K Nearest Neighbours should be avoided as much as possible since the evaluation is quite heavy if your training dataset contains several elements. KNN has a number of most important advantages: simplicity, effectiveness, intuitiveness and competitive classification performance in many domains. It is strong to noisy education data and is high-quality if the education information is large. Two despite the advantages, KNN has a few limitations. KNN can have terrible run-time performance when the education set is large. It is very touchy to beside the point or redundant points due to the fact all points contribute to the similarity and as a result to the classification. Computation value is quite high because we need to compute distance of every query occasion to all training samples.

Naive Bayes is very simple and quick to evaluate but it works strangely with unbalanced classes. The mechanism is very simple to understand, it also has a high performance and is easy to implement. Bayesian classifiers are easy to understand. Naive Bayes makes dealing with missing values a lot easier; you can drop the missing feature for all classes uniformly and evaluate the decision rule. It is a easy yet powerful model, it returns now not only the prediction but also the degree of certainty, which can be very useful. They are light to train - no problematic optimisation required and without problems updateable if new training records is received. In classification tasks you need a massive facts set in order to make reliable estimations of the chance of every class. You can use Naïve Bayes classification algorithm with a small facts set.

Artificial Neural Networks are comparatively good if you have scant knowledge on your dataset. Neural networks are gradual to converge and tough to set parameters but if accomplished with care it work wells. Relatively simple learning algorithm compared to some of the Bayesian models and scales well to larger datasets with new GPU hardware and CUDA software and can significantly out-perform other models when the conditions are right. However, it is hard to interpret the model and doesn't perform as well on small data sets (The Bayesian approaches do have an advantage here). The benefits of deep neural networks are record-breaking accuracy on a complete range of issues consisting of image and sound recognition, textual content and time series analysis.

Support Vector Machine is a new promising non-linear, non-parametric classification technique, which already showed appropriate consequences in scientific diagnosis, optical character recognition, electric load forecasting and different fields. It might not look very effective without using kernel as it is the key feature in SVM. Kernels are symmetric, semi-positive definite functions. They help SVM to handle non-linear data. However, using kernel makes it computationally expensive, hence slow. Support Vector Machines work very well in many circumstances and performs very well with large amounts of data. SVM also provides good accuracy.SVM (Support Vector Machine) [1] is a pattern recognition method used most widely in the fields of medical research like breast cancer etc. to classify benign and malignant masses. A benign mass indicates that there is no harm and the infected cells don't spread around. It also indicates that, the harmful mass can be removed by clinical operations as such. On the other hand malignant masses are the exact opposite; they tend to be contagious and cannot be removed by any medical practices.The process of extracting information from a dataset is called data mining. One of its major components is classification/clustering [1].

In classification, we analyse a set of records and generate a set of grouping guidelines which can be used to classify future data. There are many techniques for classifying data. Some of them are: k nearest neighbours, naïve bayes, ANN algorithm, SVM algorithm. However, SVM is more preferred for because of the advantages it offers – flexibility in selecting a similarity function, sparseness of solution when dealing with massive datasets, the potential to manage massive feature area and the capability to identify outliers.

SVM depends on supervised learning algorithm [2] and the aim of using it is to correctly classify unseen data. It also finds its applications in fields like text categorization, hand-written character recognition, image classification, bio-sequences evaluation and so on. It is also important to note that data classification using SVM is an advanced field and to yield proper results, it must be performed under expert supervision. It is not just any typical classification. SVM is a supervised machine mastering algorithm [3] which can be used for classification and regression. The main reason behind the success and popularity of SVM is its ability to model complex non-linear relationships by selecting suitable kernel function. Data or relationships are said to be non-linear when there is no straight line in the hyper plane that can clearly classify the datasets. That is when SVM plays a prominent role. SVM provides a way to classify even such data by transforming it using a kernel function. Then, based on these transformations, figures out how to separate data from previously defined labels or outputs.In machine learning, kernel strategies are a class of machine studying algorithms that end up being an increasingly popular tool to get to know tasks such as pattern recognition, classification or detection of novelty. They are highly recognized for their function in Support Vector Machine (SVM)[4]. The kernel methods incorporate kernel functions that allow them to operate in an excessive dimensional characteristic space barring constantly computing information coordinates in that space, but as an alternative by virtually computing the internal merchandise between the photos of all fact pairs. It is stated and proved that, the usage of a acceptable kernel function, one can remedy any problem. The trick then again is to find the appropriate kernel function.A kernel's function is to take input data and transform it into the form required. There are various kernel types[5]. They are: Linear, Polynomial, Gaussian, RBF, and Sigmoid kernel. Choosing the kernel function to map non-linear input space into linear feature space is highly dependent on data nature. The RBF kernel, for example, is a popular kernel function used in various learning algorithms that are kernelized. It is the most preferred kernel function when the data has many dimensions on which it needs to be classified.Gene expression is the manner through which data from a gene is used in the synthesis of a useful gene product [6]. The human DNA shops genetic statistics and is said to be a blueprint of all dwelling organisms. This genetic records is preserved and handed from mother or father cell to child cells. Two DNA microarray technological know-how can reveal the expression stages of lots of genes concurrently at some point of important biological procedures and across collections of associated samples. Two Knowledge gained via microarray facts evaluation is increasingly more vital as they are beneficial for classification of diseases.With DNA microarray information, choosing a smaller subset of discriminative qualities from a great many qualities is a basic advance for precise arrangement of malady determination. Quality determination evacuates countless qualities in order to improve the characterization precision. A few quality choice strategies regularly select top-positioned qualities as indicated by their individual discriminative power in ordering tests into unmistakable classifications, without thinking about relationships among qualities. Evacuating immaterial qualities builds exactness of order as well as abatement of an opportunity to perform grouping and other such points of interest. Quality determination was proposed to use SVM [7] techniques dependent on RFE (Recursive Feature Selection). Micro array data tend to have a high size of variables and a small sample size [8]. This may lead to poor classification due to irrelevant data, noise or outliers. It may also result in over-fitting of data. Hence while handling micro array data; there could be occurrence of two issues. Firstly, How to choose genes that provide a credible and positive forecast of disease condition. Second, how to determine the ultimate collection of genes that is appropriate for ranking. The way in which these two issues are tackled decide whether our classification is efficient and accurate or not.Machine learning is basically making massive amounts of data and their analysis available to the medical profession [10]. Massive amounts of data properly analysed through algorithms and computational analysis reveal hidden and little known patterns in the data which, without machine learning, would never be available to the doctor. Machine training algorithms such as Support Vector Machine (SVM) can help doctors diagnose more properly. Correct and prompt disease detection is an important medical issue. Realisations of the above have made classification using SVM a very popular and preferred classification method.

IV. PROPOSED SYSTEM:

The basic outlook of this project is to take a gene data set as an input, pre-process it, apply SVM with an RBF(Radial Basis Function) kernel on that data set and obtain classified datasets (benign and malign) as output. The proposed system overcomes the drawbacks of the existing systems. It has better accuracy in classifying both normal and high dimensional datasets and can handle more data. What makes SVM very popular among classification methods is its ability to classify non-linear data. This also stands as its biggest advantage over other methods. The kernel trick in SVM enables it convert the hyper plane into a higher dimensional plane – converting linearly inseparable data into non-linear separable data. Gene expression datasets usually have high dimensionality. This might affect the data negatively if the information is irrelevant or error prone. Hence, gene selection mechanism is considered very important and a-must-do task in data classification. We have chosen the RBF kernel because it can handle multi dimensional data with various class distribution. The RBF kernel on two samples x and x, displayed in some entry room as function vectors, is described as:

$$k(x, x') = \exp\left(-\frac{\| x - x' \|^2}{2\sigma^2}\right) \quad \text{----------Eq.1}$$

Since The RBF valuation reduces with range and varies from 0 to 1, prepared to be interpreted as a metric of resemblance. For visualizing data throughout this process, we use Jupyter notebook. Firstly, we import the required packages for classification. Following that, we perform data pre-processing tasks for data cleaning and reduce the irrelevant data. After that, we plot training and monitoring information sets with random state variables. Import the SVM model from Sklern package and fit the train and test data and apply the kernel function for transforming the input

data into feature space and then perform cross-validation to the trained dataset for calculating to accuracy of trained data. Finally, we calculate the confusion matrix. Using that, we calculate accuracy, precision, recall, support which all make the classification report.

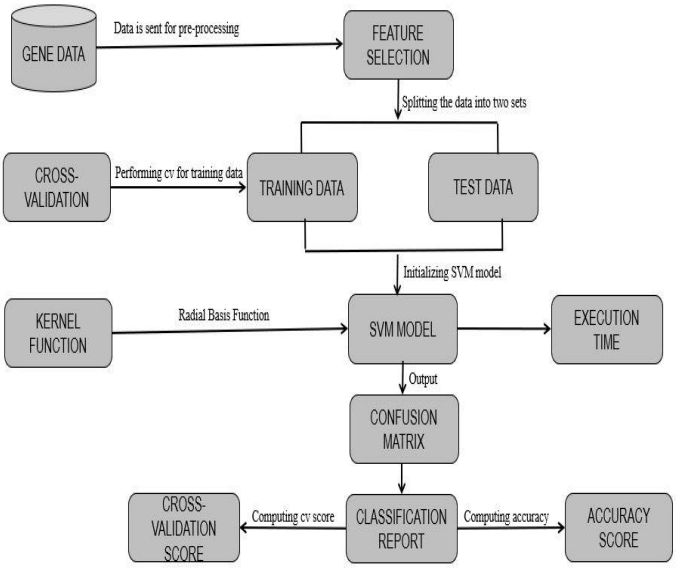

Figure 3: Architecture of the proposed model

V. COMPARATIVE ANALYSIS:

	Existing system (Drawbacks)	Proposed System (Advantages)
1	Cannot manage information that are not linearly separable	Can handle non-linearly separable data.
2	Work only with information in two dimensions.	Work with multi-dimensional data also
3	There is no such application as the kernel function	Kernel Function is the most important principle.
4	Cannot transfer into a multi-dimensional hyper plane	It can transfer into multi-dimensional spaces.
5	Cannot identify and filter the outliers	Clearly classify outliers
6	Working with extreme cases is impossible	Easily handle extreme cases
7	Cannot classify the data having uneven class distribution	Can classify the data having uneven class distribution

VI. RESULTS

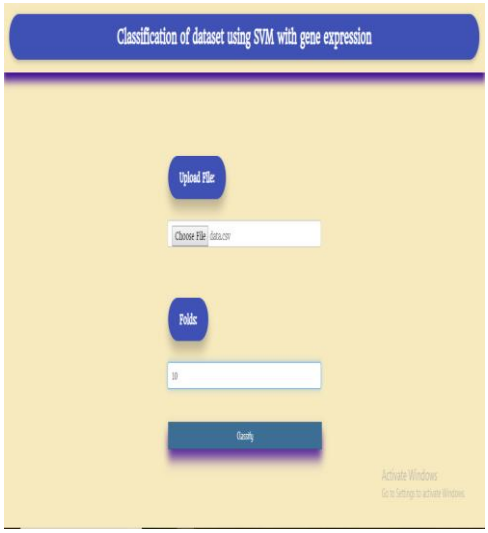

Figure4: Taking input data from dataset

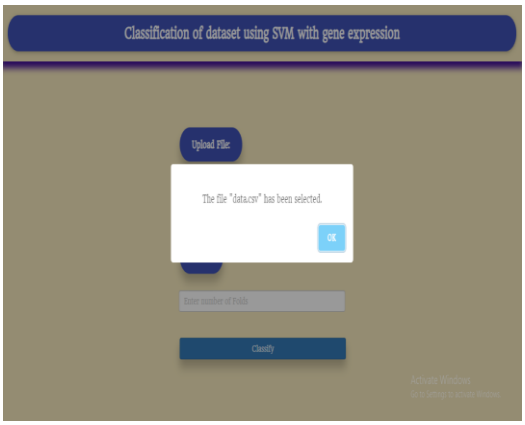

Figure 5: Input data set is selected

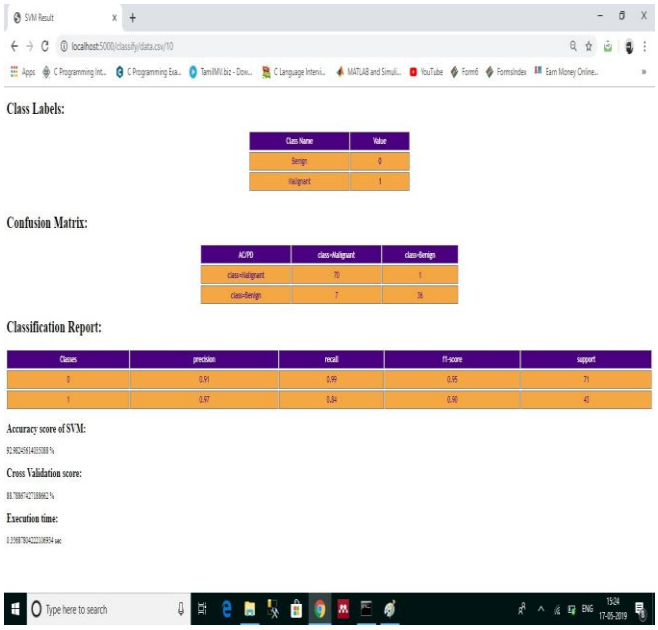

Figure 6: Resultant output

No. of Malignant	Correctly Predicted	Wrongly Predicted
71	69	2

No. of Benign	Correctly Predicted	Wrongly Predicted
43	36	7

Figure 7: Output (classification Report)

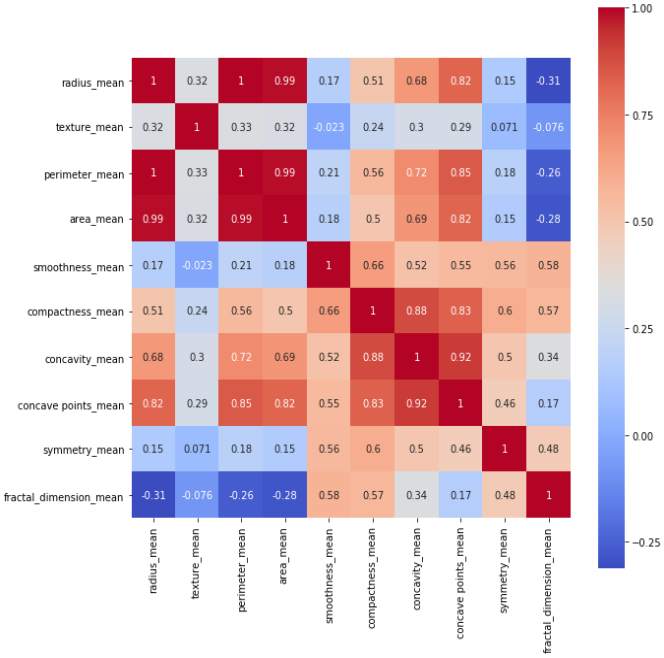

Figure 8: Heat Map

Before doing classification we have to analyse the data and choose some of the attributes that are suitable for the classification for better results for that here we have done the correlation analysis for the dataset by taking all the mean attributes That is the mean radius, the mean texture, the mean

perimeter, the mean area, the mean smoothness, the mean concavity, the mean concave, the mean symmetry and the mean fractal dimension. We used these characteristics to calculate the causes of correlation. We need to discover the connection level at which one characteristic depends on the other variable. We have to compare all the attributes with the aid of corresponding attribute through calculating the correlation we can examine that if the attributes are having identical two then dependency component is 1 pronouncing that high dependency if the attributes having unique information it results in more high quality or more bad values which shows the much less dependency while deciding on the attributes we have to take the attributes having excessive dependencies for the higher consequences this is accomplished via Machine learning method we are the use of sea born package for higher graphical visualization.

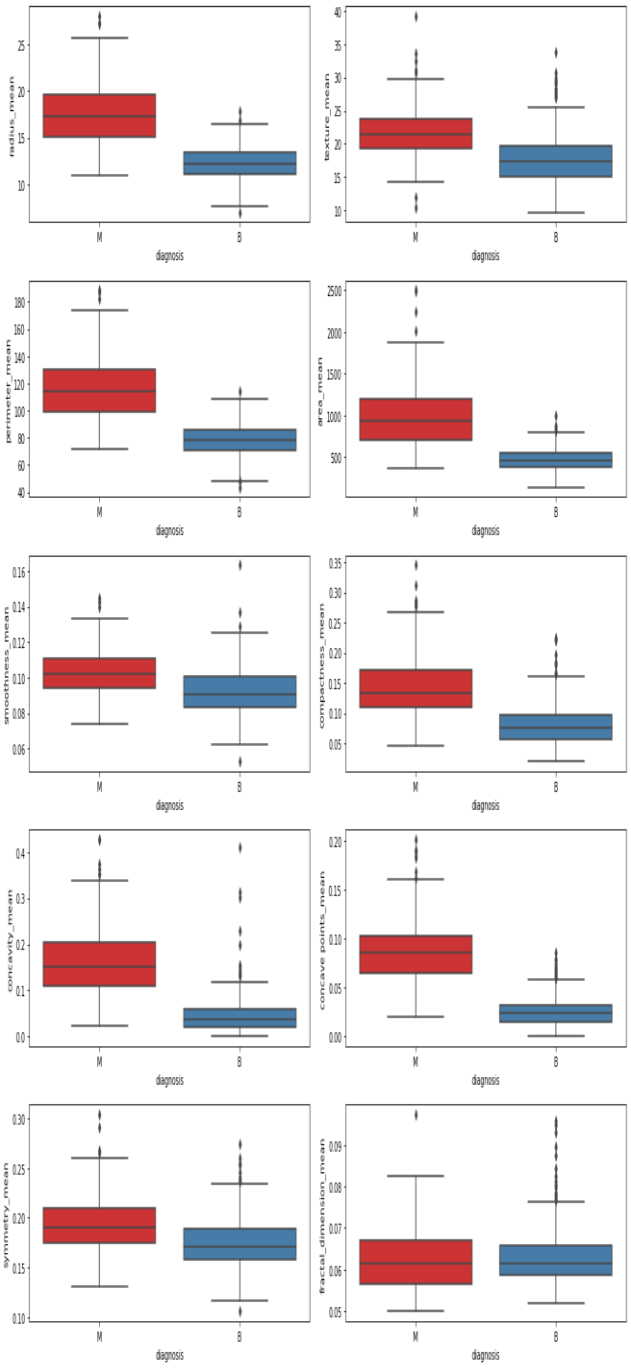

Figure 9: Box plot

For the data Pre-processing technique here we used the feature selection method that is we have to select the features based on the data class variables for every data point we chose the data points of all the mean attributes chose which is less dependent on the class variable in the above figure we can easily predict the some of the features that is texture mean and smoothness mean which shows the redundant values of malignant and benign correspondingly. As we are doing classification we have to take the attributes in such a way that these are highly independent on the class variable data. We are indicating the malignant values as red colour and benign values as blue colour after analysing the data we are selecting radius mean,

perimeter mean, concave mean taking in to consideration for feature selection attributes for data cleaning and reducing redundant data and normalization purpose. We can do many techniques not only feature selection but this is the most understandable technique and accurate than remaining and we used matplotlib and seaborn for the above graphical representation. We used the scatter matrix for the purpose of predicting whether our data is linearly separable or non-linearly separable data we can decide by using one of the machine learning package called pandas by using that we can easily see our data points when it is comparable with others for every other attribute we can observe that in most of the cases that is non linearly separable so we are using non linear SVC model for the data classification model if it is having the same attribute rather than showing graphical representation with class variables it shows the graphs for the data points. For the above implementations we used the python IDE for the programming language for the machine learning packages we have to install pip for installing other if we install anaconda then it is useful for Jupiter notebook after that we used the flask environment for connecting the python code with our front end for dataset selections and cross-validation fold selections after clicking submit button we can see the classification report as follows

SOFTWARE REQUIREMENTS:

- **Python ide** : Programming Language
- **Flask Environment:** It is used to connect Python code to the Web-Development
- **Sklearn:** Sklearn is used for importing SVM model and classification Metrices
- **Matplotlib** :It is a Machine Learning Package used for Data Visualization
- **Pandas** : Pandas is used to read the data sets(.csv , excel files)
- **Numpy** :Numpy is used for Mathematical calculations
- **Seaborn** :Seaborn is the Machine Learning Package used to represent the data in the graphical visualization eg:Heat-Map
- **Jupiter Notebook**:Jupiter Notebook is an Ide used for most of the Machine Learning Projects for the better data Visualization and analysing the formats and make the code in an understandable Manner.

Input data:

We take breast-cancer data as our input. It has different data fields varying in their types.

The following is the list of data fields and their types:

The data frame has 569 rows and 33 columns.

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 569 entries, 0 to 568, Data columns (total 33 columns): Unnamed: 32

Data types: float64(31), int64(1), object(1)

Memory usage: 146.8+ KB

S.No	DATA FIELDS	DATA TYPE
1	Id	569 non-null int64
2	Diagnosis	569 non-null object
3	radius_mean	569 non-null float64
4	texture_mean	569 non-null float64
5	perimeter_mean	569 non-null float64
6	area_mean	569 non-null float64
7	smoothness_mean	569 non-null float64
8	compactness_mean	569 non-null float64
9	concavity_mean	569 non-null float64
10	concave points_mean	569 non-null float64
11	symmetry_mean	569 non-null float64
12	fractal_dimension_mean	569 non-null float64
13	radius_se	569 non-null float64
14	texture_se	569 non-null float64
15	perimeter_se	569 non-null float64
16	area_se	569 non-null float64
17	smoothness_se	569 non-null float64
18	compactness_se	569 non-null float64
19	concavity_se	569 non-null float64
20	concave points_se	569 non-null float64
21	symmetry_se	569 non-null float64
22	fractal_dimension_se	569 non-null float64
23	radius_worst	569 non-null float64
24	texture_worst	569 non-null float64
25	perimeter_worst	569 non-null float64
26	area_worst	569 non-null float64
27	smoothness_worst	569 non-null float64
28	compactness_worst	569 non-null float64
29	concavity_worst	569 non-null float64
30	concave points_worst	569 non-null float64
31	symmetry_worst	569 non-null float64
32	fractal_dimension_worst	569 non-null float64

After classifying the data, the confusion matrix is computed to understand how good or bad our data has been classified.

Table 1: Input Data Types

		Predicted Class	
		Class = Malignant	Class = Benign
Actual Class	Class = Malignant	True Positive	False Positive
	Class = Benign	False Negative	True Negative

The confusion matrix is computed using the following performance measures:

- Accuracy – it is the ratio of correctly predicted observations to the total observations.

  i.e., TP+TN/TP+FP+FN+TN
- Precision – it is the ration of correctly predicted positive observations to the total predicted positive observations.

i.e., TP/TP+FP

- Recall – it is the ratio of correctly predicted positive observations to all observations in the actual class – yes.

i.e., TP/TP+FN

Cross validation is used to predict the accuracy of the model better because here we change the training and test data sets randomly by different folds with more iteration. Cross validation score gives the mean score of all the training data records which are cross validated.

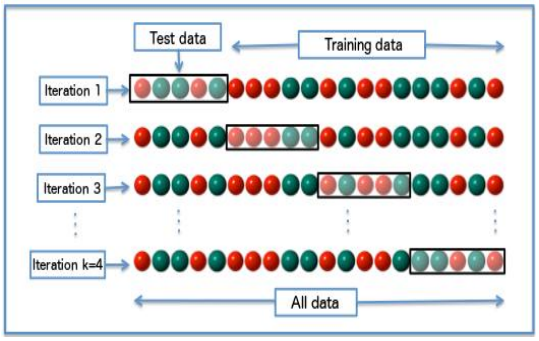

Figure 9: Separation Table Of Training And Testing Data

CONCLUSION AND FUTURE WORKS:

At the end of this project, we arrive at the conclusion that SVM is a supervised machine learning algorithm which is extensively used in the field of data classification due to the many advantages it offers such as, managing non-linear data, ability to handle high dimensional data and accuracy in classification. What makes it more preferred is its kernel methods which in turn contain kernel functions which when used appropriately can solve any problem.

REFERENCES:

1. Classification and diagnostic prediction of cancer using microarray gene expression.pdf. [2] Komura, D. (2004). Multidimensional support vector machines for visualization of gene expression data -- Komura et al_ 21 (4) 439 -- Bioinformatics. 175–179.
2. CHU, F., & WANG, L. (2005). Applications of Support Vector Machines To Cancer Classification With Microarray Data. International Journal of Neural Systems, 15(06), 475–484.
3. Pirooznia, M., & Deng, Y. (2006). SVM Classifier - A comprehensive java interface for support vector machine classification of microarray data. BMC Bioinformatics, 7(SUPPL.4).
4. Chen, X., Zhao, Y., Zhang, Y.-Q., & Harrison, R. (2007). Combining SVM Classifiers Using Genetic Fuzzy Systems Based on AUC for Gene Expression Data Analysis. Bioinformatics Research and Applications, 496–505.
5. Morelli, R. (2007). Using a support vector machine to analyze a DNA microarray. Tutorial Available at Http://2009. Hfoss. Org/Images/c/C1/ …, 1–17.
6. Wang, H., Shi, Y., Zhou, X., Zhou, Q., Shao, S., & Bouguettaya, A. (2010). Web service classification using support vector machine. Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 1, 3–6.
7. Chuang, L.-Y., Yang, C.-H., Li, J.-C., & Yang, C.-H. (2011). A Hybrid BPSO-CGA Approach for Gene Selection and Classification of Microarray Data. Journal of Computational Biology, 19(1), 68–82.
8. George, G. V. S., & Raj, V. C. (2011). Review on feature Selection Techniques and The Impact of Svm for Cancer. ArXiv Preprint ArXiv, 1109(3), 1062.
9. Liu, Q., Chen, C., Zhang, Y., & Hu, Z. (2011). Feature selection for support vector machines with RBF kernel. Artificial Intelligence Review, 36(2), 99–115.
10. KR, S. (2011). Microarray Data Classification Using Support Vector Machine. International Journal of Biometrics and Bioinformatics, (5), 10–15.
11. Zararsiz, G., Elmali, F., & Ozturk, A. (2012). Bagging Support Vector Machines for Leukemia Classification. International Journal of Computer Science, 9(6), 355–358.
12. Devi Arockia Vanitha, C., Devaraj, D., & Venkatesulu, M. (2014). Gene expression data classification using Support Vector Machine and mutual information-based gene selection. Procedia Computer Science, 47(C), 13–21.
13. Reddy, S. V. G., Reddy, K. T., Kumari, V. V., & Varma, K. V. (2014). An SVM Based Approach to Breast Cancer Classification using RBF and Polynomial Kernel Functions with Varying Arguments. International Journal of Computer Science and Information Technologies, 5(4), 5901–5904.
14. Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). SVM and SVM ensembles in breast cancer prediction. PLoS ONE, 12(1), 1–14. https://doi.org/10.1371/journal.pone.0161501
15. Ladwani, V. M. (2018). Support Vector Machines and Applications. Computer Vision, 1381–1390.
16. Nadira, T., & Rustam, Z. (2018). Classification of cancer data using support vector machines with features selection method based on global artificial bee colony. AIP Conference Proceedings, 2023(October).
17. Yahyaoui, A., & Yumuşak, N. (2018). Decision support system based on the support vector machines and the adaptive support. Biomedical Research (India), 29(7), 1474–1480.
18. Zhang, Y., Deng, Q., Liang, W., & Zou, X. (2018). An Efficient Feature Selection Strategy Based on Multiple Support Vector Machine Technology with Gene Expression Data. Biomed Research International, 2018, 1–11.

AUTHORS PROFILE

M.Ramachandra is M.Tech in Computer Science from NIT, Tiruchi, India. Since 2007 he has been working as Assistant Professor in the department of CSE, GMR Institute of Technology, Rajam, A.P, and India**.** His area of research includes Data mining & Bio Informatics.

Dr. Bhramaramba Ravi obtained her Ph.D from JNTUH in the year 2011. She has about 19 years of teaching experience and is currently Professor in the Dept.of Information Technology, GIT, GITAM, and Visakhapatnam. She has about 32 publications in reputed Journals. Her area of interest is Data Mining and Bioinformatics.