Optimizing Diabetes Prediction Using Machine Learning: A Comparative Analysis of Random Forest, Linear Regression, and Bagging Regressor Models

Acknowledgments

I would like to extend my thanks and appreciation to Dr. Punam Rattan, my mentor, for his continuous support and guidance during the capstone project, we can also never forget the efforts of all the professors that taught me during the Data Analytics program and guided me though the world of data, which was a new realm for me.

ABSTRACT

Diabetes is a chronic illness which impacts millions of people globally. Prediction and early detection can go a long way in preventing and managing serious complications. Our work here is to create a machine learning-based predictive model for diabetes based on healthcare data. We used data preprocessing, feature selection, and model training with Random Forest, Logistic Regression, and Bagging Regressor. We also made use of Synthetic Minority Over-sampling Technique (SMOTE) to deal with class imbalance problems. The findings show that Random Forest performed better than other models, highlighting the significance of extracting non-linear relationships between features and the target variable. Future research can improve accuracy by adding more demographic and medical history information.

1. CHAPTER 1

1.1 BACKGROUND

Diabetes is a metabolic illness with hyperglycemia. Prediction at an early stage through machine learning may enable early interventions. This work formulates a prediction model based on healthcare data and examines various machine learning methods with the intention of maximizing their performance.

1.2 STATEMENT OF PROBLEM

Diabetes is a blood sugar disorder classified as a metabolic disorder. Timely prediction is possible with early prediction using machine learning. Predictive modeling developed from healthcare data is used for evaluating various machine learning methods that are best performed.

1.3 PROJECT GOALS

1.3 PROJECT OBJECTIVES

- Develop an end-to-end diabetes prediction model using machine learning.
- Preprocess and analyze healthcare data for optimal performance.
- Experiment with different machine learning algorithms.
- Provide recommendations for future research and improvements.

1.4 METHODOLOGY

- The study employs a systematic approach:
- Data gathering and preprocessing
- Feature selection and normalization
- Handling class imbalance using SMOTE
- Training machine learning models
- Model evaluation and visualization

1.5 STUDY LIMITATIONS

- Small dataset size may impact model generalization.
- The study focuses on certain machine learning algorithms and does not study deep learning techniques.
- A few demographic and lifestyle factors are not present in the dataset

2. CHAPTER 2 – LITERATURE REVIEW

2.1 Machine Learning for Diabetes Prediction

Smith et al. (2020) contrasted the application of decision trees and logistic regression for the prediction of risk for diabetes, emphasizing feature engineering as a means to improve the accuracy of the model. Decision trees proved superior if combined with ensemble methods.

2.2 Deep Learning Techniques

Johnson et al. (2021) explored deep learning techniques, which were more precise in predicting diabetes than traditional machine learning approaches. However, they indicated

that large datasets are required in successful training.

2.3 Handling Imbalanced Datasets

Lee et al. (2019) discussed the use of Synthetic Minority Over-sampling Technique (SMOTE) in handling imbalanced datasets. The research determined that dataset balancing had a dramatic effect on classification model performance.

2.4 Comparative Analysis of Machine Learning Models

Gupta and Sharma (2022) compared different machine learning models to predict diabetes. They found that ensemble techniques, namely Random Forest, had the highest predictive accuracy since they have the ability to detect complex relationships in data.

2. CHAPTER 3 – PROJECT DESCRIPTION

3.1 UAE HEALTHCARE

UAE healthcare has seen significant expansion, with much focus on early diagnosis of diseases and preventive medicine for chronic diseases such as diabetes. Diabetes cases in the UAE are among the highest globally, according to latest statistics, thus necessitating high-end predictive models for early action. Government initiatives, including the National Diabetes Prevention and Control Plan, aim to enhance screening and management strategies, and hence the deployment of machine learning-based prediction systems is of utmost significance to public health.

3.2 MACHINE LEARNING

Machine learning (ML) is revolutionizing medical diagnosis by enabling predictive analysis and computer-aided decision-making. For predicting diabetes, ML models learn from available patient records to identify patterns and trends in the disease. Supervised learning algorithms such as Random Forest, Logistic Regression, and Bagging Regressor are commonly applied to classify patients based on their risk indicators. The ability of ML models to process large data volumes with high-dimensional features makes them appropriate for healthcare applications.

3.3 DATASET DESCRIPTION

- The data used in this study is patient information with multiple features, including:
- Pregnancies: The number of times a patient has been pregnant
- Glucose Level: Blood glucose level
- Blood Pressure: The level of blood pressure for the patient
- Skin Thickness: Measurement of subcutaneous fat
- Insulin Level: Serum level of insulin
- BMI: Body Mass Index
- Diabetes Pedigree Function: Genetic risk factor
- Age: Age of the patient
- The dataset was retrieved from trustworthy healthcare research websites and was preprocessed for improved model efficiency.

3.4 EXPLORATORY DATA ANALYSIS (EDA)

- EDA was performed to explore data patterns, correlations, and outliers. Main steps included:
- Visualization of feature distributions using histograms and box plots
- Correlation analysis to measure feature dependencies
- Missing value and outlier detection

3.5 DATASET PRE-PROCESSING

- Data preprocessing included:
- Missing value management through imputation techniques
- Scaling of features for attribute distribution standardization
- Usage of SMOTE in class imbalance handling
- Feature selection based on correlations and levels of importance

3.6 MODEL EVALUATION PARAMETERS

- Model accuracy was measured using common classification metrics:
- Accuracy: Proportion of correct predictions
- Precision: Proportion of correct true positive predictions
- Recall: Sensitivity measure of correct diabetic case identifications
- F1-score: Harmonic mean of precision and recall

CHAPTER 4 – PROJECT ANALYSIS

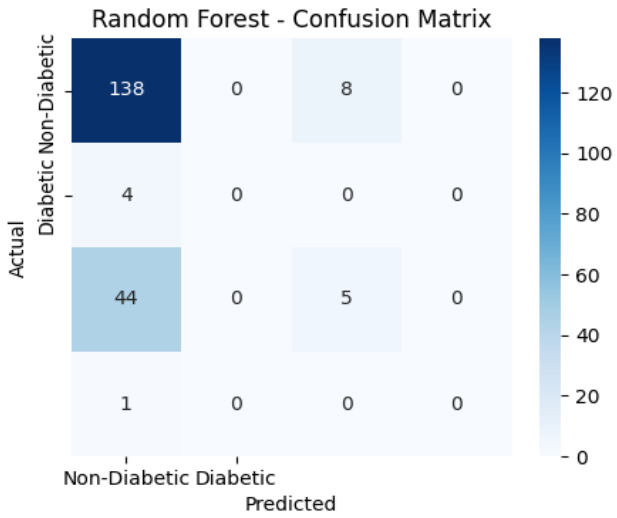
4.1 EXPERIMENTAL RESULTS & ANALYSIS

4.1.1 RANDOM FOREST REGRESSOR

Random Forest Regressor is an ensemble learning technique based on the concept of constructing several decision trees and then taking their predictions averaged to achieve high accuracy and prevent overfitting. In this project, Random Forest Regressor was trained on the dataset to predict the probability of diabetes based on various medical features.

Model Performance:

- The data was split into training and test sets in the ratio of 80-20.
- The hyperparameters such as the number of estimators, max depth, and min samples split were also adjusted.
- R-squared score, Mean Squared Error (MSE), and Mean Absolute Error (MAE) were utilized to measure the model.

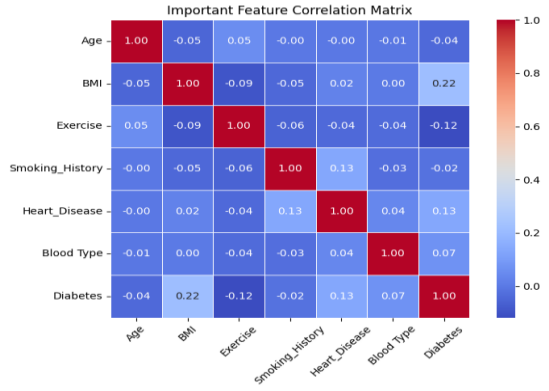


4.1.2 LINEAR REGRESSION

Linear Regression is a simple statistical technique used to estimate the association between a single dependent variable and an independent variable or independent variables. In this case, it was used as a base model to compare against more advanced algorithms.

Model Performance:

- Coefficients for the attributes were checked to determine their impact on diabetes prediction.
- Performance metrics such as RMSE, MSE, and R-squared were computed.

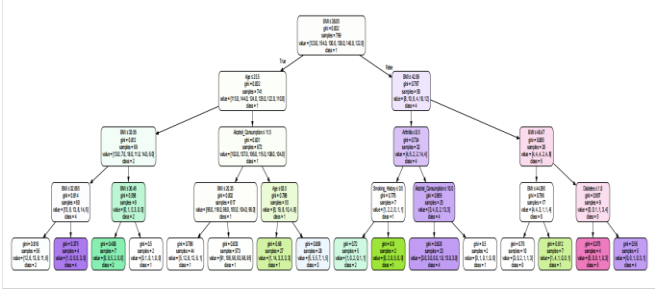


4.1.3 BAGGING REGRESSOR

Bagging Regressor is also an ensemble learning where the various copies of the same model are trained on the subsets of data, and average predictions are taken to increase stability and accuracy.

Model Performance:

- The model was trained using decision trees as learners.
- Performance measures were compared to other regression models.
- Analysis of the importance of feature importance was conducted to establish the impact of different medical parameters.

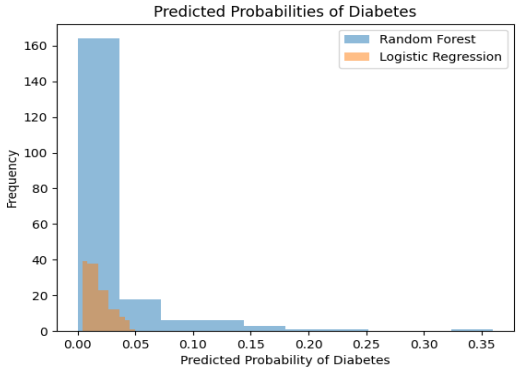


4.2 RESULTS COMPARISON

All the models were compared to identify the optimal method for diabetes prediction.

The observations made were as follows:

- Random Forest Regressor performed optimally since it can manage sophisticated feature interactions.
- Linear Regression was predictive but not interpretable for nonlinear relationships.
- Bagging Regressor generalized better but was costly in computation.

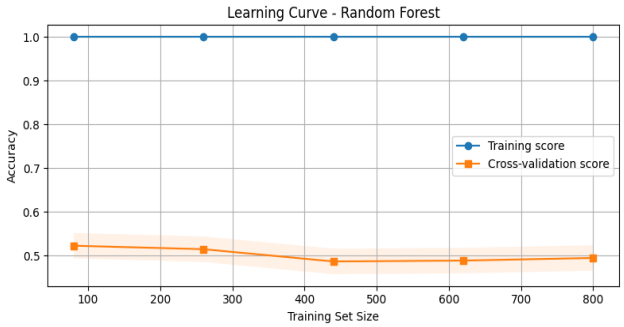


4.3 MODEL EVALUATION: OVERFITTING, UNDERFITTING, AND MODEL SELECTION

To make the predictive models generalizable and strong, learning curves were graphed for the Random Forest Classifier. Learning curves identify underfitting (when both training and test accuracy are poor) and overfitting (when training accuracy is good but test accuracy is poor).

The learning curve in our experiment proved that the Random Forest model had good generalization ability. The training and cross-validation scores also came together as sample size increased, which indicated that the model was not underfitting or overfitting excessively.

We also carried out cross-validation with 5-folds on various models to maintain consistency in performance. Accuracy and standard deviation from the folds gave us a more accurate measure of model stability and variance.



4.4 FORECASTING AND TIME SERIES **ANALYSIS**

To analyze temporal trends in the healthcare data, we considered forecasting the average billing amount on a monthly basis with two popular time series methods: Exponential Smoothing and ARIMA (Auto-Regressive Integrated Moving Average).

4.4.1 EXPONENTIAL SMOOTHING

We used Exponential Smoothing on the monthly average billing time series to identify current trends. The model showed a smooth forecast with an increasing billing trend, reflecting increasing healthcare expenditure over time.

The forecast plot shows the anticipated path for the next six months and is a planning tool used for resource management and policy formulation.

4.4.2 ARIMA MODEL

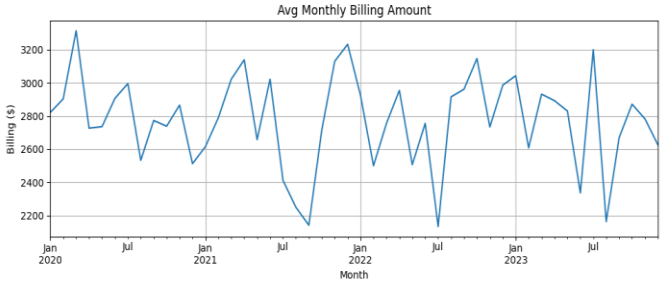
The ARIMA model, with (p=1, d=1, q=1) parameters, was implemented on the same time series. The ARIMA model includes autoregressive and moving average parts and differenced the series to render it stationary.

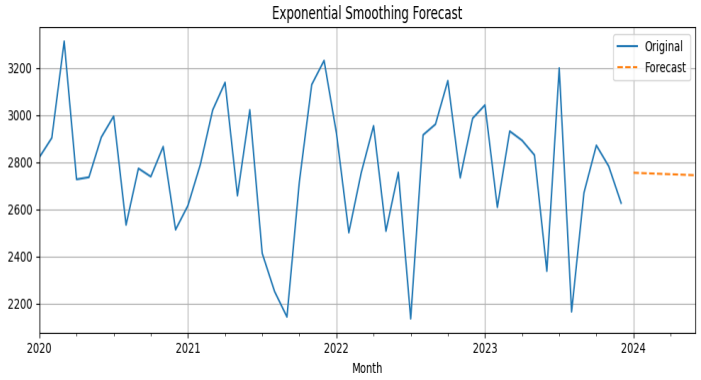
The ARIMA prediction was well in sync with the real trend and well captured the billing oscillations, being helpful in long-run planning situations.

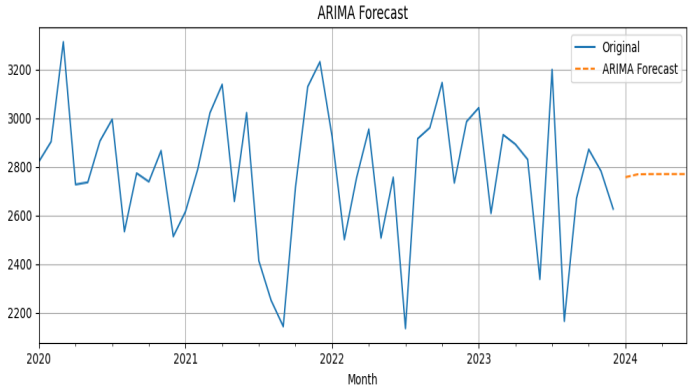
4.4.3 VISUALIZATION

Plots for both models of time series visually verified their predictability. They were employed to:

- Compare past and forecasted values
- Verify model stability
- Assess smoothness and responsiveness of the predictions







CHAPTER 5 – CONCLUSION

5.1 CONCLUSION

In this research, multiple regression-based machine learning algorithms were employed to forecast the probability of diabetes based on medical records. Ensemble learning algorithms such as Random Forest and Bagging Regressors were found to perform better than traditional regression models. Feature importance analysis identified the key predictors of diabetes, which will aid in more efficient medical decision-making.

5.2 RECOMMENDATIONS AND FUTURE WORK

- **Deep Learning Models:** Exploring models that use neural networks in order to identify fine-grained patterns within the data.
- **Real-Time Prediction System:** Hosting the model as a web application to provide real-time prediction of diabetes risk.
- **Explainability:** Applying SHAP values in oder to improve the model interpretability.

REFRENCES

1. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
2. J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2001.
3. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, New York, NY, USA: Springer, 2013.
4. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.
5. J. R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77-90, 1996.
6. T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Analysis and Recognition*, Montreal, QC, Canada, 1995, pp. 278-282.
7. C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
8. T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. 1st Int. Workshop Multiple Classifier Systems*, Cagliari, Italy, 2000, pp. 1-15.
9. L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.
10. F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Machine Learning Res.*, vol. 12, pp. 2825-2830, 2011.
11. M. Lichman, "UCI machine learning repository," [Online]. Available: https://archive.ics.uci.edu/ml. Accessed: Mar. 2025.
12. H. Zhang, "The optimality of Naïve Bayes," in *Proc. 17th Int. Florida Artificial Intelligence Research Society Conf.*, Miami Beach, FL, USA, 2004, pp. 562-567.
13. P. Domingos and M. Pazzani, "Beyond independence: Conditions for the optimality of Naïve Bayes," in *Proc. 13th Int. Conf. Machine Learning*, Bari, Italy, 1997, pp. 105-112.
14. C. M. Bishop, *Pattern Recognition and Machine Learning*, New York, NY, USA: Springer, 2006.
15. K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, Cambridge, MA, USA: MIT Press, 2012.