Diabetes Detection

Abhishek Tibrewal (MT23009), Akhil P. Dominic (MT23013), Davinder Singh (MT23031), Vani Mittal (MT23102), Rajith Ramachandran (MT23132)

Problem Statement

Diabetes is a prevalent chronic disease affecting millions in the U.S., leading to serious health complications and significant economic burdens-approximately $400 billion annually. Early diagnosis and predictive modelling of diabetes risk are critical for improving treatment outcomes and reducing healthcare costs. As of 2018, over 34 million Americans have diabetes, with 88 million at risk of developing it, highlighting the urgent need for effective monitoring and intervention strategies.

Therefore, having a machine learning model that could give a few features regarding the person to predict whether the person can contract diabetes or not can be a lifesaver and economically viable. A government can introduce policies and habit-building exercises that will eventually prevent diabetes.

Dataset Used

The dataset is derived from t**he Behavioural Risk Factor Surveillance System (BRFSS)**, an annual CDC survey collecting health-related data from over 400,000 Americans.
The Dataset contains healthcare statistics and lifestyle survey information about people in general, along with their diagnosis of diabetes. The 35 features consist of demographics, lab test results, and answers to survey questions for each patient. The target variable for classification is whether a patient has diabetes, is pre-diabetic, or healthy.

Parameter	Values
Size	70,692 rows and 22 columns
Target Column	Diabetes (Binary, 0 means no, 1 means diabetic or pre-diabetic)
Split Uses	50-50 for Training and 70-30 for Testing

Most of the features in the dataset are categorical. One of the most prominent ones is Smoker (0 if a person hasn't smoked at least 100 packets of cigarettes in the past two months), PhysicalHealth (1 if a person has not contracted any diseases recently, 0 otherwise), Stroke (1 if a person suffers from frequent strokes and 0 otherwise), and many other.

Challenges Faced and their solutions:

1. **Large Size:** The dataset is quite large; it has more than 70k rows, and each row is a tuple of 22 features. During data analysis and visualisation, we faced many challenges; for example, it was very hard to plot heat maps of covariances among different features. Since the model's performance depended upon choosing the right features and feature engineering, it was crucial to visualise the whole dataset.
   Solution: We solved this problem by intuitively checking the columns and reading their descriptions and values, building some hypotheses and running **hypothesis**

testing, selecting a few of them, and finally, using our insights to visualise the plots selectively. This helped us create a new feature and better understand the overall dataset.

2. **Missing Values:** Most of the dataset's columns/features are of a categorical nature, meaning that they only take a fixed number of values (0 or 1, for example). Handling the missing values in these columns is crucial, as assigning values randomly can result in a biased dataset, resulting in a biased model performing poorly on the test dataset. **Solution:** we replaced the missing values with their median. Why median? since the median, unlike the mean, is not sensitive to the outliers and scales well.
3. **Reduction of Dataset:** Since the dataset is large, the models were overfitting and took too long to train. Classical Machine Learning models are based on statistics and hence don't scale well to extremely large datasets; therefore, we needed to reduce the data's size while keeping enough contextual information to learn the statistical information and scale well. **Solution:** Random Sampling randomly selected 70 per cent of the rows out of data; this ensured that the dataset's size decreased and that enough statistical properties remained for the model to learn from.

Hypothesis Tests and Conclusions:

1. **Test for Association Between Diabetes and Smoking:** The first hypothesis testing that we performed was whether there exists some significant relationship between Diabetes and Smoking. The rationale behind this test is that Smoking consists of many chemicals, and some research supports this relation (especially this link to the article by CDC, USA). We performed a *Chi-Square test* for independence, with degree n-1 for n = 100. $\mathbf{H}_0$: *"Diabetes and Smoking status are independent",* and the alternative hypothesis is $\mathbf{H}_1$: *"Diabetes and Smoking status are not independent".* **Conclusion:** A p-value less than 0.05 indicates that smoking status has a statistically significant association with diabetes, meaning we failed to reject the **null hypothesis**.
2. **Test for Physical Activity Impact on Smoking Status:** The physical activity column is again a categorical feature, which is 1 if a person has done some physical activity in the past week; otherwise, it's 0. For this, we used the **T-test.** Since this test tests for the independence of the population's mean, the test is **two-tails** with a degree of n-1, n = 100. $\mathbf{H}_0$: *"There is no significant difference between the general health between physical activity and smoking status",* and the alternative hypothesis is $\mathbf{H}_1$: *"There is a significant difference between the general health between physical activity and smoking status".* **Conclusion:** A p-value less than 0.05 indicates that Physical Activity status has a measurable impact on Smoking status. We failed to reject the **null hypothesis**.
3. **Test for Association between Heavy Alcohol Consumption and Smoking:** This test's rationale is the presence of addiction in general. It is often seen that those who are addicted to alcohol followed by smoking. We performed the Chi-Square test with the same degree n-1, n = 100. $\mathbf{H}_0$: *"Heavy Alcohol Consumption and Smoking status are independent",* and the alternative hypothesis is $\mathbf{H}_1$: *"Heavy Alcohol Consumption and Smoking status are not independent".* **Conclusion:** A p-value less than 0.05 indicates that smoking status has a statistically significant association with heavy alcohol, meaning we failed to reject the **null hypothesis**.

Comparing the performance of models on Scaled and un-scaled datasets:

We tested many models, including Logistic Regression, Decision Tree, SVM, and ensembles Random Forest and XGBoost. First, we trained the model on the whole dataset, then selected the best-performing models on the un-scaled dataset and trained them on the scaled dataset (Logistic Regression, Random Forest, and XGBoost). The scaling techniques that we tried were PCA, Random Sampling and CUR. The table before shows the difference between

Accuracy:

Scaled/Unsclaed	Logistic Regression(%)	XGBoost(%)	Random Forest(%)
Scaled	1. Train Set: 89.5 2. Test Set: 88.3 3. Running Time: 0.15s	1. Train Set: 94.7 2. Test Set: 93.8 3. Running Time: 1.05s	1. Train Set: 92.3 2. Test Set: 91.2 3. Running Time: 0.72s
Unscaled	1. Train Set: 68.45 2. Test Set: 68.38 3. Running Time: 0.10s	1. Train Set: 69.57 2. Test Set: 68.8 3. Running Time: 0.85s	1. Train Set: 70 2. Test Set: 68.9 3. Running Time: 0.50s

From the above statistics, we can see some drop in the scaled dataset-trained model's performance, but we saved considerably on time. Therefore, we need to find a middle ground where time is saved while the model performance is also good.

Conclusion:

In conclusion, our project shows the potential of using machine learning models for diabetes risk prediction using a large feature-rich dataset. We overcame some problems: dealing with missing values, managing large datasets, and optimising feature selection to develop robust models that balance accuracy and efficiency. Hypothesis testing gives valuable insights into how diabetes is related to most lifestyle factors, highlighting targeted interventions. Our findings highlight the importance of integrating machine learning into public health strategies, which allows for early detection and proactive measures to combat diabetes effectively.

Future Scope:

This project has a lot of potential for further development. Incorporating real-time data from wearable devices and electronic health records could facilitate continuous monitoring and dynamic diabetes risk assessment. Adding genetic data to lifestyle and demographic features can improve predictive accuracy further. Another extension is that the dataset would be broadened to incorporate global populations and diverse demographics, enhancing generalisability. Collaboration with healthcare providers will help validate and implement the model in real-world settings, filling the gap between research and practical healthcare applications and advancing public health strategies.

Citations:

1. Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev Chronic Dis 2019;16:190109. DOI: http://dx.doi.org/10.5888/pcd16.190109 (this includes the dataset also).

2. **GitHub Link**: https://github.com/davinder23031/StatsGroup