iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** X **Month of publication:** October 2022

DOI: https://doi.org/10.22214/ijraset.2022.47114

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: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

Prediction of Diabetes Using Ensemble Learning

V. Joe Nithin[1], Prof. S. Pallam Setty[2]
[1]*Department of Computer Science and Systems Engineering (A), Andhra University, Visakhapatnam*
[2]*Professor, Department of Computer Science and Systems Engineering (A), ANDHRA UNIVERSITY, Visakhapatnam*

Abstract: Diabetes mellitus is a chronic condition that influences everyday life of the individual having this disease. Diabetes can only be treated to maintain controlled blood glucose levels than to achieve a permanent cure to lead a normal life. As the proverb goes, "prevention is better than cure", this model aims at "predicting the probability", of getting this condition, which help early prognosis enough to either avoid it or delay it. Ensemble method is used for prediction of probability of getting diabetes. Classification models in machine learning are used for decision making and enlisted in sequence of accuracy. Hyperparameters are tuned for top five accurate models.
Comparison of different classifiers are carried out and then subjected to voting to choose the best possible method of prediction. Voting is carried out in hard voting and soft voting procedures. The results obtained are better compared to general classifiers individually.
Keywords: Ensemble, stacking, hard voting, soft voting.

I. INTRODUCTION

Diabetes mellitus is a chronic condition where the pancreas loses the ability to produce enough insulin to breakdown glucose. Over the long-term high glucose levels are associated with damage to the body and failure of various organs and tissues. According to the World Health Organization, the population with diabetes rose from 108 million in 1980 to 422 million in 2014. Moreover, in 2016, it was the primary cause of 1.6 million deaths [1].
Approximately, 537 million adults, between 20 - 79 years, are living with diabetes. Undiagnosed adults account to 50% (240 million) of those living with diabetes. The total number of people living with diabetes is projected to rise to 643 million by 2030 and 783 million by 2045 [2].
Prediction of a desired outcome can be achieved through machine learning algorithms through analysis of available deciding factors. Based on this there were many predictions like weather forecasting can be made. To achieve this, classification algorithms like decision trees, regression models are used. Pre-existing techniques for diabetes prediction include classification methods and manual choice of any one of those methods depending on their accuracy of prediction. This new method aims at choosing an appropriate method by itself by stacking models and selection of the best method by voting, known as ensemble of models. Ensemble is a combination of different classifiers, results of which are then used as a classification model for the purpose of choosing the best model, which always yields higher accuracy. Large database is maintained by healthcare sectors that can be used by these kinds of techniques as a part of big data analytics, that contribute significantly to make healthcare better. Health care initiatives like Ayushman Bharat [3], Aarogyasri [4] and family doctor [5] etc., by governments can benefit with this approach with the slightest changes.

II. METHODOLOGY

Machine learning techniques are existent for basic classification of data. These classifiers are used for complex learning of parameters and predict the possible outcome. Existing methods are simple use of a classifier, believed to be better by the developer. Proposed technique uses ensemble technique, which is a collective use of different predefined classifiers, to choose the best from them by the algorithm itself. Ensemble learning [6] use evaluation through different models of classifier. Here, linear discriminant analysis, logistic regression, catboost classifier, random forest classifier, gradient boost classifier, extra trees classifier and ada boost classifier are used for comparison.
The constructed models are then stacked to create a model for the models which chooses the best model based on voting criteria. Stacking is done for the afore mentioned models in this proposed technique but can be used for any number of models which may increase the runtime.

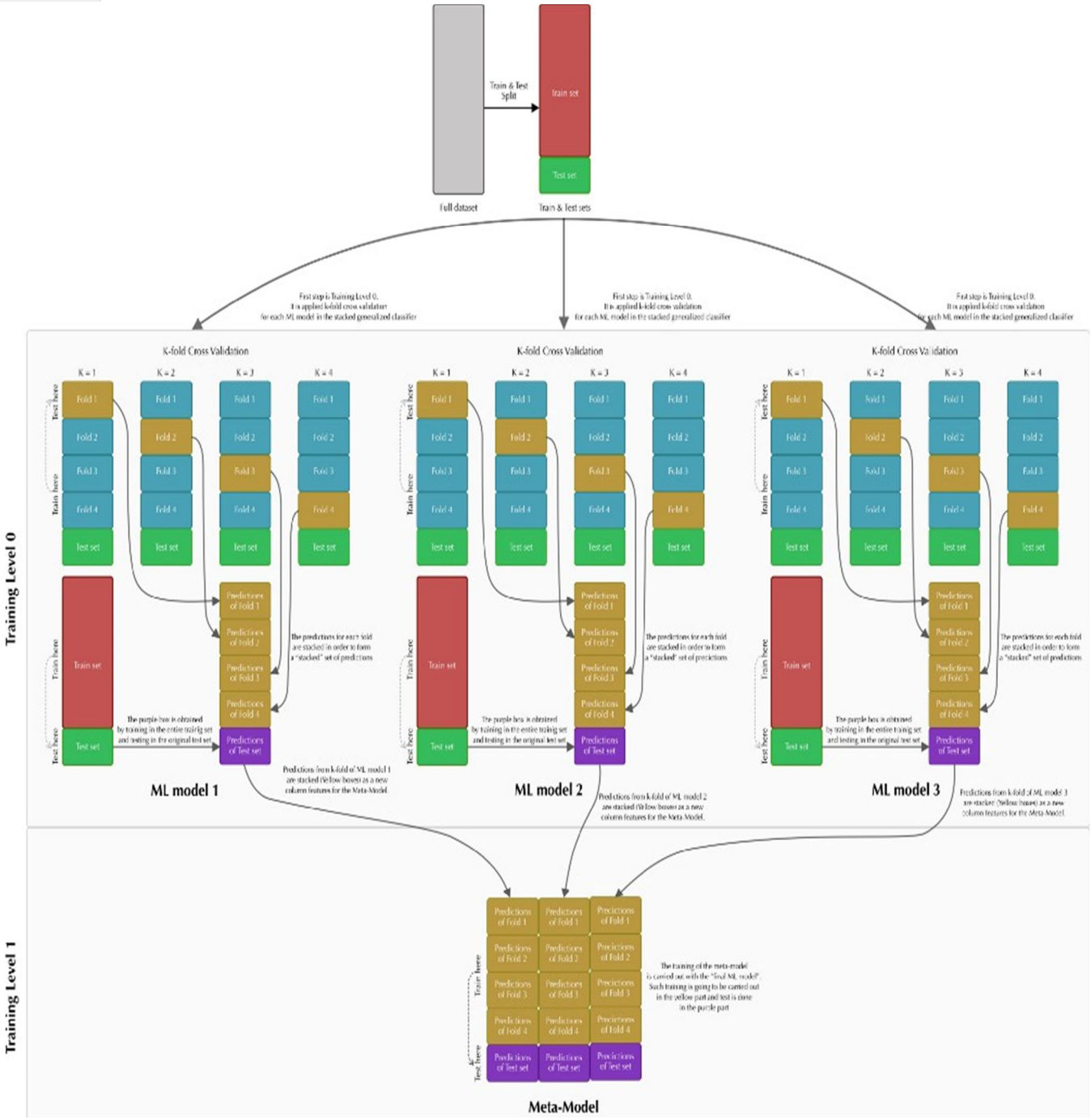


Figure-1: Stack Building for classification Models [7]

Stacked models are ranked according to accuracy and hyperparameters are tuned to improve accuracy of each model using cross validation. Top five accurate models are considered for voting the best method for the provided data. Voting methods used are

1) Soft Voting
2) Hard Voting

Soft voting is a method of choosing the best class of classifiers based on the average probability given to that class. Hard voting is a method of choosing the best class of classifiers if majority of them yielded similar outcome. The choice of model may differ based on exploratory data analysis of the trained dataset. Dataset available for testing the model was Pima Indians diabetes data of 768 women. 75% of the data is used to train the model and 25% of the records were used to test the model.

III. RESULTS

Individual classifiers returned around 77% accuracy and the proposed model returned 90% accuracy. Increasing the training data can yield better results but may lead to overfitting. Comparative advantage of the model with the same training set was considered, thus taking the significant increase in accuracy into account.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lda	Linear Discriminant Analysis	0.7711	0.8333	0.5480	0.6970	0.6090	0.4514	0.4607	0.018
lr	Logistic Regression	0.7656	0.8303	0.5369	0.6869	0.5991	0.4377	0.4465	0.512
catboost	CatBoost Classifier	0.7561	0.8182	0.5585	0.6557	0.5992	0.4261	0.4316	3.271
rf	Random Forest Classifier	0.7636	0.8152	0.5761	0.6733	0.6181	0.4483	0.4534	0.516
gbc	Gradient Boosting Classifier	0.7522	0.8123	0.5696	0.6423	0.6002	0.4221	0.4264	0.176
et	Extra Trees Classifier	0.7522	0.8027	0.5467	0.6489	0.5881	0.4140	0.4204	0.467
ada	Ada Boost Classifier	0.7466	0.7924	0.5634	0.6337	0.5914	0.4098	0.4142	0.111

Figure-2: Accuracy of individual classifiers before tuning and stacking

The accuracy of the best model was 77.11% with area under curve 83.3%, with very low precision and recall.

```
#prediction
pred = final_model.predict(X_test)
#Accuracy
final_model = confusion_matrix( y_test, pred)
accuracy = accuracy_score(y_test , pred)
precision = precision_score(y_test , pred)
recall = recall_score(y_test , pred)
f1 = f1_score(y_test,pred)
print('accuracy: {0:.4f}, precision: {1:.4f}, recall: {2:.4f},\
F1: {3:.4f}'.format(accuracy, precision, recall, f1))
```

```
accuracy: 0.9010, precision: 0.9032, recall: 0.8116,F1: 0.8550
```

Figure-3: Accuracy, precision and recall for the finalized model

The model returned comparatively better results for the same dataset with similar proportions of training and test data. The confusion matrix of the finalized model produced least false positive and false negative outcomes compared to the true positive and true negative outcomes, making the model reliable.

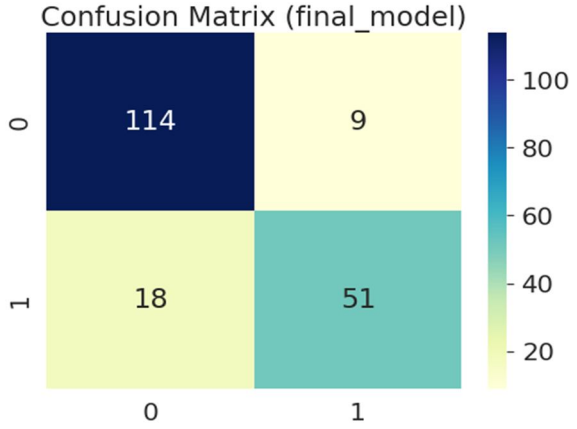


Figure-4: Confusion matrix for finalized model

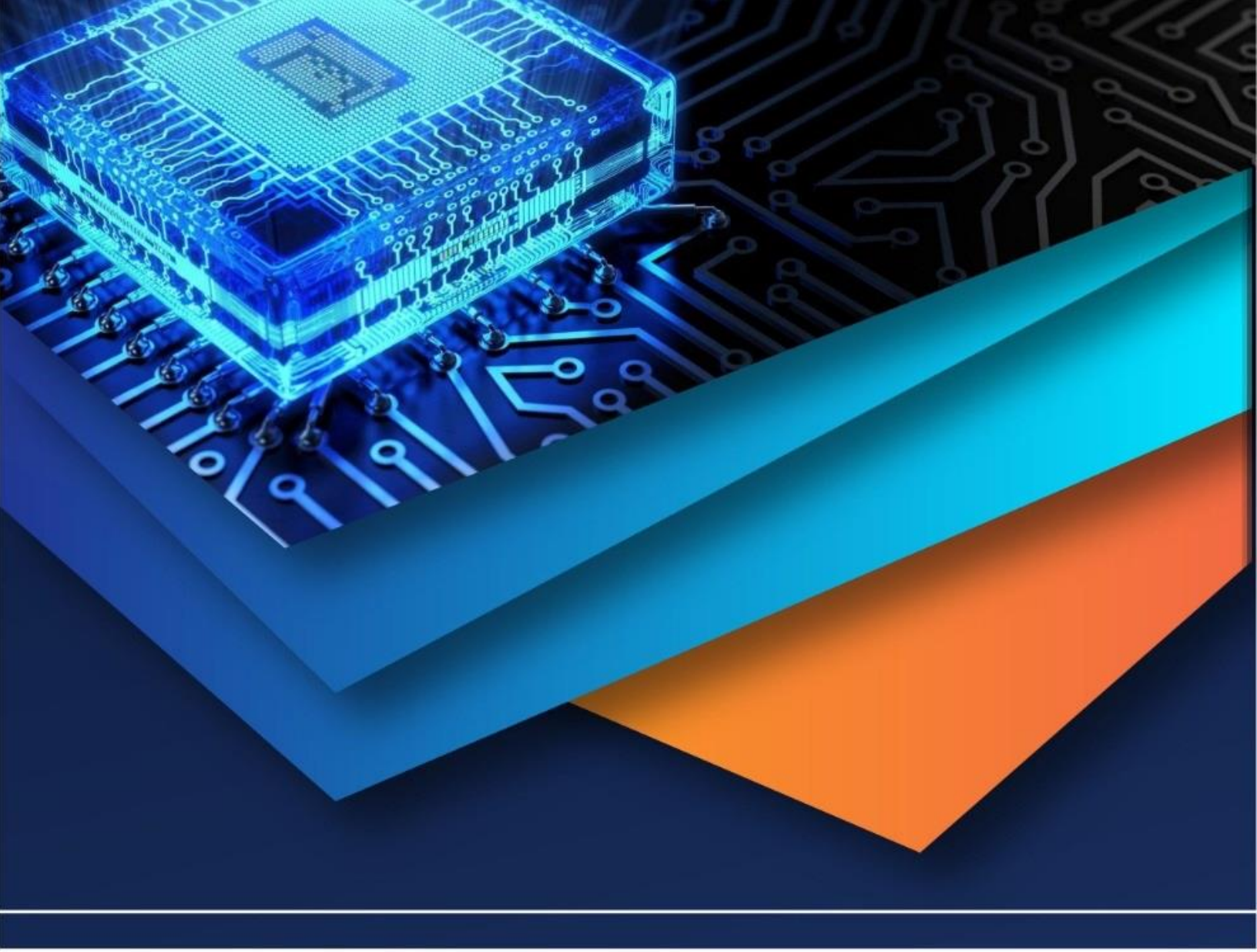
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

IV. CONCLUSION

The final model was chosen by selecting a suitable model from the stacked generalization, and then by performing voting, which in the end yielded 90% accuracy, compared to various individual classifiers which accounted to less than around 78% accuracy. This ensemble method proved to be better than individual classifiers which are to be manually checked and anticipated for results at every outcome for comparison.

REFERENCES

[1] World Health Organization. Diabetes (who.int) (Accessed on 16 December 2021)
[2] International Diabetes Federation, https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures (Accessed 16 December 2021)
[3] Ayushman Bharat, Official Website Ayushman Bharat | HWC (nhp.gov.in)
[4] Aarogyasri Scheme, Aarogyasri Health Care Trust - Quality Medicare For All (telangana.gov.in)
[5] Family doctor, Andhra Pradesh CM launches Family Doctor system to provide better health services (medicaldialogues.in)
[6] A Gentle Introduction to Ensemble Learning Algorithms (machinelearningmastery.com)
[7] Ensemble Learning: Stacking, Blending & Voting | by Fernando López | Towards Data Science



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY