PROJECT STAGE 1

GROUP- 49

Member 1: Unikey- akur0860
Student ID: 550217239

Member 2: Unikey- ysab0639
Student ID: 540958494

Member 1: Unikey- akur0860

1. Topic and Research Question

Problem Description:

Diabetes, is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. If left untreated, diabetes can cause many complications. In this study we will investigate how machine learning models can be leveraged to accurately predict the presence or absence of diabetes in patients based on available clinical data?

Research Question:

To what extent can lifestyle factors and health history serve as reliable predictors for the likelihood of developing diabetes in patients?

Key Factors to Investigate:

- BMI and its relationship with Diabetes
- Age as a risk factor
- Physical activity levels
- Health risk behaviours (smoking, alcohol consumption)
- Socioeconomic factors (income, education)

Business/ Research Need:

Diabetes is a leading cause of preventable illness and death worldwide. According to the World Health Organization, the global prevalence of diabetes has nearly quadrupled in the last 40 years. This study addresses the urgent need for predictive models that can help identify individuals at high risk of developing diabetes. By leveraging lifestyle factors and health history, early interventions can be implemented to reduce healthcare costs and improve patient outcomes.

Stakeholder Impact:

- **Patients**: Early identification enables preventative measures and lifestyle changes to reduce diabetes risk.
- **Healthcare Providers**: Enhanced diagnostic tools allow for personalized care plans based on individual risk factors.
- **Public Health Organizations**: Health agencies can design targeted prevention programs and allocate resources efficiently.
- **Insurance Companies**: Accurate risk assessment supports tailored health plans and potentially lower premiums for healthier individuals.

2. Data Description

Dataset Overview:

- **Source:** diabetes_diagnosis.csv
- **Instances:** 264,802 total observations
- **Attributes:** 23 variables, with 17 retained after data cleaning (see Appendix for the full data dictionary).

Key Variables Retained:

- **BMI:** Body Mass Index, float data type.
- **Smoker:** Smoking status, float data type.
- **Stroke:** Stroke history, float data type.
- **HeartDiseaseorAttack:** Binary indicator (Yes/No) for heart disease or attack, float data type.
- **PhysActivity:** Physical activity level, float data type.
- **Age:** Age of the patient in years, float data type.

- **BloodPressure:** Blood pressure status (Normal/High), object data type.
- **Diabetes:** Diabetes diagnosis status (No/Prediabetes/Diabetes), object data type.
- ## 3. Data Ingestion and Cleaning

3.1 Data Ingestion

- **Tool**: Pandas DataFrame
- **Structure**: The data is in CSV format, consisting of demographic, clinical, and behavioral variables. The dataset includes multiple columns with both numerical and categorical data.

3.2 Data Quality Assurance and Cleaning

Initial Steps:

- **Dropped rows** with missing values in the **Diabetes** (target) column.
- **Dropped rows** where more than 50% attributes in a row is missing
- **Dropped columns** with >50% missing values (e.g., NoDocbcCost (65%), PhysActivity (61%), AnyHealthcare (58%), Veggies (52%), Fruits (51%)).

1. **Handling Missing Values**:

    - **Categorical**: Imputed using **mode** for columns like **Sex**, **BloodPressure**, **CholCheck**, **GeneralHealth**, **Diffwalk**.
    - **Numerical**:
        - Filled **Age and Income** with **mean**.
        - Filled **BMI** based on **gender** and **age group** (mean BMI for each group).
        - Imputed **Physical (days)** and **Mental (days)** with **median**.
        - Formula: df[col].fillna(df[col].mode()/median())

2. **Outlier Handling**:

        - **Age**: Values less than 0 or greater than 100 were set to NaN, as these are biologically implausible ages.
        - **Smoker**: Values outside the range [0, 1] were set to NaN.
        - **Fruits and Physical (days)**: Values outside the range [0, 1] for **Fruits** and [0, 31] for **Mental (days)** and **Physical (days)** were set to NaN.
        - **BMI**: Values outside the range [0, 60] were set to NaN, as BMI values beyond this range are considered outliers.

3. **Encoding**:

        - Ordidal variables like **GeneralHealth, Age**, **Diabetes** and so on were encoded using **Label Encoding and manual Dictionaries**.

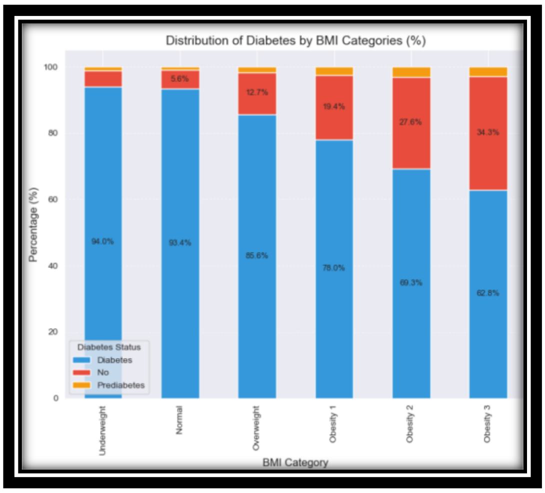
4. **Key Reasons**:

        - **Reliability** – Dropped highly missing columns (>50%) and ensured a complete target variable (**Diabetes**).
        - **Bias Prevention** – Used mode/median imputation to retain data structure.
        - **Outlier Removal** – Capped **Age, BMI, Mental/Physical days** to valid human ranges.
        - **Interpretability** – Encoded categorical variables and categorized **Income & BMI** for better analysis.
        - **Data Integrity** – Filled missing values to prevent data loss and improve model performance.

5. **Tools Used**: Pandas (drop, fillna, map), NumPy (np.nan), Scikit-learn (LabelEncoder)

4. Exploratory Data Analysis (EDA)

Figure 1: Distribution of Diabetes by BMI Categories

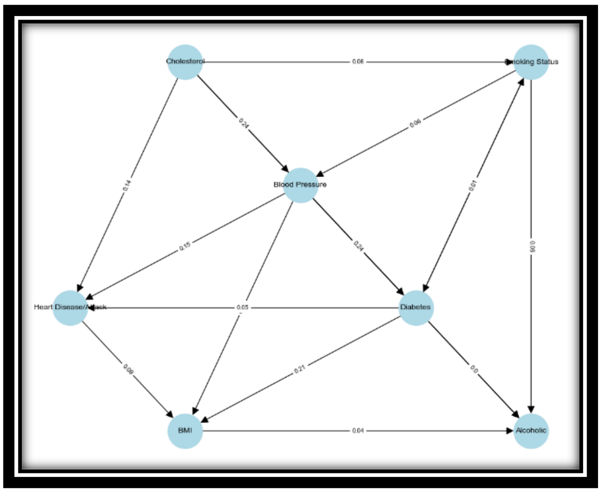


Insights:

Risk Factor Definition: Risk is defined as having either Prediabetes or Diabetes, categorized by BMI.

- **BMI Group Analysis:**
- Underweight/Normal: ~94% have no diabetes/prediabetes.
- Overweight: 12.7% have prediabetes/diabetes.
- Obesity 3: 34.3% have diabetes.
- **Trend:** Diabetes and prediabetes prevalence increases with rising BMI.
- **Interpretation:** BMI categorizes populations into distinct diabetes risk tiers, with higher BMI significantly increasing the likelihood of diabetes and prediabetes.

Figure 2: Relationship Between Blood Pressure and Diabetes Prevalence



Insights:

- Strong associations are observed between Blood Pressure and Diabetes (0.24), Blood Pressure and Cholesterol (0.24), and BMI and Diabetes (0.21).
- Moderate associations exist between Blood Pressure and Heart Disease/Attack (0.15), and Blood Pressure and BMI (0.18).
- Alcoholic consumption shows relatively weak associations with other factors.

Member 2: Unikey- ysab0639

1. Topic and Research Question

Problem Description: The primary task is to predict the forest cover type (the predominant kind of tree cover) from cartographic variables. Our task is to classify test samples by cover type (i.e. this is a multi-class classification task).
There are seven possible forest cover types:
1. Spruce/Fir
2. Lodgepole Pine
3. Ponderosa Pine
4. Cottonwood/Willow
5. Aspen
6. Douglas-fir
7. Krummholz

Research Question:
How topological factors such as elevation, distance to water sources, etc. influence the type of forest cover in a given patch of land.

Business/ Research Need:
With climate change and deforestation on the rise it is important to raise conservation and rehabilitation efforts. With predictive models like this we can identify types of forest cover in a region based on factors like elevation, slope and soil types. This study can help a lot of key people to make impactful decisions to speed up the effort from conceptualisation to implementation.

Stakeholder Impact:

- **Forest Department/Ministry of Environment:** To monitor the forests and watch out of anomalies.
- **Researchers:** Using this model, they can help the government to find sites for rehabilitation of lost forests based on topology we can restore the exact type of forest lost.
- **Agricultural and land use experts**: To asses the quality of land and to educate people about sustainable farming practices near forested areas so it can be done without harming the local environment.

2. Data Description

Dataset Overview:

- Source: Forest Cover dataset (forest_cover.csv)
- Instances: 30,860 total observations, all of them retained after data ingestion.
- Attributes: 55 features initially 49 retained after cleaning, including terrain characteristics, soil types (34 retained), and hydrological factors.

Key Variables Retained:

- **Elevation**: Elevation above sea level in meters.
- **Aspect**: Direction of the slope faces (Ranging from 0-360 degrees).
- **Slope**: Steepness of terrain (Ranging from 0 to 90 degrees).
- **Soil Type:** (40 types) 34 soil types retained for soil classification.
- **Wilderness Areas**: One-hot encoded values of Neota, Rawah, Comanche Peak, and Cache la Poudre
- **Forest Cover** (Target Variable): Categorical variable (1-7) indicating the type of forest cover.

3. Data Ingestion and Cleaning

Data Ingestion

- **Tool**: Pandas DataFrame, Numpy
- **Structure**: The data is in CSV format, consisting of demographic, clinical, and behavioral variables. The dataset includes multiple columns with both numerical and categorical data.

Data Quality Assurance and Cleaning

Initial Steps: Dropped columns with no significant data that would contribute meaningfully towards building a reliable algorithm.

Dropped columns: Soil_Type7, Soil_Type8, Soil_Type36Soil_Type37, Soil_Type15, Soil_Type25

Handling Missing Values, **Numerical**:

- Imputed using **mode** for columns Soil_Type1 to Soil_Type40
- Filled slope, elevation, aspect, hillshade_9am to hillshade_3pm, Horizontal_Distance_To_Hydrology, Vertical_Distance_To_Hydrology, Horizontal_Distance_To_Roadways, Horizontal_Distance_To_Fire_Points with **mean**.

Outlier Handling:

- Slope- Capped from 0 to 90, source: https://pro.arcgis.com/en/pro-app/latest/help/analysis/raster-functions/slope-function.htm
- Aspect- Capped from 0 to 360, source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/how-aspect-works.htm
- Elevation- Capped from 0 to 4000 source: https://lpdaac.usgs.gov/resources/data-action/role-terrain-data/
- Hillshade- Capped from 0 to 255, source https://www.usu.edu/geospatial/tutorials/introductory/raster-data

Encoding: Ordidal variables like **Forest_Cover,** split them into unique individual columns.

Key Reasons:

- **Reliability** – Dropped highly missing columns (>50%) and ensured a complete target variable (**Diabetes**).
- **Bias Prevention** – Used mode/median imputation to retain data structure.
- **Outlier Removal** – Capped **Age, BMI, Mental/Physical days** to valid human ranges.
- **Interpretability** – Encoded categorical variables and categorized **Income & BMI** for better analysis.
- **Data Integrity** – Filled missing values to prevent data loss and improve model performance.

Reliability – Dropped highly missing columns and ensured a complete target variable (Forest_Cover).

Bias Prevention – Used mode/median imputation to retain data structure.

Outlier Removal – Capped Hillshade, Aspect, Slope, Elevation to valid scientifically accurate ranges.

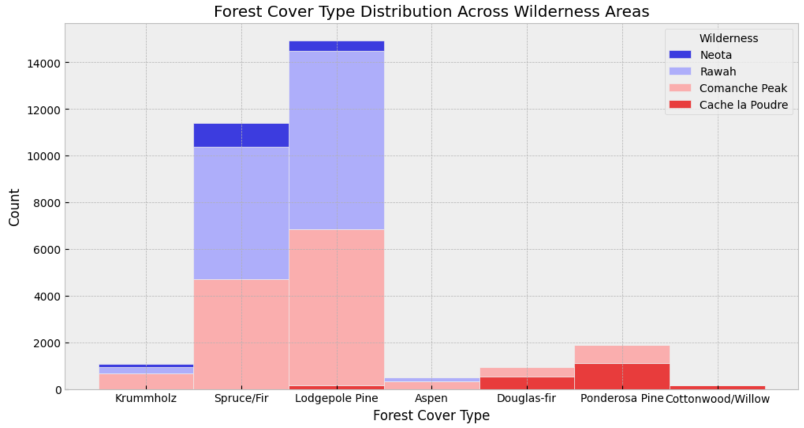
Interpretability – Encoded categorical variables and categorized Forest_Cover for better analysis.

Data Integrity – Filled missing values to prevent data loss and improve model performance.

Tools Used: Pandas (drop, fillna, map), NumPy (np.nan), Scikit-learn (OneHotEncoder).

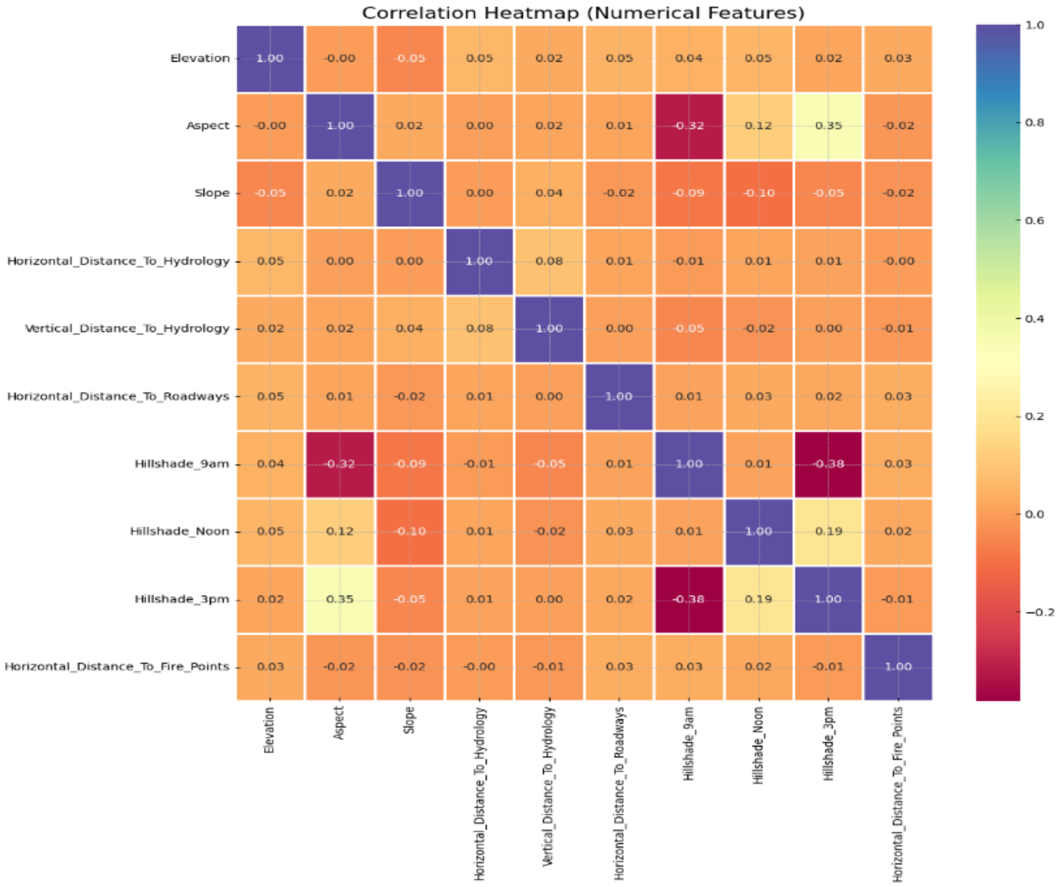
4. Exploratory Data Analysis (EDA)

Figure 1: Forest Cover across wilderness area



Key insights: The stacked histogram reveals that forest cover distribution varies across wilderness areas, with some cover types being more dominant in specific regions. Overlapping colours indicate shared forest types, while distinct colours suggest unique vegetation patterns. Certain forest types are rare, while others are widespread, potentially making "Wilderness Area" a useful predictor in classification models.

Figure 2: Correlation Heatmap of numerical features



Key Insights: The heatmap highlights feature relationships, showing strong correlations that may indicate redundancy and weak correlations that suggest independent variables. High correlations could lead to multicollinearity in models, requiring feature selection or dimensionality reduction. Unique, low-correlation features are valuable for predictive modelling.

Discussion:

The diabetes dataset presents both strengths and limitations when analysed for the research question. One key strength is the dataset's relevance, as it contains essential clinical and demographic variables that are typically associated with diabetes diagnosis and progression. Additionally, the dataset offers a sufficient sample size for meaningful statistical analysis. However, there are significant limitations, primarily due to missing values, including those found in critical predictor variables and even the target column.

The exploratory data analysis (EDA) performed highlighted various patterns and inconsistencies in the dataset. One of the main strengths of the approach was the systematic identification of missing values, enabling the group to determine the extent of data loss. Summary statistics and visualization techniques, such as histograms and boxplots, effectively illustrated the distribution of variables, highlighting possible outliers and inconsistencies. Correlation matrices and pairwise comparisons further helped in understanding variable relationships.

Nevertheless, certain limitations were noted in the EDA approach. First, missing values in predictor variables, such as blood glucose levels and BMI, led to reduced sample sizes in subsequent analyses when imputation techniques could not be properly implemented as imputation of such variables would lead to bias. Additionally, missing values in the target column introduced bias in the classification process, as keeping such rows may lead to an underrepresentation of specific patient groups. Hence the decision was made to remove such rows. Furthermore, the reliance on statistical imputation methods, such as mean or median imputation, may have introduced errors by assuming normality in the data distribution.

The forest cover dataset has a lot of strength in its organized structure and carefully curated columns which are relevant to the research question. In terms of size however it is not exactly enough data in there for any impactful feature engineering or extraction. Another shortcoming was its lack of organization in the dataset, during the cleaning process, I had to go out of my way to ensure all the values given were in the acceptable range so I had to verify it from external sources such as research papers and other institutional websites. Problems like negative values or missing values in key features another major issue in the dataset, although it was handled by imputation it does decrease the quality of the data as imputed data is always prone to under/overfitting and biases.

The key strengths of this dataset included systematically identifying missing values, allowing the group to assess data loss. Summary statistics and visualization techniques, effectively showcased variable distributions, revealing potential outliers and inconsistencies. Additionally, correlation matrices and pairwise comparisons provided deeper insights into variable relationships.

Due to these shortcomings we had to cap all the affected fields to the accepted range and then impute the rest using the mean, this would lead to increase in bias and also result in overfitting or underrepresentation of specific forest types. Additionally, using statistical imputation methods like mean or median imputation may have introduced errors by incorrectly assuming a normal data distribution.

Conclusion:

The exploratory data analysis (EDA) conducted on the diabetes dataset yielded crucial insights into its structure and quality. The most important outcome of the EDA was the identification of significant missing values across various predictor variables and, critically, within the target column. This posed challenges in building a reliable model, as missing target values required careful handling to prevent bias. Furthermore, data cleaning and preprocessing, including outlier detection and imputation strategies, were instrumental in improving the dataset's usability.

The cleaned dataset and the outcomes of EDA will help answer the research question by ensuring that the data is more reliable and representative of the population. By addressing missing values and normalizing data distributions, we can improve the quality of statistical analyses and predictive modelling. The chosen approach to handling missing data—through appropriate imputation techniques and potential removal of highly incomplete records—aims to strike a balance between maintaining sample size and data integrity.

After discussions and evaluations, our group has agreed to proceed with the cleaned diabetes dataset for Project Stage 2. This decision is justified as the dataset, despite its challenges, remains highly relevant to the research question. Moreover, by applying refined preprocessing techniques, such as advanced imputation and feature engineering, we can enhance the dataset's predictive power and ensure more reliable outcomes in subsequent modelling efforts. When we used XGBoost for feature importance, our diabetes dataset achieved an accuracy of 83%, whereas the forest fire dataset only achieved 55%. This further reinforced our decision to move forward with the diabetes dataset, as it demonstrated superior predictive potential and reliability for our research objectives.

Appendix: Diabetes Dataset

Data Dictionary-

Column Name	Description	Data Type	Encoding/ Categories
CholCheck	Cholesterol Check Status	float	0 = No, 1 = Yes
BMI	Body Mass Index	float	Continuous value (e.g., 18.5, 24.9, etc.)
Smoker	Smoking Status	float	0 = No, 1 = Yes
Stroke	History of Stroke	float	0 = No, 1 = Yes
HeartDiseaseOrAttack	History of Heart Disease or Heart Attack	float	0 = No, 1 = Yes
GeneralHealth	Self-reported health rating (Ordinal)	object	Poor, Fair, Good, Very Good, Excellent
Mental (days)	Days of mental health issues in the past 30 days	float	Continuous value (e.g., 0, 1, 5, etc.)
Physical (days)	Days of physical health issues in the past 30 days	float	Continuous value (e.g., 0, 1, 5, etc.)
DiffWalk	Difficulty walking	float	0 = No, 1 = Yes
Sex	Gender of the patient	object	Male, Female, Not Specified
Age	Age of the patient in years	float	Continuous value (e.g., 18, 24, 52, etc.)
Education	Education level	object	College graduate, some college, High school graduate, some high school, Elementary, never attended school
Income	Annual income range	object	0 = 0-20k, 1 = 20k-40k, 2 = 40k-60k, 3 = 60k-80k, 4 = 80k-100k, 5 = 100k-120k, 6 = 120k-140k, 7 = 140k+
BloodPressure	Blood pressure status (Normal/High)	object	0- Normal, 1- High
Cholesterol	Cholesterol level status (Normal/High)	object	0- Normal, 1- High
Alcoholic	Alcohol consumption status (Yes/No)	object	Yes, No
Diabetes	Diabetes diagnosis status	object	No, Prediabetes, Diabetes

Derived Variable:

Derived variables were created using one-hot encoding and custom categorization. These include:

- **NewBMI** categories: Underweight, Normal, Overweight, Obesity 1, Obesity 2, Obesity 3 (based on BMI).
- **GeneralHealth** categories: Poor, Fair, Good, Very Good, Excellent (based on self-reported health).
- **Age Group** categories: Child (0-12), Teen (12-18), Young Adult (18-35), Adult (35-50), Middle-Aged (50-65), Senior (65-80), Elderly (80+).
- **Income Classes categories:** 0-20k, 20k-40k, 40k-60k, 60k-80k, 80k-100k, 100k-120k, 120k-140k, 140k+

These were included to enhance model performance and insights.

Appendix: Forest Cover

Column Name	Data Description	Data Type	Encoding if applicable
Elevation	Height above the sea level	int	NA as it is numerical data
Aspect	Direction of slope faces	int	NA as it is numerical data
Slope	Angle of steepness of the terrain	int	NA as it is numerical data
Horizontal_Distance_To_Hydrology Vertical_Distance_To_Hydrology	Distance to nearest source of water	int	NA as it is numerical data
Horizontal_Distance_To_Fire_Points	Distance to nearest fire place	int	NA as it is numerical data
Horizontal_Distance_To_Roadways	Distance to nearest road	int	NA as it is numerical data
Hillshade_9am Hillshade_Noon Hillshade_3pm	Area covered by the sunlight at different times through the day	int	NA as it is numerical data
Forest_Cover	Types of forests	int	One hot encoding used, now split according to its unique values i.e., 7 separate columns
Neota Rawah Comanche Park Cache la Poudre	Wilderness Areas	bool	NA
Forest_Cover_		object	NA
Forest_Cover_Aspen Forest_Cover_Cottonwood/Willow Forest_Cover_Dougla-fir Forest_Cover_Krummholz Forest_Cover_Lodgepole Pine Forest_Cover_Ponderosa Pine Forest_Cover_Spruce fir	Unique Types of Forest Cover	bool	1= yes 0= no