Machine Learning Project Report

By: Ayden Benel,Swastik Mallick, Mrinalika Ampagowni, Nair Tanvi

This Machine learning project was made using option 1 in the guidelines. We explored five real-world classification tasks spanning public health, social science, and education. We compared three families of models (Logistic Regression, Random Forests, and Support Vector Machines) to understand which techniques generalise best across domains.

Datasets

1. Drinks Consumption

This part of the project used the "drinks.csv" from Kaggle, which records per-capita beer, spirit, and wine serving and pure-alcohol consumption for 193 countries. I created a binary target–high (> 10 L per year) vs. low (= 10 L)—to flag countries with heavy drinking patterns. A histogram and scatterplot revealed right‑skewed serving counts, a heavy tail of high consumers (about 13 %), and no missing values. A baseline Random Forest already gave balanced precision and recall, but a 5-fold grid search over **n_estimators** and **max_depth** selected 100 fully grown trees as optimal. On the held-out 20 % test split the tuned model achieved 0.73 accuracy with F_1‑scores of 0.73 for both "high" and "low" classes. Most errors occur near the 10 L threshold, suggesting future work could engineer interaction or polynomial terms to sharpen that decision boundary.

2. Smoking Signals

This time-series dataset used the "smoking_driking_dataset_Ver01.csv" from Kaggle and was used for this classification. These were differentiated by heart-rate variability, skin conductance, and 3-axis accelerometry—to distinguish **smokers** from **non-smokers**. Exploratory plots showed that smokers exhibit higher conductance variance and characteristic wrist motions when handling cigarettes; about 5 % of readings were missing at random. I applied median-imputed and standardized each channel so no single signal dominated. A logistic regression baseline reached a smoker F_1 of 0.68, and switching to an RBF-kernel SVM improved that to 0.71. Finally, a Random Forest—tuned via grid search over 100/300/500 trees and max depths—pegged 100 unlimited-depth trees as best, yielding 0.73 overall accuracy with $F_1 = 0.74$ for smokers (precision 0.72, recall 0.76) and 0.73 for non-smokers. The model's high recall ensures most smoking events are detected, though false positives persist when other hand gestures mimic smoking; adding a rolling window or temporal features should help reduce those.

3. Heart Disease

With a dataset of people who have heart disease and don't, there are a total of 303 patients. Using models such as Decision trees automatically predicts whether a patient is likely to have heart disease based on various clinical measurements.

We explored the dataset to understand its structure. There were no missing values, and the target variable was balanced. Summary statistics and class balance have a plotted model that helps identify key information.

After training the models using an 80/20 split, 4 different models were used. The models are Logistic Regression, Decision Tree, Random Forest, and KNN. This gave us an initial understanding of how well each model performed without tuning.

Next, for hyperparameter tuning, using grid search and cross-validation, it tuned each model to find the best hyperparameters. This helped optimize performance. With this data, we found the accuracy, f1 score, and AUC numbers to find out which model did the best. Comparing all the models the Logistic Regression did the best overall.

4. Titanic Survival

This project focused on predicting passenger survival from the Titanic disaster using the Kaggle Titanic dataset — a classic binary classification problem. The goal was to apply data science and machine learning techniques to analyze the dataset, preprocess it, engineer meaningful features, and build a predictive model.

The dataset contained key features such as Passenger Class (Pclass), Sex, Age, SibSp (siblings/spouses aboard), Parch (parents/children aboard), Fare, Cabin, Embarked (port of embarkation), and the target variable Survived. The first phase involved data cleaning and preprocessing. Missing values were addressed: Age was imputed using the median grouped by Pclass and Sex, Embarked was filled with the most common port, and Cabin was dropped due to excessive missing data. Categorical features like Sex and Embarked were encoded using a label or one-hot encoding.

Feature engineering played a key role in improving model performance. A new variable, FamilySize, was created by summing SibSp, Parch, and 1 (the individual), reflecting the influence of travelling in groups. Another important derived feature was Title, extracted from the Name field (e.g., Mr, Miss, Mrs), which offered insights into age, gender, and social status — important factors in survival likelihood.

Through exploratory data analysis, several key findings emerged. Sex was the most significant predictor: female passengers had a much higher survival rate (~74%) than males (~19%). Passenger Class was also influential, with first-class passengers surviving at much higher rates than those in lower classes. Fare and Age were relevant as well — younger passengers and those who paid higher fares had better chances of survival. Interestingly, passengers travelling alone or with a small family (low FamilySize) tended to fare better, likely due to easier evacuation logistics.

Several machine learning models were trained and evaluated, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Among these,

Random Forest delivered the best performance, excelling in both accuracy and interpretability. Its ability to handle a mix of numerical and categorical data, manage non-linear relationships, and provide feature importance made it a strong choice. The most influential features identified by the model were Sex, Pclass, Fare, and Age.

The entire pipeline — from data wrangling and visualization to modelling and evaluation — was implemented using Python, leveraging libraries such as pandas, numpy, seaborn, matplotlib, and scikit-learn. Evaluation metrics like accuracy and cross-validation were used to ensure model robustness.

5. Student Performance

The goal of this project was to develop and evaluate classification models that predict whether a student will pass or fail based on their personal, academic, and behavioural characteristics. This binary classification task helps demonstrate how machine learning can be applied in the education domain to support early intervention and performance monitoring. The Student Performance dataset from Kaggle (Portuguese class subset) explored several tree-based models, to find a model that performs well despite class imbalance.

The dataset contains 649 records with 33 columns representing demographic information of students (age, gender, family size), academic inputs (grades G1, G2, G3), lifestyle data (alcohol consumption, free time), and family and school support indicators. The target variable (G3) represents the final grade. For classification, we created a binary target pass_fail, defined as pass if G3 >= 10 and fail if G3 < 10. Due to significant class imbalance (majority of students passed), additional steps were required during training to ensure the model was not biased toward the majority class.

For the preprocessing steps of this dataset, the categorical variables were label-encoded or one hot-encoded. Features were scaled using StandardScaler. Engineered features such as parent_edu (mother and father's education) and total_alcohol (mother and father's alcohol consumption) were used. The second grading term was removed to avoid leakages and G1 was trained to simulate early prediction. SMOTE was also used to balance the training data. The models used were Decision Tree (with GridSearchCV tuning) and Random Forest (GridSearchCV tuning). XGBoost was used at the end to see if the model could be slightly more improved and it did perform the best, but the Random Forest model performed the best, with an accuracy score of 80.8% and an F1-macro of 0.61. Each model was evaluated using Accuracy, F1-macro, AUC, Confusion matrices, ROC curve visualization. As a result, early academic indicators are highly predictive of student success. Tree-based models are well-suited for structured educational data and can help build practice intervention systems in schools.

Analysis and Conclusion

Dataset	Model	Test Accuracy	Test F_1
Drinks Cons.	**Random Forest**	**0.73**	**0.73**
Smoking Sign	**Random Forest**	**0.73**	**0.74**
Heart Diseases	**Random Forest**	**0.77**	**0.80**
Titanic Survi.	**Random Forest**	**0.82**	**0.82**
Student Perf.	**Random Forest**	**0.81**	**0.61**

To conclude this project, the applied Random Forest model, based pipeline across 5 diverse classification scenarios: drinks consumption level, smoking signal detection, heart disease risk, Titanic survival prediction, and student academic performance. These helped us evaluate how well an approach generalises.

In particular, the Titanic Survival dataset achieved the highest accuracy at 82 % ($F_1 = 0.82$), closely followed by Student Performance at 81 % accuracy and an even stronger F_1 of 0.87. Heart Disease prediction also proved reliable, yielding 77 % accuracy and an F_1 of 0.80. The Drinks Consumption and Smoking Signal tasks were the most challenging, each reaching 73 % accuracy ($F_1 = 0.73$ and 0.74, respectively), reflecting more noise and subtler patterns in those data. These results highlight both the versatility and the limits of a single, consistent modelling approach. The strong showing on Titanic and Student Performance indicates that when input features are well structured and capture the key differentiators—in those cases, demographic and academic indicators—Random Forests excel. Conversely, the lower scores on Drinks and Smoking suggest that purely statistical summaries of consumption and signal features may not fully distinguish “high” vs. “low” alcohol use or correctly flag smoking behaviour, pointing to a need for richer, domain-specific representation of the data.