Diabetes
Exploratory Data Analysis (EDA)

Prepared by

[Team 10 – Mellitus]

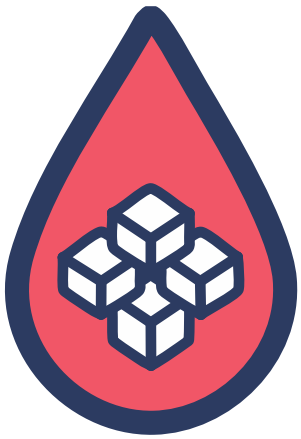

Table of Contents

1 Overview

1.1 Purpose

With the increasing awareness of diabetes, this project aims to analyse the relationship between diabetes and different factors to support our client, a drug manufacturer, to launch a novel product segment for diabetes. The project has been divided into 2 phases, with phase 1 analysing the internal dataset from the company database in the drug manufacturing industry and phase 2 combining the external dataset to deliver a comprehensive outcome.

1.2 Dataset Overview

In phase 1 of the research, the initial dataset received from the company database has the listed 16 attributes, which are categorised into numerical, categorical, and binary attributes (please refer to the attached table). Binary attributes can be considered as categorical or numerical depending on the analysis.

Table 1: Dataset Overview

#	Attribute	Data Type	Attribute Type	Non-Null Count	Null Count	Unique Count
1	gender	object	Categorical	79954	20046	3
2	age	float64 (10, 0)	Numerical	80145	19855	75
3	hypertension	float64 (10,0)	Binary	80169	19831	3
4	diabetes_pedigree_function	float64 (10,2)	Numerical	80120	19880	62
5	diet_type	object	Categorical	79939	20061	13
6	star_sign	object	Categorical	79806	20194	13
7	BMI	float64 (10,1)	Numerical	79934	20066	447
8	weight	float64 (10,1)	Numerical	80126	19874	2002
9	family_diabetes_history	float64 (10,0)	Binary	79863	20137	3
10	social_media_usage	object	Categorical	79968	20032	5
11	physical_activity_level	object	Categorical	80032	19968	6
12	sleep_duration	float64 (10,1)	Numerical	80063	19937	103
13	stress_level	object	Categorical	80024	19976	5
14	pregnancies	float64 (10,0)	Numerical	80033	19967	7
15	alcohol_consumption	object	Categorical	79896	20104	5
16	diabetes	float64 (10,0)	Binary	80242	19758	3

- Dataset shape & volume:

The original dataset has 10000 rows x 16 columns; the data volume is 7.5 MB.

- Missing values:

The attached dataset information table also highlights the missing values in each attribute in the original dataset.

- Data sensitivity:

The data provided is health data which is sensitive information. Whilst there is no direct PII data, there are indirect PII variables. A person can be indirectly identified from that information in combination with other information. As such, GDPR policies will apply in handling the dataset.

2 Types of Analysis

This research has gone through 3 types of analyses, including univariate analysis, bivariate analysis, and multivariate analysis to interpret the data from three dimensions:

- The distribution and characteristics of a single variable
- The relationship between two variables.
- The relationship between multiple (2+) variables.

3 Objective

The primary objective of this project is to analyse the internal dataset provided by the drug company to gain meaningful insights into the various factors influencing diabetes. This research aims to explore the relationships between key attributes among others, to better understand their impact on diabetes prevalence and progression. By identifying patterns, correlations, and trends within the dataset, the study seeks to uncover actionable insights that can inform targeted interventions, preventive measures, and treatment strategies. Ultimately, the findings will provide valuable data-driven recommendations to support the drug company in developing more effective solutions for diabetes management and prevention.

2 Univariate Analysis

2.1 Categorical Variables

a) Distribution Statistics

Attributes and unique counts: (please check the full list of categorical attributes in Appendix – Table 1)

- **gender**: 2 - Male, Female
- **diet_type**: 12 - Pescatarian, Atkins, Vegetarian, Mediterranean, Raw Food, Paleo, Ketogenic, Gluten Free, Weight Watchers, Carnivore, Vegan, Low Carb
- **star_sign**: 12 – Cancee, Sagittarius, Scorpio, Virgo, Aries, Aquarius, Gemini, Libra, Taurus, Capricorn, Pisces, Leo
- **social_media_usage**: 4 - Never, Occasionally, Moderate, Excessive
- **physical_activity_level**: 5 - Sedentary, Lightly Active, Moderately Active, Very Active, Extremely Active
- **stress_level**:4 - Low, Moderate, Elevated, Extreme
- **alcohol_consumption**: 4 - None, Light, Moderate, Heavy

b) Distribution Visualisation

The below charts show the following:

1. Hypertension: Most of the population (64%) does not have hypertension, while 16.2% are hypertensive, and 19.8% of the data is missing.
2. Family Diabetes History: A significant portion (55.7%) has no family history of diabetes, 24.1% have a family history, and 20.1% of the data is missing.
3. Diabetes: Most of the population (76.6%) has diabetes, while only 3.6% are non-diabetic, with 19.8% of the data missing.

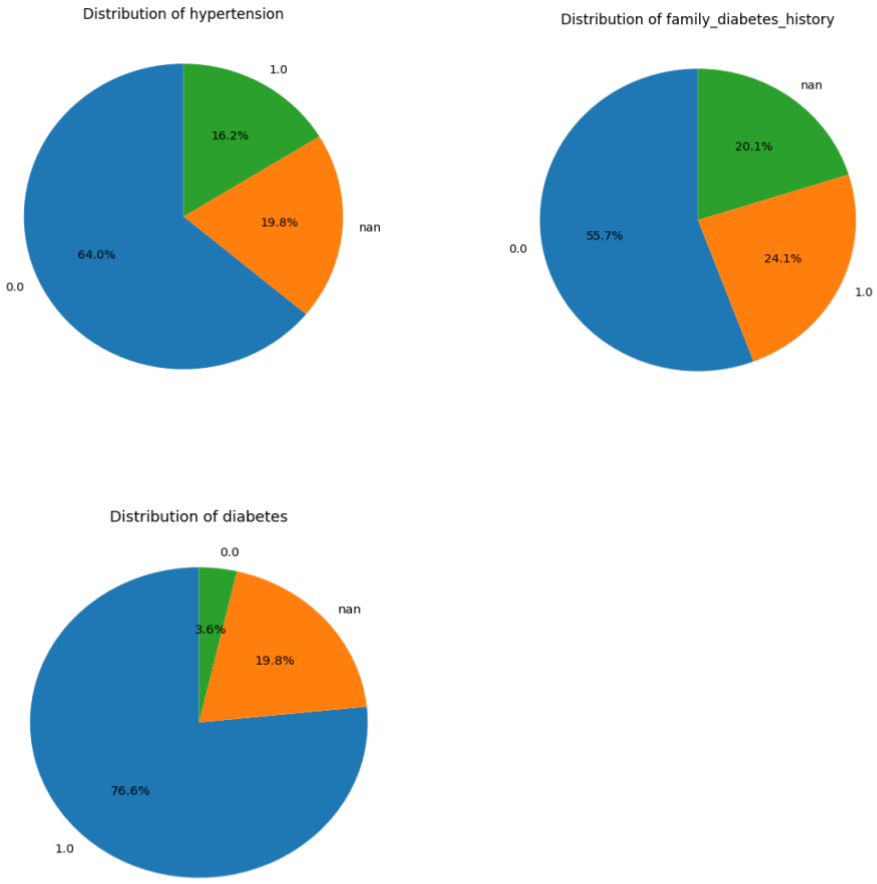

Figure 1: Set of bar charts for distribution of diabetes, family_diabetes_history, and hypertension

c) Key Insights

- Most of the population is **male**.
- Diet type, star sign, stress level, and social media usage are evenly distributed
- A **sedentary physical activity level** is predominant.
- **Heavy alcohol consumption** is the most prevalent behaviour.
- Approximately a quarter of population has **family diabetes history**
- Most of the population has **Diabetes**

These insights suggest potential areas for targeted health interventions or deeper analysis to identify correlations between these characteristics and health outcomes.

2.2 Numerical Variables

a) Distribution Statistics

Table 2: Distribution of numerical variables

Attributes	Count	Mean	StdDev	Min	25%	Median	75%	Max	Mode
age	80145	45.10731	18.55043	18	27	45	60	91	18.0
diabetes_pedigree_function	80120	0.500877	0.173783	0.2	0.35	0.5	0.65	0.8	0.77
BMI	79934	26.97854	6.005039	1.8	22.9	27	31	53.1	26.8
weight	80126	150.5266	57.73154	50	100.3	150.9	200.4	250	231.1
sleep_duration	80063	5.295149	2.842133	0	3.3	5.3	7	12	4.5
pregnancies	80033	0.758212	1.281326	0	0	0	1	5	0.0

b) Distribution Visualization

The histogram below shows that BMI is generally higher among individuals with diabetes (red), peaking around 30, while the non-diabetic population (blue) is more evenly distributed around a slightly lower BMI.

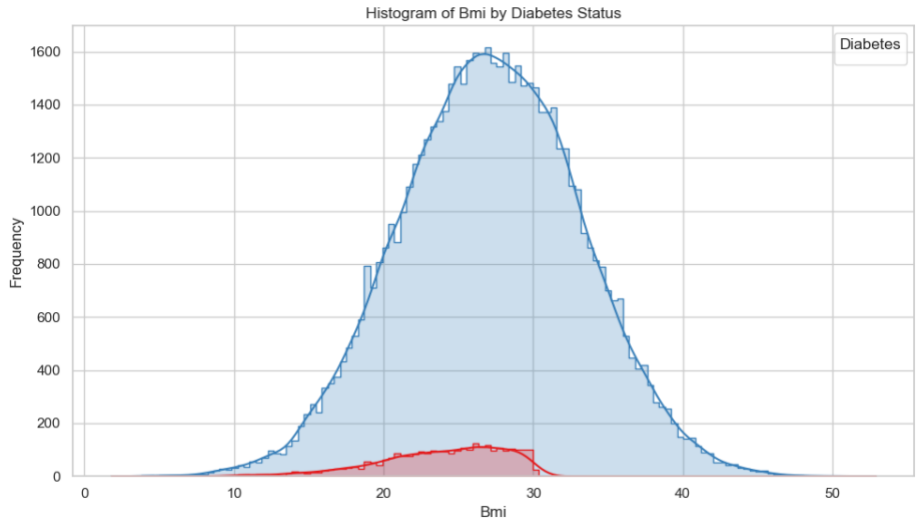

Figure 2: BMI histogram distribution

c) Key Insights

- The mode of age is 18, suggesting a significant proportion of younger individuals in the dataset.
- The average value of diabetes pedigree function is 0.50, indicating moderate genetic predisposition to diabetes, with a most common value (mode) of 0.77.
- The mean BMI is 26.98, suggesting a tendency toward being overweight according to health standards.
- The average weight is 150.5 lbs, with a wide spread from 50 to 250 lbs. A notable mode of 231.1 lbs suggests that some individuals are significantly above average weight.

- The mean sleep duration is 5.3 hours, which is below the recommended 7–8 hours for healthy adults. The mode is 4.5 hours, indicating that insufficient sleep is common.
- Most individuals (mode = 0) have not had any pregnancies, with an average slightly below 1 (0.76). The maximum value of 5 highlights a smaller subset with multiple pregnancies.

2.3 Binary Variables

Treat binary variables as numerical variables for this exercise.

a) Distribution Statistics

Table 3: Binary variables statistics

Attributes	Count	Mean	StdDev	Min	25%	Median	75%	Max	Mode
hypertension	80169	0.202248	0.401678	0	0	0	0	1	0.0
family_diabetes_history	79863	0.302167	0.4592	0	0	0	1	1	0.0
diabetes	80242	0.954936	0.207445	0	1	1	1	1	1.0

b) Distribution Visualization

This bar chart shows that heavy alcohol consumption is the most common across both diabetic and non-diabetic groups, though the proportion of diabetics is consistently smaller across all alcohol consumption levels.

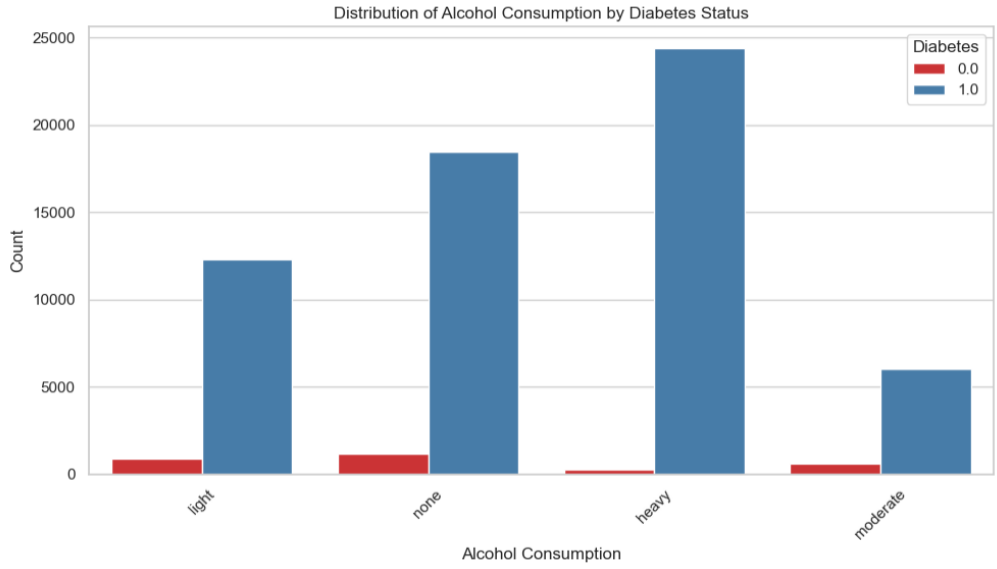

Figure 3: Alcohol consumption by diabetes status

c) Key Insights

- Hypertension, with 20.2% of the population in the sample, is relatively uncommon in this dataset but remains an important factor for the minority affected.
- While not the majority, 30.2% of the population has a family history of diabetes, which may contribute to an increased risk for the condition.

3 Bivariate Analysis

3.1 Numerical vs. Numerical

a) Variable Relationships

The violin plot below of BMI by diabetes status shows that diabetic individuals tend to have a higher median BMI and greater variability compared to non-diabetic individuals. The distribution suggests a stronger association between higher BMI and the likelihood of diabetes in this dataset.

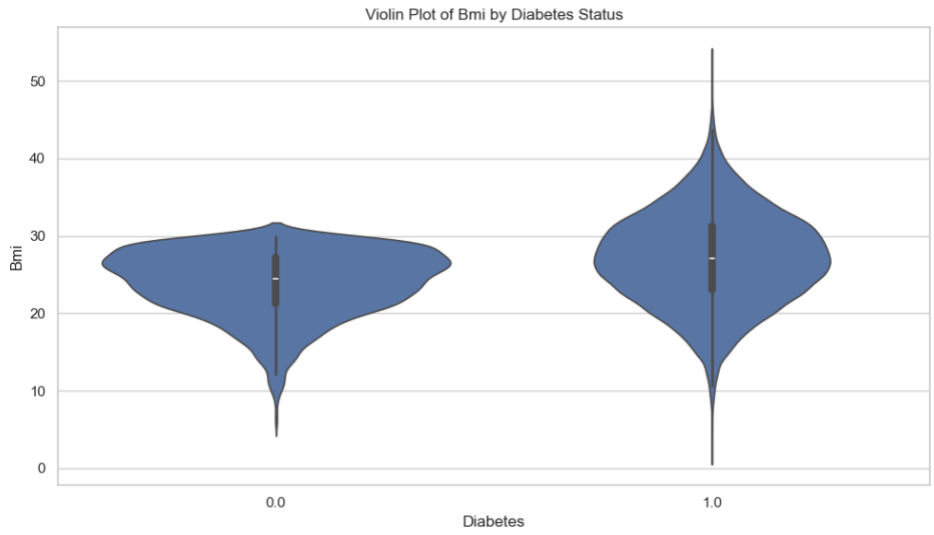

Figure 4: BMI violin plot

b) Key Insights

• Correlation: Higher BMI is linked to a higher likelihood of diabetes, due to increased insulin resistance.

• Prevalence: Diabetes is more common in individuals with higher BMI, especially in overweight and obese categories.

3.2 Numerical vs. Categorical

a) Variable Relationships

This strip plot shows the age distribution by diabetes status, with both diabetic and non-diabetic individuals spanning a wide age range. There is no clear age clustering, suggesting that diabetes affects individuals across all age groups in this dataset.

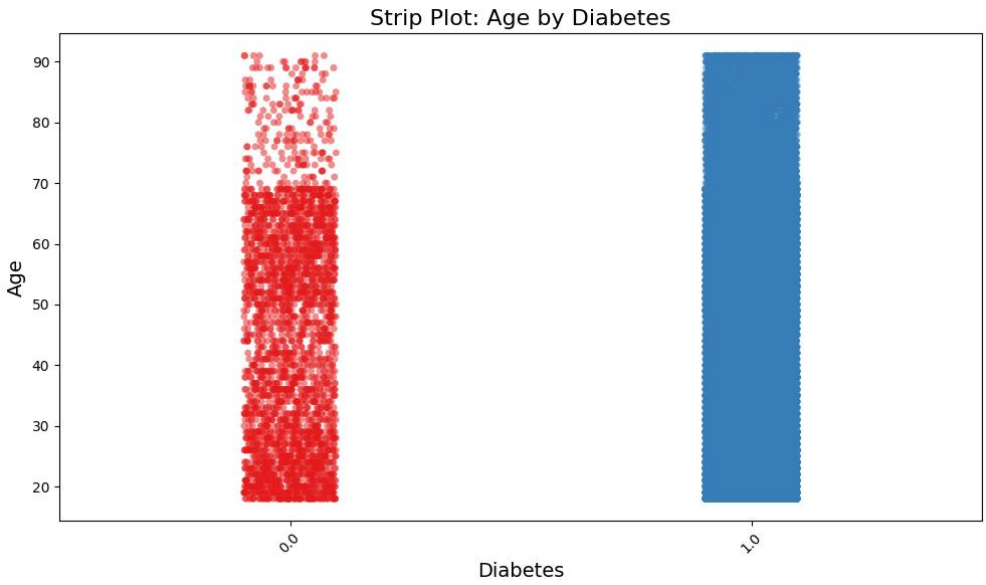

Figure 5: Age strip plot

The box plot below shows that the weight distributions across different diet types are relatively similar, with overlapping medians and ranges. This indicates that diet type alone does not significantly differentiate weight in this dataset.

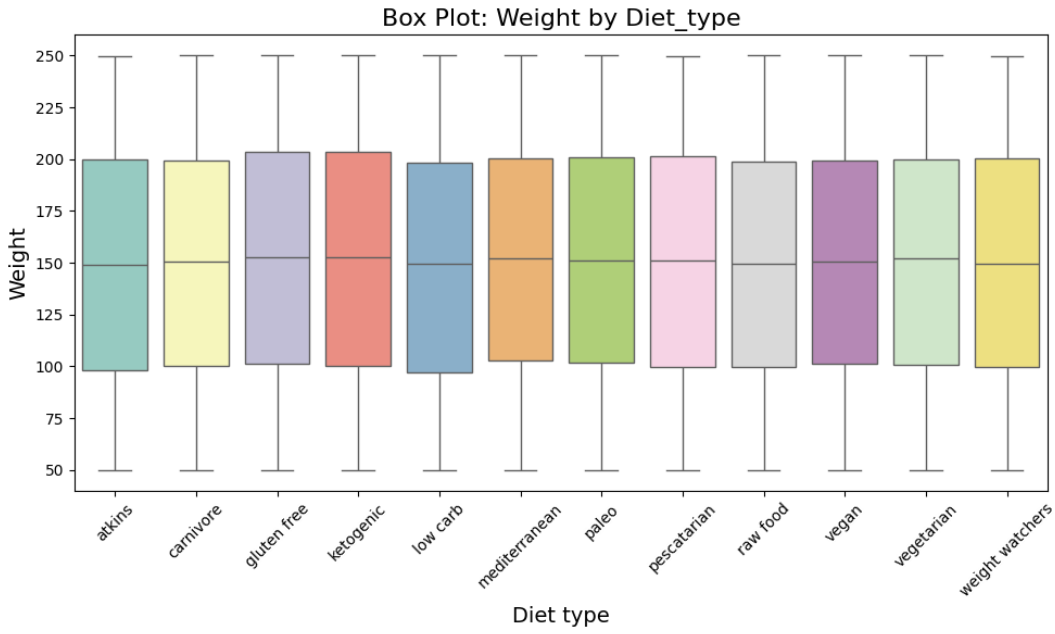

Figure 6: Weight and Diet type box plot

a) Key Insights

- Trend: Certain diet types (e.g., pescatarian, vegetarian) may correlate with lower BMI, while others (e.g., omnivore, heavy meat-based diets) might be associated with higher BMI.
- Variation: BMI values can vary significantly across different diet types, indicating potential dietary influences on weight management.

3.3 Categorical vs. Categorical

a) Variable Relationships

The violin plot below shows the distribution of age by gender, with both males and females having similar age ranges but slightly different density patterns. Males exhibit a more uniform age distribution, while females show higher density in the younger and middle age ranges.

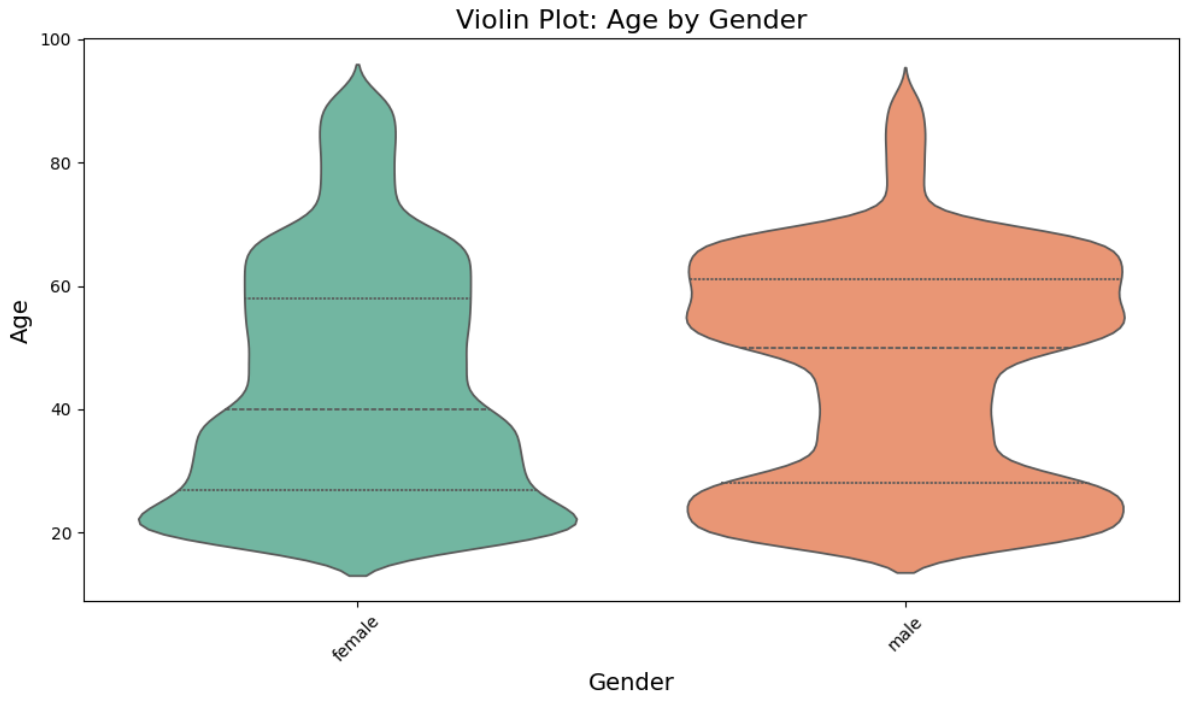

Figure 7: Violin Plot by gender and age

The stacked bar plot below shows that heavy alcohol consumption is consistently high across all stress levels, with elevated and extreme stress having the largest counts. Additionally, individuals with no alcohol consumption remain a significant group, indicating varied drinking behaviors across different stress categories.

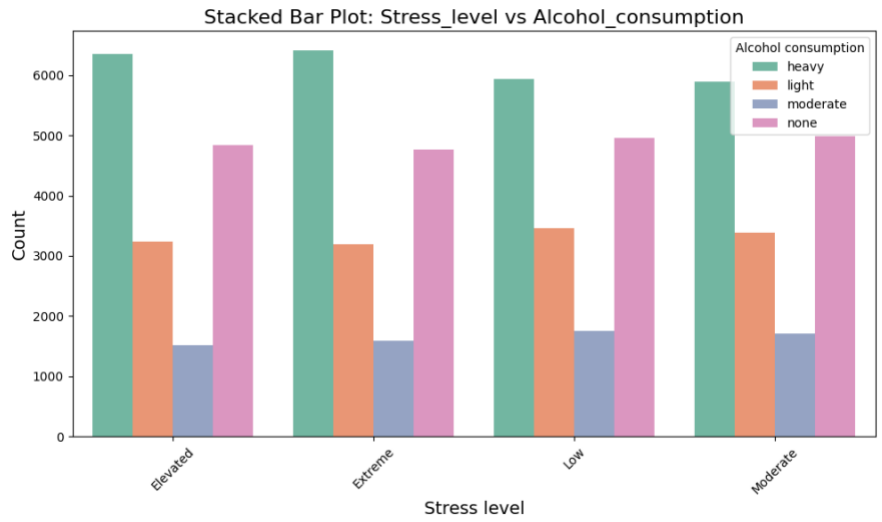

Figure 8: Stacked bar chart for stress and alcohol consumption

The heatmap below shows the distribution of individuals across combinations of stress levels and alcohol consumption, revealing that heavy alcohol consumption is most common across all stress levels, particularly among those with elevated or extreme stress. Additionally, individuals with no alcohol consumption also display high counts, suggesting diverse patterns in how stress and alcohol behaviours coexist.

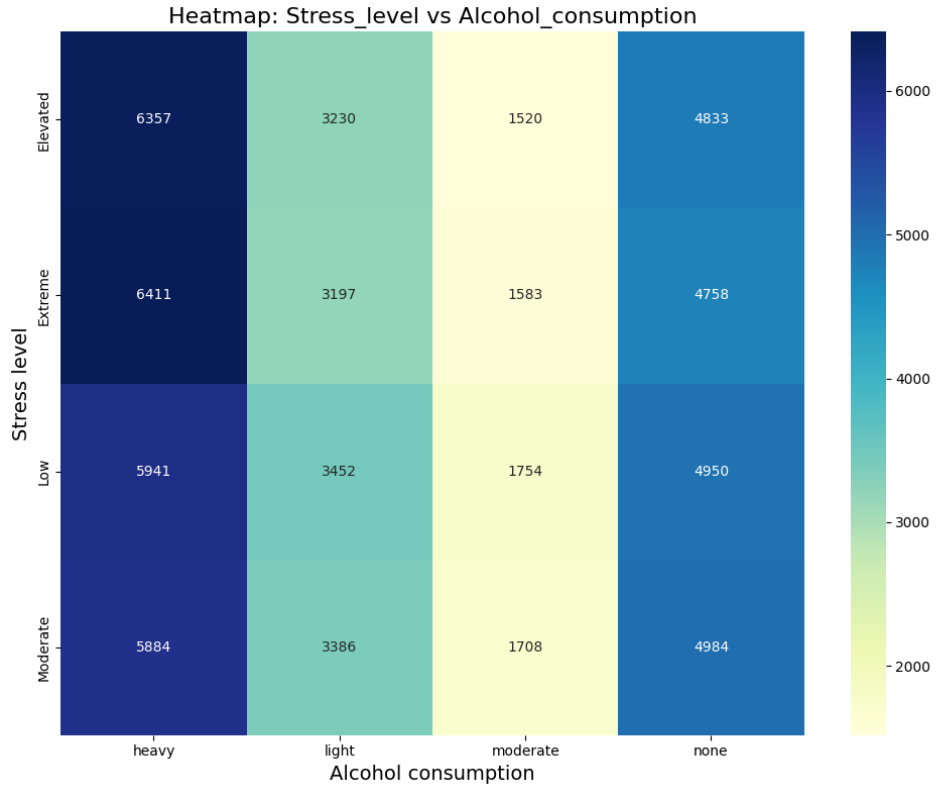

Figure 9: Heatmap showing stress level and alcohol consumption

The chart below shows that while the overall population is predominantly male, the proportion of diabetes cases is relatively similar between males and females, indicating diabetes prevalence is not strongly gender-dependent in this dataset.

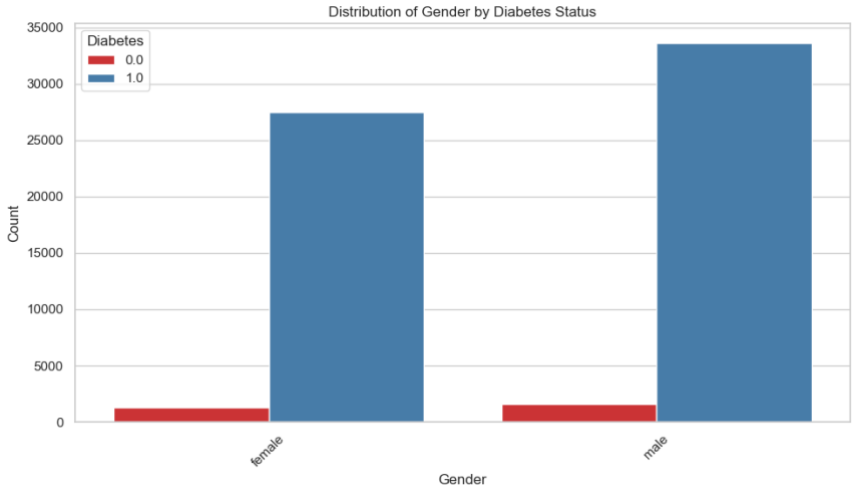

Figure 10: Gender and diabetes bar chart

b) Key Insights

- Imbalanced gender and diabetes prevalence
- Low prevalence of Non-diabetes
- Imbalanced distribution between the number of male and female individuals in the dataset

4 Multivariate Analysis

a) Variable Relationships

The pair plot below shows relationships among age, BMI, weight, and sleep duration, with diabetes status highlighted. Diabetic individuals (orange) dominate across all features, while non-diabetic individuals (blue) cluster within specific ranges, particularly at lower BMI and weight values.

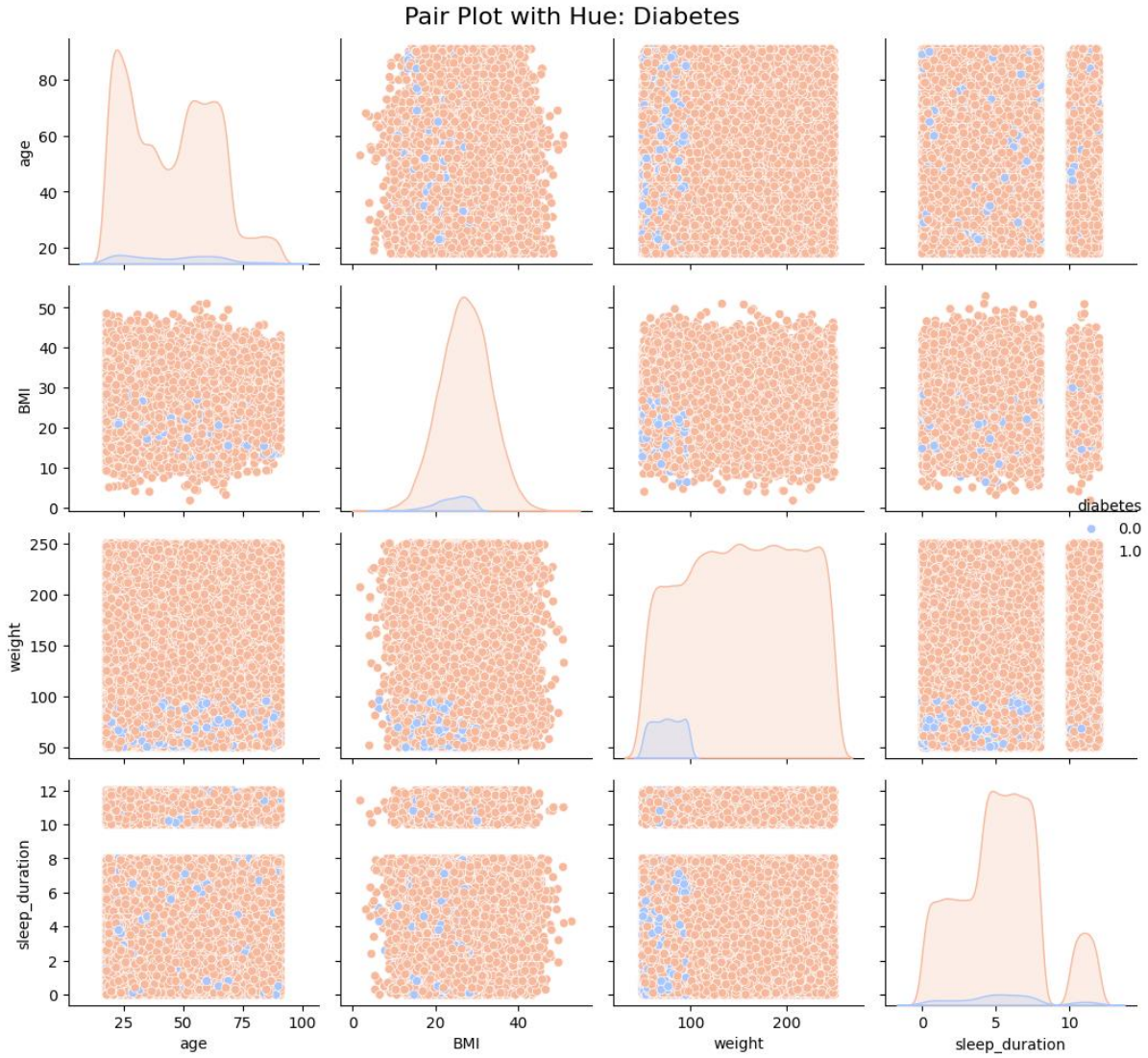

Figure 11: Pair plot

The below facet grid of BMI versus age by diabetes status shows that non-diabetic individuals (blue) predominantly maintain a BMI below 30, regardless of age. In contrast, diabetic individuals (orange) are more evenly distributed across all BMI levels, suggesting a stronger association between higher BMI and diabetes prevalence across age groups.

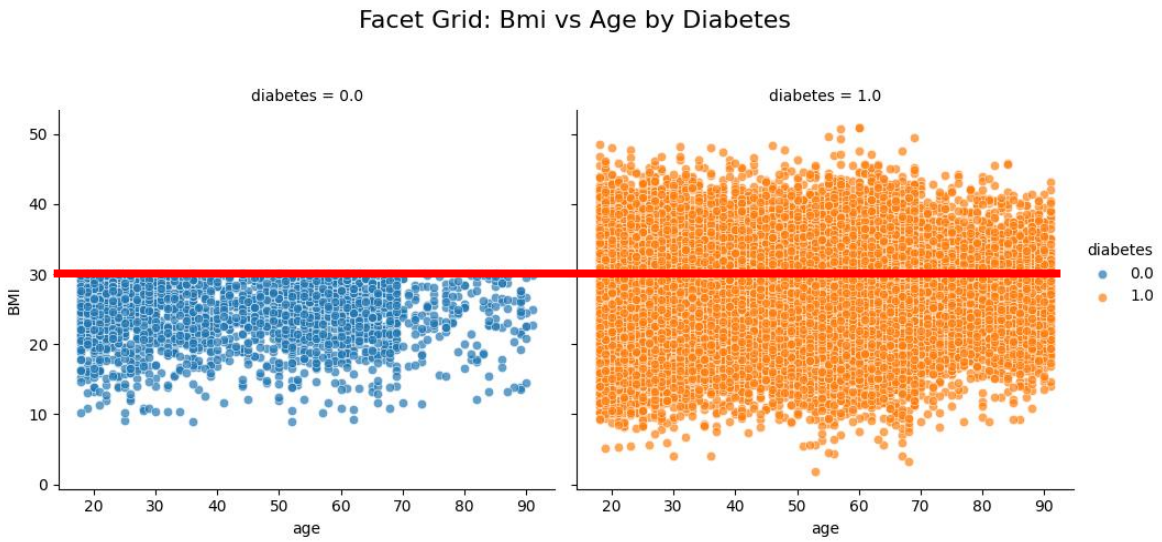

Figure 12: Fact grid showing BMI, Age, and Diabetes

The below scatter plot of weight versus BMI reveals a distinct cluster where most non-diabetic individuals are concentrated, defined by BMI less than 30 and weight less than 100. However, within this same range, there is a notable presence of diabetic cases (red points), suggesting unique risk factors or underlying conditions contributing to diabetes in this subgroup.

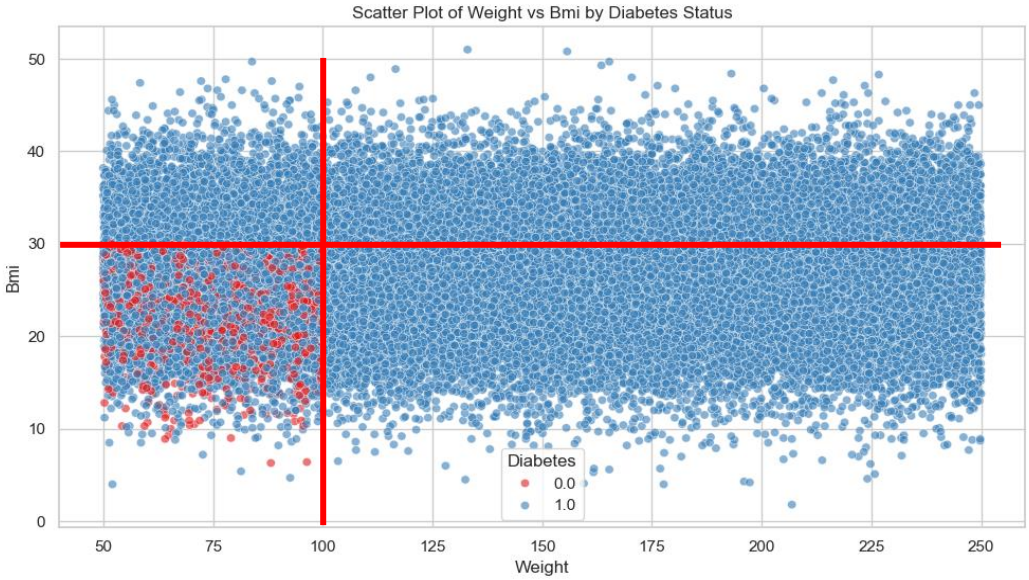

Figure 13: Scatter graph for weight and bmi against diabetes

b) Key Insights

Higher BMI appears to correlate with a higher likelihood of diabetes, with most diabetic individuals having BMI values above 20. This highlights the importance of weight management and BMI in diabetes prevention and management.

5 Advanced Analytics

In this section, we delve into the advanced analytics performed on the preprocessed diabetes dataset. The goal is to build predictive models to understand the factors influencing diabetes and identify the most significant features contributing to the disease. We will cover model preparation, training, tuning, and feature importance analysis using multiple machine learning algorithms.

5.1 Model Preparation

a. Understanding Dataset

The dataset has been thoroughly preprocessed and analysed in previous steps, ensuring data quality and readiness for modelling. It includes encoded and scaled features derived from the original variables, making it suitable for machine learning algorithms that require numerical inputs. Additionally, initial insights have been collated from above variable analysis.

b. Data Splitting

To evaluate the models effectively, the dataset was split into training and testing sets in an 80:20 ratio while maintaining the class distribution (stratification). This approach ensures that both sets are representative of the overall data, allowing for reliable model performance assessment.

c. Feature Selection

All encoded and scaled variables were selected as features for modeling. This selection focuses on variables transformed to have zero mean and unit variance, which is essential for algorithms sensitive to feature scaling.

Selected Feature
age_scaled
diabetes_pedigree_function_scaled
BMI_scaled
weight_scaled
sleep_duration_scaled

pregnancies_scaled
gender_encoded_scaled
diet_type_encoded_scaled
physical_activity_level_encoded_scaled
alcohol_consumption_encoded_scaled
star_sign_encoded_scaled
social_media_usage_encoded_scaled
stress_level_encoded_scaled
family_diabetes_history_encoded_scaled
hypertension_encoded_scaled

d. Class Imbalance

The dataset contains an imbalance, with more individuals diagnosed with diabetes than those without. To address this, Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data to balance the classes. This method generates synthetic examples of the minority class, improving the model's ability to learn from underrepresented cases.

e. Model Training and Tuning

Multiple machine learning algorithms were employed to build predictive models. These include:

- Random Forest Classifier
- Extreme Gradient Boosting (XGBoost) Classifier
- Logistic Regression

Hyperparameter tuning was conducted using RandomizedSearchCV to optimize each model's performance.

5.2 Feature Importance

The below table shows the outcome of the direct model response.

Table 4: ML Model response ranking

	Feature Importance - ML Model Ranking				
Model	1	2	3	4	5
Random Forest	weight	stress	alcohol	physical activity	bmi
XGBoost	family_diabetes	weight	alcohol	physical activity	stress
Logistic	family_diabetes	weight	stress	bmi	alcohol

The following are the common attributes that form the top 5 across all models:

family_diabetes	weight	stress	alcohol	physical activity	bmi

The bar chart below displays the Pearson correlation coefficients between features and diabetes, showing positive correlations with weight, family diabetes history, and physical activity, while stress level and alcohol consumption are negatively correlated.

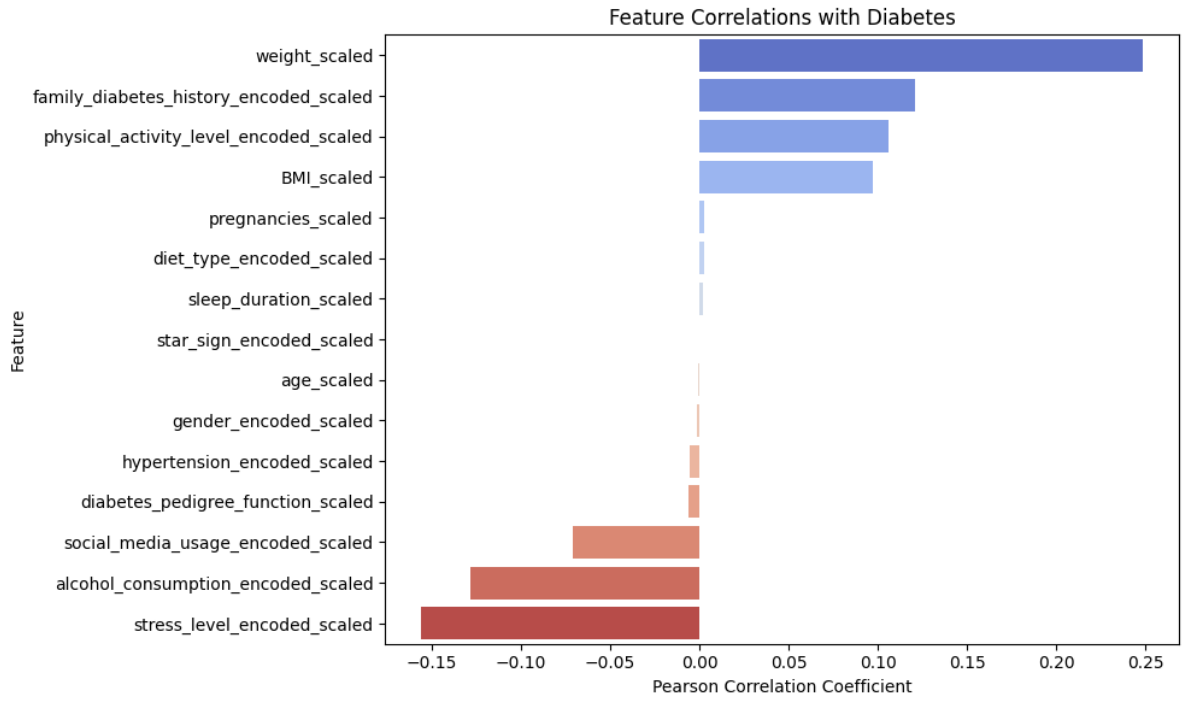

Figure 14: Pearson Correlation of Diabetes features

5.3 Model Evaluation

Performance Evaluation Metrics

Table 5: Evaluation Metrics

Model	Accuracy	ROC AUC	Precision	Recall	F1-Score
Random Forest	0.984	0.977	0.990	0.993	0.991
XGBoost	0.979	0.977	0.991	0.987	0.989
Logistic Regression	0.871	0.944	0.993	0.871	0.928

1. Accuracy: The proportion of correctly classified instances out of all instances.
2. ROC AUC (Receiver Operating Characteristic Area Under the Curve): Measures the model's ability to distinguish between classes. A higher value indicates better discriminative ability.
3. Precision (Positive Predictive Value): The proportion of true positive predictions out of all positive predictions made.
4. Recall (Sensitivity or True Positive Rate): The proportion of true positive predictions out of all actual positive instances.
5. F1-Score: The harmonic mean of precision and recall, providing a balance between the two.

Cross Fold Evaluation

Cross-validation was performed to assess the robustness of the models across different subsets of data.

Table 6: Cross Fold Evaluation

Model	Average_ROC_AUC	Std_Dev	Min_ROC_AUC	Max_ROC_AUC
Random Forest	0.9801	0.0042	0.9746	0.9838
XGBoost	0.9812	0.0024	0.9778	0.9839
Logistic Regression	0.9463	0.0032	0.9441	0.9517

Average_ROC_AUC: The mean ROC AUC score across all 5 folds, providing an overall performance metric.

Std_Dev: The standard deviation of the ROC AUC scores, indicating the variability or consistency of the model's performance across different folds.

Min_ROC_AUC: The lowest ROC AUC score achieved in any of the 5 folds.

Max_ROC_AUC: The highest ROC AUC score achieved in any of the 5 folds.

5.4 Visualisations

Confusion matrices, ROC curves, and precision-recall curves were generated to visualise performance.

5.4.1 Confusion Matrix

Logistic Regression model struggles with true negatives (749) and has high false negatives (2468), indicating it's not very good at detecting the minority class (0).

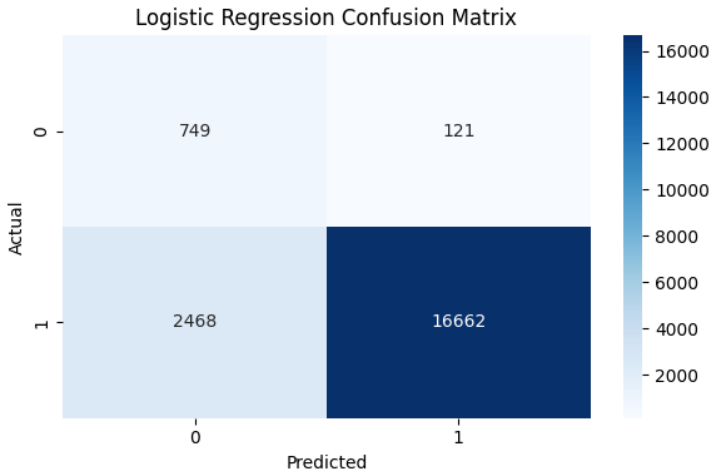

Figure 15: Confused matrix for logistic regression model

Random Forest model performs significantly better with true positives (18990) and lower false negatives (140), showcasing strong predictive power overall but slightly weaker at capturing true negatives (681).

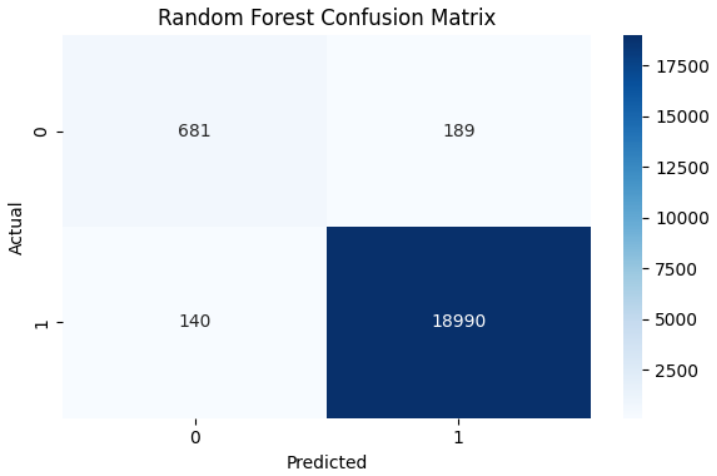

Figure 16: Confused matrix for Random Forest model

Similar to Random Forest, it demonstrates high accuracy for true positives (18890) with slightly higher false negatives (240) and moderate performance in true negatives (693). It offers a balanced trade-off between sensitivity and specificity.

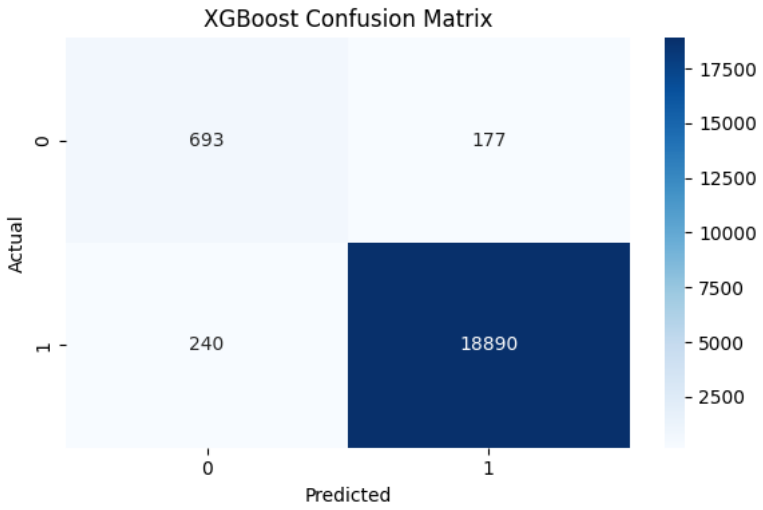

Figure 17: Confused matrix for XGBoost model

5.4.2 Precision-Recall Graph

All three models (Logistic Regression, Random Forest, XGBoost) demonstrate high precision across varying recall levels, with nearly perfect area under the precision-recall curve (AP ≈ 1.00), indicating excellent performance in balancing false positives and false negatives.

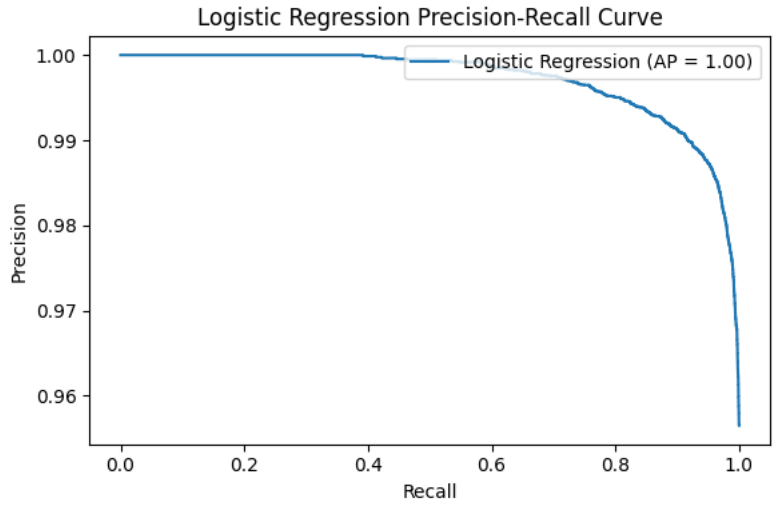

Figure 18: Logisitic Precision Recall

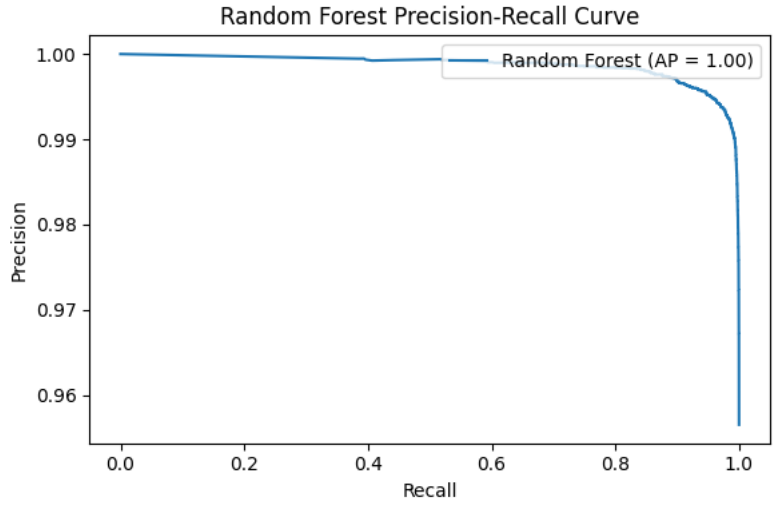

Figure 19: Random Forest Precision Recall

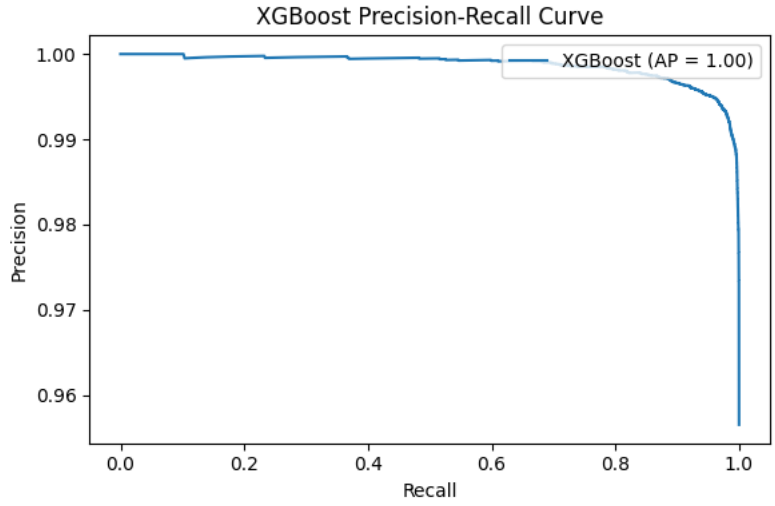

Figure 20: XGBoost Precision Recall

5.4.3 Receiver Operating Characteristic (ROC) Curve

The ROC curves show that Logistic Regression (AUC = 0.94) has slightly lower discriminatory ability compared to Random Forest and XGBoost (both AUC = 0.98), with the latter two models demonstrating near-perfect classification performance.

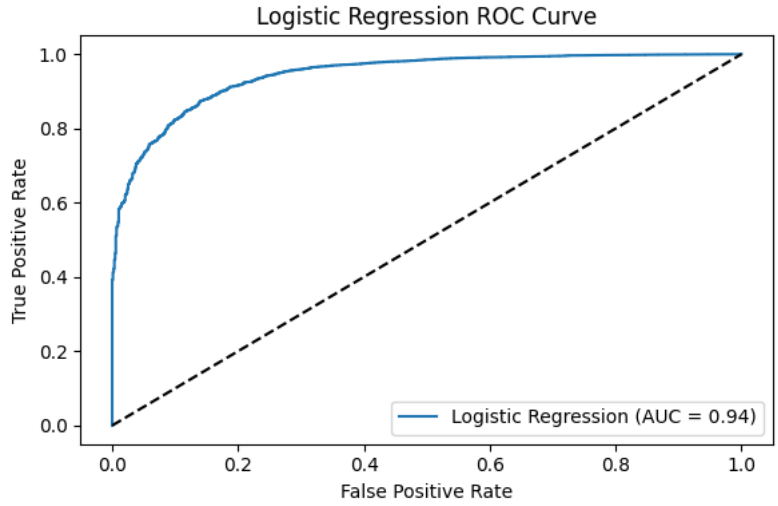

Figure 21: Logistic ROC Curve

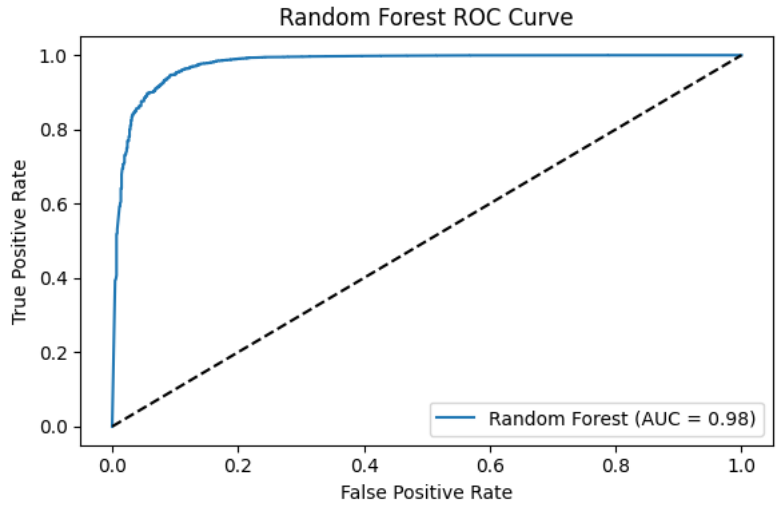

Figure 22: Random Forest ROC Curve

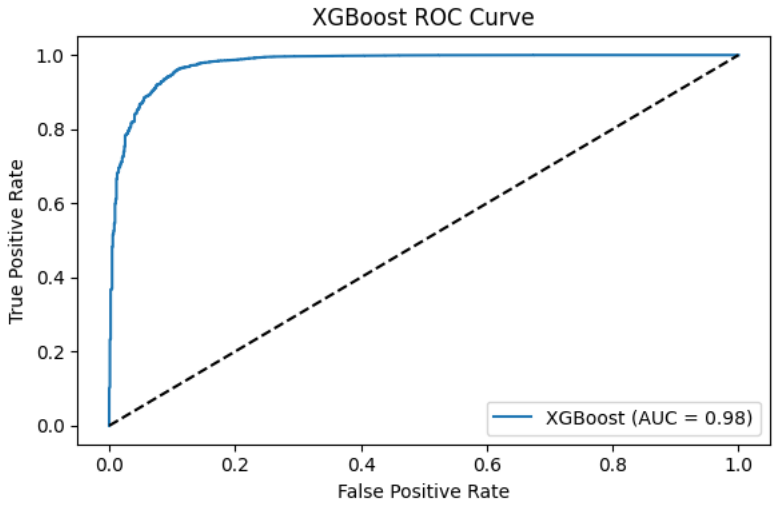

Figure 23: XGBoost ROC Curve

5.5 SHAP Analysis

To understand the impact of each feature on the model's predictions, SHAP (SHapley Additive exPlanations) values were calculated. This method provides insights into how each feature contributes to the prediction for each individual sample.

XGBoost Model is used for the SHAP analysis. This is because as shown in the model evaluation, the performance was the best across all the metrics measured. SHAP analysis also represents global data patterns better, and great to use if there is no limitation on computational resources.

5.5.1 Summary

This SHAP summary plot shows the impact of each feature on the model's output, with the color gradient representing feature values (high in red, low in blue), highlighting how variations in features like physical activity and weight influence predictions.

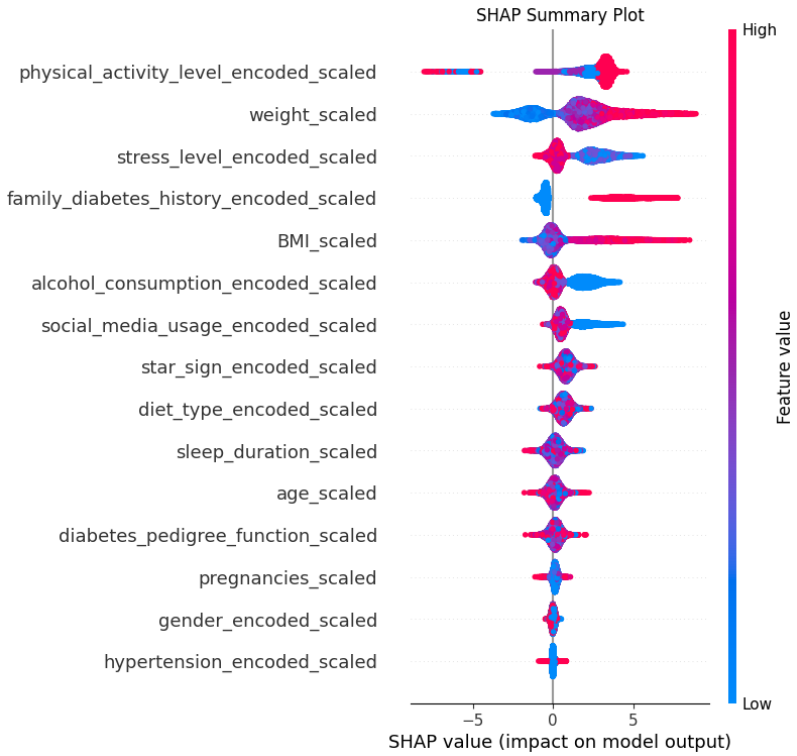

Figure 24: SHAP Summary

5.5.2 Feature Importance

The bar graph below displays the SHAP (SHapley Additive exPlanations) feature importance values, highlighting how much each feature contributes to the model's predictions. The most impactful features include physical activity level, weight, and stress level, indicating their strong influence on the output of the diabetes prediction model.

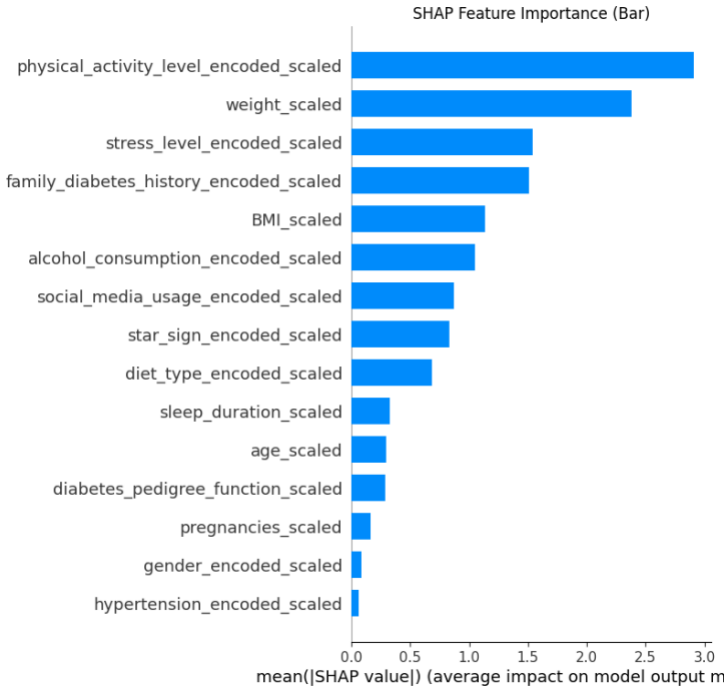

Figure 25: SHAP Importance

5.5.3 Key Insights

1. Performance Comparison:
    - XGBoost has the highest average ROC AUC (0.9812), closely followed by Random Forest (0.9801).
    - Logistic Regression has a lower average ROC AUC (0.9463) compared to the ensemble methods.

2. Consistency:
    - XGBoost exhibits the lowest standard deviation (0.0024), indicating the most consistent performance across folds.
    - Random Forest and Logistic Regression have slightly higher standard deviations but still demonstrate good consistency.
3. Range of Performance:
    - Random Forest and XGBoost have similar ranges between their minimum and maximum ROC AUC scores, showcasing reliable performance.
    - Logistic Regression has a narrower range but operates at a lower performance level.

5 Summary of Insights

5.5 Key insights

Base Data Insights

- Most of the population is **male**.
- Diet type, star sign, stress level, and social media usage are evenly distributed so very little variability is observed.
- A **sedentary physical activity level** is predominant.
- **Heavy alcohol consumption** is the most prevalent behaviour.
- Most of the population has **diabetes**.
- Diabetes affects individuals across **all age groups**

Advanced Data Insights

Top 6 Predictors of Diabetes

1. Physical activity level
2. Weight
3. Stress level
4. Family history of diabetes
5. BMI
6. Alcohol consumption

Model Evaluation Insights

- For the advanced techniques, Random Forest and XGBoost dominate in performance, making them ideal choices.
- Logistic Regression is only suitable if simplicity and computational efficiency are prioritised over accuracy.

5.6 Potential Implications

Based on the analysis, there are significant implications for the diabetes drugs company. The high prevalence of diabetes among males with sedentary lifestyles, elevated stress, and heavy alcohol consumption presents a substantial market opportunity.

By focusing on the top predictors, the company can develop targeted medications and personalised treatment plans tailored to this demographic.

Leveraging advanced predictive models like Random Forest and XGBoost will enable more accurate identification of high-risk individuals, allowing for effective intervention

strategies. This data-driven approach can enhance patient outcomes, expand market reach, and reinforce the company's position as a leader in diabetes care.

5.7 Recommendations

The following are recommendations for next actions:

1. **Enhance Predictive Model Accuracy with Advanced Techniques**
   Integrate advanced algorithms like deep learning and hybrid models to improve the model's precision in identifying high-risk individuals.

2. **Integrate Comprehensive Clinical and Nutritional Data**
   Incorporate vital clinical measurements and detailed nutritional data to enrich the model's insights for more personalized interventions.

3. **Develop an Intuitive Interface**
   Design a user-friendly interface that seamlessly integrates with electronic health records, enabling practitioners to utilize the predictive tool efficiently.

4. **Foster Strategic Collaborations and Ensure Regulatory Compliance**
   Engage with healthcare professionals and comply with medical regulations (PII data) to enhance the tool's credibility and expand its market reach.

6 Appendix

Table 1 – categorical attributes summary

Attribute	Value	Count
gender	male	44023
gender	female	35931
gender		20046
diet_type		20061
diet_type	pescatarian	6801
diet_type	atkins	6736
diet_type	vegetarian	6705
diet_type	mediterranean	6697
diet_type	raw food	6663
diet_type	paleo	6639
diet_type	ketogenic	6639
diet_type	gluten free	6639
diet_type	weight watchers	6637
diet_type	carnivore	6631
diet_type	vegan	6628
diet_type	low carb	6524
star_sign		20194
star_sign	Cancer	6770
star_sign	Sagittarius	6736
star_sign	Scorpio	6708
star_sign	Virgo	6703
star_sign	Aries	6657
star_sign	Aquarius	6656
star_sign	Gemini	6652
star_sign	Libra	6607
star_sign	Taurus	6604
star_sign	Capricorn	6599
star_sign	Pisces	6583
star_sign	Leo	6531
social_media_usage	Moderate	20140
social_media_usage	Never	20054
social_media_usage		20032
social_media_usage	Excessive	19893
social_media_usage	Occasionally	19881
physical_activity_level	Sedentary	46679
physical_activity_level	Lightly Active	23683
physical_activity_level		19968
physical_activity_level	Moderately Active	8237
physical_activity_level	Very Active	726

physical_activity_level	Extremely Active	707
stress_level	Elevated	20087
stress_level	Extreme	20021
stress_level	Low	19984
stress_level		19976
stress_level	Moderate	19932
alcohol_consumption	heavy	30624
alcohol_consumption	none	24409
alcohol_consumption		20104
alcohol_consumption	light	16523
alcohol_consumption	moderate	8340

Table 2 -

Attribute	Value	Count
hypertension	0	63955
hypertension		19831
hypertension	1	16214
family_diabetes_history	0	55731
family_diabetes_history	1	24132
family_diabetes_history		20137
diabetes	1	76626
diabetes		19758
diabetes	0	3616