**PREDICTIVE ANALYTICS PROJECT REPORT**

(Project Semester August-December 2024)

***PREDICTION AND CLUSTERING OF INSURANCE CHARGES***

Submitted by

Raman Goyal

Registration No 12216883

Programme and Section K22GX

Course Code INT234

Under the Guidance of

**Anchal Kaundal 29612**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Raman Goyal bearing Registration no. 12216883 has completed INT 234 project titled, **“Prediction and Clustering of Insurance Charges”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science**

Lovely Professional University

Phagwara, Punjab.

Date: 17th November, 2024

**DECLARATION**

I, Raman Goyal, student of Computer Science and Engineering under CSE/IT 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.

Raman Goyal

Date:17th November, 2024 Signature

Registration No. 12216883 Name of the student

**1.** **Introduction**

This report investigates the insurance dataset to identify key factors influencing medical charges and to develop a predictive model for classifying individuals based on their medical expenses. By leveraging statistical and machine learning methods, the report aims to provide insights into demographic and behavioral drivers of healthcare costs.

**2. Scope of the Analysis**

The primary goals of this analysis are:

1. To explore relationships between demographic, behavioral, and health factors (e.g., age, BMI, smoking habits) and medical charges.

2. To classify individuals into high or low insurance cost groups using machine learning models.

3. To visualize the trends and results for interpretability.

4. To identify limitations and propose future improvements.

**3. Existing System**

**i. Drawbacks or Limitations of the Existing System**

- Insufficient Attributes: The dataset lacks medical history or lifestyle details that might strongly influence healthcare costs.

- Imbalanced Data: The smoker category and high-cost individuals are disproportionately represented, impacting the model's performance.

- Limited Predictive Models: Traditional models like KNN are used without exploring advanced techniques for cost prediction.

**4. Source of Dataset**

The dataset, `insurance.csv`, is sourced from a public repository and consists of 1338 entries. The variables included are:

- **age**: Individual's age.

- **sex**: Gender of the individual (male or female).

- **bmi**: Body Mass Index, indicating the individual's health based on weight and height.

- **children**: Number of dependents.

- **smoker**: Whether the individual is a smoker (yes or no).

- **region**: Geographical location of residence.

- **charges**: Medical insurance charges.

**5. ETL Process**

**1. Extraction:**

- Data was imported using R's `read.csv()` function.

**2. Transformation:**

- Missing values were removed (`na.omit`).

- Categorical variables were encoded numerically (e.g., `sex`, `smoker`, `region`).

- A new binary target variable, `charges\_bin`, was created to classify individuals into high or low cost groups based on the median value of `charges`.

**3. Loading:**

- The cleaned and processed dataset was split into training (70%) and testing (30%) sets for modeling.

**6. Analysis on Dataset**

The following analyses were conducted:

**i. Introduction**

The analysis focuses on relationships between variables and their impact on insurance charges. Predictive models were developed to classify individuals into high- or low-cost categories.

**ii. General Description**

- Variables include demographic (age, sex, children), behavioral (smoking status), and health-related attributes (BMI).

- The target variable for predictive modeling is `charges\_bin`.

**iii. Specific Requirements, Functions, and Formulas**

**- Data Preprocessing:**

- `mutate()` and `na.omit()` for feature engineering and cleaning.

- Splitting into training/testing sets with `createDataPartition`.

**- Predictive Model:** K-Nearest Neighbors (KNN) was implemented using `knn()` for binary classification.

**- Visualization:**

- Scatterplots for continuous variables (e.g., BMI vs. charges).

- Boxplots to analyze smoker and region-wise distributions.

- Dashboards for holistic data visualization.

**iv. Analysis Results**

**Key findings:**

- Smokers: Higher charges on average compared to non-smokers.

- BMI: A positive correlation with charges, indicating that higher BMI often leads to higher healthcare costs.

- Region: No significant impact on charges.

- Children: Marginal influence on costs, with slight increases for individuals with more children.

**v. Visualization**

1. Scatterplot: Relationship between BMI and charges.

2. Boxplot: Charges distribution based on smoking status.

3. Barplot: Average charges by gender and region.

4. Dashboard: Combining key insights into a single visual interface.

**7. List of Analysis with Results**

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| K-Nearest Neighbor (KNN) | 79.75% |
| Support Vector Machine (SVM) | 99.25% |
| Naïve Bayes | 84.25% |

Support Vector Machine has best accuracy.

**8. Future Scope**

**1. Advanced Modeling:**

- Explore machine learning models such as Random Forests, Gradient Boosting, or Neural Networks.

- Address class imbalance using techniques like SMOTE.

**2. Feature Engineering:**

- Incorporate external data (e.g., medical conditions or exercise habits).

- Investigate interactions between variables (e.g., BMI × smoker status).

**3. Visualization:**

- Develop interactive dashboards for stakeholder use.

- Use geospatial maps to analyze regional trends.

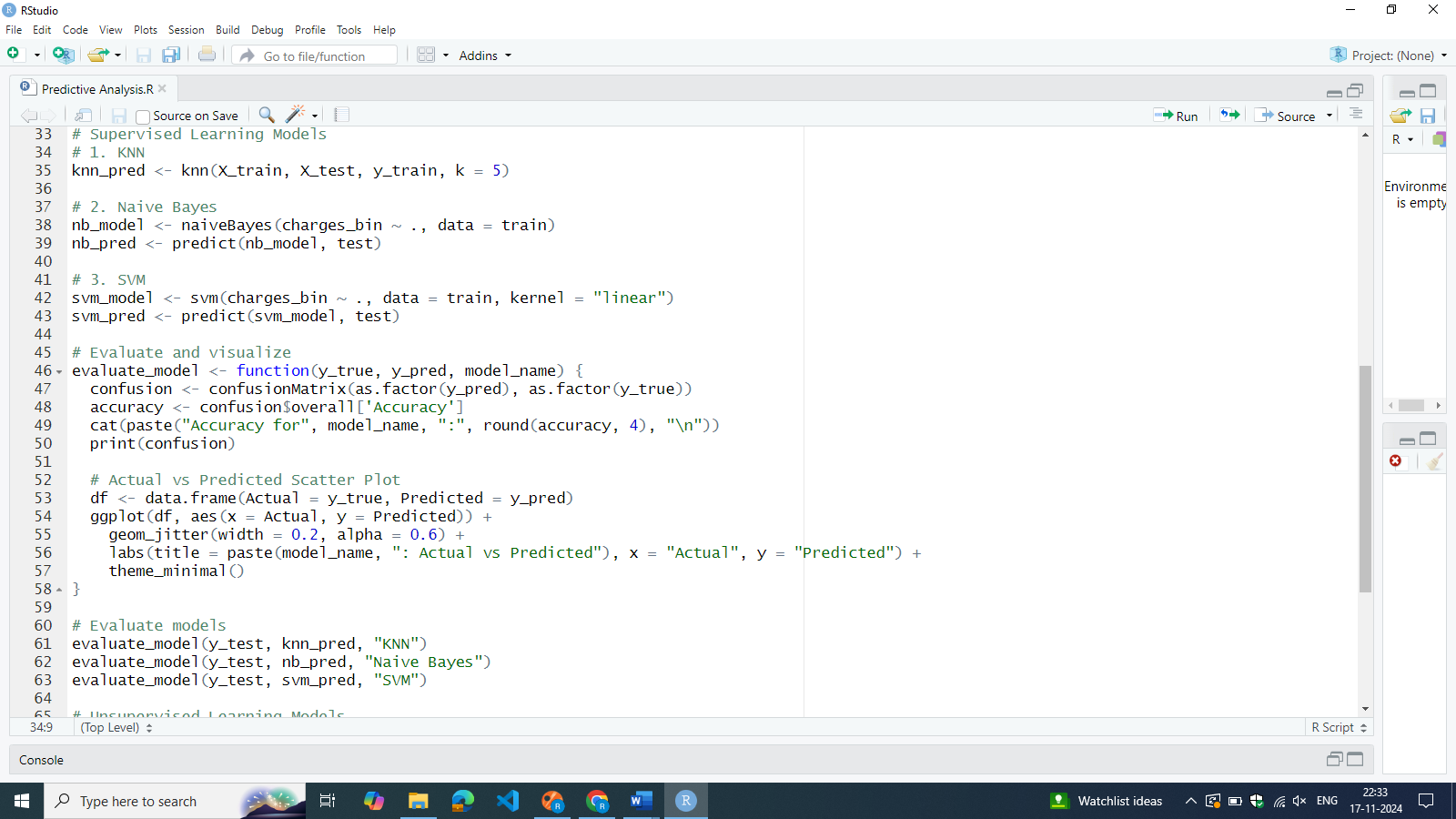
**4. Scalability:**

- Apply analysis to larger datasets for greater generalizability.

**R Code**

**A screenshot of a computer

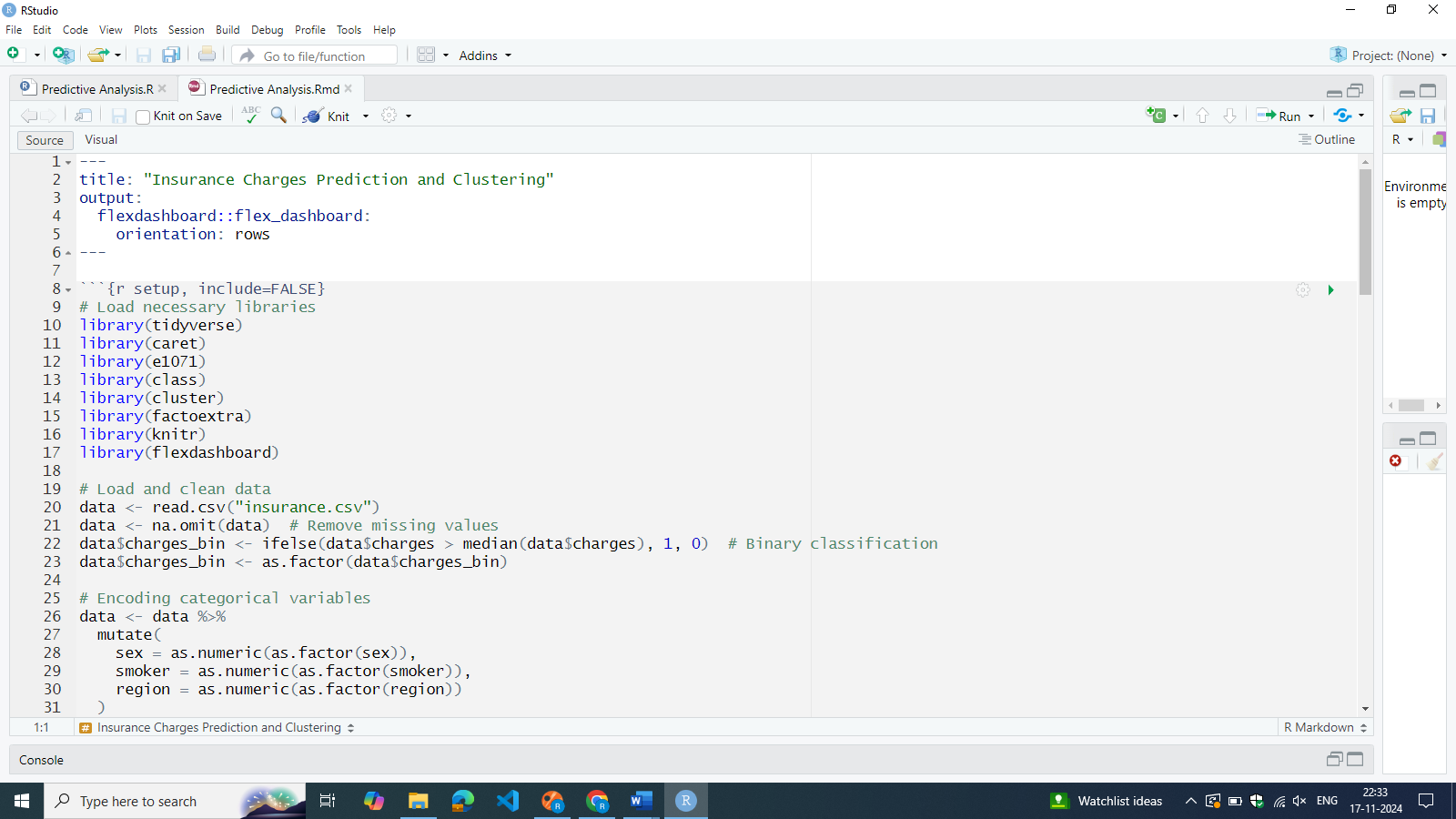
Description automatically generated**

****

**A screenshot of a computer

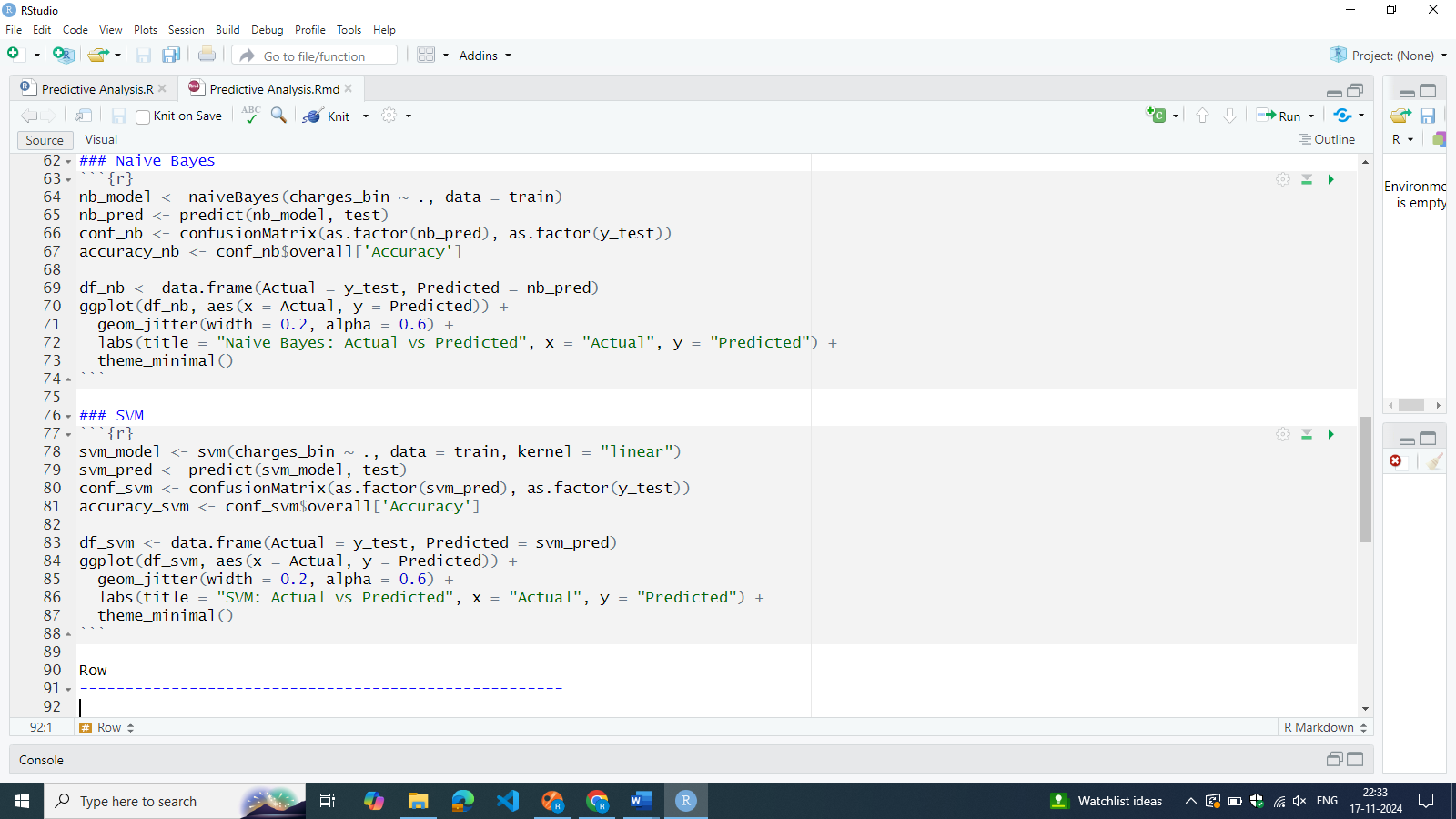
Description automatically generated**

**R Markdown Code**

****

**A screenshot of a computer

Description automatically generated**

****

**A screenshot of a computer

Description automatically generated**

**Dashboard**

**A screenshot of a computer

Description automatically generated**

**9. References**

1. Tidyverse Documentation - [https://tidyverse.tidyverse.org](https://tidyverse.tidyverse.org)

2. Caret Library Documentation - [https://topepo.github.io/caret/](https://topepo.github.io/caret/)

3. KNN Algorithm Overview - [https://scikit-learn.org/stable/modules/neighbors.html](https://scikit-learn.org/stable/modules/neighbors.html)

**10. Bibliography**

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). \*An Introduction to Statistical Learning\*. Springer.

2. Kuhn, M., & Johnson, K. (2013). \*Applied Predictive Modeling\*. Springer.

3. Wickham, H. (2017). \*R for Data Science: Import, Tidy, Transform, Visualize, and Model Data\*. O'Reilly Media.