**Naive Bayes Classification Analysis on Credit Default by Demographic Attributes**

**Author: Beakal Zekaryas  
Date: June – 17 - 2025**

**1. Introduction**

This report provides an analysis of credit card default prediction using a Naive Bayes classifier applied to a dataset of credit card clients. The model was trained to identify patterns of default based on key demographic and financial variables, with a focus on **education level**, **marital status**, **sex**, and **age group**. The findings aim to uncover disparities in default risk and inform data-driven credit risk assessment.

**2. Model and Methodology**

A Gaussian Naive Bayes model was trained on a dataset consisting of 30,000 clients, using financial and demographic predictors. The dataset was cleaned and transformed, including one-hot encoding of categorical features (SEX, EDUCATION, MARRIAGE) and binning of age into defined groups (AGE\_GROUP). The target variable was binary (default: 1 = default, 0 = no default).

The dataset was split into training and testing sets using a 70:30 ratio, and model performance was evaluated using accuracy and a classification report (precision, recall, F1-score).

3. **Model Performance Summary**

* Overall Accuracy: 0.378 (or 37.8%)
* Classification Report:

| Class | Precision | Recall | F1-score |
| --- | --- | --- | --- |
| No Default (0) | 0.88 | 0.24 | 0.37 |
| Default (1) | 0.24 | 0.88 | 0.38 |

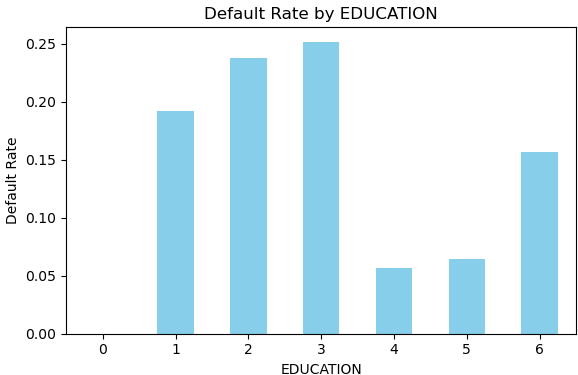
The classifier shows high recall for the default class (1), meaning it captures most of the actual defaulters (88%). However, precision is low (24%), which indicates many false positives. The model performs poorly at predicting **non-defaulters**.

This pattern reflects the conservative nature of Naive Bayes, which prioritizes catching positive cases (defaults) but shows weak ability to detect actual **non-defaulters** or **false alarms**.

**4. Demographic Model Analysis**

**4.1 Education**

The default rate seems to decrease with higher education levels. Clients with only high school education or lower showed significantly higher default rates than those with university or graduate degrees. This aligns with the hypothesis that financial literacy and income stability may increase with education**.**

****

**4.2 Marital Status**

Divorced\Widowed clients showed slightly higher default rates than married or single clients. This could reflect different levels of financial responsibility or dual income benefits in married households.

A graph of a marriage

AI-generated content may be incorrect.

**4.3 Sex**

The difference in default rates between male and female clients looks marginal. However, females had a slightly lower default rate, indicating potential gender-related behavioral or income differences.

A graph with blue rectangles

AI-generated content may be incorrect.

**4.4 Age Group**

Older adults (71–80) show a surprisingly high default rate, which could be due to fixed income, retirement, or small sample bias. Older adults have a moderately high default rate, this is consistent with financial maturity increasing with age, as younger clients may be more credit inexperienced. Younger adults (21–30) have a moderate default rate, possibly due to limited credit experience.

A graph of blue bars

AI-generated content may be incorrect.

**5. Summary and Conclusion**

The Naive Bayes model, while simple and interpretable, made significant assumptions (e.g., conditional independence between predictors) which may not be held in real-world credit scenarios. According to the authors Agresti and Kateri (2021), Naive Bayes often underperforms when features are highly correlated, which is likely in this financial context (e.g., payment history variables).

Moreover, the model prioritizes high recall over precision, The classifier shows recall is high for defaulters (important in credit risk), but precision is low, meaning many predicted defaulters aren't defaulters.

* The model correctly identified 88% of defaulters but misclassified many non-defaulters as high risk.
* Demographically, low education, older adults, and being divorced or widowed were associated with higher default rates.
* Sex had little effect, though women showed marginally better repayment behavior but lower default rate.
* The overall accuracy is 37.8%, is too low for deployment, suggesting model limitations in handling complex dependencies. The model is too conservative & has a weak ability to detect actual defaulters.

**7. Recommendations**

* **Improve recall on defaulters or Model Upgrade**: Use advanced models (e.g., Random Forest, XGBoost, or Logistic Regression) to capture interactions and improve accuracy.
* **Threshold Tuning**: Adjust the classification threshold from the default 0.5 to balance precision and recall.
* **Feature Engineering**: Incorporate more informative features like income, job stability, or credit history duration.
* **Bias Auditing**: Regularly evaluate model fairness to avoid discrimination based on age, sex, or marital status.
* **SMOTE or Resampling**: Balance class distribution to improve prediction on minority classes**.**

**Reference**

Agresti, A., & Kateri, M. (2021). *Foundations of statistics for data scientists: With R and Python*. CRC Press. https://doi.org/10.1201/9781003218135