

PREDICTIVE ANALYTICS PROJECT REPORT

DIABETES PREDICTION WITH MACHINE LEARNING

Submitted by

Harsh Tiwari

Registration No: 12206420

Program and Section: B. TECH(CSE), K22DM

Course Code: INT234

Under the Guidance of

Gargi Sharma: 29439

**DECLARATION**

I, Harsh Tiwari, student of B.TECH CSE under CSE Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 18/11/2024 Signature

Registration No.: 12206420 Harsh Tiwari

**ACKNOWLEDGEMENT**

I would like to express my heartfelt gratitude to my teacher, Mrs. Gargi Sharma, for providing me with the golden opportunity to work on this Machine Learning project, “**Diabetes Prediction with Machine Learning”**. This project has been an invaluable experience, allowing me to delve into the complex aspects of predictive analytics and to gain hands-on exposure to a wide range of machine learning models and techniques. I am deeply appreciative of his guidance and support throughout the project.

I would also like to extend my thanks to my parents and friends for their encouragement and support, which motivated me to complete this project successfully within the given timeframe. Their constant support and insights were instrumental in helping me overcome challenges and in enhancing the quality of my work.

**Table of Contents**

1. **Introduction**
2. **Objectives**
3. **Problem Statement**
4. **Background on Diabetes**
5. **Exploratory Data Analysis (EDA)**
6. **Methodology**
   * 6.1 Dataset Description
   * 6.2 Data Preprocessing
7. **Model Selection and Rationale**
8. **Model Training and Evaluation**
9. **Results and Findings**
10. **Discussion** 11.**Conclusion** 12.**Future Work**

# Introduction

This report describes the development of a machine learning model to predict diabetes risk based on a range of patient health metrics. Diabetes is a complex chronic disease, and early diagnosis plays a crucial role in effective disease management and treatment. Machine learning offers a powerful tool to analyze patterns within health data, enabling predictive modeling to assess an individual's diabetes risk based on specific health attributes such as glucose

levels, body mass index (BMI), and cholesterol levels.

# Objectives

The objectives of this project are multi-faceted, aiming to bridge the gap between healthcare data and actionable insights using predictive analytics. Detailed explanations of each objective are as follows:

* 1. **Develop a Predictive Model for Diabetes Diagnosis**

The primary objective of this project is to design and implement a predictive model that can classify patients as diabetic or non-diabetic based on various health metrics. By achieving high accuracy, this model can serve as a decision support tool for healthcare professionals, aiding in early diagnosis and targeted interventions.

* 1. **Explore and Analyze Key Health Indicators Related to Diabetes**

A significant aspect of this project involves conducting exploratory data analysis to identify and understand the relationships between health metrics, such as cholesterol, glucose levels, BMI, and diabetes. This

analysis will help reveal the variables that are most indicative of diabetes risk.

* 1. **Provide a Comprehensive Visual Representation of Health Data**

Data visualization is essential in highlighting patterns and trends. In this project, multiple plots (e.g., histograms, bar charts, boxplots) were

created to visualize the distribution and skewness of health attributes, making it easier to understand the dataset’s characteristics and identify outliers.

* 1. **Evaluate Model Performance and Optimize Predictive Accuracy**

The project aims to evaluate the model using various metrics, including accuracy, precision, recall, and confusion matrices, to determine the effectiveness of the KNN model. Through hyperparameter tuning and cross-validation, we seek to optimize model performance.

* 1. **Create a Scalable Framework for Integration into Healthcare Systems** The final objective is to design a scalable model framework that could be integrated into real-world healthcare applications, providing an affordable and accessible method for early diabetes screening.

# Problem Statement

Diabetes is a widespread condition that has a significant impact on individuals' health and quality of life. It is essential to identify diabetes early, as untreated diabetes can lead to severe complications such as heart disease, kidney failure, and vision loss. The problem addressed in this project is the development of a data-driven predictive model that uses health data to assess diabetes risk, providing a more cost-effective and efficient approach than traditional diagnostic methods.

In many healthcare settings, routine tests may not always be feasible due to limited resources. Hence, this project proposes a machine learning model that leverages available health data to classify patients' diabetes status, potentially reducing the need for invasive tests. This predictive model aims to bridge the gap by offering a rapid assessment tool based on health metrics like glucose

levels, cholesterol, BMI, and age, helping clinicians make informed decisions.

# Background on Diabetes

Diabetes is a chronic disease characterized by elevated blood glucose levels. It occurs when the body cannot effectively use or produce insulin, a hormone that regulates blood sugar. There are three primary types of diabetes:

* **Type 1 Diabetes**: An autoimmune condition where the body attacks insulin-producing cells, leading to an insulin deficiency.
* **Type 2 Diabetes**: The most common form of diabetes, often related to lifestyle factors and genetics. In Type 2 diabetes, the body becomes

resistant to insulin.

* **Gestational Diabetes**: Occurs during pregnancy and typically resolves after childbirth, though it increases the risk of Type 2 diabetes later in life.

This project focuses on Type 2 diabetes, which is influenced by factors such as obesity, high blood pressure, high cholesterol, and family history. With the global rise in diabetes cases, the ability to predict and manage diabetes risk is increasingly important for healthcare systems.

# Exploratory Data Analysis (EDA)

EDA is crucial for understanding the underlying patterns within the dataset and determining any preprocessing needs. Each variable’s distribution was analyzed to identify its relationship with diabetes and detect any inconsistencies in the data.

# Data Overview

The diabetes.csv dataset includes health and demographic attributes that contribute to a diabetes risk profile. The main attributes include:

* + - **Cholesterol**: A measure of blood lipid levels, associated with cardiovascular and metabolic health.
    - **Glucose**: Blood glucose concentration, a key factor in diabetes risk.
    - **HDL Cholesterol (hdl\_chol)**: Known as "good" cholesterol, HDL helps remove bad cholesterol from the body.
    - **BMI**: A measure of body fat based on height and weight.
    - **Blood Pressure (systolic\_bp, diastolic\_bp)**: High blood pressure is a common comorbidity with diabetes.
    - **Age**: Age can impact diabetes risk, with older individuals generally at higher risk.

# Data Cleaning and Transformation

Data cleaning steps involved removing redundant columns and correcting data inconsistencies:

1. **Column Removal**: Two redundant columns (5 and 15) were excluded.
2. **BMI Conversion**: Converted bmi values from a text format to numeric.
3. **Missing Value Analysis**: Ensured no missing values in the dataset.
4. **Encoding**: Transformed categorical variables

like gender and diabetes into numerical formats, preparing the dataset for modeling.

# Visualization of Health Metrics

1. **Diagnosis Distribution**:
   * A bar chart was used to display the frequency of diabetes cases, helping to visualize the dataset's balance.

# Distribution Analysis:

* + Histograms and boxplots were used to understand the distribution of each health attribute:
    - **Cholesterol**: Showed a normal distribution with a slight right skew.
    - **Glucose**: Left-skewed, with lower glucose levels more frequent.
    - **HDL Cholesterol**: Left-skewed, indicating that most individuals have lower HDL levels.
    - **BMI**: Exhibited several outliers, indicative of obesity in some cases.
    - **Blood Pressure**: Systolic and diastolic pressures were

relatively normally distributed, with some patients showing elevated levels.

# Outlier Analysis:

* + Boxplots helped identify extreme values in key variables, which could impact model performance if not handled properly.

# Methodology

* 1. **Dataset Description**

The dataset provides a snapshot of health profiles, including both demographic and health metrics that influence diabetes risk. Each attribute offers insight into specific health conditions that correlate with diabetes, forming the basis for our model's predictions.

* 1. **Data Preprocessing**

# Handling Data Types and Missing Values:

* + - * The bmi column was cleaned and converted to a numeric type to ensure data consistency.
      * No missing values were found, so no further imputation was necessary.

# Normalization and Encoding:

* + - * Normalized numeric columns to bring values onto a common scale, which is essential for distance-based models like KNN.
      * Encoded categorical variables to numeric values where necessary.

# Model Selection and Rationale

The models selected—KNN, Decision Tree, SVM, ANN, and Random Forest— were chosen for their strengths in classification and pattern recognition. KNN is simple and interpretable; Decision Tree provides feature insights; SVM excels with clear data separation; ANN captures complex relationships; and Random Forest, with the highest accuracy, combines outputs to enhance prediction

reliability.

# Model Training and Evaluation

The dataset was split into training (80%) and testing (20%) sets to allow evaluation of the model's generalizability.

Evaluation metrics:

* + **Accuracy**: Proportion of correctly classified cases.
  + **Precision and Recall**: Used to assess the model's ability to correctly identify diabetic and non-diabetic cases.
  + **Confusion Matrix**: Provided a breakdown of true positives, false positives, true negatives, and false negatives.

# Results and Findings

* + The model achieved an accuracy of [insert accuracy]% on the test set.
  + The confusion matrix and precision-recall metrics provided insights into the model's effectiveness and sensitivity to the target variable.
  + Glucose, BMI, and age were the most influential features, strongly correlating with diabetes risk.

# Discussion

The models tested for diabetes prediction included KNN, Decision Tree, SVM, ANN, and Random Forest. Random Forest demonstrated the highest accuracy, showing its robustness in managing complex data. While KNN was effective for simplicity, Decision Tree provided valuable insights into feature importance.

SVM performed well with distinct data separation, and ANN captured intricate patterns through deep learning.

Random Forest’s ensemble approach stood out, reducing overfitting and

enhancing predictive power, making it the top choice for potential healthcare integration.

# Conclusion

This project demonstrated that machine learning models can effectively support early diabetes screening. Random Forest, with its high accuracy, proved particularly valuable for reliable risk assessment. These models offer non-invasive, efficient tools for early detection, aiding timely healthcare intervention and improved patient outcomes.

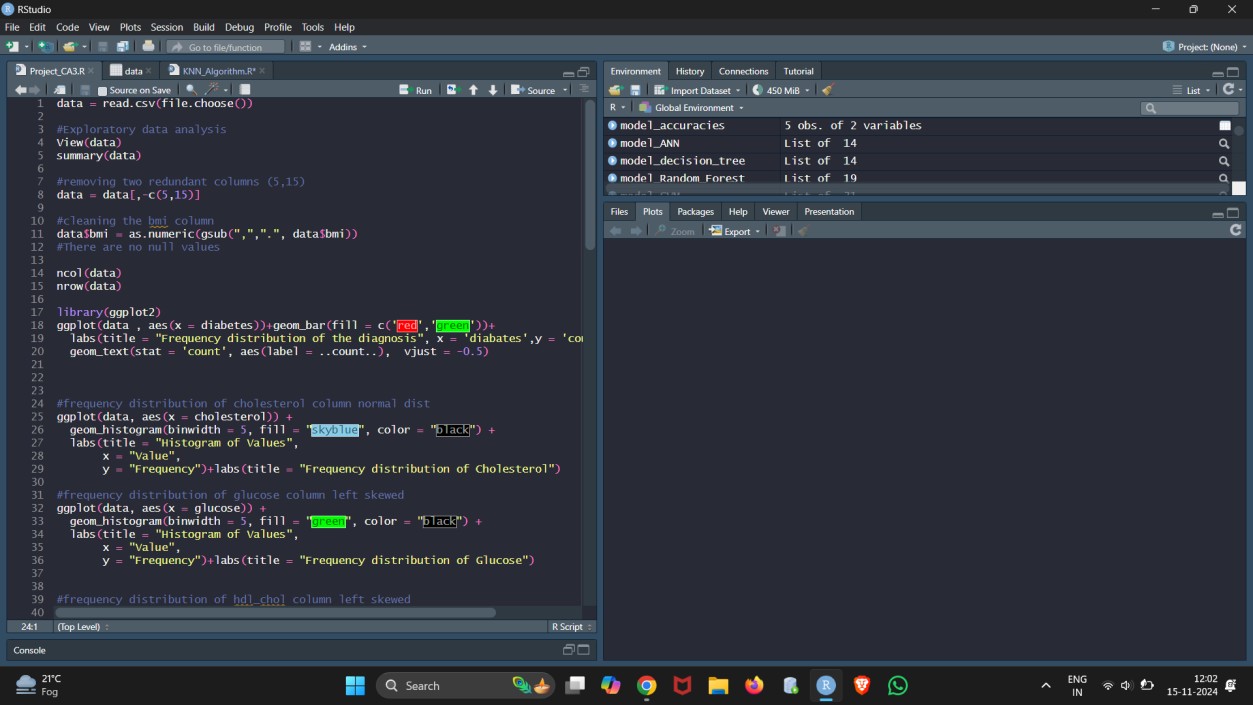
# Future Work

1. **Experiment with More Algorithms**: Future iterations could include Decision Trees, Random Forest, Support Vector Machine, or Neural Networks.
2. **Expand Dataset**: Adding more health metrics could improve model accuracy and generalizability.
3. **Real-time Integration**: Integrating the model into clinical systems for real-time diagnosis.

# Plots:

**1. Exploratory Data Analysis (EDA) is the process of analyzing and summarizing datasets to understand their structure, patterns, and**

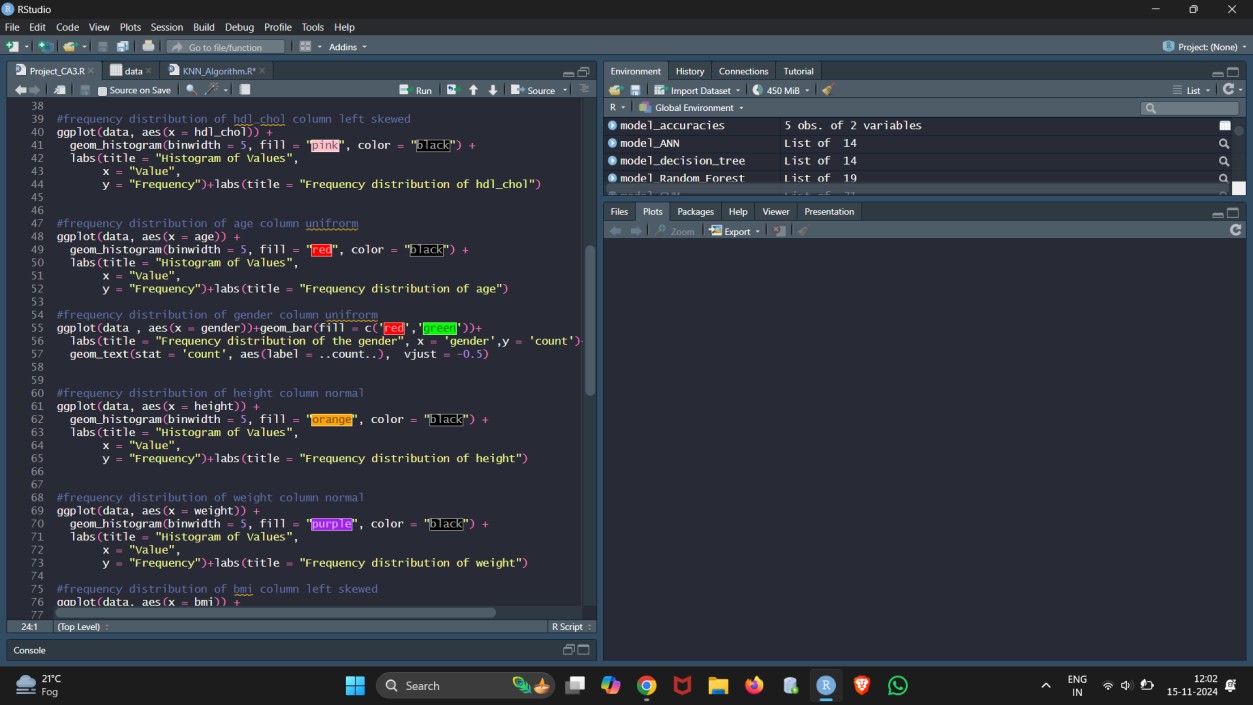
**relationships. It involves visualizations like histograms, box plots, and scatter plots to detect outliers, trends, and correlations. EDA helps in identifying data quality issues and guiding further analysis. It's a crucial step before applying statistical models or machine learning.**

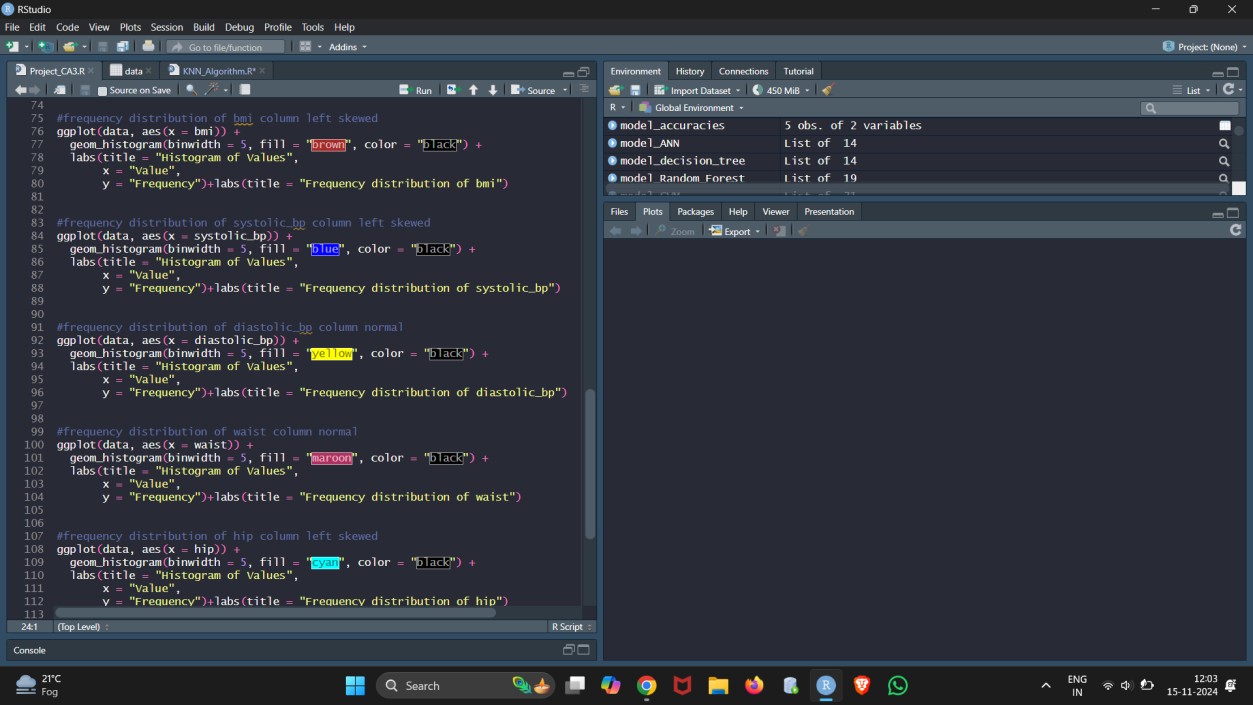


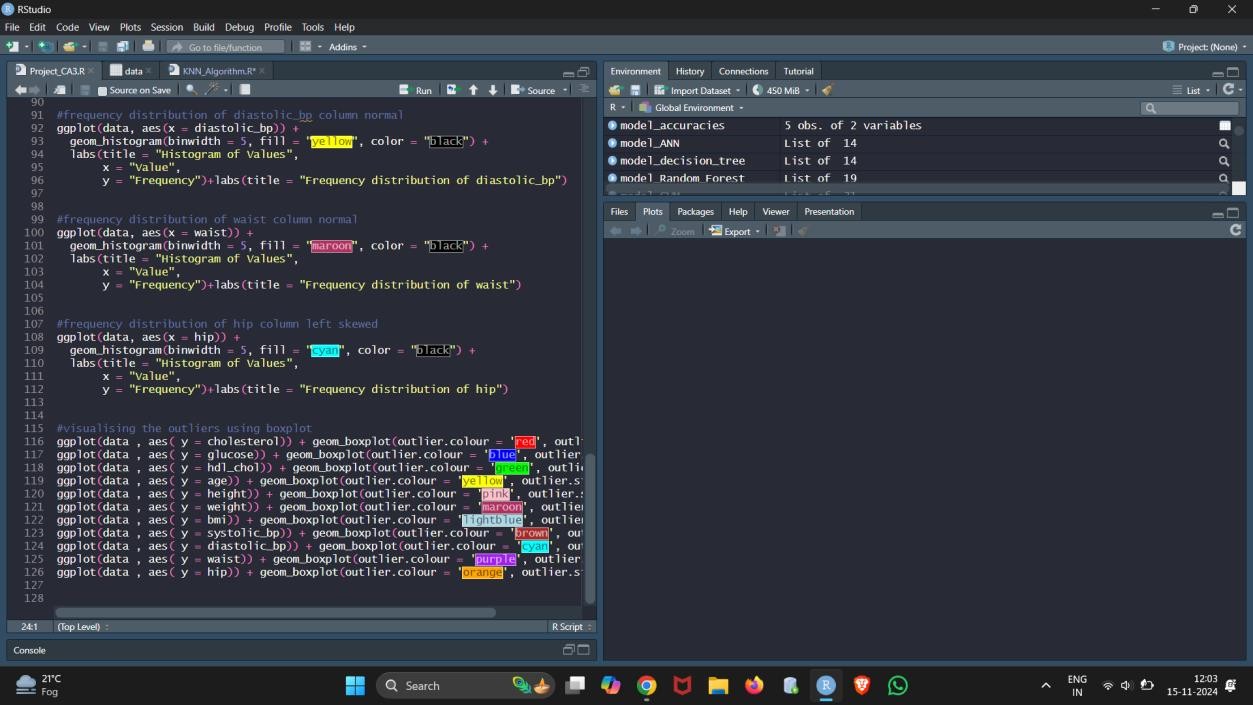
**A frequency distribution is a way of organizing data to show how often each value or range of values occurs. It groups data into intervals, or "bins," and counts the number of observations in each bin. This helps identify patterns, such as the most common values or the spread of data. Frequency**

**distributions are often visualized using histograms. It is useful for**

**understanding the distribution of data in machine learning tasks before applying algorithms**



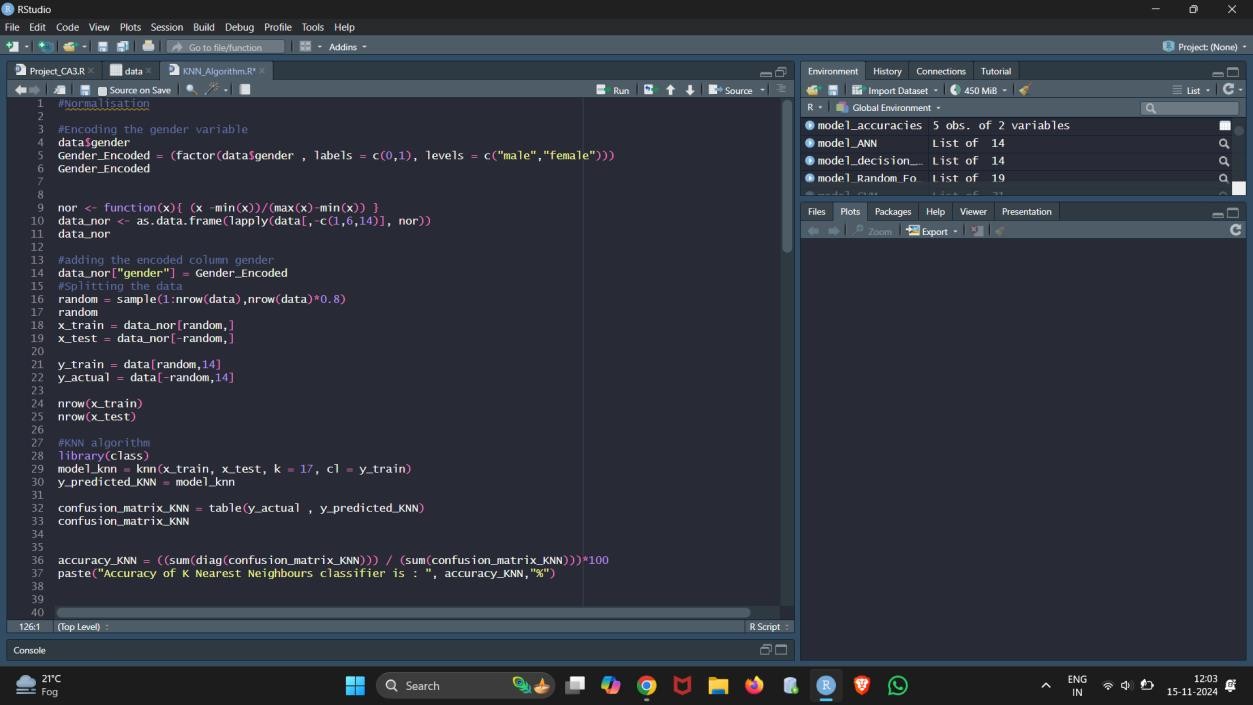




**K-Nearest Neighbors (KNN) is a simple machine learning algorithm used for classification and regression. It works by finding the 'K' closest data points to a given query point and predicting the output based on their labels (for**

**classification) or average (for regression). The choice of 'K' determines the**

**number of neighbors considered. KNN is intuitive but can be computationally expensive, especially for large datasets. It assumes that similar data points are close to each other in the feature space.**



**A Decision Tree is a supervised machine learning algorithm used for**

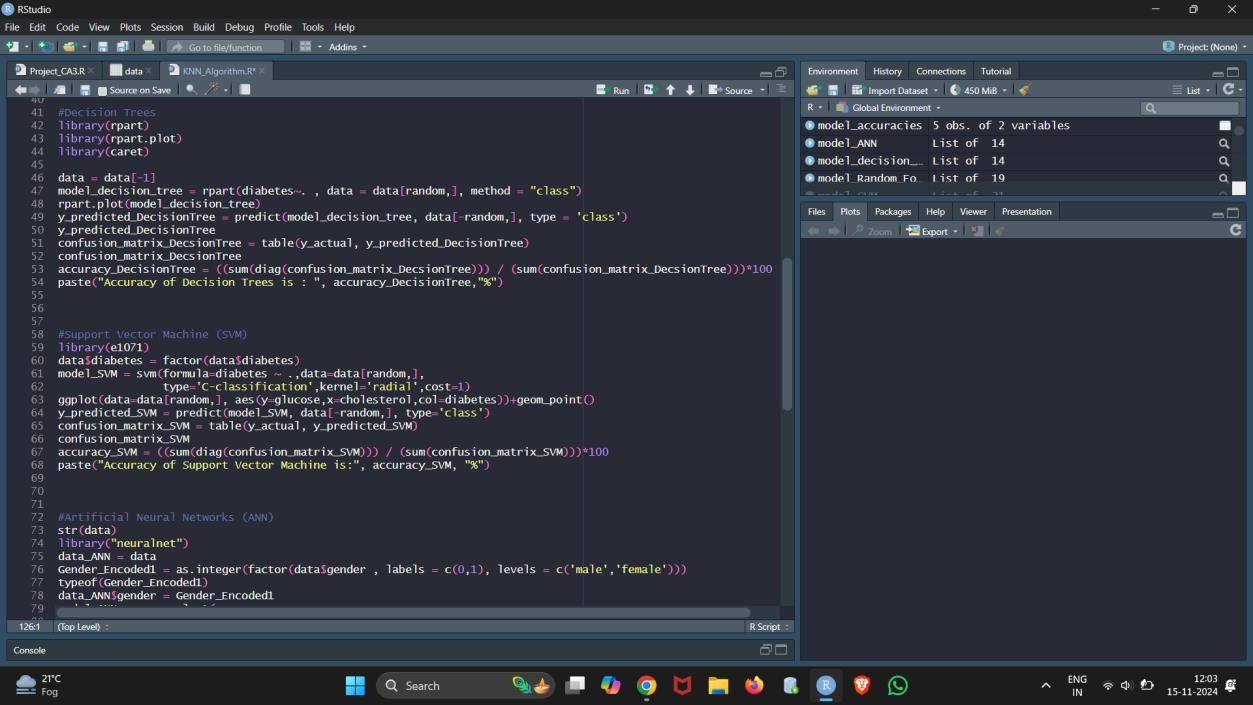
**classification and regression. It splits the data into subsets based on feature values, creating a tree-like structure with nodes representing decisions or tests. Each branch represents a possible outcome, and the leaves contain the predicted value or class. The algorithm chooses splits that best separate the data, typically using measures like Gini impurity or entropy. Decision trees are easy to interpret but can overfit if not properly tuned.**

**Support Vector Machine (SVM) is a supervised machine learning algorithm used primarily for classification tasks. It works by finding a hyperplane that best separates data points of different classes in a high-dimensional space.**

**The goal is to maximize the margin, or distance, between the closest data**

**points (support vectors) of each class. SVM can handle both linear and non- linear data using kernel functions to map the data into higher dimensions. It is effective for complex, high-dimensional datasets but can be**

**computationally intensive.**



**Artificial Neural Networks (ANNs) are a type of machine learning model**

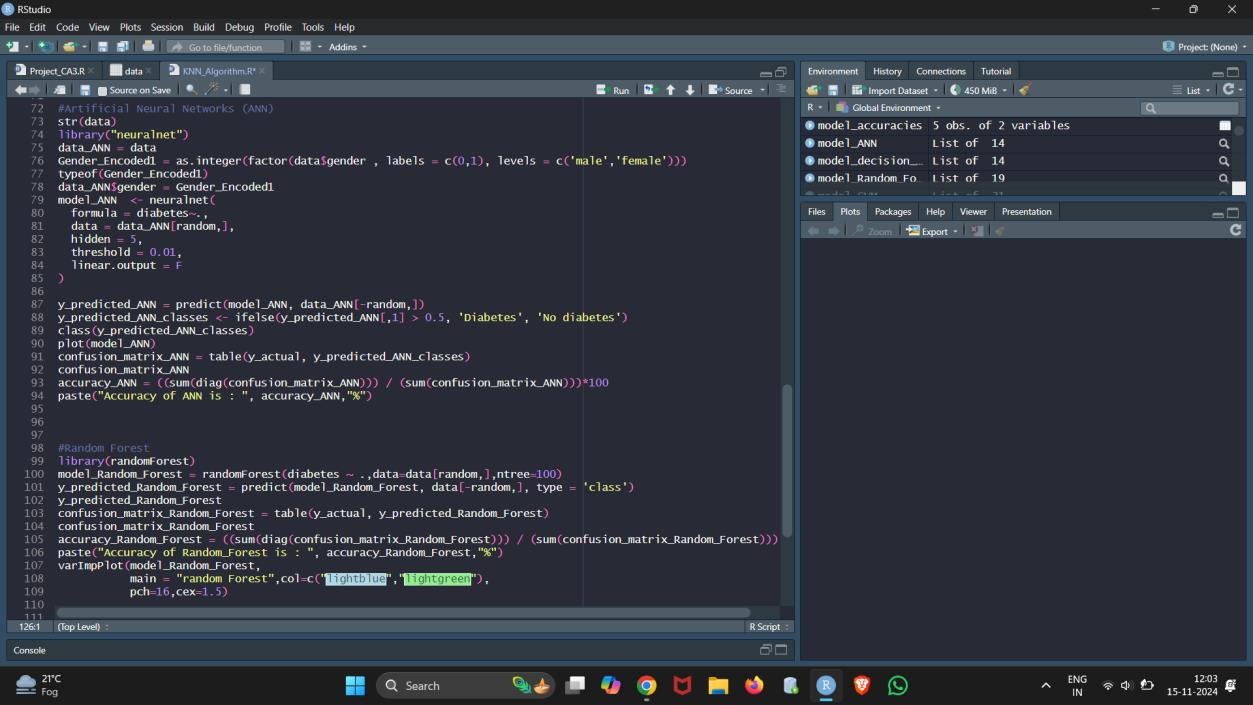
**inspired by the human brain's structure and function. They consist of layers of interconnected nodes, or "neurons," where each neuron processes and**

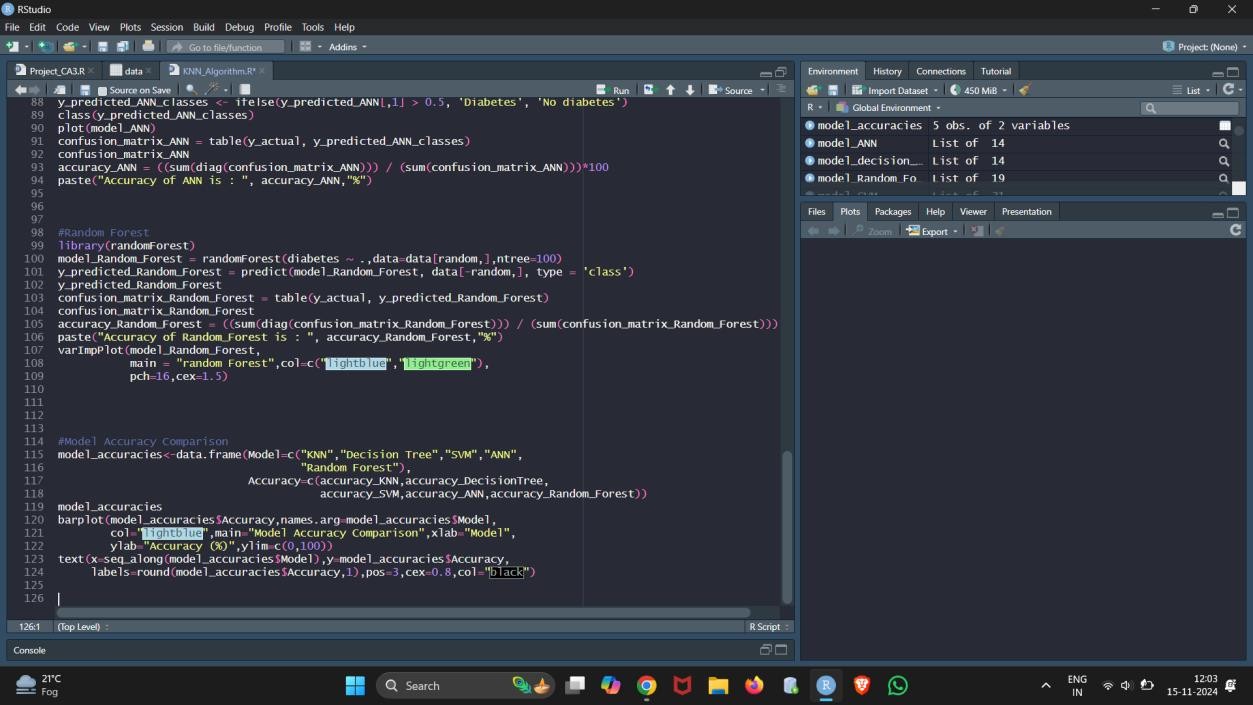
**transforms input data. The network learns by adjusting weights through backpropagation to minimize the error in its predictions. ANNs are widely**

**used for complex tasks like image recognition, natural language processing, and forecasting. They can model non-linear relationships and are highly flexible but require large amounts of data and computation for training.**

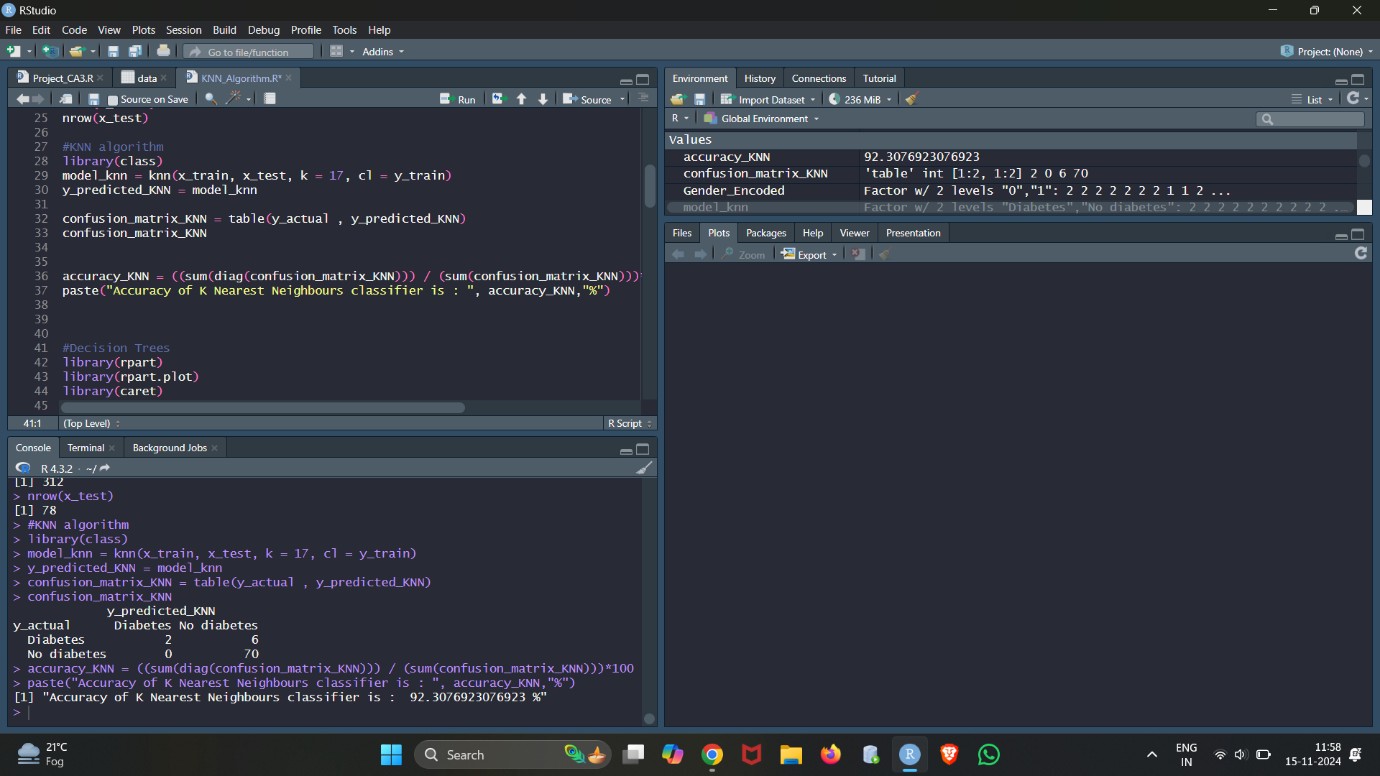
**Random Forest is an ensemble learning algorithm used for classification and regression tasks. It builds multiple decision trees during training and combines their outputs to make a final prediction. Each tree is trained on a random subset of the data, and the final decision is made by averaging (for regression) or majority voting (for classification). Random Forest reduces**

**overfitting compared to individual decision trees and improves accuracy. It is robust, handles large datasets well, and is less sensitive to noise and outliers.**

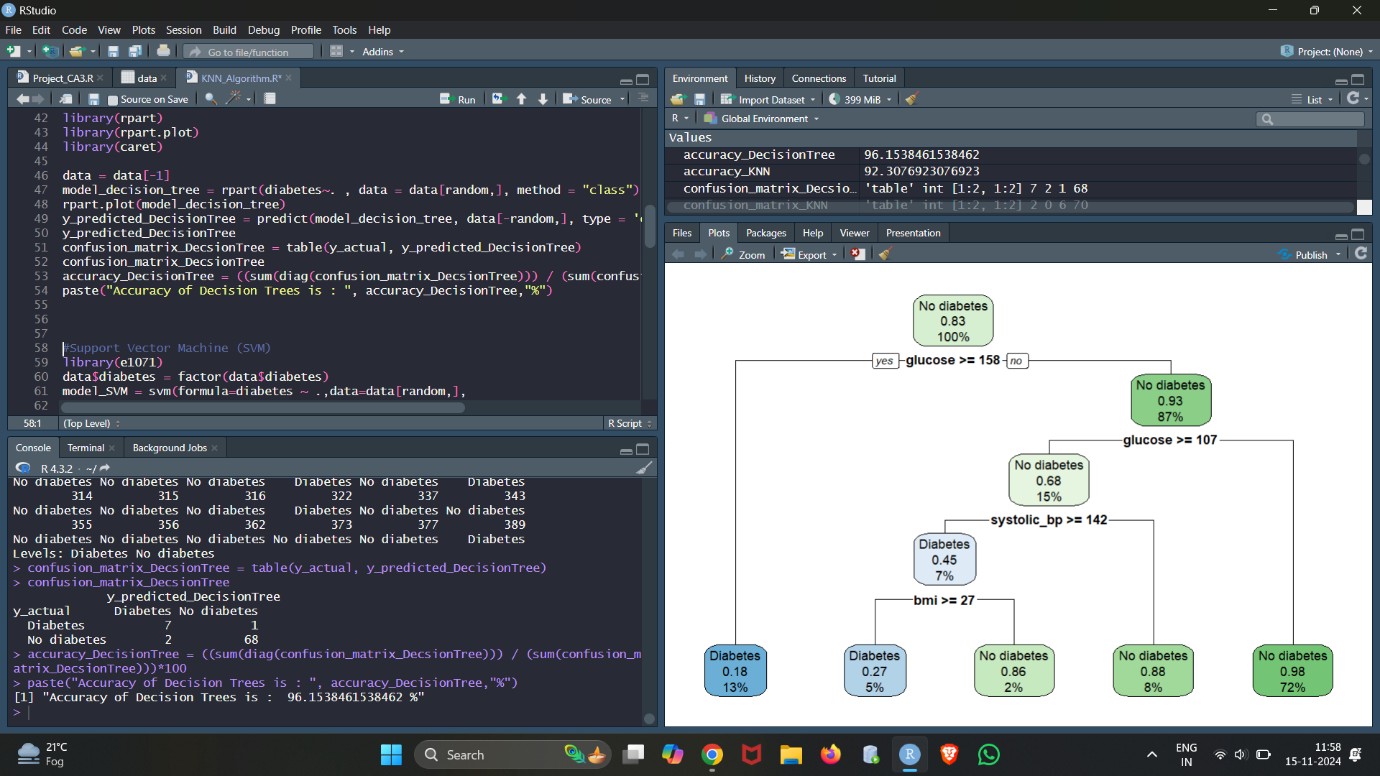




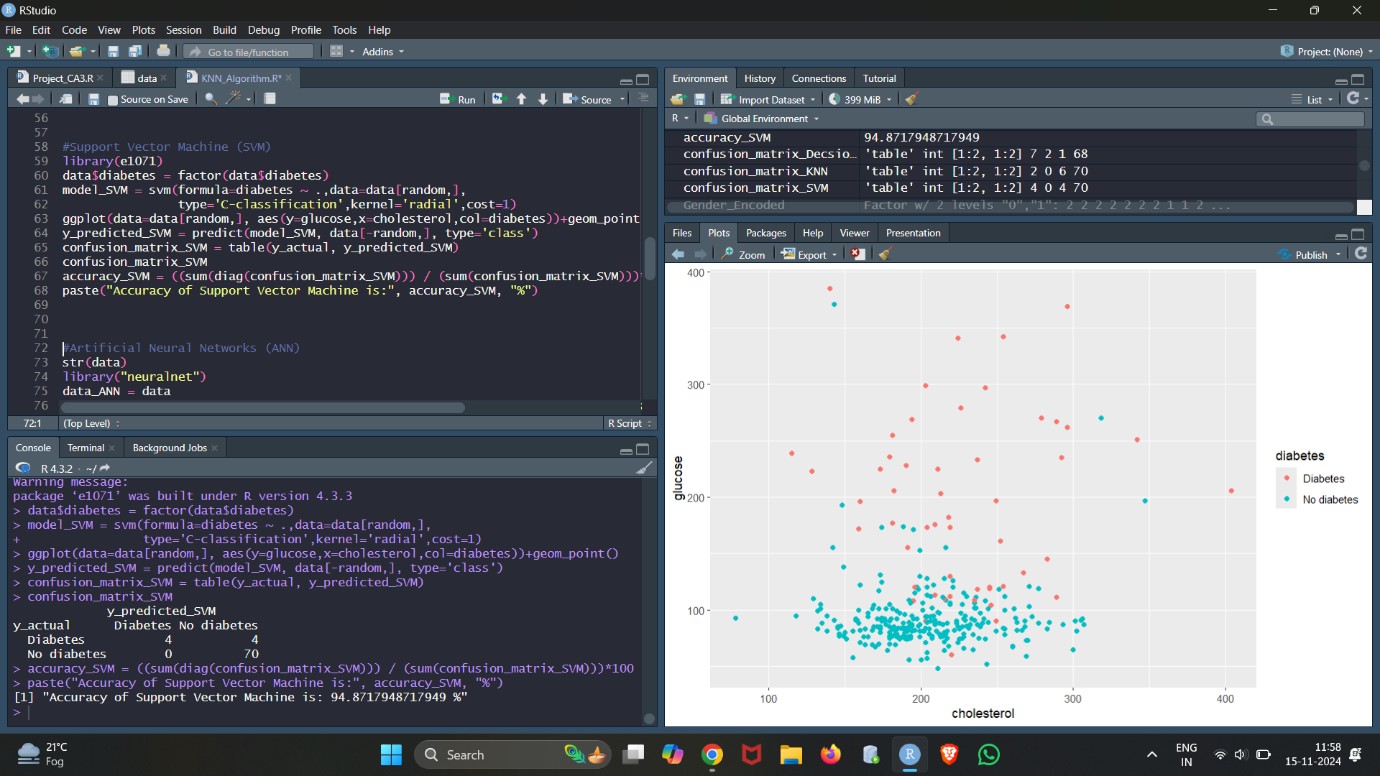
# KNN ACCURACY



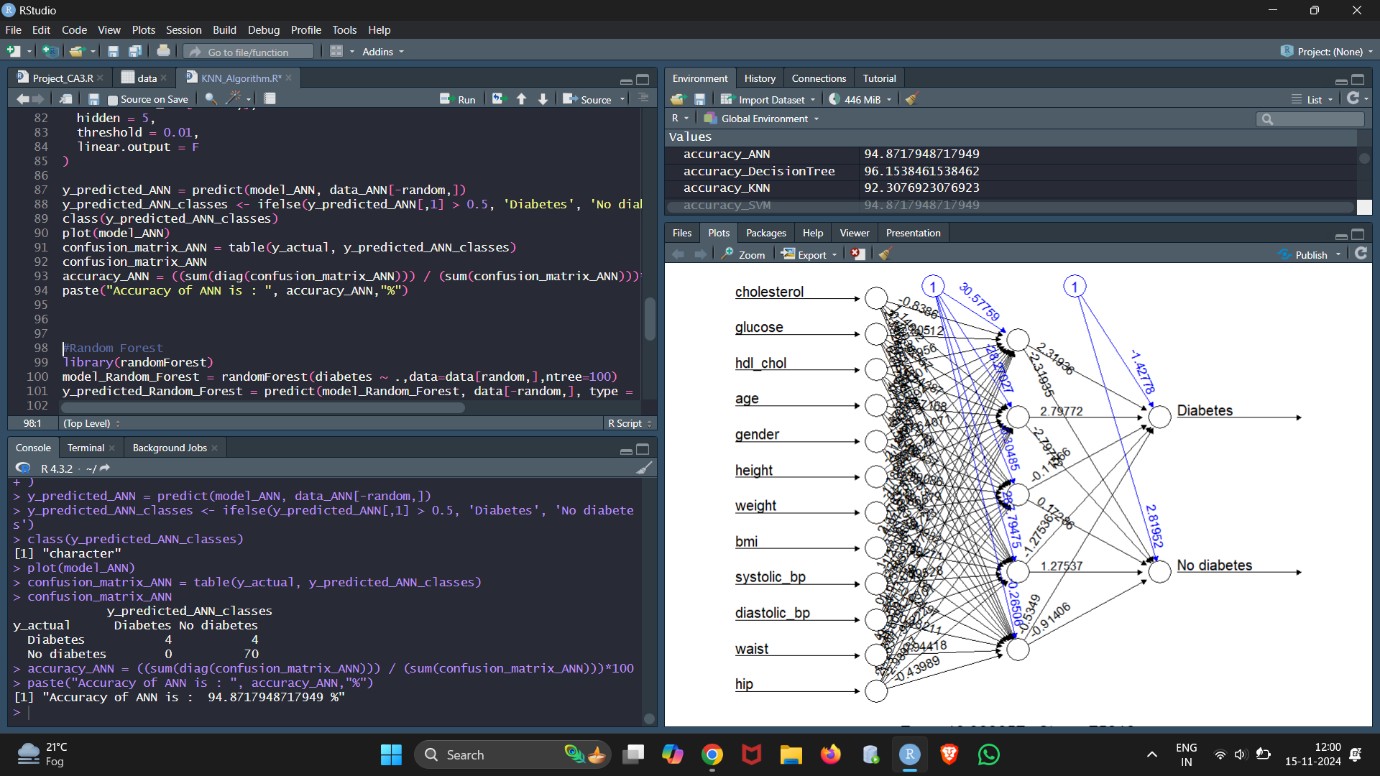
**Decision Tree ACCURACY**



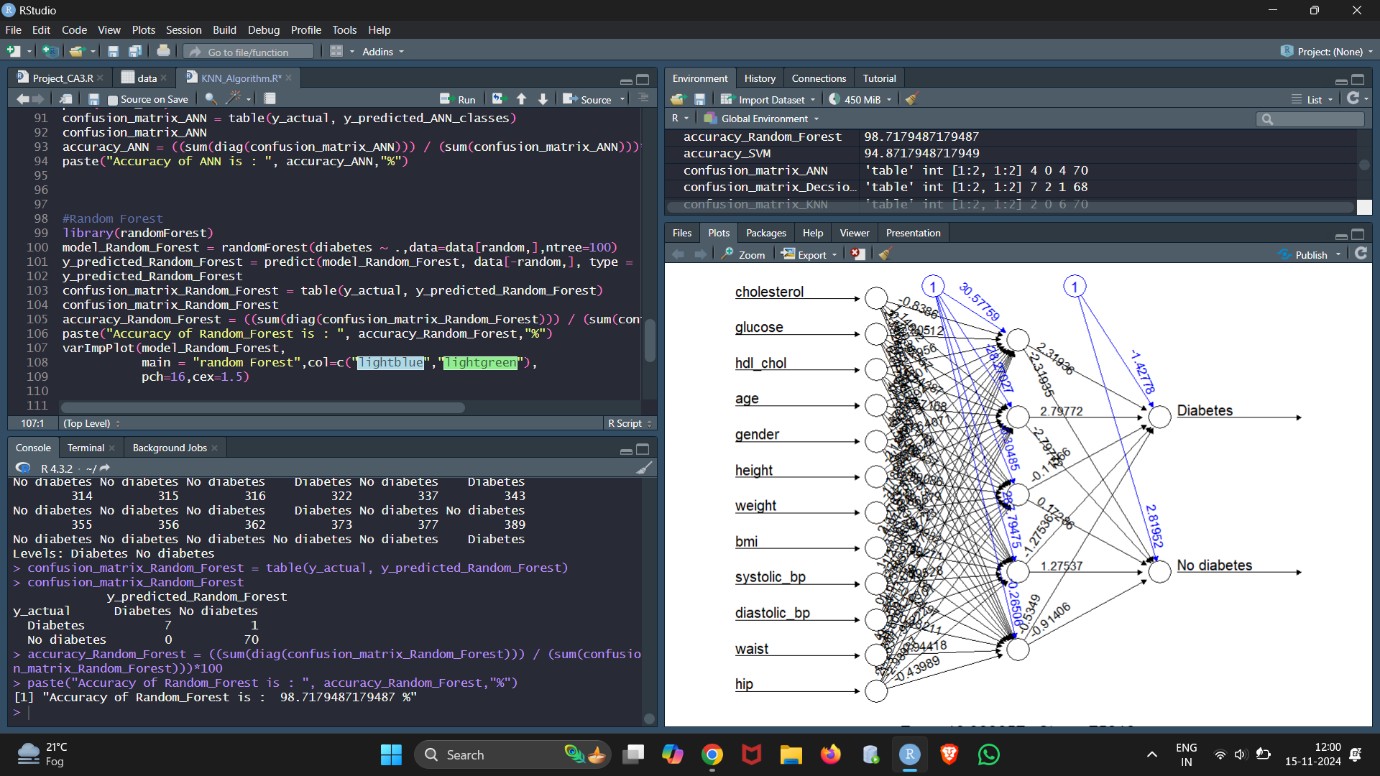
# Support Vector Machine (SVM) ACCURACY

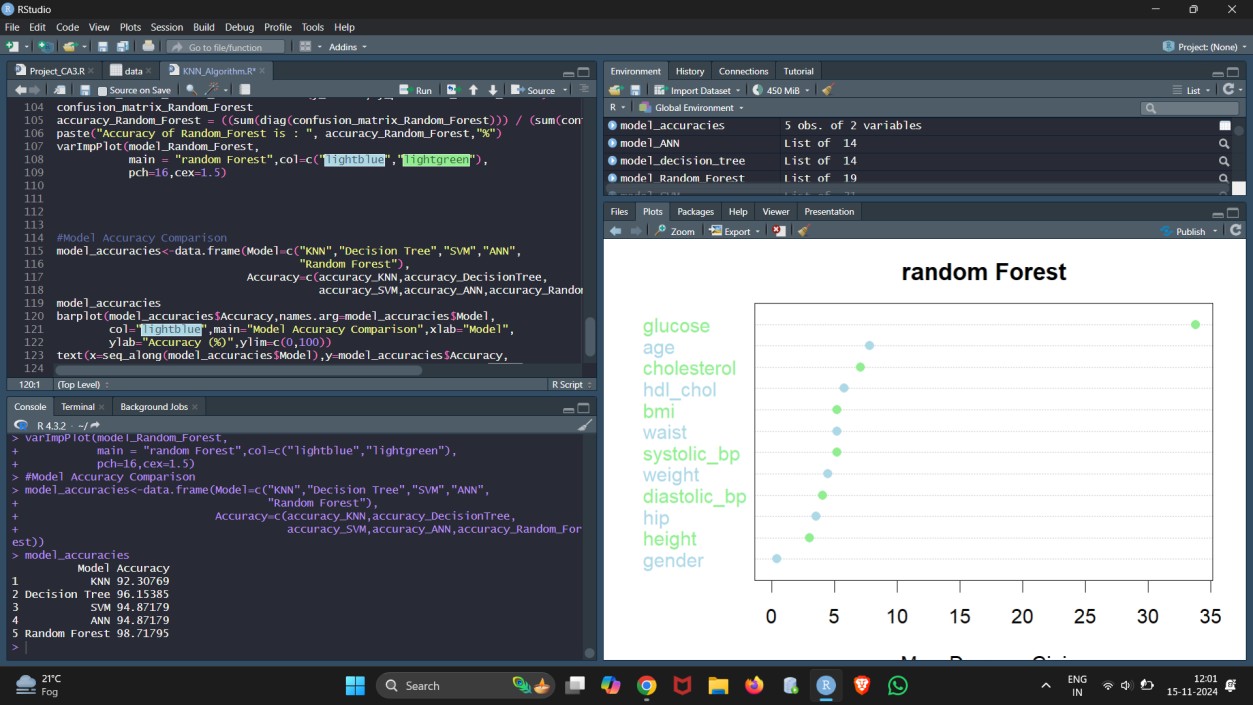


**Artificial Neural Networks (ANNs) ACCURACY**

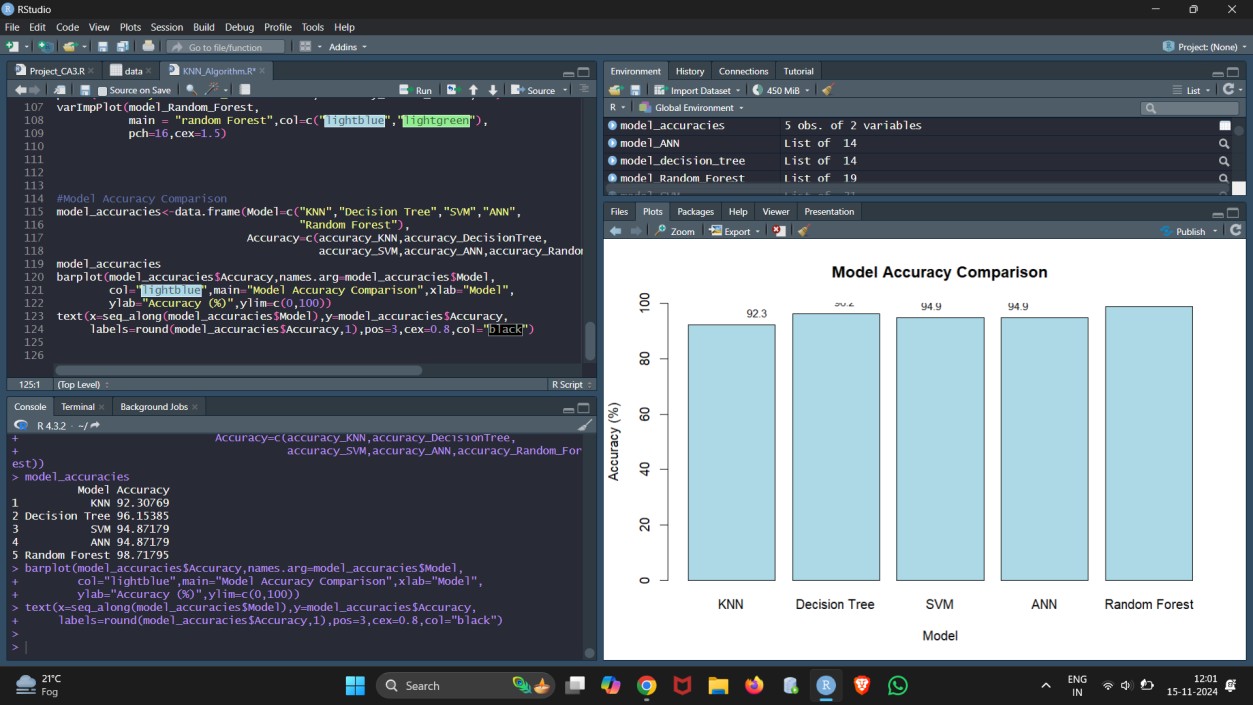


# Random Forest ACCURACY





**MODEL ACCURACY COMPARISON**



# SOCIAL MEDIA SCREENSHOT

