Ensemble Learning on Diabetes Data Set and Early Diabetes Prediction

Gulam Gaus Warsi[1]
Amity institute of Information Technology
Amity University, Uttar Pradesh,
Noida, U.P., India
gulamgauswarsi@gmail.com,

Sonia Saini[2]
Amity institute of Information Technology
Amity University, Uttar Pradesh,
Noida, U.P., India
ssaini@amity.edu

Kumar Khatri[3]
sunilkkhatri@gmail.com

Abstract **– Ensemble Learning is a process through which multiple classifiers are used to find solution to a problem. With the use of Ensemble Learning on a Diabetes data set a model is trained to predict the early onset of diabetes. This model targets for two major objectives: whether the person has a risk of diabetes in the near future or not and the risk probability of having diabetes associated with the person. The dataset consists of a basic set of questionnaires which would be helpful in training the model based on various classifiers or experts that strategically train the model. The libraries used in the system are Scikit-Learn, Pandas and Numpy. Once the model is trained it requires no retraining and could be helpful in predicting the early onset of diabetes in a person based on a basic questionnaire on the lifestyle of the subject such as regular lifestyle habits, eating habits, family history and medical history. Since predicting diabetes before it occurs is a problem complex to solve as the disease is basically a lifestyle disease which requires self-assessment of the patient and prevention of the disease at every step, so for reducing the uncertainty ensemble learning is used for prediction of diabetes risk in future as ensemble learning is a collection of various machine learning models.**

Keywords – Ensemble Learning, Early Diabetes Prediction, Risk Probability of Diabetes

I. INTRODUCTION

Diabetes is the condition of human body which impairs the ability of the body to process and maintain blood glucose levels, this condition is also known as blood sugar. Mainly there are three major kinds of diabetes: Diabetes Type I, Diabetes Type II and Gestational diabetes. Type I diabetes is the type of Diabetes where the human body fails in the production of enough insulin as per the requirement of the body. Type II diabetes is the one in which the human body produces insulin but is not effectively able to utilize it and Gestational diabetes occurs in pregnant women when the body becomes less sensitive to insulin. Diabetes is basically a lifestyle disease which is very hard to diagnose at its early stages and by the moment it gets caught then it is mostly advanced and can be only managed with medications and in cases patients are given shots of insulin to manage the blood sugar levels in the body. In serious cases where blood sugar levels stay unmanaged for a longer period of time cause a serious of organ damage including diabetic retinopathy (loss of vision), diabetic neuropathy (nerve damage), diabetic foot and other major organ damage such as the heart, pancreas, kidneys and many more.

Blood sugar in the human body can be managed by proper eating and lifestyle habits. Once a person is diabetic then a person has to follow proper living style apart from the medicine to manage the blood sugar otherwise high blood sugar levels can cause a lot of damage to the body. The best way to manage a chronic problem such as diabetes is to get regular health check-ups to detect any signs of irregular blood sugar levels in the body. Even after all these measures it is hard to detect diabetes in its early stages or to predict the early onset of the disease.

To help with the prediction of diabetes this system is developed which works on machine learning algorithms, which is trained with a dataset of questionnaires related to normal living and eating habits, family and medical history of more than 10,000 people. This dataset helps the model to train itself for the prediction of diabetes. At the time if the model is trained it can precisely and accurately predict the early onset of diabetes along with the possibility of a person to have diabetes in the future using multiple classifiers in ensemble learning. Using various algorithms the model can predict the onset of diabetes in near future and with this prediction a person can cautiously adapt to a healthier lifestyle to stay clean from this chronic disease. Ensemble learning is used in the model as it provides multiple sets of outputs for the same sets of inputs by using different classifiers which is helpful in the calculation of variance, error percentages and deviation of the predicted output.

II. LITERATURE SURVEY

Han et. al [1] in their paper used support vector machine to evaluate diabetes and added ensemble learning model which was used for presentation of the support vector machine decisions into transparent and comprehensible rules. They also used this model for solving unbalanced problems. Results of their paper according to China Health and Nutrition Survey's data concludes that the proposed model that has used ensemble learning creates rule sets with average precision 94.2% and average recall 93.9% for all the classes. Addition to it a hybrid system could be helpful in diagnosing diabetes and act as a second opinion for the patients.

Kavakiotis et. al [2] in their study conducted a research to get an overview of machine learning techniques that could be useful in the field of diabetic research mainly on diabetes mellitus in relation to the prediction, diabetic complications, genetic background, diagnosis, environment, healthcare and management. They used a wide variety of machine learning algorithms to carry out the computations, mainly 85% of supervised learning algorithms and 15% of unsupervised learning algorithms. This study was to provide knowledge that lead to a brand new hypothesis that targets deeper understanding of diabetes mellitus.

Alghamdi et. al [3] applied various machine learning techniques to uncover and find various predictors of diabetes. The study that was done, used a dataset of 32,555 patients who had no signs and symptoms of any coronary or heart disease and had a follow-up of 5-year. After 5-years follow-up completion 5,099 patients had developed the signs of diabetes or the disease itself. The data obtained from the study contained 62 attributes which were later on classified into four categories. 13 attributes were used in ensemble learning models to achieve an accuracy of 0.92.

Perveen et. al, [4] conducted a study following the bagging and adaboost ensemble learning models that used J48 decision tree serving a basic bootstrap with data mining. This study took place and the results of the experiment show that adaboost ensemble learning method is better than bagging and J48 decision tree.

Ozcift and Gulten [5] in their study have constructed a Rotation Forest ensemble classifier on as much as 30 machine learning algorithms to calculate their performances on the data set of Parkinson's, diabetes and heart diseases. Base classifiers had succeeded in achieving the accuracies of 72.15%, 77.52% and 84.43% on an average for diabetes, heart and Parkinson’s data respectively. The accuracy of RF classifier was 74.47%, 80.49% and 87.13% respectively.

Ali et. al, [6] conducted a study that revolves on the prediction of diabetes on the basis of their clinical and personal data with the use of boosting ensemble learning technique. A data set of 100 records was used generating the accuracy of 81%.

Chen et. al[7] aimed to present importance of finding the possibility of the connection in between diabetes mellitus and hair/urine specimens by chemometrics. A dataset was used that involved 211 specimens. On an average, the sensitivity, accuracy and specificity of the ensemble learning model was found to be 100%, 99%, 99% and 89%, 97%, 99% for hair and urine samples respectively.

Dewangan and Agrawal [8] in their study proposed the diagnosis of diabetes to be done using Artificial Neural Network, K-fold cross classification and validation etc. These techniques were used for the creation of an ensemble model for the accuracy, sensitivity and specificity measures of diagnosis of diabetes-mellitus.

Ozcift [9] proposed a classification method that uses support vector machine (SVM) using some features to teach the rotation forest classifier that is used for a diagnosis of the Parkinson’s disease. The dataset that was used contained records of voice measurements from 31 people, 23 with Parkinson’s disease. The step one of the diagnosis uses a linear support vector machine for the selection of ten most relevant features from the overall list of 22 features that were previously selected. This takes to the second step of the classification model that is used with six classifiers that are trained with the subset of the overall 22 features. Finally, on the last step, the accuracy of classifiers is improved with ensemble learning techniques.

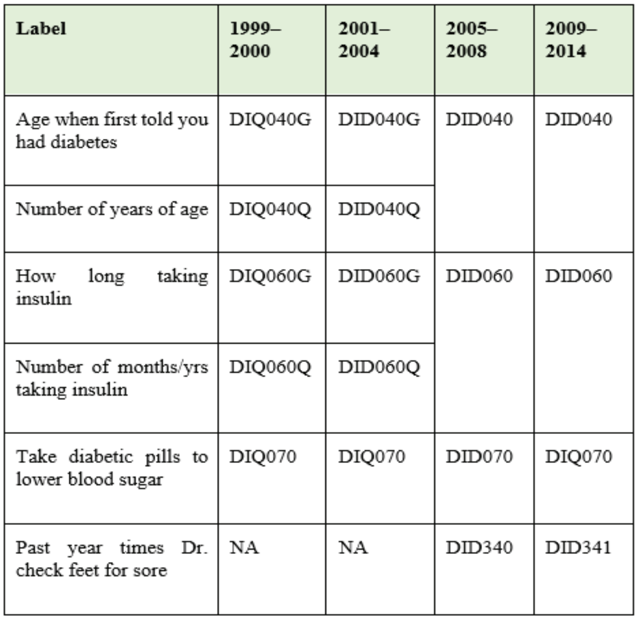
Zolfaghari [10] uses PIDD data set of 8 variables to predict the accuracy at 88.08% using the ensemble learning that is based on support vector machine (SVM), BP and Artificial Neural Networks and was successful in classification of the diabetic patient data provided by the Pima Indian diabetic database.

III. DATASET DESCRIPTION

The National Health and Nutrition Examination Survey is used as a dataset which contains a questionnaire of 10,176 patients. The diabetes section (prefix DIQ) of the survey contains personal interview data of the patients on diabetes and usage of insulin shots and/or oral hypoglycemic (low blood sugar) medications along with diabetes of the eye. The dataset also provides us with the information reported by the patients on general awareness of the risk factors of diabetes and risks for diabetes, general information of diabetic complications that occur and medical care or personal care that are connected to diabetes. All the participants included in the survey were aged 1 years or above and the questions asked to the patients varied upon the age and history of diabetes. The interviews were Computer Assisted Personal Interviews so there is a minimal chance of data errors.

Table 1 shows the cross reference for the variable names in the dataset from 1999-2000 dataset to 2013-2014 dataset.

TABLE I. VARIABLE NAMES ACROSS CYCLES



Label	1999–2000	2001–2004	2005–2008	2009–2014
Age when first told you had diabetes	DIQ040G	DID040G	DID040	DID040
Number of years of age	DIQ040Q	DID040Q		
How long taking insulin	DIQ060G	DID060G	DID060	DID060
Number of months/yrs taking insulin	DIQ060Q	DID060Q		
Take diabetic pills to lower blood sugar	DIQ070	DIQ070	DID070	DIQ070
Past year times Dr. check feet for sore	NA	NA	DID340	DID341

Table 2 shows some of the variable name, SAS label, code and its description which were targeted for both males and females in the age group of 12-150 years.

TABLE II VARIABLE NAMES ALONG WITH CODE AND DESCRIPTION IN THE DATASET

Variable Name	SAS Label	Code	Description
DIQ175A	Family history	10	Family history
		77	Refused
		99	Don't know
		.	Missing
DIQ175B	Overweight	11	Overweight
		.	Missing
DIQ175G	Lack of physical activity	16	Lack of physical activity
		.	Missing
DID250	Past year how many times visited doctor	1 to 24	Range of Values
		0	None
		7777	Refused
		9999	Don’t know
		.	Missing

There are such 21 more variables with each having at an average 4 possible values. These values were taken for a set of 10,176 patients. This dataset therefore acts as a training dataset for the ensemble learning model for the purpose of prediction and possibility of diabetes in the future.

IV. CASSIFIERS

A. Gradient Boosting Classifier

Gradient boosting is a system of machine learning used for regression and classification issues that delivers a forecast as an ensemble of feeble expectation simulations, regularly decision trees. It assembles the model in a phase insightful manner like other strategies used for boosting do, and it sums them up by permitting streamlining of a discretionary differentiable loss function.

Like other boosting strategies, gradient boosting joins frail "learners" into a solitary strong learner in an iterative style. It is most straightforward to clarify in the least-squares regression defining, where the objective is to "teach" a model **F** for the prediction of values in the form of **ŷ=F(x)** by minimizing $\mathbf{1/n \sum_i (y - yi)^2}$, where **i** indexes over some training set of size **n** of actual values of the output variable **y**.

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i))\right) \quad (1)$$

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}_{R_{jm}}(x), \quad \gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma). \quad (2)$$

Eq. 1 and 2 are gradient boosting algorithms that are used on the data set for classifying the values and thus helping in generating the outputs.

B. K Nearest Neighbors – Classification

K nearest neighbors classification is an algorithm that is used to store all the available options and possibilities there upon classifies new option that is mainly based on a measure that is similar. This classification has been used for models requiring statistical estimation and pattern recognition in datasets.

A new case is classified by the majority of votes of the neighbors, with the case is being assigned that is most common amongst all its K nearest neighbors which is measured by a distance function as shown in eq. 3, 4 and 5. If the value of K = 1, then the case is simply allocated to the class of its neighbor that is nearest to it.

$$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \quad (3)$$

$$\sum_{i=1}^{k} |x_i - y_i| \quad (4)$$

$$\left(\sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{1/q} \quad (5)$$

Equation 3, 4 and 5 are Euclidean, Manhattan and Minkowski’s distance functions respectively.

C. Voting Classifier

The Ensemble Vote Classifier is a meta-classifier for joining comparable or adroitly unique machine learning classifiers for classification by means of larger part or majority voting. The Ensemble Vote Classifier executes "hard" and "soft" voting. In hard voting, we anticipate the final class name as the class name that has been anticipated most every now and again by the classification models. In soft voting, we anticipate the class names by averaging the class-probabilities.

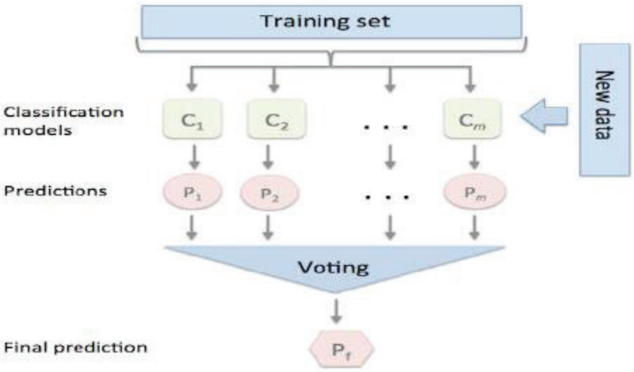


Fig. 1. Ensemble Voting Classifier[11]

Fig. 1 shows the working ensemble voting classifier where training dataset is provided to different classification models and predictions are made on those models. After the predictions are done the models are voted upon and a final prediction is made.

D. *Random Forest Classification*

Random forests are ensemble learning methods for classification and regression that works by developing a huge number of decision trees at the time of training and yielding the class which is the method of the classification or regression of the individual trees that are present in the forest. Random decision forests are right for decision trees' propensity for overfitting to their training dataset provided.

$$\hat{y} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} W_j(x_i, x')\, y_i = \sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} W_j(x_i, x')\right) y_i. \qquad (6)$$

Equation 6 shows the relationship of a decision tree with its nearest neighbors. Which shows that the complete forest is again follows a weighted neighborhood scheme of classification, with the weights which average those weights of the individual trees in the forest.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \qquad (7)$$

Equation 7 is used for bagging of the data again and again opts for a random sample with reallocation of the training dataset and it fits the trees to these samples of data.

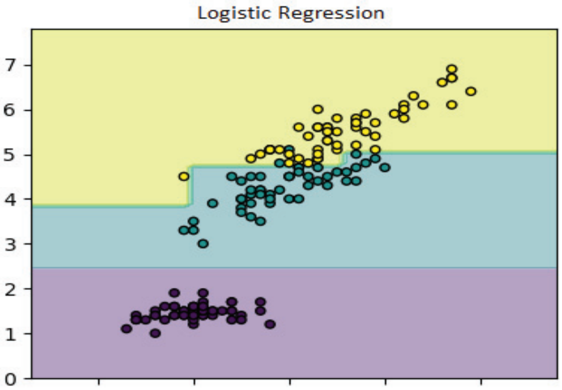
V. RESEARCH AND DEVELOPMENT

The ensemble learning model is created with the use of python libraries such as:

a. NumPy which is used for supporting large and multi-dimensional matrices or arrays along with higher level of support of mathematical function that helps to operate on these multi-dimensional data.
b. SciPy which is used in python for the computation of scientific and technical data.
c. Pandas is a software library which is written for Python mainly used for data analysis and manipulation on data structures and time series.
d. Matplotlib is a plotting library for Python that is used for providing APIs for plotting graphs and graphical representation of data.
e. Scikit-learn is a software machine learning library for the Python. It includes various algorithms for classification, regression and clustering which can be used for computations in python.

VI. SIMULATION & IMPLEMENTATION RESULTS

The model is trained using the dataset of diabetic survey of 10,176 patients that contains a set of questionnaires based on the lifestyle, eating habits, family and medical history. Once the model is trained with the dataset then there is no need of retraining of the model. After training the model it can very accurately provide the results as whether the patient is at risk of early onset of diabetes or not. If so what is the probability of having the early onset. Along with these results the model also shows the mean error rate between the actual and predicted values and mean square error.



Fig, 2. Logistic Regression[12]

The regression values form figures 2, 3 and 4 are used to initialize a soft voting classifier with weights as shown in figure 5. With the help of this data average probabilities are calculated.

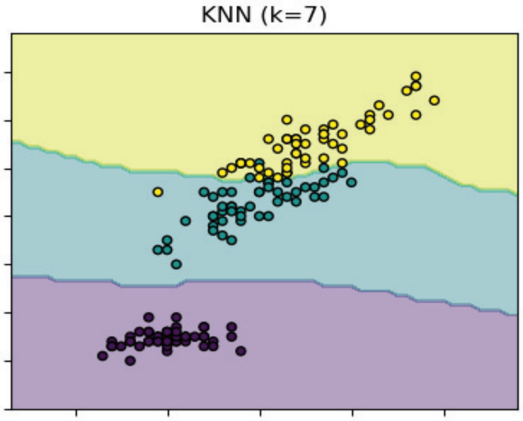


Fig. 3. K Nearest Neighbors Classifier[12]

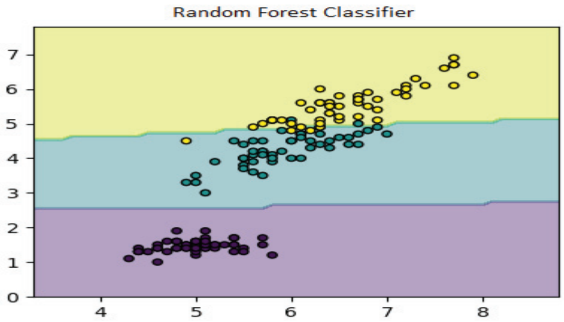


Fig. 4. Random Forest Classifier[12]

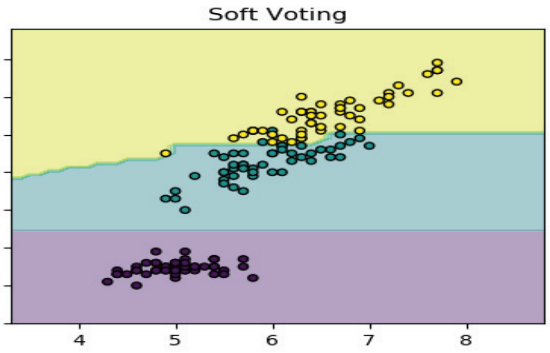


Fig. 5. Soft Voting Classifier initiated from the values of the initial three classifiers Logical Regression, KNN and Random Forest[12]

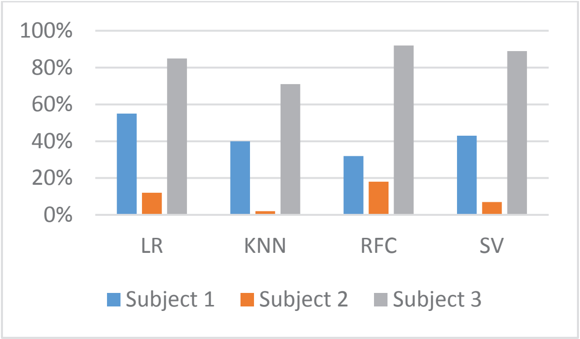


Fig. 6. Comparison of values obtained from various classifiers on 3 subjects

TABLE III. DATA OBTAINED FROM CLASSIFIERS OF 3 SUBJECTS

Classifiers	Subject 1	Subject 2	Subject 3
Logistic Regression	55%	12%	85%
K-Nearest Neighbor	40%	2%	71%
Random Forest Classifier	32%	18%	92%
Soft Voting	43%	7%	89%

Fig. 6 shows the result analysis of the implemented system for 3 subjects whose data was.

VII. CONCLUSION

In this paper, an ensemble learning model has been proposed which would be helpful in the early prediction of diabetes accurately based on a set of questionnaires about the lifestyle of a person. Once the person correctly provides the inputs to the machine or model which has been previously been trained by a large dataset of more than 10,000 sets of data, then the machine would compare the user's input to the existing plotted values, compare the outputs and provide an accurate possibility of the person having diabetes in the future. The model can accurately predict the early onset of diabetes in a person based on a set of questionnaires. The purpose of using ensemble learning is that it uses multiple classifiers for the same sets of values and thus providing multiple sets of results for the same sets of inputs which can be helpful in calculating the deviation value, error percentages, and variance of the outputs.

VIII. FUTURE WORKS

A variety of works can be done in the near future with the help of this model as this model can be used for the prediction of chronic and lifestyle diseases that slowly affect a person and can have disastrous effect on the body when they are advanced such as hypotension, hypertension, stress, obesity and others. Such diseases are hard to diagnose in their earlier stages as they are usually missed due to the patient's carelessness, but with the help of machine learning model they can be easily predicted by non-evasive medical procedures, just by asking a few questions about a person's regular lifestyle, family history, medical history and eating habits. If machine learning models are properly trained then these models can accurately predict the onset of such chronic diseases even before the first symptoms appear on the patient.

This model can be also trained with the lifestyle, eating and living habits of healthier sections of the population whose habits can be analyzed by the model and once the model is trained then after the diagnosis of a person the model can accurately guide the person with a good set of lifestyle habits to lead a healthy life just by implementing a few specific lifestyle changes and eating habits. Thus this model can be successfully used in the medicine sector, not just in the form of educative medicine, also preventive medicine as the model will be successfully able to predict future diagnosis and also provide advice on minor changes of lifestyle that could help a person to avoid those diseases in their life.

ACKNOWLEDGEMENT

We express our deep sense of gratitude to the Founder President of Amity Group, Dr. Ashok K. Chauhan, for his keen interest in promoting research in Amity University.

REFERENCES

[1] Han, L., Luo, S., Yu, J., Pan, L., & Chen, S. (2015). Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. *IEEE journal of biomedical and health informatics*, *19*(2), 728-734.

[2] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, *15*, 104-116.

[3] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PloS one*, *12*(7), e0179805.

[4] Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, *82*, 115-121.

[5] Ozcift, A., & Gulten, A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *Computer methods and programs in biomedicine*, *104*(3), 443-451.

[6] Ali, R., Siddiqi, M. H., Idris, M., Kang, B. H., & Lee, S. (2014, December). Prediction of diabetes mellitus based on boosting ensemble modeling. In *International conference on ubiquitous computing and ambient intelligence* (pp. 25-28). Springer, Cham.

[7] Chen, H., Tan, C., Lin, Z., & Wu, T. (2014). The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis. *Computers in biology and medicine*, *50*, 70-75.

[8] kumar Dewangan, A., & Agrawal, P. (2015). Classification of diabetes mellitus using machine learning techniques. *International Journal of Engineering and Applied Sciences*, *2*(5).

[9] Ozcift, A. (2012). SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease. *Journal of medical systems*, *36*(4), 2141-2147.

[10] Zolfaghari, Rahmat. "Diagnosis of diabetes in female population of pima indian heritage with ensemble of bp neural network and svm." *Int. J. Comput. Eng. Manag* 15 (2012): 2230-7893.

[11] http://rasbt.github.io/mlxtend/user_guide/classifier/EnsembleVoteClassifier/

[12] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html