Predicting Diabetes in Pima Indian Women Using Logistic Regression

Inder Khera, Jenny Zhang, Jessica Kuo, Javier Martinez (alphabetically ordered)

2023-12-08

Table of contents

Summary

This study evaluated logistic regression for predicting diabetes in Pima Indian women using features such as glucose levels, BMI, and pregnancies. The model achieved 75% accuracy on the test set, outperforming the baseline dummy classifier's 67%. Glucose was the most significant predictor, followed by BMI and pregnancies, while blood pressure and insulin had weaker impacts. The model misclassified 54 cases, with 41 false negatives and 13 false positives, highlighting areas for improvement.

The results indicate that logistic regression is a promising tool for diabetes screening, providing an efficient way to identify potential cases. However, the high number of false negatives is concerning, as they could lead to delayed diagnoses and treatments. Future improvements could

include feature engineering to address misclassifications, testing alternative machine learning models, and incorporating additional data, such as lifestyle or genetic factors. Adding probability estimates for predictions could also enhance its clinical usability by helping prioritize further diagnostic tests. These steps could make the model more reliable and practical for real-world healthcare applications.

Introduction

Diabetes is a serious chronic disease characterized by high levels of glucose in the blood, caused by either insufficient insulin production by the pancreas or the body's inability to effectively use insulin. It has become a significant global health issue, with its prevalence nearly doubling since 1980, and in 2022, 14% of adults aged 18 and older were diagnosed with diabetes, doubling from 7% in 1990 (World Health Organization n.d.). Diabetes can lead to severe complications, including blindness, kidney failure, heart attacks, strokes, and lower limb amputations. Early detection enables timely interventions, reduces complications, lowers healthcare costs, and improves quality of life and long-term outcomes (Marshall and Flyvbjerg 2006).

Artificial intelligence (AI) leverages computer systems and big data to simulate intelligent behavior with minimal human intervention, and within it, machine learning (ML) is a subset of AI methodologies. Since the rise of AI, Machine learning has increasingly been applied in various areas of disease detection and prevention in the healthcare field (Bini 2018). Numerous machine learning techniques have been deployed to develop more efficient and effective methods for diagnosing chronic diseases (Battineni et al. 2020). Utilizing machine learning methods in diabetes research has been proven to be a critical strategy or harnessing large volumes of diabetes-related data to extract valuable insights (Agarwal and Vadiwala 2022). Therefore, The goal of this report is to leverage a supervised machine learning model, logistic regression (LR), to evaluate its predictive performance in diagnosing diabetes using a real-world dataset focused specifically on Pima Indian women aged 21 and older.

Methods and Results

Data

The dataset that was used for the analysis of this project was created by Jack W Smith, JE Everhart, WC Dickson, WC Knowler, RS Johannes and sourced from the National Librabry of Medicine database from the National Institues of Health. Access to their respective analysis can be found here and access to the dataset can be found via kaggle (Dua and Graff 2017). The primary objective of the dataset is to enable diagnostic prediction of whether a patient has diabetes based on specific diagnostic measurements. To ensure consistency and relevance,

several constraints were applied to the selection of data instances. Specifically, the dataset includes only female patients who are at least 21 years old and of Pima Indian heritage.

Each row/obersvation from the dataset is an individual that identifies to be a part of the Pima (also known as The Akimel O'odham) Indeginous group, located mainly in the Central and Southern regions of the United States. Each observation recorded has summary statistics regarding features that include the Age, BMI, Blood Pressure, Number of Pregnancies, as well as The Diabetes Pedigree Function (which is a score that gives an idea about how much correlation is between person with diabetes and their family history). The dataset offers comprehensive features for machine learning analysis.

Analysis

Logistic Regression was employed to develop a classification model for predicting whether the patient is diabetic or not (as indicated in the `outcome` column of the dataset). All variables from the original dataset were used to train the model. The data was split into 70% for the training set and 30% for the testing set. Hyperparameter tuning was performed using `RandomizedSearchCV`, with the accuracy score serving as the classification metric. All variables were standardized just before model fitting. The analysis was conducted using the Python programming language (Van Rossum and Drake 2009) and several Python packages: numpy (Harris et al. 2020), Pandas (McKinney 2010), altair (VanderPlas 2018), altair_ally (Ostblom 2021) and scikit-learn (Pedregosa et al. 2011). The code used for this analysis and report is available at: https://github.com/UBC-MDS/diabetes_predictor_py

We do see 49 observations being dropped post our preliminary data validation as the rows dropped contain meaningless and / or invalid data entries that would unlikely provide useful information but instead introduce noise or spurious relationships into our model. Details of the observations dropped can be found at here under `validation_errors.log`. It is clearly shown from the log that majority of data points that are dropped contained values of 0 where levels cannot possibly be, hence, we suspect that the values recorded at 0 are likely missing values recorded in such way, a critical point we will confirm with data collectors if we have access.

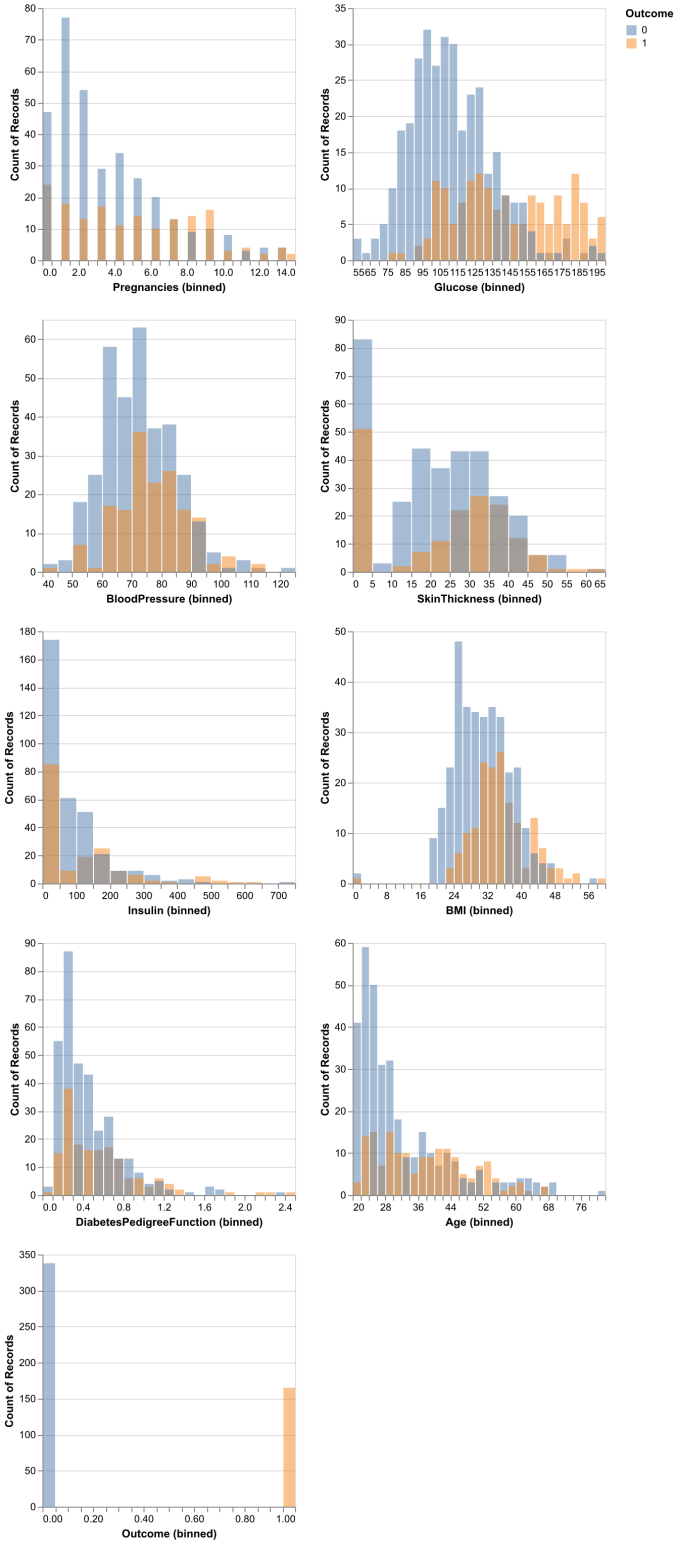

Figure 1: Comparison of the empirical distributions of training data predictors between those non-diabetic and diabetic.

Figure 1 illustrates the distribution of each feature, categorized based on the Outcome variable: 0 (Non-Diabetic) and 1 (Diabetic). This visualization provides insights into the relationships between individual features and the target variable.

For the `Glucose` levels, Non-Diabetic class exhibits a roughly normal distribution, whereas the Diabetic class shows a pronounced shift toward the middle-to-higher range of glucose levels.

The `BMI` distribution resembles a normal distribution but skews slightly toward higher values. Similar to Glucose levels, the Diabetic class displays a decent shit towards the middle-to-higher ranges when compared to Non-Diabetic class, suggesting the potential for distinct differences between target groups within this category.

The `Age` distribution reveals that individuals aged 20 to 32 are predominantly Non-Diabetic. Beyond age 32, the counts of Diabetic and Non-Diabetic individuals become comparable, with some bins showing a higher count for the Diabetic class, despite fewer overall observations in this group. The Non-Diabetic class leans toward younger ages, while the Diabetic class has a more even distribution across its age range.

For `Pregnancies`, `Insulin`, and `DiabetesPedigreeFunction` -genetic risk of diabetes based on family history ranging from 0 to 2.5, the lower range of pregnancies is dominated by the Non-Diabetic class, whereas whereas higher numbers are more common in the Diabetic class.

For `Blood Pressure` and `Skin Thickness`, both the Diabetic and Non-Diabetic classes approximates a normal distribution; however, the Non-Diabetic distribution skews slightly towards lower values, while the Diabetic class skews more towards higher values.

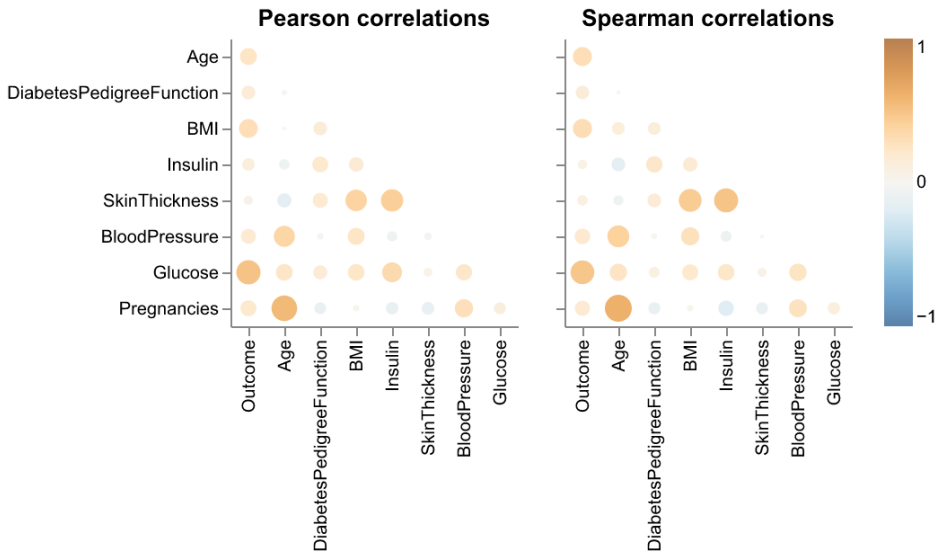

Figure 2: Pearson and Spearman correlations across all features.

Figure 2 shows the correlation between all of the respective features. The main reasoning to analyze this is to see if there is any multicollinearity between any of the features which could be problematic when conducting a Logistic Regression. We see that highest level of correlation is between Age and Pregnancies (0.626 via Spearman, and 0.566 by Pearson). Since this is below the threshold of 0.7, we can conclude that all features' coefficients are suitable and will not cause any multicollinearity in our model.

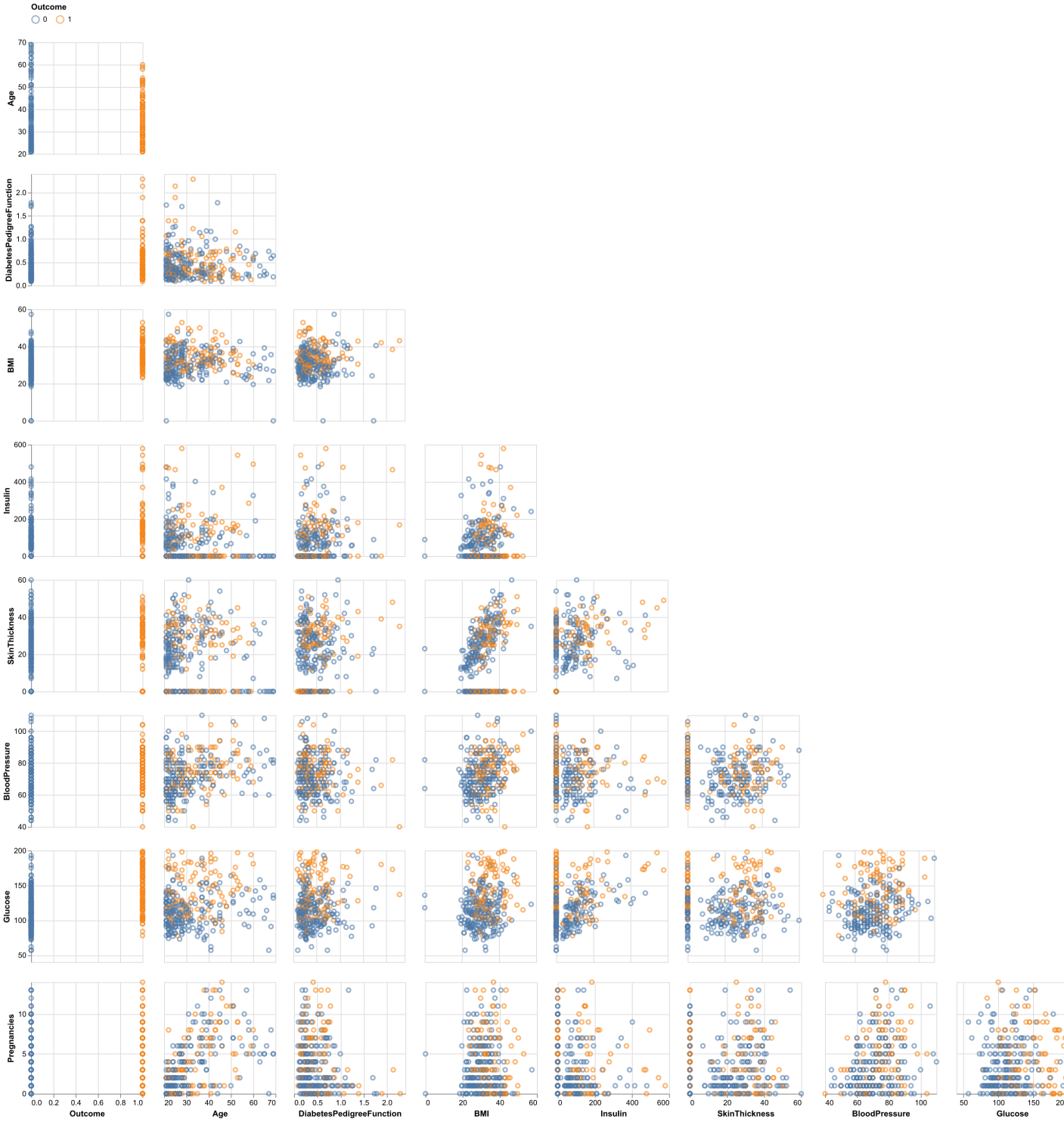

Figure 3: Pairwise scatterplots between each of features in dataset to visualize relationship.

Figure 3 illustrates the relationships between the features. For the most part, the features do not display noticeable trends. However, Skin Thickness and BMI show a moderate visual relationship, which is intuitive since higher body mass is generally associated with increased skin thickness.

Referring back to the correlation graph, Skin Thickness and BMI have a Spearman correlation of 0.446. This value is below the multicollinearity threshold of 0.7, indicating that these features do not pose a risk of multicollinearity in our model.

We then further split our dataset into X and y for both the training and test.

As a result, the Dummy Classifier acts as our baseline for conducting our initial analysis. The Dummy Baseline gives us a score of around 0.67.

We will use a Logistic Regression model for classification. Given the presence of outliers in our features, it is advisable to apply StandardScaler() to normalize the feature values before fitting the model. This ensures that all features are on a similar scale, improving the model's performance and stability. We optimize the hyperparameter `C` for our Logistic Regression model using a random search approach and have identified C = 0.027 as the optimal C to be used in our Logistic Regression model.

Table 1: Logistic regression feature importance measured by coefficients.

	Features	Coefficients
1	Glucose	0.723708
5	BMI	0.3887
0	Pregnancies	0.228731
7	Age	0.194047
6	DiabetesPedigreeFunction	0.160928
2	BloodPressure	0.0480461
4	Insulin	0.00196834
3	SkinThickness	-0.00650682

Having determined the best Logistic Regression model for our analysis, we further explore feature importance with coefficients. Based on the heatmap and Table 1 above, the feature importance coefficients for the logistic regression model predicting diabetes reveal that `Glucose` (0.39) is the strongest positive influence, followed by `BMI` (0.05), `Pregnancies` (0.72), `Age` (-0.01), and `DiabetesPedigreeFunction` (0.0). The negative influence `SkinThickness` (0.19) along with the remaining positive features `BloodPressure` (0.23) and `Insulin` (0.16), have weak impacts on the prediction, with their effects being less pronounced.

We then evaluate the best Logistic Regression model, obtained from the hyperparameter search, on the test set. In addition, to enhance the model's practical use in a clinical setting, we are providing and reporting probability estimates for the predictions of diabetes.

Offering probability estimates would allow clinicians to gauge the model's confidence in its predictions. This would give clinicians the opportunity to conduct additional diagnostic tests if the predicted probability for the outcome (i.e. diagnosis of prediction) is not sufficiently high.

Our prediction model performed decent on test data, with a final overall accuracy of 75%. In addition, looking through the prediction results dataframe, there are a total of 54 mistakes. Of which, 41 mistakes were predicting diabetic as non-diabetic (false negatives) and 13 mistakes were made predicting diabetic as non-diabetic (false positives). Considering implementation in clinic, there is room for improvement in the algorithm as false negatives are more harmful than false positives, and we should aim to lower false positives even further.

Moreover, visualizing prediction probabilities alongside the prediction accuracy for each test sample provides a clearer understanding of the model's performance. This approach allows us to easily assess how well the model predicts, while also highlighting patients who were misdiagnosed. Particularly, it helps us focus on false negatives, as the consequences of these errors are more critical in a clinical context.

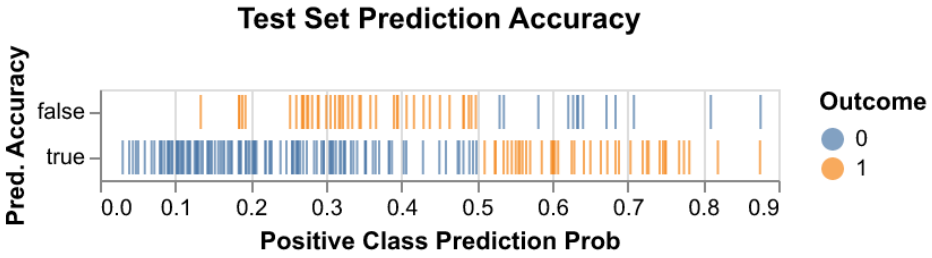

Figure 4: Test Set Prediction Accuracy by Prediction Probability.

Discussion

While the performance of this model may be valuable as a screening tool in a clinical context, especially given its improvements over the baseline, there are several opportunities for further enhancement. One potential approach is to closely examine the 54 misclassified observations, comparing them with correctly classified examples from both classes. The objective would be to identify which features may be contributing to the misclassifications and investigate whether feature engineering could help the model improve its predictions on the observations it is currently struggling with. Additionally, we would try seeing whether we can get improved predictions using other classifiers. Other classifiers we might try are 1) random forest because it automatically allows for feature interaction, 2) k-nearest neighbours (k-NN) which usually provides easily interpretable and decent predictions, and 3) support vector classifier (SVC) as it allows for non-linear prediction using the rbf kernel. Finally, there runs the possibility that the features offered from this dataset alone are not sufficient to predict with high accuracy.

In this case, conducting additional conversations with data collectors for additional useable information or explore additional datasets that can be joined so our set of features can be expanded for more complicated analysis might be beneficial.

At last, we recognize the limitation with this dataset, as it focuses solely on Pima Indian women aged 21 and older, which limits its generalizability to other populations. To improve the analysis, it would be valuable to combine this data with other datasets representing different age groups, genders, and ethnicities, enabling more comprehensive insights and broader applicability of the findings.

Conclusion

In conclusion, this study demonstrated the effectiveness of logistic regression in predicting diabetes among Pima Indian women using diagnostic features such as glucose, BMI, and pregnancies. With an accuracy of 75% on the test set, the model outperformed the baseline Dummy Classifier's 67%. Glucose was identified as the most influential predictor, followed by BMI and pregnancies, while features like blood pressure, insulin, and skin thickness had weaker impacts. However, the model's 54 misclassifications, particularly the 41 false negatives, underscore the need for further refinement to minimize the risk of undiagnosed cases.

These findings highlight logistic regression's potential as an initial screening tool in clinical settings, offering a data-driven approach to early diabetes detection. Nevertheless, improvements are essential to enhance its accuracy and practical utility. Strategies such as feature engineering, alternative machine learning models, and the incorporation of additional data, such as lifestyle or genetic factors, could further optimize performance. Additionally, providing probability estimates for predictions could enhance clinical decision-making by identifying cases requiring further diagnostics. With these refinements, the model could become a valuable tool for reducing complications and improving outcomes in diabetes care.

References

Agarwal, Neetu, and Ronak Vadiwala. 2022. "Machine Learning and Data Mining Methods in Diabetes Research." *Asian Journal of Organic & Medicinal Chemistry*.

Battineni, Gopi, Getu Gamo Sagaro, Nalini Chinatalapudi, and Francesco Amenta. 2020. "Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis." *Journal of Personalized Medicine* 10 (2): 21.

Bini, Stefano A. 2018. "Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?" *The Journal of Arthroplasty* 33 (8): 2358–61.

Dua, Dheeru, and Casey Graff. 2017. "Pima Indians Diabetes Database." https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data.

Harris, Charles R, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62.

Marshall, Sally M, and Allan Flyvbjerg. 2006. "Prevention and Early Detection of Vascular Complications of Diabetes." *Bmj* 333 (7566): 475–80.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, =51–56.

Ostblom, Joakim. 2021. "Altair_ally: Enhancing Altair for Statistical Visualization." https://github.com/jostblom/altair_ally.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *The Journal of Machine Learning Research* 12: 2825–30.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

VanderPlas, Jake. 2018. "Altair: Interactive Statistical Visualizations for Python." *Journal of Open Source Software* 3 (7825, 32): 1057. https://doi.org/10.21105/joss.01057.

World Health Organization. n.d. "Diabetes." https://www.who.int/news-room/fact-sheets/detail/diabetes.