ML-Project

**Classification on Multiple Disease**

1. Introduction
2. Problem Analysis
3. Feasibility Analysis
4. Possible Solutions
5. Proposed Design
6. Design Description
7. Implementation Details
8. Experimental Results
9. Performance Analysis
10. Future Scope (Optional)
11. Conclusion

References

The main objective of this project is to use machine learning to predict the likelihood of diabetes in individuals based on various health-related factors. Diabetes is a major global health issue, and early detection can help in managing the disease more effectively, reducing complications. The dataset we are using contains several medical features, such as Body Mass Index (BMI), smoking habits, physical activity levels, heart disease history, and more, that are known to be important indicators of diabetes risk.

The project will involve using machine learning classification models, including logistic regression, decision trees, and random forests, to identify patterns in the data that can predict diabetes risk. These models are suited for classification tasks, where the goal is to categorize individuals into two groups: those at risk of developing diabetes and those who are not.

A key concept in this project is class imbalance, which occurs when there are significantly fewer instances of one class (in this case, people with diabetes) compared to the other class (those without diabetes). Class imbalance can lead to biased predictions, where the model tends to predict the majority class more accurately while failing to predict the minority class correctly. To address this issue, techniques like oversampling the minority class or adjusting class weights will be employed.

This project is significant because it demonstrates how machine learning can be applied to real-world healthcare problems, helping to predict diabetes risk and improve early detection efforts. By accurately predicting diabetes, healthcare providers can take timely steps to manage the disease, improving patient outcomes and reducing the overall healthcare burden

The problem at hand involves predicting whether an individual is at risk of developing diabetes based on various health-related features. Diabetes is a chronic condition affecting millions worldwide, and its early detection and management are essential for preventing severe complications such as heart disease, kidney failure, and nerve damage. This project aims to develop a machine learning model that classifies individuals into categories such as "Yes," "No," "No Borderline," and "Yes during pregnancy," with the goal of identifying individuals at risk for diabetes.

**Key Features in the Dataset:**

The dataset provided contains multiple health attributes that are relevant to diabetes risk. These features are crucial in predicting whether an individual is at risk of developing diabetes or not. Key features in the dataset include:

**1. Heart Disease:** Individuals with a history of heart disease are at higher risk of developing diabetes due to shared risk factors like high blood pressure and poor circulation.

**2. BMI (Body Mass Index):** A high BMI, often associated with obesity, is a significant risk factor for type 2 diabetes. Obesity can lead to insulin resistance, raising the risk of diabetes.

**3. Smoking**: Smoking adversely affects insulin sensitivity and can contribute to inflammation, both of which increase the risk of diabetes.

**4. Alcohol Drinking**: Excessive alcohol consumption can damage the liver and increase insulin resistance, both of which are linked to diabetes.

**5. Stroke:** A history of stroke is associated with underlying health conditions like poor circulation and hypertension, which can increase the risk of diabetes.

**6. Physical Health:** Chronic illness and limited mobility can impair the body’s ability to regulate blood sugar, thereby increasing diabetes risk.

**7. Mental Health:** Mental health issues such as chronic stress and depression can lead to unhealthy lifestyle choices, further impacting blood sugar regulation.

**8. Physical Activity**: Lack of physical activity is a key contributor to obesity and insulin resistance, increasing the likelihood of diabetes.

**9. General Health**: Overall poor health, including comorbidities and chronic conditions, can impair the body’s ability to maintain stable blood sugar levels.

**10. Sleep Time:** Inadequate or poor-quality sleep is linked to insulin resistance, contributing to a higher risk of developing diabetes.

**Approach to Solving the Problem:**

To solve this multi-class classification problem, the following steps will be taken:

1**. Data Preprocessing:** The dataset will undergo preprocessing, which includes handling missing values, encoding categorical variables, and scaling numerical features. This step ensures that the data is in a format suitable for machine learning algorithms.

2. **Class Imbalance Handling:** Since the dataset may exhibit an imbalance in class distribution (e.g., more individuals without diabetes), techniques like SMOTE (Synthetic Minority Over-sampling Technique) or class weighting will be applied to address this imbalance and ensure the model does not favor the majority class.

1. **Model Selection and Training:** Various machine learning models will be used to train the data, including:

* **Logistic Regression**
* **Decision Trees**
* **Random Forests**
* **SVM (Support Vector Machines)**
* **K-Nearest Neighbors (KNN)**
* **Naive Bayes**
* **Neural Networks**
* **XGBoost**

1. **Evaluation Metrics:** The models will be evaluated using appropriate multi-class classification metrics, including accuracy, precision, recall, and F1-score. These metrics will help assess the performance of the models and ensure they generalize well to unseen data.
2. **Model Tuning**: Hyperparameter tuning will be performed to optimize the performance of each model. This process helps to find the best parameters that improve accuracy and other performance metrics.
3. **Goal:**

The goal is to develop an effective predictive model that can help healthcare professionals identify individuals at risk for diabetes, allowing for timely intervention and better management of the disease. By accurately predicting the likelihood of diabetes, the model can assist in early detection and help individuals make lifestyle changes that may reduce the risk of developing diabetes in the future.

By addressing the multi-class nature of the problem and employing appropriate techniques for data preprocessing and model evaluation, this project aims to deliver a robust solution for predicting diabetes risk based on health-related features.

The feasibility analysis for the diabetes risk prediction project evaluates its practicality within a 20-day timeframe, focusing on time and cost management. The project aims to develop a machine learning model for predicting diabetes risk, leveraging health-related features from a provided dataset.

**Time Management Feasibility:**

The project will be completed in phases with specific tasks allocated for each:

1. **Data Collection and Preprocessing (3-4 days):**

This phase includes loading the dataset, handling missing values, and encoding categorical features. Preprocessing is essential for model training and ensures the data is in a usable format.

1. **Model Selection and Training (5-6 days):**

Various machine learning algorithms (logistic regression, decision trees, etc.) will be trained on the dataset, with proper splitting of the data into training and testing sets.

1. **Model Evaluation and Hyperparameter Tuning (4-5 days):**

After training the models, performance will be evaluated using metrics like accuracy, precision, recall, and F1-score. Hyperparameter optimization will be done to enhance the model's performance.

1. **Final Model Selection and Integration (3-4 days):**

The best-performing model will be chosen, and final testing will ensure it works well in a real-world scenario.

1. **Report Generation and Presentation (2-3 days):**

A report will summarize findings, results, and visualizations, and will be prepared for presentation.

**Cost Management Feasibility:**

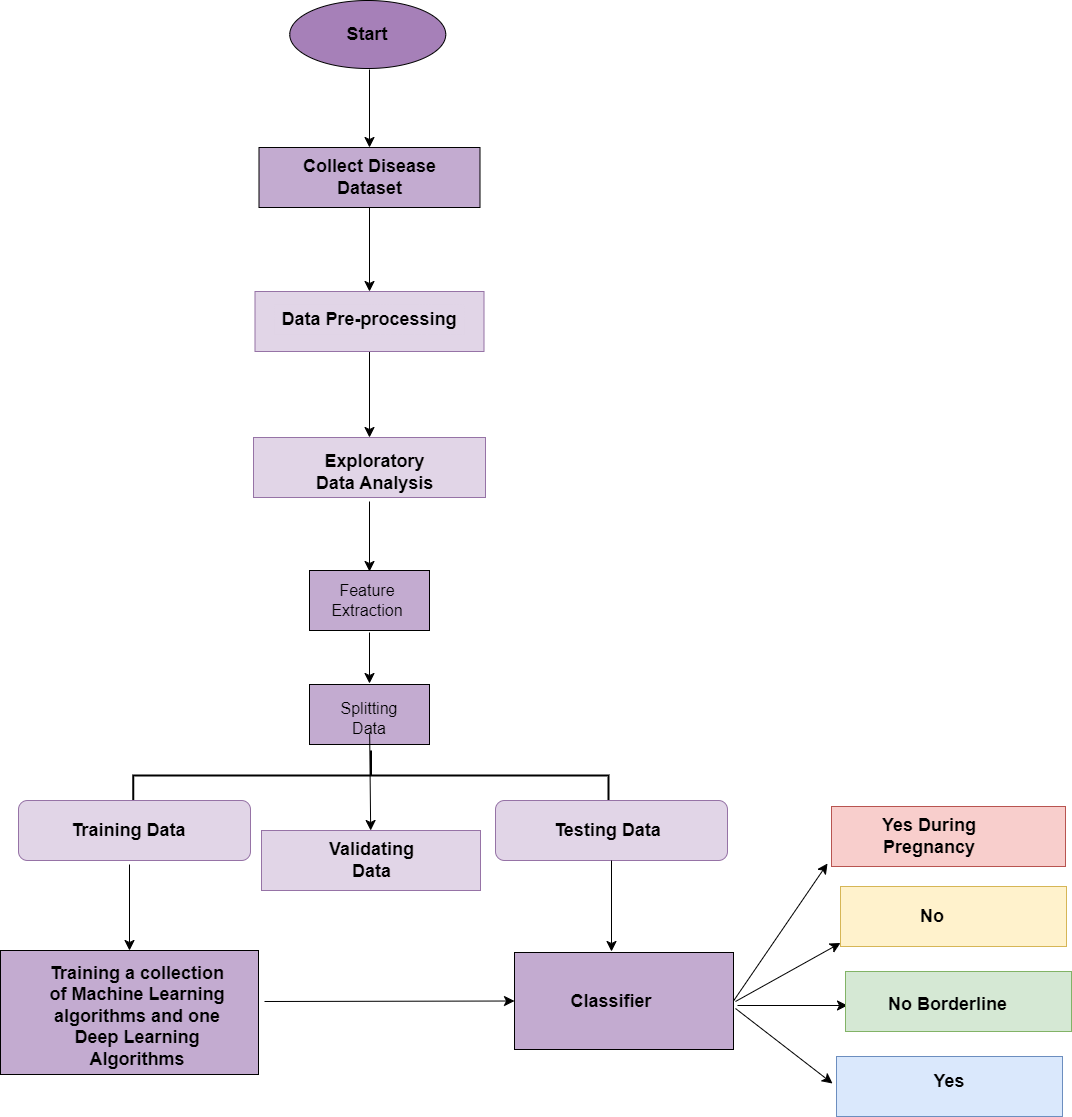
The project will incur minimal costs due to the use of open-source tools and platforms. Tools like Python, scikit-learn, and pandas are free and will be used throughout the project. For cloud computing, free-tier resources such as Google Colab or AWS can be utilized for model training without additional costs. The project requires basic hardware, and no expensive equipment or software licenses are needed.

**Conclusion:**

With efficient time management and the use of free tools, the diabetes risk prediction project is highly feasible within the 20-day period. The tasks are broken down into manageable phases, and the costs are minimal, mainly comprising time investment. The project can be successfully completed with proper planning, ensuring it meets the goals within the given constraints.

[Note: Discuss possible solutions if there are multiple solutions available for the problem. Choose the best amongst all possible solutions and give a comparison to justify why the chosen one is best. Remove this content while writing your report.]

**Flowchart:**



**Diabetic Prediction Model Design**

1. **Data Collection**

The process begins by collecting a dataset containing patient health-related features, such as age, gender, medical history, symptoms, and test results, relevant to the disease being predicted.

1. **Data Preprocessing**

* **Missing Values:** Handle missing data by imputation or removal.
* **Normalization:** Standardize numerical features to ensure equal scaling across the dataset.
* **Categorical Encoding:** Convert categorical features (e.g., gender, pregnancy status) into numerical form using one-hot encoding or label encoding.
* **Data Cleaning:** Remove erroneous or outlier entries.

1. **Exploratory Data Analysis (EDA)**

* **Statistical Summary:** Understand the dataset’s distribution using mean, median, etc.
* **Correlations:** Use a correlation matrix to identify relationships between features.
* **Visualizations:** Create histograms, box plots, and scatter plots to explore data patterns.

1. **Feature Extraction**

* **Feature Selection:** Use statistical tests and feature importance methods (e.g., decision trees) to select relevant features.
* **Dimensionality Reduction:** Apply techniques like PCA to reduce the feature space, if necessary.

1. **Splitting Data**

Divide the dataset into:

* Training Data: Used to train models.
* Validation Data: Helps tune hyperparameters.
* Testing Data: Evaluates final model performance.

1. **Model Training (ML & DL)**

* ML Algorithms: Train various algorithms (e.g., Logistic Regression, Decision Trees, SVM, Random Forest, KNN) on the training set.
* DL Algorithm: Train a deep learning model, such as a neural network, for potentially higher accuracy with complex data.

1. **Model Evaluation & Classification**

* Testing: Test models on the testing dataset and classify into:
  + Yes, During Pregnancy: Disease likely during pregnancy.
  + No: Disease unlikely.
  + No, Borderline: Condition is borderline, requiring further analysis.

1. **Conclusion**

Performance Metrics: Evaluate model accuracy, precision, recall, and F1 score.

Model Tuning: Fine-tune the best-performing model using hyperparameter optimization.

Deployment: Deploy the final model for real-time predictions.

1. **Exploratory Analysis**
   1. **Visualization of Categorical Features**

**Interpretation of Diabetes Distribution by Gender:**

The x-axis represents gender, where 1 refers to men and 0 refers to women.

* **Non-Diabetic (No):**

Among females, 21,869 individuals fall into the "No" category, indicating that the majority of women in the dataset do not have diabetes.

For males, 23,180 individuals are non-diabetic, representing a slightly higher number than females.

* **Borderline Diabetes (No, Borderline):**

A total of 739 females are classified as borderline diabetic, which suggests that a smaller portion of women are at risk but do not yet have full diabetes.

In comparison, 766 males fall into the same category, showing a similar trend of borderline diabetes risk across both genders.

* **Diabetic (Yes):**

5,335 females are diagnosed with diabetes, indicating that a significant portion of women in the dataset live with the condition.

For males, 6,853 individuals are diagnosed with diabetes, which is higher than the number of females diagnosed, suggesting that men in this dataset are more likely to be diabetic.

* **Gestational Diabetes (Yes During Pregnancy):**

326 females are categorized under "Yes, During Pregnancy," reflecting those diagnosed with diabetes during pregnancy. This is a condition exclusive to women and affects a small portion of the female population.

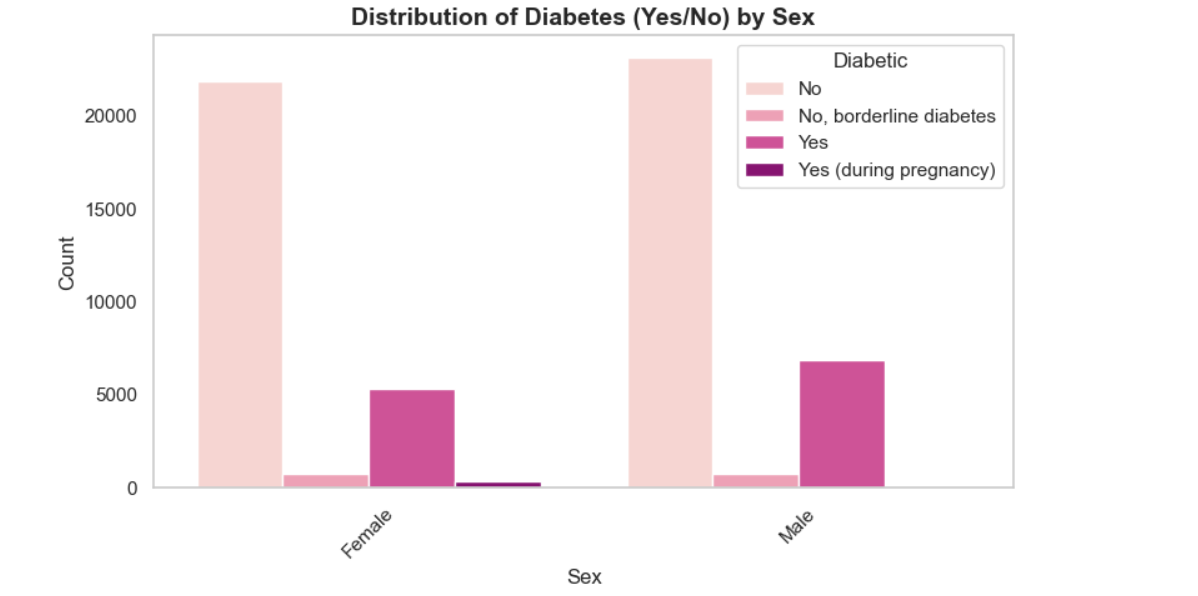


Figure 1 Distribution of Diabetes (Yes/No) according to Sex

**Interpretation of Diabetes Distribution by Smoking Status:**

Figure 1 shows the distribution of diabetes categories (No, No Borderline, Yes, Yes during Pregnancy) based on smoking status. The x-axis represents smoking status, where 'Yes' indicates smokers and 'No' represents non-smokers. The y-axis displays the number of individuals in each diabetes category.

* **Non-Smokers:**
* The largest group among non-smokers (23,980 individuals) is categorized as having No Diabetes.
* A smaller proportion (752 individuals) falls under the No, Borderline Diabetes category.
* Yes Diabetes is observed in 5,346 non-smokers.
* Yes during Pregnancy is found in 189 non-smokers.
* **Smokers:**
* Similarly, the majority of smokers (21,069 individuals) belong to the No Diabetes category.
* No, Borderline Diabetes is seen in 753 smokers, which is almost identical to non-smokers.
* Smokers show a higher number of individuals with Yes Diabetes (6,842), suggesting a stronger prevalence of diabetes among smokers compared to non-smokers.
* The number of smokers with Yes during Pregnancy is 137, lower than in non-smokers.

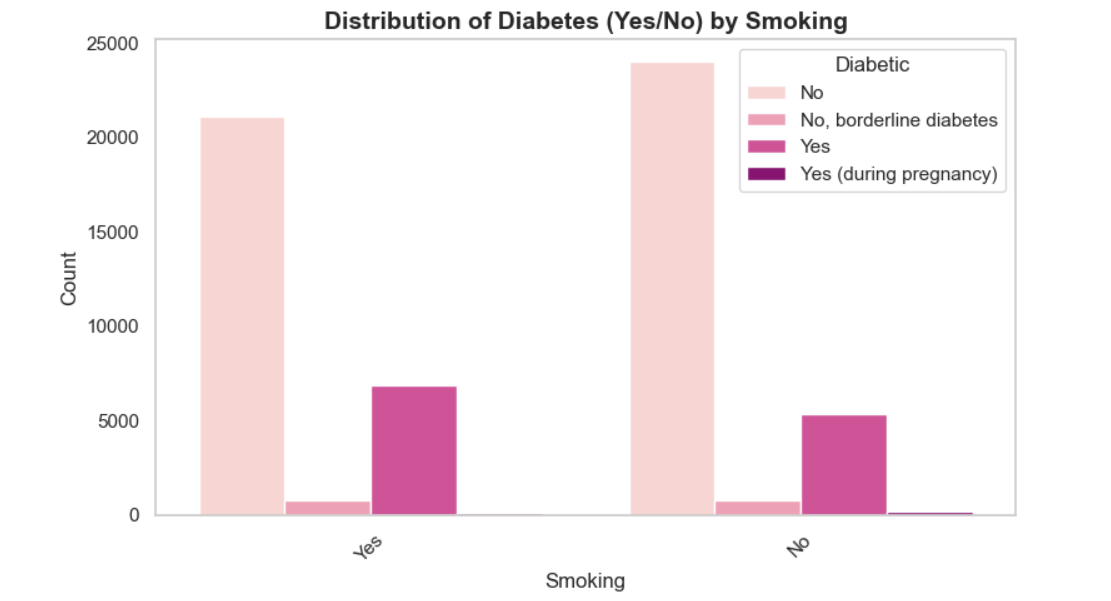


Figure 2 Interpretation of Diabetes Distribution by Smoking Status

In this analysis, the higher incidence of Yes Diabetes among smokers (6,842) compared to non-smokers (5,346) suggests a potential association between smoking and a higher prevalence of diabetes. Additionally, the occurrence of No, Borderline Diabetes is fairly consistent across both groups. Yes during Pregnancy remains relatively low in both groups, indicating that smoking status may not have a significant impact on this category.

**Interpretation of Diabetes Distribution by Race:**

* White individuals exhibit the highest prevalence of diabetes across all categories, with particularly elevated rates of Yes Diabetes and Yes During Pregnancy, emphasizing a significant racial disparity.
* Black individuals also demonstrate a notable burden of diabetes, ranking second for Yes Diabetes and showing moderate occurrences of Yes During Pregnancy.
* Hispanic individuals display a considerable proportion of Yes Diabetes cases, along with a notable presence of Yes During Pregnancy, highlighting a potential area of concern.
* American Indian/Alaskan Native and Asian populations report lower rates of diabetes across all categories, with Yes During Pregnancy being especially rare among these groups.

.

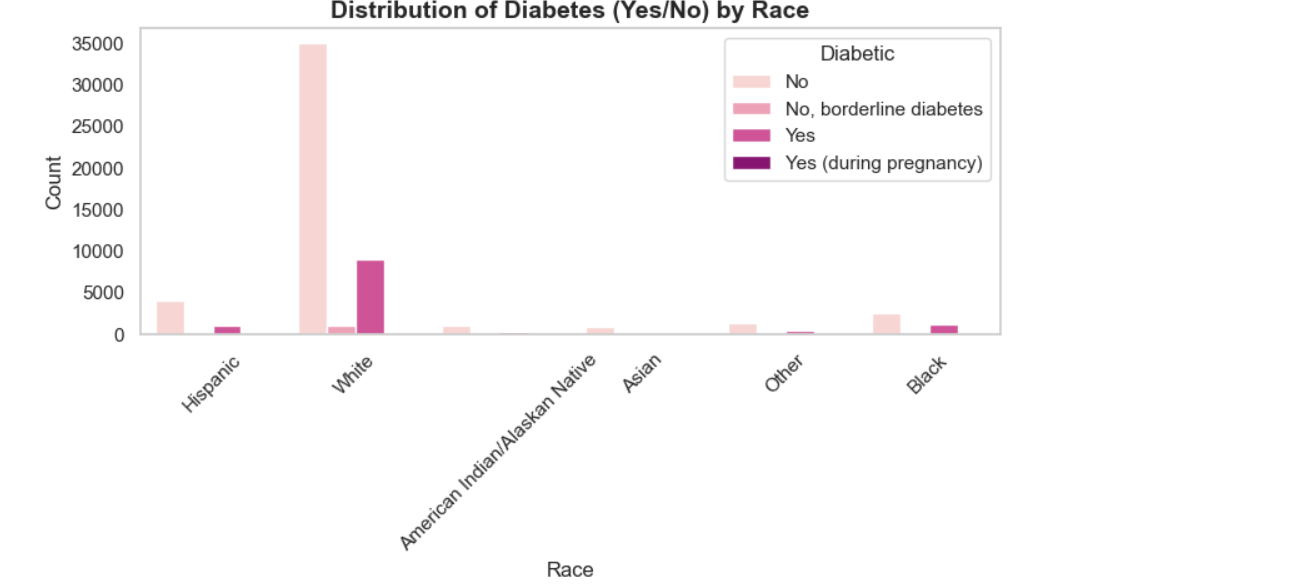


Figure 3 Interpretation of Diabetes Distribution by Race

The analysis reveals significant racial disparities in diabetes prevalence. White individuals have the highest rates across all categories, followed by Black and Hispanic populations, with notable occurrences of pregnancy-related diabetes. In contrast, American Indian/Alaskan Native and Asian groups show lower diabetes rates. These findings underscore the need for targeted interventions to address diabetes among disproportionately affected racial groups.

**Interpretation of Diabetes Distribution by Age Category:**

* **Age 18-24**: Minimal diabetes cases, with 23 "Yes Diabetes" and 5 "Yes During Pregnancy"; majority are "No Diabetes" (2213).
* **Age 25-29:** Slight increase in "Yes Diabetes" (22) and "Yes During Pregnancy" (9), while 1972 are "No Diabetes."
* **Age 30-34:** Noticeable rise in "Yes Diabetes" (71) and "Yes During Pregnancy" (35); 2144 remain "No Diabetes."
* **Age 35-39:** "Yes Diabetes" increases to 125, "Yes During Pregnancy" peaks at 38; majority still "No Diabetes" (2226).
* **Age 40-44:** Sharp rise in "Yes Diabetes" (238), slight drop in "Yes During Pregnancy" (31); 2265 are "No Diabetes."
* **Age 45-49:** Significant increase in "Yes Diabetes" (419), stable "Yes During Pregnancy" (32); "No Diabetes" at 2527.
* **Age 50-54**: Major surge in "Yes Diabetes" (694), stable pregnancy cases (32); "No Diabetes" still high at 3074.
* **Age 55-59:** "Yes Diabetes" soars to 1143, "Yes During Pregnancy" drops to 27; 3731 remain "No Diabetes."
* **Age 60-64**: Diabetes peaks with 1631 "Yes Diabetes" and 34 pregnancy-related cases; "No Diabetes" at 4712.
* **Age 65-69:** Highest "Yes Diabetes" cases (1996), minimal pregnancy-related (21); 5166 are "No Diabetes."
* **Age 70-74:** "Yes Diabetes" peaks at 2292, rare pregnancy cases (28); "No Diabetes" still significant at 5252.
* **Age 75-79:** "Yes Diabetes" decreases to 1784, rare pregnancy cases (12); "No Diabetes" at 4157.
* **Age 80+:** "Yes Diabetes" remains high (1750), pregnancy cases rare (22); "No Diabetes" dominate at 5610.

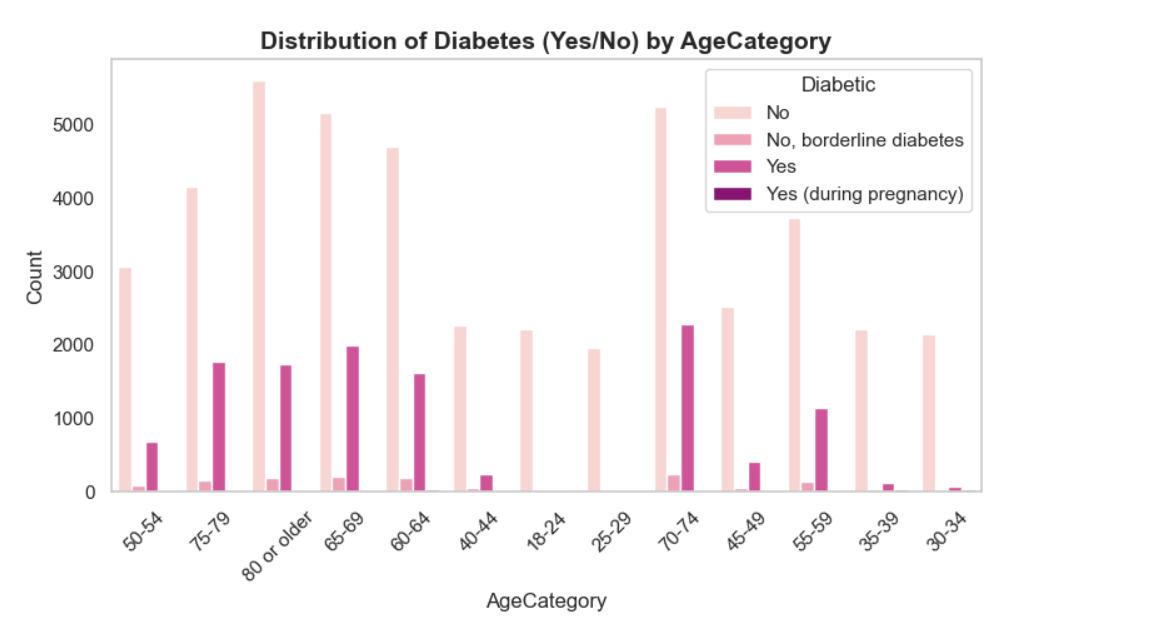


Figure 4 Interpretation of Diabetes Distribution by Age Category

**Interpretation of Diabetes Distribution by KidneyDisease**

* The majority of individuals without kidney disease (42,901) do not have diabetes.
* A smaller proportion of individuals without kidney disease (1,393) are classified as borderline diabetes.
* A moderate number of individuals without kidney disease (10,080) are diagnosed with diabetes.
* There are 302 cases of diabetes occurring during pregnancy in individuals without kidney disease.
* Among individuals with kidney disease, 2,148 do not have diabetes.
* Only 112 individuals with kidney disease fall into the borderline diabetes category.
* A similar number of individuals with kidney disease (2,108) are diagnosed with diabetes, as compared to the population without kidney disease.
* There are 24 cases of diabetes during pregnancy in individuals with kidney disease.
* While the majority of individuals without kidney disease fall into the "No Diabetes" category, the prevalence of diabetes is notably higher among those with kidney disease, although the total population with kidney disease is much smaller.

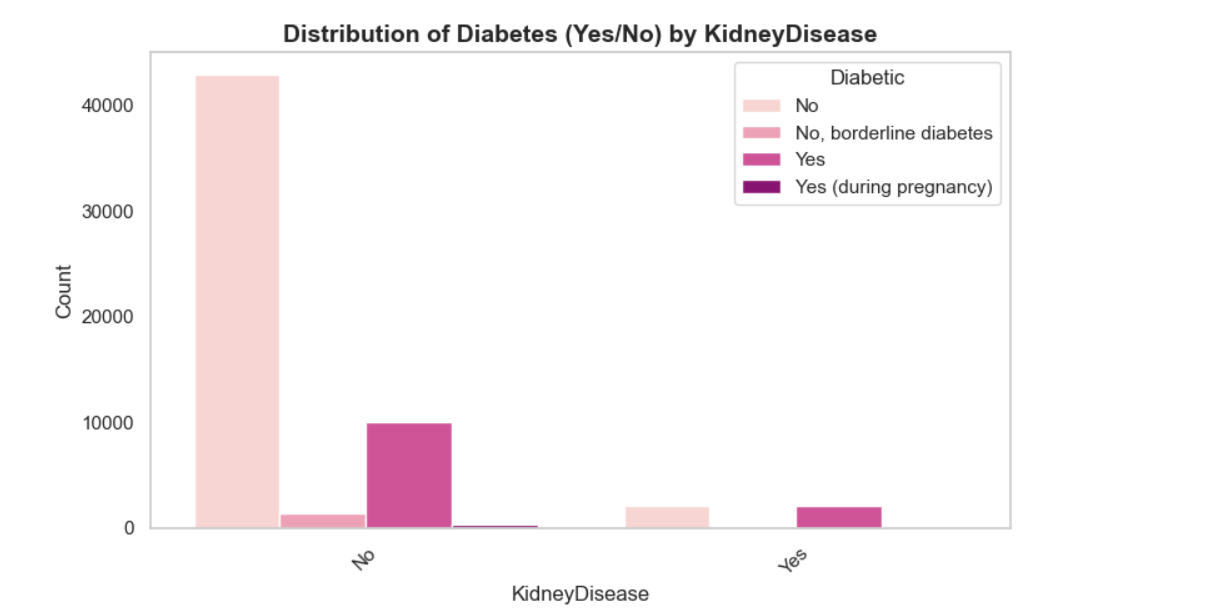


Figure 5 Interpretation of Diabetes Distribution by KidneyDisease

The analysis indicates a clear association between kidney disease and higher diabetes prevalence, with individuals suffering from kidney disease showing a greater likelihood of being diagnosed with diabetes. Although the population with kidney disease is smaller, the number of diabetes cases is comparable to those without kidney disease. The incidence of diabetes during pregnancy is minimal in both groups, but slightly higher in individuals without kidney disease. Overall, these findings emphasize the relationship between kidney disease and diabetes, highlighting the importance of monitoring and managing diabetes in individuals with kidney conditions.

**Interpretation of Diabetes Distribution by SkinCancer**

* + - Among individuals without skin cancer, the majority (39,049) do not have diabetes, while 1,288 are classified as borderline diabetes and 10,301 are diagnosed with diabetes. There are 298 cases of diabetes during pregnancy.
    - For individuals with skin cancer, a smaller number (6,000) do not have diabetes, and only 217 are borderline diabetes. A moderate number (1,887) are diagnosed with diabetes, and there are 28 cases of diabetes during pregnancy.
    - The data shows that individuals with skin cancer have fewer cases of "No Diabetes" and "Borderline Diabetes" compared to those without skin cancer. However, the number of people with diabetes is still significant in the skin cancer group, although lower than in the group without skin cancer.
    - The occurrence of diabetes during pregnancy is low in both groups, with slightly more cases in individuals without skin cancer.

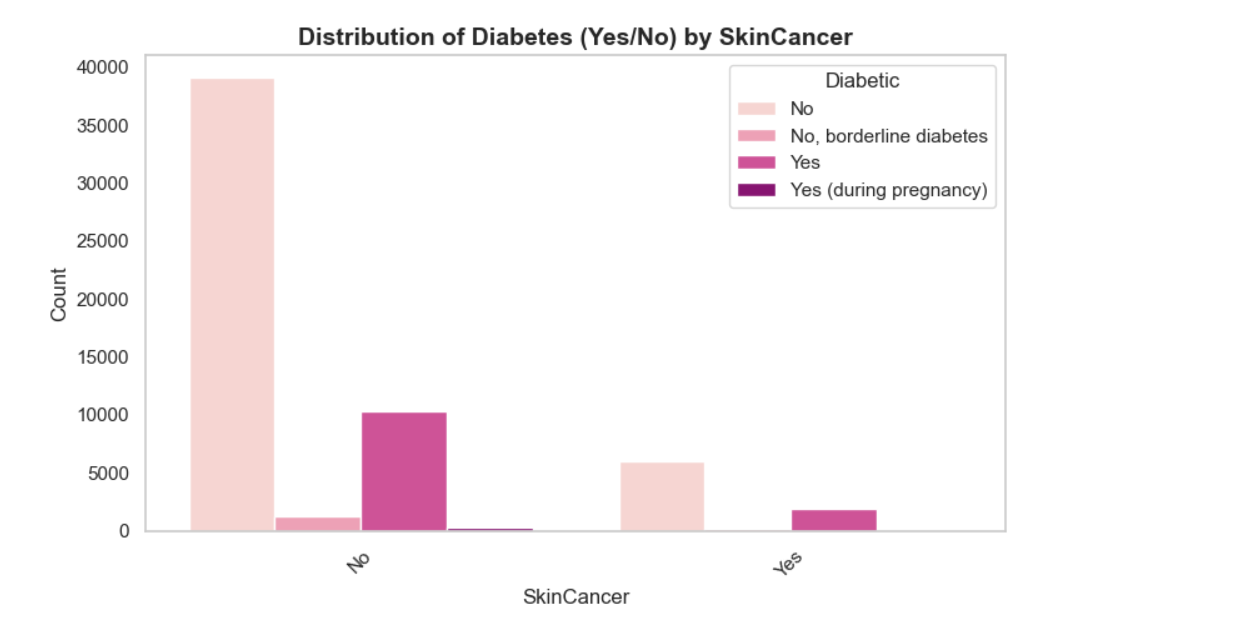


Figure 6 Interpretation of Diabetes Distribution by SkinCancer

The data reveals a notable difference in the distribution of diabetes between individuals with and without skin cancer. While the majority of individuals without skin cancer do not have diabetes, the proportion of those diagnosed with diabetes is still significant. In contrast, individuals with skin cancer have a lower number of "No Diabetes" and "Borderline Diabetes" cases, with a relatively higher occurrence of diabetes, though still fewer than in the non-skin cancer group. The incidence of diabetes during pregnancy is low in both groups, but slightly higher in those without skin cancer.

These findings suggest a potential link between skin cancer and diabetes, with individuals diagnosed with skin cancer being more likely also to have diabetes. However, the smaller population of individuals with skin cancer limits the extent of this association. Overall, the data emphasizes the importance of monitoring diabetes in individuals with skin cancer, as there appears to be a higher likelihood of co-occurrence between these two conditions.

**Interpretation of Diabetes Distribution by Stroke**

* Among individuals without a history of stroke, the majority (41,994) do not have diabetes, with 1,367 classified as borderline diabetes and 10,203 diagnosed with diabetes. There are 305 cases of diabetes during pregnancy.
* In individuals with a history of stroke, 3,055 do not have diabetes, and 138 are classified as borderline diabetes. A significant portion (1,985) is diagnosed with diabetes, with 21 cases of diabetes during pregnancy.
* The data indicates that individuals without a history of stroke have a higher proportion of "No Diabetes" and "Borderline Diabetes" cases compared to those with a stroke, although the number of diabetes cases is still substantial in both groups.
* The incidence of diabetes during pregnancy is low in both groups, with slightly more cases among those without a stroke.
* These findings suggest a potential association between stroke and diabetes, with individuals having a history of stroke appearing more likely to be diagnosed with diabetes.
* Overall, the data highlights the importance of monitoring diabetes in individuals with a history of stroke, given their higher prevalence of the condition.

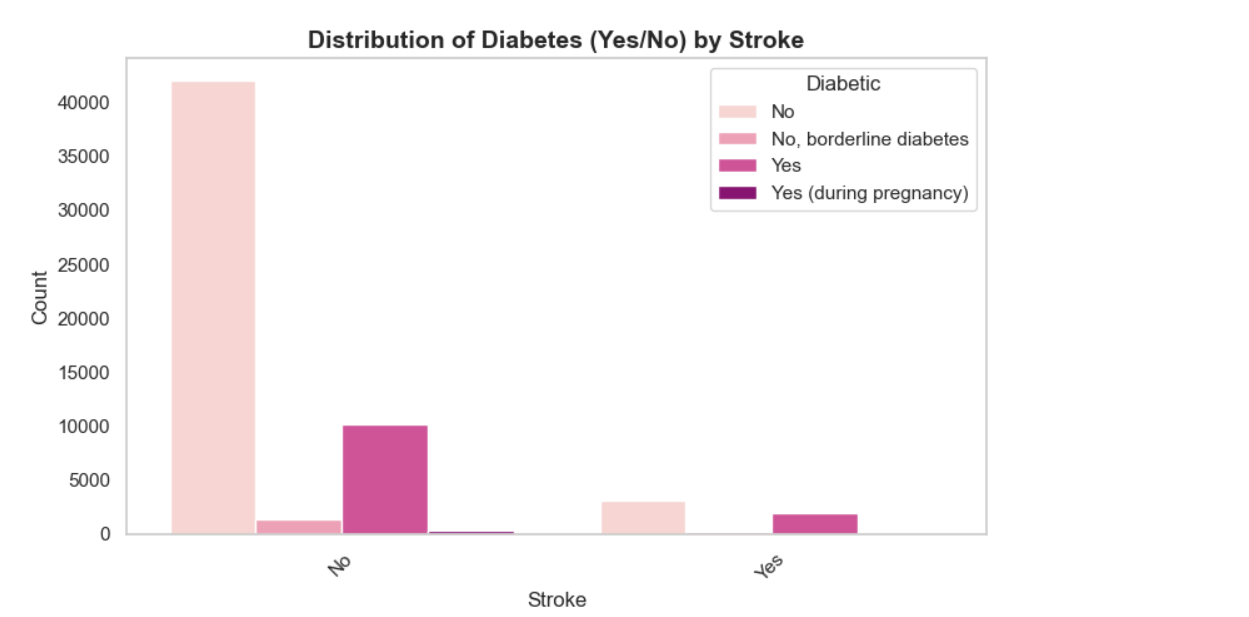


Figure 7 Interpretation of Diabetes Distribution by Stroke

The data shows a higher prevalence of diabetes in individuals with a history of stroke compared to those without. While most individuals without a stroke are free from diabetes, those with a stroke are more likely to be diagnosed with the condition. The incidence of diabetes during pregnancy is low in both groups, with slightly more cases in those without a stroke. These findings highlight the importance of monitoring diabetes in individuals with a history of stroke

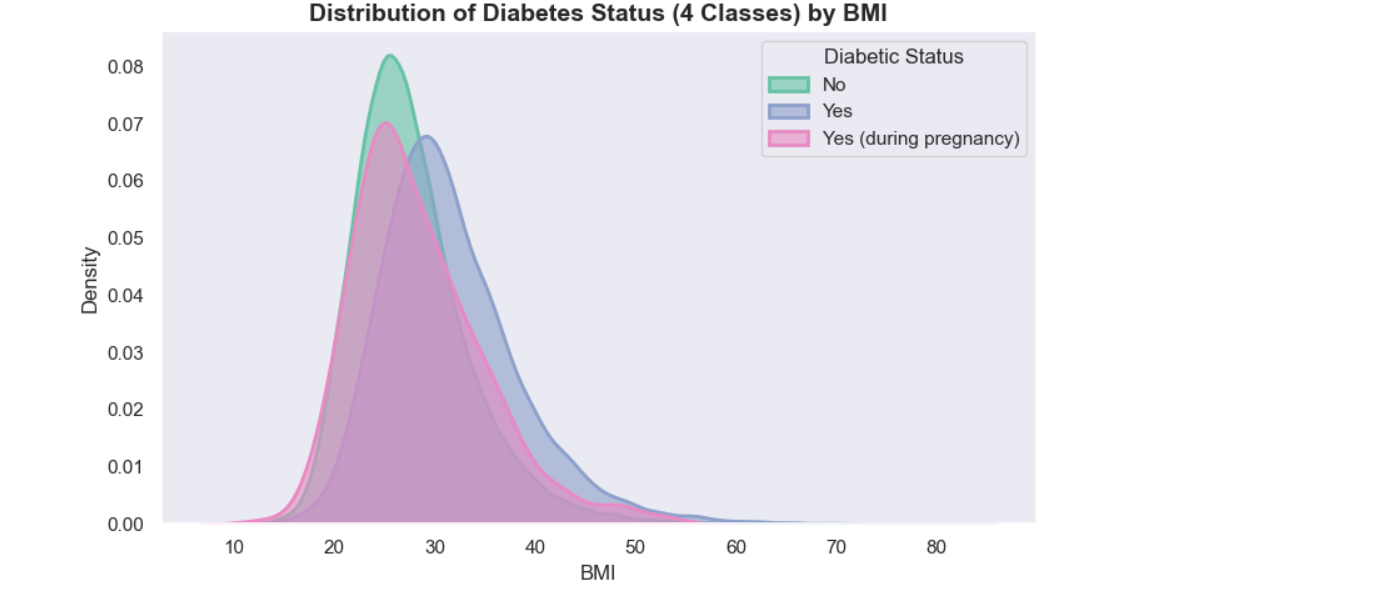
* 1. **Visualization of Numerical Features**

The visualizations aim to analyze the distribution of numerical features in the dataset and their relationship with the target variable, Diabetic Status, which consists of four classes:

* No: Individuals without diabetes.
* Yes: Individuals with diabetes.
* Yes (during pregnancy): Individuals diagnosed with gestational diabetes.

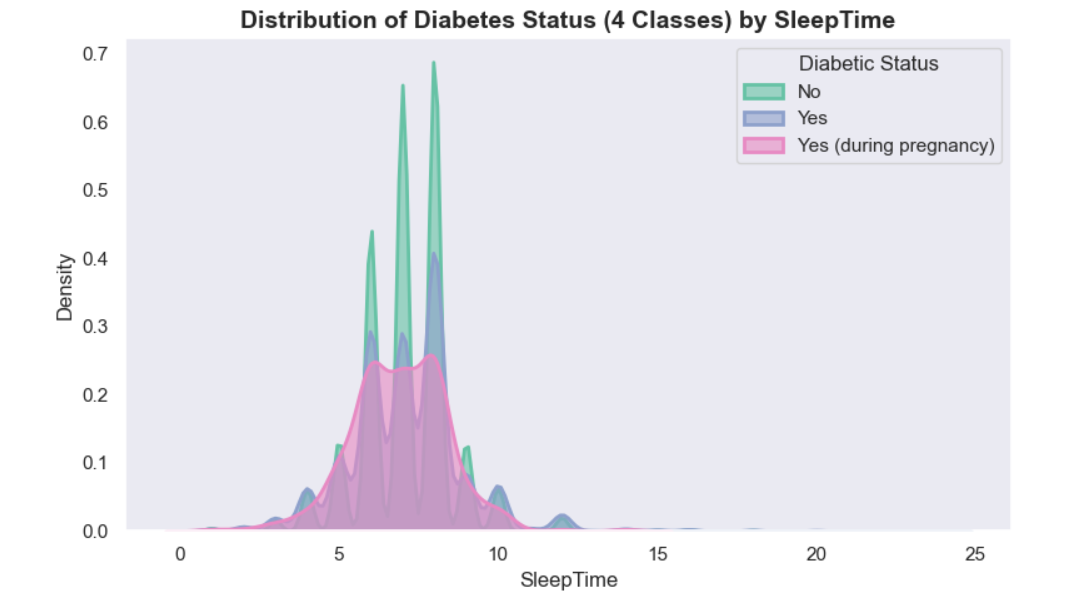
**Distribution of Diabetes Status by BMI**

* The BMI distribution shows significant overlap among the classes.
* Higher BMIs are more frequently associated with individuals having diabetes (both general and gestational).
* Non-diabetic individuals are concentrated around a lower BMI range, indicating that BMI could serve as an influential predictor.

****

**Distribution of Diabetes Status by Sleep Time**

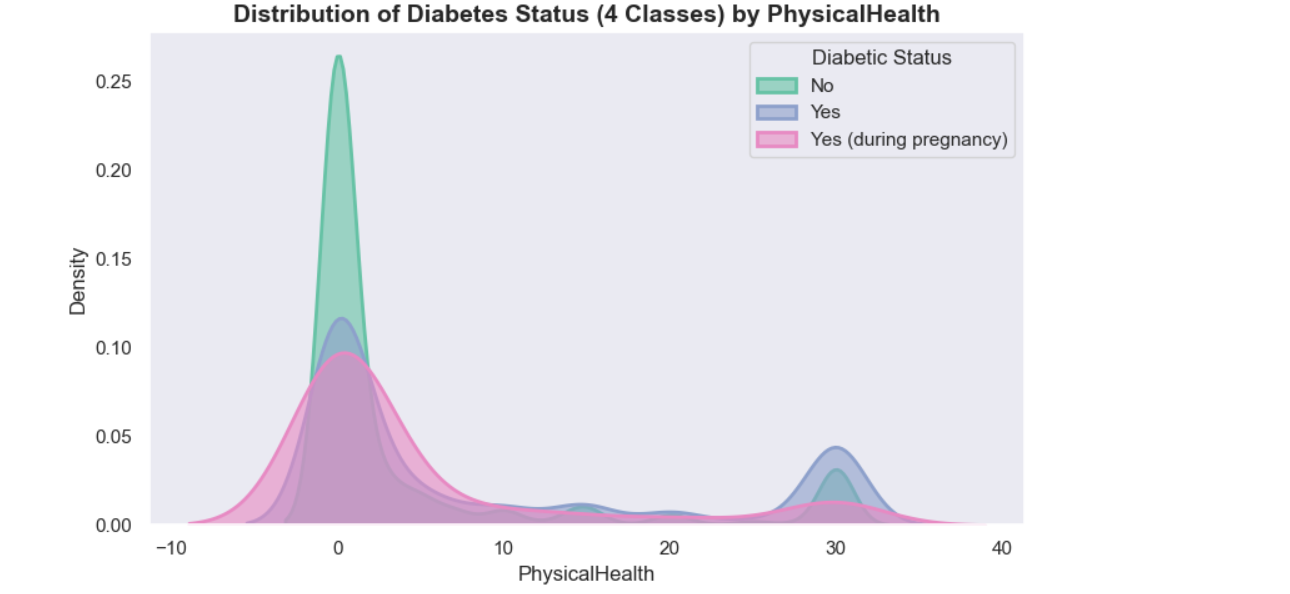
* Sleep time shows a sharp peak around 7–9 hours for non-diabetic individuals.
* Individuals with diabetes (both general and gestational) exhibit more variability in sleep time, with a noticeable presence in the extremes (both short and long sleep durations).
  + - The trends suggest a possible association between irregular sleep patterns and diabetes risk.



**Distribution of Diabetes Status by Physical Health**

* Physical health, measured on an unspecified scale, reveals distinct patterns.
* Non-diabetic individuals cluster around better physical health values.
* Individuals with diabetes display a wider distribution skewed toward poorer physical health.
* Gestational diabetes shows a distribution similar to non-diabetic individuals but with slightly more variability.

.



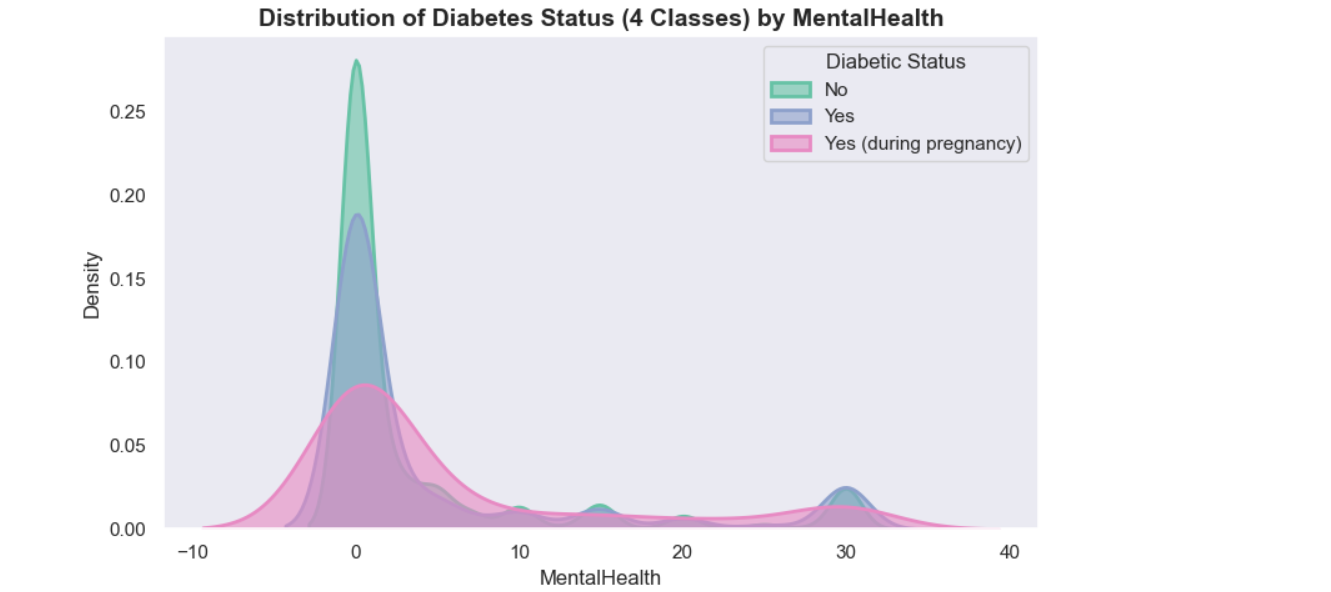
**Distribution of Diabetes Status by Mental Health**

Mental health distributions also vary significantly across the classes.

Non-diabetic individuals exhibit better mental health scores with smaller variance.

Both diabetic and gestational diabetic classes show broader distributions with lower mental health scores, hinting at a possible link between diabetes and psychological well-being.

Conclusion



*The numerical feature analysis provides valuable insights:*

* ***BMI and sleep time*** *demonstrate noticeable associations with diabetic status, indicating their predictive potential.*
* ***Physical and mental health*** *distributions suggest that poor health in these domains may be indicative of diabetes risk*

*.*

*These observations will inform feature selection and data preprocessing for the subsequent machine learning modeling.*

**Experiment 1**

The first experiment was conducted without handling class imbalance in the dataset. Various machine learning algorithms and a deep learning model implemented from scratch were trained on the dataset using popular machine learning libraries such as scikit-learn for the machine learning models and TensorFlow for the neural network. The performance of each algorithm was evaluated based on accuracy, precision, recall, F1-score, and the time taken for training. The results of each model are shown in Table below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | P(0) | R(0) | F(0) | P(1) | R(1) | F(1) | P(2) | R(2) | F(2) | P(3) | R(3) | F(3) |
| Logistic Regression | |  | | --- | | 0.78 | | |  | | --- | | 0.80 | | |  | | --- | | 0.95 | | |  | | --- | | 0.87 | | |  | | --- | | 0.00 | | |  | | --- | | 0.00 | | |  | | --- | | 0.00 | | |  | | --- | | 0.55 | | |  | | --- | | 0.27 | | |  | | --- | | 0.36 | | |  | | --- | | 0.00 | | |  | | --- | | 0.00 |  |  | | --- | |  | | 0.00 |
| |  | | --- | |  |   Decision Tree | |  | | --- | | 0.68 | | |  | | --- | | 0.81 | | |  | | --- | | 0.79 | | |  | | --- | | 0.80 | | |  | | --- | | 0.04 | | |  | | --- | | 0.04 | | |  | | --- | | 0.04 | | |  | | --- | | 0.35 | | |  | | --- | | 0.36 | | |  | | --- | | 0.35 | | |  | | --- | | 0.00 | | |  | | --- | | 0.00 |  |  | | --- | |  | | 0.00 |
| Random Forest | 0.77 | 0.81 | 0.93 | 0.86 | 0.00 | 0.00 | 0.00 | 0.51 | 0.30 | 0.38 | 0.00 | 0.00 | 0.00 |
| K-Nearest Neighbors | 0.75 | 0.80 | 0.91 | 0.85 | 0.05 | 0.00 | 0.01 | 0.44 | 0.28 | 0.34 | 0.00 | 0.00 | 0.00 |
| Naive Bayes | 0.40 | 0.83 | 0.42 | 0.56 | 0.00 | 0.00 | 0.00 | 0.44 | 0.38 | 0.40 | 0.01 | 0.96 | 0.02 |
| Neural Network | 0.77 | 0.81 | 0.93 | 0.86 | 0.00 | 0.00 | 0.00 | 0.51 | 0.31 | 0.39 | 0.00 | 0.00 | 0.00 |
| XGBoost | 0.78 | 0.81 | 0.95 | 0.87 | 0.00 | 0.00 | 0.00 | 0.57 | 0.29 | 0.38 | 0.00 | 0.00 | 0.00 |

**Experiment 2**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | P(0) | R(0) | F(0) | P(1) | R(1) | F(1) | P(2) | R(2) | F(2) | P(3) | R(3) | F(3) |
| Logistic Regression | |  | | --- | | 0.5048 | | |  | | --- | | 0.55 | | |  | | --- | | 0.34 | | |  | | --- | | 0.42 | | |  | | --- | | 0.39 | | |  | | --- | | 0.23 | | |  | | --- | | 0.29 | | |  | | --- | | 0.48 | | |  | | --- | | 0.59 | | |  | | --- | | 0.53 | | |  | | --- | | 0.55 | | |  | | --- | | 0.85 |  |  | | --- | |  | | 0.67 |
| |  | | --- | |  |   Decision Tree | |  | | --- | | 0.8396 | | |  | | --- | | 0.80 | | |  | | --- | | 0.76 | | |  | | --- | | 0.78 | | |  | | --- | | 0.86 | | |  | | --- | | 0.88 | | |  | | --- | | 0.87 | | |  | | --- | | 0.72 | | |  | | --- | | 0.73 | | |  | | --- | | 0.73 | | |  | | --- | | 0.97 | | |  | | --- | | 0.98 |  |  | | --- | |  | | 0.97 |
| Random Forest | 0.9135 | 0.86 | 0.85 | 0.85 | 0.96 | 0.96 | 0.96 | 0.84 | 0.85 | 0.84 | 0.99 | 1.00 | 0.99 |
| K-Nearest Neighbors | 0.8575 | 0.87 | 0.58 | 0.70 | 0.86 | 0.98 | 0.92 | 0.75 | 0.87 | 0.81 | 0.96 | 1.00 | 0.98 |
| Naive Bayes | 0.4636 | 0.49 | 0.35 | 0.41 | 0.42 | 0.12 | 0.19 | 0.50 | 0.39 | 0.44 | 0.45 | 0.98 | 0.61 |
| Neural Network | 0.7324 | 0.71 | 0.51 | 0.59 | 0.66 | 0.79 | 0.72 | 0.65 | 0.64 | 0.64 | 0.90 | 0.98 | 0.94 |
| XGBoost | 0.7931 | 0.79 | 0.86 | 0.82 | 0.81 | 0.70 | 0.75 | 0.70 | 0.65 | 0.67 | 0.85 | 0.96 | 0.90 |
| SVM | 0.7332 | 0.68 | 0.53 | 0.60 | 0.71 | 0.74 | 0.72 | 0.69 | 0.67 | 0.68 | 0.82 | 0.99 | 0.90 |

The second experiment was performed with techniques applied to handle class imbalance in the dataset. In this case, we implemented SMOTE (Synthetic Minority Over-sampling Technique) and under-sampling methods to balance the dataset before training the machine learning models. The models trained were the same as those in the previous experiment, but the data now contained a more balanced distribution of classes. The results were recorded using the same evaluation metrics: accuracy, precision, recall, F1-score, and the time taken for training. The outcomes of each model are shown in Table below.

**Experiment 1 Analysis:**

In the first experiment, no class imbalance handling techniques were applied. The performance of each model was evaluated based on accuracy, precision, recall, and F1-score for each class. Below are the key observations:

1. Logistic Regression showed moderate accuracy (0.78), but its performance for class 1 (positive class) was very poor, with precision, recall, and F1-score all close to zero. This indicates that the model failed to identify the positive class well, possibly due to class imbalance.
2. Decision Tree performed relatively worse than some other models with an accuracy of 0.68. It struggled with precision and recall for the positive class (class 1), but it performed better for the majority class (class 0), reflecting the typical bias towards the majority class in an imbalanced dataset.
3. Random Forest had a decent performance (accuracy of 0.77), with good results for the majority class (precision and recall around 0.81 and 0.93, respectively). However, the performance for class 1 was poor, which indicates that the model could not identify the minority class well.
4. K-Nearest Neighbors performed similarly to Random Forest with an accuracy of 0.75, showing decent performance for the majority class, but struggled with identifying class 1 (low precision and recall).
5. Naive Bayes exhibited very poor performance with an accuracy of 0.40. It had a very low precision and recall for class 1 and class 3, suggesting it failed to make useful predictions for the minority classes.
6. Neural Network had good accuracy (0.77) and performed similarly to Random Forest, with high precision and recall for the majority class, but low performance on the minority classes.
7. XGBoost had the best performance with an accuracy of 0.78, showing a good balance in class prediction, but still struggled with the minority class, showing low performance for class 1.

Overall, the models in Experiment 1 suffer from class imbalance, as evident from the poor performance for the minority classes (class 1 and class 3).

**Experiment 2 Analysis:**

In the second experiment, class imbalance was addressed by applying SMOTE and under-sampling techniques, leading to a more balanced dataset. Here are the observations:

1. Logistic Regression showed a significant drop in performance compared to Experiment 1, with accuracy reduced to 0.50. The precision and recall for class 1, however, improved (compared to the previous experiment), indicating that the handling of class imbalance helped the model make better predictions for the minority class.
2. Decision Tree showed substantial improvement in accuracy (0.84) and performed well across all classes, especially with high recall and precision for both classes 0 and 1. This shows that the decision tree model benefited significantly from the class balancing techniques.
3. Random Forest exhibited outstanding performance with an accuracy of 0.91. The precision, recall, and F1-scores were high across the board, with excellent performance in all classes, making this model the best performer in Experiment 2.
4. K-Nearest Neighbors improved compared to Experiment 1, with an accuracy of 0.86 and significantly better recall for the minority class. However, its precision and recall for class 0 were slightly lower than some of the other models.
5. Naive Bayes still showed poor performance, with an accuracy of 0.46. Despite class balancing, its performance remained weak, particularly for the minority class (class 1).
6. Neural Network showed a significant improvement in accuracy (0.73) and better precision, recall, and F1-scores for class 1. The neural network benefited from the more balanced data, but it was not as strong as Random Forest or Decision Tree.
7. XGBoost also saw an improvement in performance, with an accuracy of 0.79 and better metrics for both classes. While it did better than Experiment 1, it still lagged behind Random Forest and Decision Tree in terms of overall performance.

.

The current design demonstrates the ability to handle class imbalance and provide reliable predictions for diabetic classification using various machine learning models. However, there are several ways in which this design can be expanded and improved for real-world applications, particularly in healthcare.

1. **Scalability for Large Datasets:** While the current model works effectively for the given dataset, healthcare data in real-life scenarios is typically larger and more complex. The system can be improved by implementing scalable architectures, such as cloud-based solutions, to handle large volumes of data efficiently. This can include integrating distributed computing systems to speed up model training and evaluation.
2. **Real-Time Predictive Systems:** Deploying the trained models into real-time systems for continuous monitoring of diabetic patients is a potential improvement. This would allow for immediate intervention and early detection of diabetes-related complications. Integrating the models with wearable devices that track patient vitals can enhance their application.
3. **Model Interpretability and Explainability:** Although accuracy is a crucial metric, healthcare professionals need to understand the reasoning behind a model's prediction. Future improvements can include implementing model interpretability tools such as SHAP or LIME to explain predictions, making the system more transparent and trusted by healthcare providers.
4. **Federated Learning for Privacy Preservation:** Given the sensitivity of healthcare data, preserving patient privacy is essential. Federated learning can be explored to allow model training across decentralized devices (like smartphones or health monitoring systems) without sharing patient data. This would ensure that the model's performance improves without compromising individual privacy.
5. **Incorporating More Features:** The current model uses features like age, BMI, and blood pressure, but adding more relevant features like genetic information, lifestyle factors, and environmental influences can improve prediction accuracy. Advanced feature engineering can also help identify hidden patterns within the data.
6. **Handling Data Drift and Model Retraining:** Over time, the data may evolve due to changes in demographics, healthcare practices, or technology. The system can be designed to adapt to these changes by continuously monitoring performance and implementing regular model retraining to avoid model decay.
7. **Integration with Electronic Health Records (HER):** The model could be integrated with EHR systems for seamless prediction and decision support. This would allow healthcare providers to access real-time predictions directly within their workflow, improving the efficiency of diabetic care.

In conclusion, current design offers valuable insights into diabetic prediction, its future scope includes making it more adaptable, scalable, and interpretable for real-world healthcare settings. These improvements will ensure that the model not only provides accurate predictions but also aligns with the privacy, scalability, and usability requirements of healthcare systems.

[Note: Give conclusive remarks in the light of results obtained above and all the discussion regarding your developed solution. Remove this content while writing your report.]