**FIRST RECIPE**
**1st Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Target distribution `adult$income` visualised using `ggplot` histograms and bar plots |
| Sampling type | Random | Used `sample.split()` with `SplitRatio = 0.7` from `caTools` |
| Outliers removal | No | Visualised using `ggplot`, but no removal or capping performed |
| Check for duplicates | No | No `duplicated()` or similar used |
| Imputation of missing values | drop the missing value rows | Replaced ? with `NA`, then removed rows with `na.omit()` |
| Drop columns | No | Renamed columns (`native.country` → `region`), but no columns were dropped |
| Encoding | mixture of encoding | Applied grouping and merging of factor levels (e.g., workclass, marital.status, region) |
| Create new columns | No | Transformed values within existing columns but didn't generate new ones |
| Feature selection | No | All available columns used in logistic regression formula (`income ~ .`) |
| Data scaling/standardisation | No | No scaling applied to numeric features |
| Hyperparameter tuning | No | Basic logistic regression used (`glm`) without tuning |

**2nd Prompt**
**Accuracy 6/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|

| | | |
|---|---|---|
| Check for balanced data | No | No distribution checks for target variable (`Y house price of unit area`) found |
| Sampling type | Random | `set.seed()` and `sample()`-based random splitting used (`sample.split()` from `caTools`) |
| Outliers removal | No | No IQR, Z-score, or filtering applied |
| Check for duplicates | No | No `duplicated()` or similar logic used |
| Imputation of missing values | None | No `is.na()` checks or imputation performed |
| Drop columns | Yes | `'No'` column dropped using `df$No <- NULL` |
| Encoding | None | All features are numeric; no factor encoding applied |
| Create new columns | No | No new variables derived from existing ones |
| Feature selection | No | All columns used without elimination or correlation filtering |
| Data scaling/standardisation | Yes | `scale()` used to standardize numeric variables |
| Hyperparameter tuning | No | Linear regression used as-is via `lm()` with no tuning |

**3rd Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | No | No frequency/count plot or table of `income` distribution was used. |
| Sampling type | Stratified | Used `sample.split(adult$income, SplitRatio = 0.7)` from `caTools`, which performs stratified split. |
| Outlier removal | No | No outliers were removed (histograms used but no filtering applied). |

| | | |
|---|---|---|
| Check for duplicates | No | No check for duplicate rows (e.g., via `duplicated()` or `anyDuplicated()`). |
| Imputation of missing values | drop the missing value rows | Replaced `"?"` with `NA`, then used `na.omit()` to drop rows with missing values. |
| Drop columns | No | No columns were removed entirely. Only renamed `native.country` to `region`. |
| Encoding | mixture of encoding | Merged levels (e.g., grouping `Self-emp-*` into `Self-employed`) and converted to factors. |
| Create new columns | No | No entirely new features were engineered. |
| Feature selection | No | Used all features (`income ~ .`) in the logistic regression. |
| Standardization | No | No standardization or normalization applied. |
| Hyperparameter tuning | No | Logistic regression run with default settings, no tuning performed. |

**4th Prompt**
**Accuracy 8/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | `table(adult$income)` used indirectly through prediction threshold and performance metrics |
| Sampling type | Stratified | `sample.split(adult$income, SplitRatio = 0.7)` from `caTools` is stratified on `income` |
| Outliers removal | No | Outliers explored with histograms but not removed |
| Check for duplicates | No | No use of `duplicated()` or related methods |
| Imputation of missing values | drop the missing value rows | Replaced ? with `NA`, then applied `na.omit()` |

| | | |
|---|---|---|
| Drop columns | No | No column dropped except factor level recoding |
| Encoding | none | No use of LabelEncoder or one-hot encoding (factors used directly) |
| Create new columns | Yes | Region categorisation from `native.country` and grouped levels |
| Feature selection | No | All columns used in `glm(income ~ ., ...)` formula |
| Data scaling/standardisation | No | No scaling or standardisation applied |
| Hyperparameter tuning | No | Logistic regression fitted directly without tuning |

**5th Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Class distribution of `income` checked via multiple `table(...)` calls and visualised using histograms (`ggplot(adult, aes(...))`). |
| Sampling type | Random | `sample.split(...)` used from `caTools` without stratification → random sampling. |
| Outliers removal | No | No Z-score, IQR, or visual/manual outlier filtering performed. |
| Check for duplicates | No | No duplicate check using `duplicated()` or equivalent observed. |
| Imputation of missing values | drop the missing value rows | Replaced `"?"` with `NA`, visualised missing data (`missmap`), then removed rows using `na.omit(...)`. |
| Drop columns | No | No feature removed permanently; renaming only (`native.country` to `region`). |
| Encoding | Label Encoder | Categorical columns (`workclass`, `marital.status`, `native.country`) |

| | | |
|---|---|---|
| | | converted to factors, which R treats similarly to label encoding. |
| Create new columns | No | Only recoding/relabelling was done (e.g. grouped countries into continents); no new column added. |
| Feature selection | No | No variables dropped post-EDA; model used all features (`income ~ .`). |
| Data scaling/standardisation | No | No numeric feature scaled or standardised. |
| Hyperparameter tuning | No | Default logistic regression (`glm`) used without any hyperparameter tuning. |

**Ground Truth**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | No | No explicit distribution check of the target (`income`) such as `table(income)`/count plot of `income` alone. |
| Sampling type | Stratified | `caTools::sample.split(adult$income, SplitRatio = 0.7)` performs class-preserving (stratified) train/test split. |
| Outliers removal | No | No IQR/quantile/z-score filtering or row removal for outliers. |
| Check for duplicates | No | No use of `duplicated()` / `unique()` / de-duplication steps. |
| Imputation of missing values | drop the missing value rows | Replaced `"?"` with `NA`, visualised with `Amelia::missmap`, then removed missing rows via `na.omit(adult)`. |
| Drop columns | No | Columns were recoded/renamed (`setnames(..., "native.country", "region")`) but not dropped. |
| Encoding | dummy | Categorical variables converted to `factor`; `glm()` will encode factors into dummy/contrast columns internally. |

| | | |
|---|---|---|
| Create new columns | No | Values recoded within existing columns; temporary `region.ordered` not kept as a new feature. |
| Feature selection | No | Model fitted as `income ~ .` using all remaining predictors; no correlation/model-importance or post-EDA drops. |
| Data scaling/standardisation | No | No scaling (`scale()`, centering/standardising) applied. |
| Hyperparameter tuning | No | Single `glm(family = binomial)` without grid/random/Bayesian search. |

**SECOND RECIPE**

**1st Prompt**

**Accuracy 8/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | No | No `value_counts()` or visualisation for class distribution |
| Sampling type | Random | Used `train_test_split(..., random_state=0)` without stratification |
| Outliers removal | No | No logic to identify or remove outliers |
| Check for duplicates | Yes | Used `data.drop_duplicates()` |
| Imputation of missing values | replace with text | Replaced `"?"` with `"None"` in multiple columns |
| Drop columns | Yes | Dropped `"fnlwgt"` using `data.drop()` |
| Encoding | Label Encoder | Used `LabelEncoder` for most categorical columns |
| Create new columns | No | No feature engineering or derived columns |
| Feature selection | Yes | Manually selected a subset of columns for modeling |
| Data scaling/standardisation | Yes | Applied `StandardScaler()` to features |

| | | |
|---|---|---|
| Hyperparameter tuning | Yes | Tuned K in KNN manually (`for i in 1 to 10`) |

**2nd Prompt**
**Accuracy 8/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | No | No check or plot of class balance for `income` |
| Sampling type | Random | `train_test_split` used with `random_state=0` |
| Outliers removal | No | No IQR, Z-score, or filtering logic applied |
| Check for duplicates | Yes | `data.drop_duplicates()` used |
| Imputation of missing values | replace with text | `"?"` values replaced with `"None"` in multiple columns |
| Drop columns | Yes | `fnlwgt` dropped from the dataset |
| Encoding | Label Encoder | All categorical columns encoded with `LabelEncoder` |
| Create new columns | No | No columns created or engineered |
| Feature selection | Yes | Only selected features passed to models (excluded `education`, used `educational-num`) |
| Data scaling/standardisation | Yes | `StandardScaler` used for model inputs |
| Hyperparameter tuning | Yes | KNN `k` manually tuned via cross-validation (1–10) |

**3rd Prompt**
**Accuracy 8/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|

| | | |
|---|---|---|
| Check for balanced data | No | No `value_counts()` or visual analysis on target balance. |
| Sampling type | Random | Used `train_test_split(..., random_state=0)` without stratification. |
| Outlier removal | No | EDA performed but no outlier filtering. |
| Check for duplicates | Yes | Explicitly removed using `data.drop_duplicates()`. |
| Imputation of missing values | replace with text | Replaced `"?"` with `"None"` in categorical columns. |
| Drop columns | Yes | Dropped `"fnlwgt"` column. |
| Encoding | mixture of encoding | Applied `LabelEncoder` on most categorical columns, and `map()` for marital status. |
| Create new columns | No | No new derived features added. |
| Feature selection | Yes | Only selected a subset of features for training. |
| Standardization | Yes | Used `StandardScaler` on features before model training. |
| Hyperparameter tuning | Yes | KNN evaluated with different K values to choose the best K (=10). |

**4th Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | No | No countplot or value_counts for label distribution prior to modeling |
| Sampling type | Random | `train_test_split(..., random_state=0)` used without stratify |
| Outliers removal | No | Outliers visualised via plots but not removed or capped |
| Check for duplicates | Yes | `data.drop_duplicates()` used |

| | | |
|---|---|---|
| Imputation of missing values | replace with text | `"?"` values replaced with `"None"` in multiple columns |
| Drop columns | Yes | `"fnlwgt"` dropped as unneeded feature |
| Encoding | Label Encoder | `LabelEncoder()` used for all categorical columns |
| Create new columns | No | No derived features added |
| Feature selection | Yes | Subset of columns manually selected for x before training |
| Data scaling/standardisation | Yes | `StandardScaler()` applied to both training and test data |
| Hyperparameter tuning | No | K values (1–10) tested manually; no automated tuning method used |

**5th Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Class distribution of `income` inspected using `value_counts()` and visualised during EDA via plots. |
| Sampling type | Random | `train_test_split` used without `stratify`. |
| Outliers removal | No | No IQR/Z-score filtering or outlier adjustment seen, though capital-gain/loss histograms and violin plots were shown. |
| Check for duplicates | Yes | `data.drop_duplicates()` was called to remove duplicate rows. |
| Imputation of missing values | replace with text | `"?"` replaced with `"None"` in multiple columns (`workclass`, `occupation`, `native-country`). |
| Drop columns | Yes | `'fnlwgt'` column was dropped from the dataset during early cleaning. |
| Encoding | Label Encoder | Applied `LabelEncoder` across all categorical columns including target (`income`). |

| | | |
|---|---|---|
| Create new columns | No | No new derived features were created. |
| Feature selection | Yes | Only a subset of features selected for modeling: 11 input features chosen explicitly. |
| Data scaling/standardisation | Yes | `StandardScaler` used to scale `x_train` and `x_test` before applying KNN and logistic regression. |
| Hyperparameter tuning | Yes | K for KNN tuned manually (1–10), and cross-validation (`cross_val_score`) used to pick best `k=10`. Logistic regression used default. |

**Ground Truth**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Explicit check on the target via `data['income'].value_counts()` (target distribution inspected). |
| Sampling type | Random | `train_test_split(..., test_size=0.3, random_state=0)` without `stratify`. |
| Outliers removal | No | No IQR/quantile/z-score filtering or row deletion performed. |
| Check for duplicates | Yes | Duplicates removed with `data.drop_duplicates()`. |
| Imputation of missing values | replace with text | Placeholder `'?'` replaced with `"None"` in `workclass`, `occupation`, `native-country`. |
| Drop columns | Yes | Identifier/weight column `fnlwgt` dropped (`data.drop('fnlwgt', axis=1)`) and not reused elsewhere. |
| Encoding | Label Encoder | Multiple `LabelEncoder()` applications (e.g., `workclass`, `education`, `occupation`, `relationship`, `race`, `gender`, `native-country`, `income`) plus a manual map for `marital-status`. |

| | | |
|---|---|---|
| Create new columns | No | No truly new features created; all changes are in-place recodes/encodes. |
| Feature selection | No | Fixed subset chosen for X; no correlation/variance/model-importance pruning or post-EDA drops. |
| Data scaling/standardisation | Yes | `StandardScaler` fit on `x_train` and applied to `x_test`. |
| Hyperparameter tuning | No | Manual sweep of K (1–10) with `cross_val_score`; no systematic search (`GridSearchCV`/`RandomizedSearchCV`/Optuna). |

**THIRD RECIPE**
**1st Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Used `countplot(y="predclass")` and comments on class imbalance |
| Sampling type | Random | `train_test_split(..., random_state=2)` used without stratification |
| Outliers removal | No | Only KDE and distribution plots; no filtering logic applied |
| Check for duplicates | No | No use of `duplicated()` or related functions |
| Imputation of missing values | drop the missing value rows | Used `dropna()` to remove all NA values from `income_df` |
| Drop columns | Yes | Dropped `'education'`, `'native-country'`, and engineered columns |
| Encoding | Label Encoder | Applied `LabelEncoder()` to the entire dataframe using `.apply()` |

| | | |
|---|---|---|
| Create new columns | Yes | Created `age_bin`, `hours-per-week_bin`, `age-hours`, and binned columns |
| Feature selection | Yes | Selected and dropped multiple columns before model input |
| Data scaling/standardisation | Yes | Used `StandardScaler()` before PCA and model training |
| Hyperparameter tuning | Yes | Used `GridSearchCV` for RandomForest tuning |

**2nd Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Class imbalance of target (`predclass`) was visualised using countplot |
| Sampling type | Random | `train_test_split` used with `random_state=2` |
| Outliers removal | No | No explicit filtering or clipping logic shown |
| Check for duplicates | No | No use of `.drop_duplicates()` or equivalent |
| Imputation of missing values | drop the missing value rows | `dropna()` used to remove rows with missing values |
| Drop columns | Yes | Columns like `income`, `fnlwgt`, and `educational-num` were dropped |
| Encoding | Label Encoder | `LabelEncoder()` applied to all features using `.apply()` |
| Create new columns | Yes | Features such as `age-hours`, `age_bin`, `hours-per-week_bin` created |
| Feature selection | Yes | Specific features dropped before modeling (e.g., `education`, `native-country`) |
| Data scaling/standardisation | Yes | `StandardScaler()` applied before PCA and model training |

| | | |
|---|---|---|
| Hyperparameter tuning | Yes | `GridSearchCV` applied to optimize Random Forest parameters |

**3rd Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Commented and plotted class distribution of `predclass` (above vs. below 50K income). |
| Sampling type | Random | Used `train_test_split(..., random_state=2)` for both training/testing and model evaluation. |
| Outlier removal | No | No explicit filtering or removal of outliers despite distributional plots. |
| Check for duplicates | No | No check like `drop_duplicates()` observed. |
| Imputation of missing values | drop the missing value rows | Dropped all rows with missing values using `dropna()`. |
| Drop columns | Yes | Dropped `income`, `educational-num`, and other derived columns during preprocessing. |
| Encoding | mixture of encoding | Label encoding applied to entire DataFrame; category regrouping (e.g., education) also done manually. |
| Create new columns | Yes | Created `age-hours`, `age_bin`, `age-hours_bin`, etc. |
| Feature selection | Yes | Dropped non-predictive and redundant features before modeling. |
| Standardization | Yes | Applied `StandardScaler` before PCA. |
| Hyperparameter tuning | Yes | Used `GridSearchCV` for Random Forest, also evaluated multiple models via `cross_val_score`. |

**4th Prompt**

**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | `countplot(y="predclass")` used; imbalance noted in markdown |
| Sampling type | Random | `train_test_split(..., random_state=2)` used without stratification |
| Outliers removal | No | Outliers explored visually (e.g., violin plots, distplots), but not explicitly removed |
| Check for duplicates | No | No use of `duplicated()` or `.drop_duplicates()` |
| Imputation of missing values | drop the missing value rows | Missing values dropped via `dropna()` on categorical attributes |
| Drop columns | No | Dropped columns: `income`, `educational-num`, later `predclass`, `native-country`, etc after EDA. |
| Encoding | Label Encoder | Full dataset encoded with `LabelEncoder()` after engineering |
| Create new columns | No | `'age-hours'` is derived from `'age'` and `'hours-per-week'` |
| Feature selection | Yes | Manual selection before modeling (excluded engineered bins, redundant features, etc.) |
| Data scaling/standardisation | Yes | `StandardScaler` used prior to PCA and modeling |
| Hyperparameter tuning | Yes | `GridSearchCV` used for tuning `RandomForestClassifier` |

**5th Prompt**
**Accuracy 10/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|

| | | |
|---|---|---|
| Check for balanced data | Yes | Target class `predclass` was checked for imbalance with `countplot()` and explained in markdown notes. |
| Sampling type | Stratified (Implied) | Although not explicitly stratified in `train_test_split`, stratified KFold cross-validation (`KFold`) used for final evaluation. |
| Outliers removal | No | Visual distribution plots created, but no outlier filtering (Z-score, IQR, etc.) applied. |
| Check for duplicates | No | No check using `duplicated()` or equivalent observed. |
| Imputation of missing values | drop the missing value rows | NA rows dropped via `dropna()` early in the pipeline. |
| Drop columns | No | Columns like `income`, `educational-num`, and `fnlwgt` were dropped/replaced. PCA-based reductions also apply. |
| Encoding | Label Encoder | `LabelEncoder()` applied to entire DataFrame using `.apply(...)`, covering all categorical fields. |
| Create new columns | No | Created `age-hours` from existing columns |
| Feature selection | Yes | Manual dropping of non-informative columns (`native-country`, bins, etc.), PCA applied to reduce dimensionality. |
| Data scaling/standardisation | Yes | `StandardScaler()` used before PCA and modeling. |
| Hyperparameter tuning | Yes | `GridSearchCV` used with a detailed param grid on Random Forest; also used 10-fold `cross_val_score` on multiple classifiers. |

**Ground Truth**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|

| | | |
|---|---|---|
| Check for balanced data | Yes | Target distribution inspected via `sns.countplot(y="predclass", data=my_df)`. |
| Sampling type | Random | `train_test_split(X, y, test_size=0.2, random_state=2)` with no `stratify`. |
| Outliers removal | No | No IQR/quantile/z-score masking or row deletion present. |
| Check for duplicates | No | No use of `.duplicated()` / `.drop_duplicates()`. |
| Imputation of missing values | drop the missing value rows | `my_df = income_df.dropna()` removes rows with NA after type fixes. |
| Drop columns | No | Columns removed for modeling are dropped **after EDA** (see `drop_elements`), so counted under Feature selection (not Drop columns). |
| Encoding | Label Encoder | `my_df = my_df.apply(LabelEncoder().fit_transform)` encodes all columns. |
| Create new columns | No | New fields (`age_bin`, `hours-per-week_bin`, `age-hours`, `age-hours_bin`) are **derived** from existing columns, which does not count as “new” per rubric. |
| Feature selection | Yes | Post-EDA manual pruning before training: `drop_elements = ['education','native-country','predclass','age_bin','age-hours_bin','hours-per-week_bin']`; model uses `X = my_df.drop(drop_elements, axis=1)`. |
| Data scaling/standardisation | Yes | `StandardScaler()` applied (for PCA and standardized train matrix). |
| Hyperparameter tuning | Yes | Systematic search via `GridSearchCV` on `RandomForestClassifier` (cv=5, parameter grid). |

## FOURTH RECIPE

**1st Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Used `value_counts()` and pie chart on `income` column |
| Sampling type | Random | Used `train_test_split(..., random_state=43)` without stratification |
| Outliers removal | Yes | Dropped `capital-gain` and `capital-loss` due to skewed distributions with mostly zeroes |
| Check for duplicates | Yes | Applied `df.drop_duplicates()` |
| Imputation of missing values | drop the missing value rows | Replaced `'?'` with `np.nan`, then dropped rows with `df.dropna()` |
| Drop columns | Yes | Dropped `'education'`, `'capital-gain'`, `'capital-loss'` |
| Encoding | Label Encoder | Applied `LabelEncoder` on all object-type columns |
| Create new columns | No | No new features created |
| Feature selection | Yes | Dropped specific columns before model training |
| Data scaling/standardisation | Yes | Used `StandardScaler()` on all features |
| Hyperparameter tuning | No | Used default `LogisticRegression()` model without tuning |

**2nd Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Pie chart plotted showing class imbalance (approx. 75% vs 25%) |

| | | |
|---|---|---|
| Sampling type | Random | Used `train_test_split` with `random_state=43` |
| Outliers removal | Yes | Removed capital-gain and capital-loss due to right-skewed distribution |
| Check for duplicates | Yes | Used `drop_duplicates()` |
| Imputation of missing values | drop the missing value rows | Replaced `"?"` with NaN and dropped rows using `dropna()` |
| Drop columns | Yes | Dropped `education`, `capital-gain`, `capital-loss` explicitly |
| Encoding | Label Encoder | All categorical columns encoded using `LabelEncoder()` loop |
| Create new columns | No | No new features derived or engineered |
| Feature selection | Yes | Redundant or non-contributing features dropped manually |
| Data scaling/standardisation | Yes | `StandardScaler()` applied to input features before training |
| Hyperparameter tuning | No | Used default logistic regression; no `GridSearchCV` or tuning shown |

**3rd Prompt**
**Accuracy 10/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Class balance of `income` plotted and discussed (75% ≤50K vs 25% >50K). |
| Sampling type | Random | Used `train_test_split(..., random_state=43)` with 80:20 split. |
| Outlier removal | No | Outliers observed in plots (e.g., capital-gain/loss), but not explicitly removed. |
| Check for duplicates | Yes | Applied `df.drop_duplicates()`. |
| Imputation of missing values | drop the missing value rows | Replaced `"?"` with NaN and used `dropna()` to remove rows. |

| | | |
|---|---|---|
| Drop columns | Yes | Dropped `education`, `capital-gain`, `capital-loss` during preprocessing. |
| Encoding | Map to ordinal values | Used `LabelEncoder` for all object columns (ordinal mapping). |
| Create new columns | No | Did not engineer new columns; used existing features post-cleaning. |
| Feature selection | Yes | Dropped less useful columns before modeling (e.g., mentioned `education`, etc.). |
| Standardization | Yes | Used `StandardScaler()` to standardize features before logistic regression. |
| Hyperparameter tuning | No | Used default `LogisticRegression()` without tuning or cross-validation. |

**4th Prompt**
**Accuracy 10/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | `value_counts()` on `income` and pie chart used to highlight class imbalance |
| Sampling type | Random | `train_test_split(..., train_size=0.8, random_state=43)` used without `stratify` |
| Outliers removal | No | Capital-gain and capital-loss dropped due to sparsity, but not capped or filtered numerically |
| Check for duplicates | Yes | Explicit `df.drop_duplicates(inplace=True)` used |
| Imputation of missing values | drop the missing value rows | `'?'` replaced with `np.nan` and removed using `df.dropna()` |
| Drop columns | Yes | Dropped `education`, `capital-gain`, and `capital-loss` |

| | | |
|---|---|---|
| Encoding | Label Encoder | All object columns encoded with `LabelEncoder()` |
| Create new columns | No | No new features added |
| Feature selection | Yes | Dropped some features manually and selected subset for modeling |
| Data scaling/standardisation | Yes | `StandardScaler()` used before model training |
| Hyperparameter tuning | No | Logistic regression trained without tuning |

**5th Prompt**
**Accuracy 10/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Income class distribution explored using `value_counts()`, pie chart, and discussed in markdown. |
| Sampling type | Random | `train_test_split` used without `stratify`. |
| Outliers removal | No | Capital-gain and capital-loss features were dropped, but not due to statistical outlier treatment (e.g. Z-score/IQR). |
| Check for duplicates | Yes | `drop_duplicates()` used after NA handling. |
| Imputation of missing values | drop the missing value rows | '?' replaced with `np.nan`, then removed using `dropna()`. |
| Drop columns | Yes | Dropped: `capital-gain`, `capital-loss`, and `education` due to redundancy or lack of relevance. |
| Encoding | Label Encoder | Applied LabelEncoder for all object-type categorical columns. |
| Create new columns | No | No new features were created or engineered in the final model pipeline. |

| | | |
|---|---|---|
| Feature selection | Yes | `education`, `capital-gain`, and `capital-loss` dropped after EDA insights and redundancy with `education-num`. |
| Data scaling/standardisation | Yes | `StandardScaler` used on feature matrix `X` before model training. |
| Hyperparameter tuning | No | Logistic Regression used with default parameters; no GridSearch or cross-validation applied. |

**Ground Truth**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Target distribution inspected with `df['income'].value_counts()` and a pie plot. |
| Sampling type | Random | `train_test_split(X, y, train_size=0.8, random_state=43)` without `stratify`. |
| Outliers removal | No | No IQR/quantile/z-score filtering or row deletion for outliers present. |
| Check for duplicates | Yes | Duplicates removed with `df.drop_duplicates(inplace=True)`. |
| Imputation of missing values | drop the missing value rows | First converted `"?"` → `NaN`, then dropped rows via `df.dropna(inplace=True)`. |
| Drop columns | No | Columns used later for modeling were dropped **after** EDA, so counted under Feature selection (see next row). |
| Feature selection | Yes | Post-EDA removal of low-value features: `['education','capital-gain','capital-loss']` dropped before modeling; this follows earlier EDA notes justifying their removal. |
| Encoding | Label Encoder | Categorical columns mapped to integers using `LabelEncoder()` in a loop over object dtypes. |

| | | |
|---|---|---|
| Create new columns | No | No new features retained for modeling; only in-place recodes/encodes performed. |
| Data scaling/standardisatio n | Yes | Standardised predictors with `StandardScaler().fit_transform(X)`. |
| Hyperparameter tuning | No | Single `LogisticRegression` trained with default params; no grid/random/Bayesian search. |

**FIFTH RECIPE**
**1st Prompt**
**Accuracy 10/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | No | No `value_counts()` or visual check on `income` |
| Sampling type | Random | Used `train_test_split(..., train_size=0.8)` without `stratify` |
| Outliers removal | No | No statistical or visual filtering of outliers |
| Check for duplicates | No | No use of `duplicated()` or related methods |
| Imputation of missing values | none | No check or handling of missing values |
| Drop columns | Yes | Dropped `'fnlwgt'`, `'educational-num'`, `'marital-status'`, `'relationship'`, `'capital-gain'`, `'capital-loss'` |
| Encoding | dummy | Used `pd.get_dummies()` on `income` column |
| Create new columns | No | Only renamed dummy variables; no feature engineering |
| Feature selection | Yes | Used only `'age'` as the input feature |
| Data scaling/standardisation | No | No use of `StandardScaler()` or similar |
| Hyperparameter tuning | No | Logistic regression used with default settings |

**2nd Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | No | No distribution check on the `income` variable |
| Sampling type | Random | Used `train_test_split` without specifying `random_state` |
| Outliers removal | No | No outlier filtering applied |
| Check for duplicates | No | No check for duplicates performed |
| Imputation of missing values | None | No missing value handling applied |
| Drop columns | Yes | Dropped multiple columns: `fnlwgt`, `educational-num`, `marital-status`, `relationship`, `capital-gain`, `capital-loss` |
| Encoding | One hot encoding | Applied via `pd.get_dummies()` on `income` column |
| Create new columns | Yes | Created binary columns `<=50K` and `>50K` from `income` |
| Feature selection | Yes | Only `age` used for training the logistic regression model |
| Data scaling/standardisation | No | Model trained directly on raw `age` without scaling |
| Hyperparameter tuning | No | Default `LogisticRegression()` used without tuning |

**3rd Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | No | No visual or statistical balance check for `income`. |

| | | |
|---|---|---|
| Sampling type | Random | Used `train_test_split(..., train_size=0.8)` without stratification. |
| Outlier removal | No | No handling or analysis of outliers. |
| Check for duplicates | No | No `drop_duplicates()` or similar method used. |
| Imputation of missing values | No | No treatment of missing values. |
| Drop columns | Yes | Dropped `fnlwgt`, `educational-num`, `marital-status`, `relationship`, `capital-gain`, `capital-loss`. |
| Encoding | Dummy | Used `pd.get_dummies()` for the `income` column. |
| Create new columns | Yes | Created `Less_than_50K`, `More_than_50K` binary indicators. |
| Feature selection | Yes | Only `age` used as input feature; all others were ignored for model training. |
| Standardization | No | No scaling or standardisation applied. |
| Hyperparameter tuning | No | Used default `LogisticRegression()` without tuning. |

**4th Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | No | No label distribution checks like `value_counts()` or plots were performed |
| Sampling type | Random | `train_test_split(..., train_size=0.8)` used without `stratify` |
| Outliers removal | No | No filtering or IQR/z-score-based treatment of age or other features |
| Check for duplicates | No | No use of `duplicated()` or `drop_duplicates()` |
| Imputation of missing values | none | No missing value treatment applied |

| | | |
|---|---|---|
| Drop columns | Yes | Dropped `fnlwgt`, `educational-num`, `marital-status`, `relationship`, `capital-gain`, `capital-loss` |
| Encoding | dummy | Created dummy variables for `income` using `pd.get_dummies()` |
| Create new columns | No | Columns are derived from existing |
| Feature selection | No | No explicit feature selection |
| Data scaling/standardisation | No | No scaler applied on the age feature |
| Hyperparameter tuning | No | LogisticRegression used with default parameters |

**5th Prompt**
**Accuracy 10/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | No | No class balance check on `income`; no `value_counts()` or distribution plots used. |
| Sampling type | Random | `train_test_split` used without `stratify`. |
| Outliers removal | No | No treatment or filtering of outliers in any variable. |
| Check for duplicates | No | No use of `duplicated()` or similar function. |
| Imputation of missing values | none | No NA values were handled or mentioned; assumed dataset was clean. |
| Drop columns | Yes | Columns dropped: `'fnlwgt'`, `'educational-num'`, `'marital-status'`, `'relationship'`, `'capital-gain'`, `'capital-loss'`. |
| Encoding | One hot encoding | `pd.get_dummies()` used on `income` to create binary columns. |
| Create new columns | No | Columns were renamed and encoded, but no new derived features were added. |

| | | |
|---|---|---|
| Feature selection | Yes | Final model used only `'age'` as feature; other columns dropped manually. |
| Data scaling/standardisation | No | No scaling applied to feature(s) before logistic regression. |
| Hyperparameter tuning | No | `LogisticRegression` used with default parameters; no grid search or cross-validation implemented. |

**Ground Truth**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | No | No explicit check of target distribution (no `value_counts`/countplot/hist on `income`). |
| Sampling type | Random | `train_test_split(df_dummies[['age']], df_dummies[['income']], train_size=0.8)` without `stratify`. |
| Outliers removal | No | No IQR/quantile/z-score filtering or row deletion. |
| Check for duplicates | No | No use of `.duplicated()` / `.drop_duplicates()`. |
| Imputation of missing values | none | No `fillna`, imputer, or NA row/column drops performed. |
| Drop columns | Yes | Dropped `['fnlwgt','educational-num','marital-status','relationship','capital-gain','capital-loss']` without reuse. |
| Encoding | One hot encoding | `pd.get_dummies(df.income)` created dummy columns (not used downstream for y). |
| Create new columns | No | Dummies are derived from `income`; per rubric, derived features don't count as “new”. |
| Feature selection | No | No correlation/variance/model-importance pruning or post-EDA drops. |

| | | |
|---|---|---|
| Data scaling/standardisation | No | No scaler (`StandardScaler`, `MinMaxScaler`, etc.) used. |
| Hyperparameter tuning | No | `LogisticRegression()` fit with defaults; no grid/random/Bayesian search. |

**SIXTH RECIPE**
**1st Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Used `value_counts()` on target and commented on income imbalance |
| Sampling type | Random | Used `train_test_split(..., random_state=11)` without `stratify` |
| Outliers removal | Yes | Applied IQR-based capping on `age`, `fnlwgt`, and `hours-per-week` |
| Check for duplicates | Yes | Used `drop_duplicates()` after initial inspection |
| Imputation of missing values | replace with text | Replaced `'?'` with `'Unknown'` in categorical fields |
| Drop columns | Yes | Dropped `'educational-num'` and `'relationship'` due to redundancy |
| Encoding | mixture of encoding | Used `LabelEncoder` for `income`, ordinal mapping for `education`, and `get_dummies` for other categorical columns |
| Create new columns | Yes | Grouped categories (e.g. `'occupation'`, `'marital-status'`, etc.) to create meaningful levels |
| Feature selection | Yes | Removed redundant/informationally overlapping columns |
| Data scaling/standardisation | No | No `StandardScaler` or normalization used before modeling |

| | | |
|---|---|---|
| Hyperparameter tuning | Yes | Applied `GridSearchCV` for logistic regression with cross-validation |

**2nd Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Class distribution of `income` checked using value_counts and discussed in text |
| Sampling type | Random | `train_test_split` used with `random_state=11` |
| Outliers removal | Yes | IQR-based outlier capping for `age`, `fnlwgt`, and `hours-per-week` |
| Check for duplicates | Yes | `.drop_duplicates()` used |
| Imputation of missing values | replace with text | Replaced `"?"` in multiple columns with `"Unknown"` |
| Drop columns | Yes | Dropped `educational-num`, `relationship` |
| Encoding | mixture of encoding | Used `LabelEncoder` for `income`, and `pd.get_dummies()` for other categorical columns |
| Create new columns | Yes | Created aggregated categorical groups like `Blue Collar`, `Married`, `Government` etc. |
| Feature selection | Yes | Categorical grouping and column dropping informed by correlation and EDA |
| Data scaling/standardisation | No | Model trained without applying a scaler |
| Hyperparameter tuning | Yes | `GridSearchCV` used with logistic regression to optimize `penalty`, `solver`, and `tol` |

**3rd Prompt**
**Accuracy 10/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Class balance explicitly noted: `income` is skewed and imbalanced. |
| Sampling type | Random | Used `train_test_split(..., test_size=0.3, random_state=11)` without stratification. |
| Outlier removal | Yes | Applied IQR-based capping to `age`, `fnlwgt`, and `hours-per-week`. |
| Check for duplicates | Yes | Used `drop_duplicates()` on the dataset. |
| Imputation of missing values | replace with text | Replaced `"?"` values in object columns with `"Unknown"`. |
| Drop columns | Yes | Dropped `"educational-num"` and `"relationship"`. |
| Encoding | mixture of encoding | Used category grouping + ordinal encoding (education), label encoding (`income`), and one-hot encoding. |
| Create new columns | No | Did not create new features beyond encoded/recoded variables. |
| Feature selection | Yes | Dropped weak or redundant features and ran correlation checks before modeling. |
| Standardization | No | Logistic Regression applied without standardisation. |
| Hyperparameter tuning | Yes | Used `GridSearchCV` to optimise penalty, solver, and tolerance. |

**4th Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | `value_counts()` on `income`, class distribution discussed in markdown |

|  |  |  |
|---|---|---|
| Sampling type | Random | `train_test_split(..., random_state=11)` used without stratify |
| Outliers removal | Yes | IQR-based outlier capping applied to `age`, `fnlwgt`, and `hours-per-week` |
| Check for duplicates | Yes | `drop_duplicates()` used during preprocessing |
| Imputation of missing values | replace with text | Replaced `'?'` with `'Unknown'` in multiple categorical columns |
| Drop columns | Yes | Dropped `educational-num`, `relationship` explicitly |
| Encoding | mixture of encoding | Used LabelEncoder for target; ordinal mapping + `get_dummies()` for others |
| Create new columns | No | Recoding done via binning and grouping, not adding new variables |
| Feature selection | No | Final predictors selected manually from cleaned and encoded data |
| Data scaling/standardisation | No | No scaler (e.g. StandardScaler) used before model fitting |
| Hyperparameter tuning | Yes | `GridSearchCV` applied to tune `LogisticRegression` over penalty, solver, and tolerance |

**5th Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Skew in income distribution discussed in observations, confirmed in `value_counts()` and pairplot overlays. |
| Sampling type | Random | `train_test_split` used without `stratify`. |
| Outliers removal | Yes | IQR method applied for capping outliers in `age`, `fnlwgt`, and `hours-per-week`. |

| | | |
|---|---|---|
| Check for duplicates | Yes | Explicit check using `.duplicated().sum()` and removal via `drop_duplicates()`. |
| Imputation of missing values | replace with text | `"?"` entries in categorical variables replaced with `"Unknown"`. |
| Drop columns | Yes | Columns dropped: `educational-num`, `relationship`, to avoid redundancy. |
| Encoding | mixture of encoding | Ordinal encoding for `education`, label encoding for `income`, one-hot encoding for other categorical columns. |
| Create new columns | No | No new features created; all transformations were recoding or aggregation of existing features. |
| Feature selection | No | No columns dropped AFTER EDA/visualisation/correlation/model-importance |
| Data scaling/standardisation | No | No `StandardScaler` or equivalent used before logistic regression. |
| Hyperparameter tuning | Yes | `GridSearchCV` applied with multiple solvers (`saga`, `liblinear`) and penalties (`l1`, `l2`) on logistic regression. |

**Ground Truth**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Target distribution inspected via `y_train.value_counts(1)` and `y_test.value_counts(1)` after split. |
| Sampling type | Random | `train_test_split(X, Y, test_size=0.3, random_state=11)` without `stratify`. |
| Outliers removal | Yes | Outliers in `age`, `fnlwgt`, `hours-per-week` are capped (IQR winsorization). |
| Check for duplicates | Yes | Duplicates removed with `adult_data_v1.drop_duplicates().reset_index(drop=True)`. |

| | | |
|---|---|---|
| Imputation of missing values | replace with text | Placeholder `'?'` replaced with `'Unknown'` in `workclass`, `occupation`, `native-country`. |
| Drop columns | Yes | Pre-EDA drop of `['educational-num','relationship']` as overlapping/irrelevant. |
| Encoding | mixture of encoding | Manual ordinal mapping for `education`, `LabelEncoder` for `income`, and `pd.get_dummies(drop_first=True)` for remaining categoricals. |
| Create new columns | No | Dummy variables and recoded categories are **derived** from existing columns (do not count as “new” per rubric). |
| Feature selection | No | No correlation/variance/model-importance or post-EDA pruning; model uses all features (`income ~ all dummies`). |
| Data scaling/standardisation | No | No scaler (`StandardScaler/MinMax/Robust`) applied. |
| Hyperparameter tuning | Yes | Systematic search via `GridSearchCV` over `LogisticRegression` penalties/solvers/tolerances (cv=3). |

**SEVENTH RECIPE**
**1st Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Used `value_counts()` on multiple categorical variables; visual correlation plots also used |
| Sampling type | Random | Used `train_test_split(..., random_state=42)` without stratification |
| Outliers removal | No | Outliers explored visually (histograms, density plots), but no removal applied |

| | | |
|---|---|---|
| Check for duplicates | No | No use of `duplicated()` or `.drop_duplicates()` |
| Imputation of missing values | drop the missing value rows | Replaced `'?'` with `np.nan`, then dropped rows with `.dropna(how='any')` |
| Drop columns | Yes | Dropped `'fnlwgt'`, `'educational-num'`, `'capital gain'`, `'capital loss'`, `'age'`, `'hours per week'`, `'country'` |
| Encoding | mixture of encoding | Used `.map()` to apply custom numeric mapping for multiple categorical features |
| Create new columns | No | Selected existing columns only; no new columns created |
| Feature selection | Yes | Only selected: `'relationship'`, `'education'`, `'race'`, `'occupation'`, `'gender'`, `'marital'`, `'workclass'` |
| Data scaling/standardisation | No | No use of `StandardScaler` or any scaling method |
| Hyperparameter tuning | No | Used cross-validation (K-Fold) but no tuning via GridSearch or manual parameter search |

**2nd Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | `value_counts()` and class-specific bar plots used for balance checks |
| Sampling type | Random | Used `train_test_split` with `random_state=42` |
| Outliers removal | No | Visualizations (hist, density, heatmap), but no filtering applied |
| Check for duplicates | No | No explicit use of `.drop_duplicates()` |
| Imputation of missing values | drop the missing value rows | `"?"` replaced with `NaN`; rows dropped using `dropna()` |

| | | |
|---|---|---|
| Drop columns | Yes | Dropped: `educational-num`, `age`, `hours per week`, `fnlwgt`, `capital gain`, `capital loss`, `country` |
| Encoding | mixture of encoding | Used `.map()` to convert categorical variables to integers manually |
| Create new columns | No | No new feature engineering beyond renaming and selection |
| Feature selection | Yes | Manual selection of 7 features (`relationship`, `education`, `race`, etc.) informed by correlation |
| Data scaling/standardisation | No | No scaler or normalization applied |
| Hyperparameter tuning | No | Logistic regression used with default settings, but evaluated via KFold CV |

**3rd Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Class balance of `income` considered, mean income plotted by multiple features. |
| Sampling type | Random | Used `train_test_split(..., random_state=42)` without stratification. |
| Outlier removal | No | Histograms and density plots shown, but no outlier filtering or capping. |
| Check for duplicates | No | No use of `drop_duplicates()` observed. |
| Imputation of missing values | drop the missing value rows | Replaced `"?"` with np`.`nan and then used `dropna(how='any')`. |
| Drop columns | Yes | Dropped age, `capital-gain`, `capital-loss`, `fnlwgt`, `educational-num`, `hours per week`, `country`. |

|  |  |  |
|---|---|---|
| Encoding | mixture of encoding | Extensive `map()`-based manual ordinal encoding of all categorical columns. |
| Create new columns | No | Features were mapped, not derived or constructed. |
| Feature selection | Yes | Selected 7 features for modeling (`relationship`, `education`, `race`, etc.). |
| Standardization | No | No feature scaling applied. |
| Hyperparameter tuning | No | Used default `LogisticRegression()` and KFold cross-validation without tuning parameters. |

**4th Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | `value_counts()` on `income` used and class balance discussed |
| Sampling type | Random | `train_test_split(..., random_state=42)` used without `stratify` |
| Outliers removal | No | Outliers visualised via histograms/density plots but not removed |
| Check for duplicates | No | No explicit use of `drop_duplicates()` |
| Imputation of missing values | drop the missing value rows | Replaced `"?"` with `np.nan`, then applied `dropna(how='any')` |
| Drop columns | Yes | Dropped `educational-num`, `age`, `hours per week`, `fnlwgt`, `capital gain`, `capital loss`, `country` |
| Encoding | mixture of encoding | Used `.map()` with custom dictionaries for multiple columns |
| Create new columns | No | Subset of features selected but no new variables created |

| | | |
|---|---|---|
| Feature selection | Yes | Manually selected 7 features for modeling |
| Data scaling/standardisation | No | No scaler (e.g. StandardScaler) applied |
| Hyperparameter tuning | No | Only `KFold` used for cross-validation; no search across param space |

**5th Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Target (`income`) distribution reviewed via `value_counts` and multiple `groupby().mean().plot()` visualizations. |
| Sampling type | Random | `train_test_split` used without `stratify`. |
| Outliers removal | No | No statistical removal (IQR, Z-score) done; histograms and density plots created for distributional insight. |
| Check for duplicates | No | No use of `duplicated()` or `drop_duplicates()` observed. |
| Imputation of missing values | drop the missing value rows | `"?"` replaced with `NaN`, then removed using `dropna(how='any')`. |
| Drop columns | Yes | Dropped: `'educational-num'`, `'age'`, `'hours-per-week'`, `'fnlwgt'`, `'capital-gain'`, `'capital-loss'`, `'country'`. |
| Encoding | mixture of encoding | Mixed: `.map()` used for ordinal-style encoding on all categorical columns; binary and nominal handled similarly. |
| Create new columns | No | No new features added; existing features were recoded and filtered. |
| Feature selection | Yes | Only subset of transformed features (`relationship`, `education`, `race`, |

| | | |
|---|---|---|
| | | `occupation`, `gender`, `marital`, `workclass`) used in modeling. |
| Data scaling/standardisation | No | No use of `StandardScaler` or similar observed before model training. |
| Hyperparameter tuning | No | `LogisticRegression` used as-is; evaluation done using `KFold` and `cross_val_score`, but no hyperparameter grid search applied. |

**Ground Truth**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Target distribution inspected via a loop printing `value_counts()` for every column (includes `income`). |
| Sampling type | Random | `train_test_split(df_x, df_y, test_size=0.33, random_state=42)` without `stratify`. |
| Outliers removal | No | No IQR/quantile/z-score masking or row deletion for outliers present. |
| Check for duplicates | No | No use of `.duplicated()` / `.drop_duplicates()` found. |
| Imputation of missing values | drop the missing value rows | Replaced `'?'` with NaN in `country`/`workclass`/`occupation`, then `df.dropna(how='any', inplace=True)`. |
| Drop columns | Yes | Pre-EDA drop of `['educational-num','age','hours per week','fnlwgt','capital gain','capital loss','country']` with no reuse. |
| Encoding | Label Encoder | Manual integer mapping for many categoricals (`income`, `gender`, `race`, `marital`, `workclass`, `education`, `occupation`, `relationship`). |
| Create new columns | No | `df_x` is an assembly of existing columns; mappings/dummies are derived and don't count as "new" per rubric. |

| | | |
|---|---|---|
| Feature selection | Yes | Post-EDA manual subset for modelling: `df_x` built from selected predictors after correlation/plots (relationship, education, race, occupation, gender, marital, workclass). |
| Data scaling/standardisation | No | No scaler (`StandardScaler`, `MinMaxScaler`, etc.) applied. |
| Hyperparameter tuning | No | K-Fold used for evaluation only; no `GridSearchCV`/`RandomizedSearchCV`/Optuna. |

**EIGHTH RECIPE**
**1st Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Used `value_counts()` and visualized class distribution before and after oversampling |
| Sampling type | Random | Used `train_test_split(..., random_state=0, test_size=0.33)` without `stratify` |
| Outliers removal | No | Explored correlation but no IQR/Z-score filtering applied |
| Check for duplicates | Yes | Used `drop_duplicates()` explicitly |
| Imputation of missing values | mixture of imputation techniques | Replaced ? with NaN, then filled with class-wise mode for 3 categorical columns |
| Drop columns | No | Columns were retained for correlation analysis and selection |
| Encoding | mixture of encoding | Used binary `.replace()` for binary variables and `get_dummies()` for others |
| Create new columns | No | No new feature creation; only transformations and filtering |
| Feature selection | Yes | Applied both feature reduction (correlation > 0.8) and correlation-based thresholding |

| | | |
|---|---|---|
| Data scaling/standardisation | Yes | Used `StandardScaler` (z-score normalization) on all numerical features |
| Hyperparameter tuning | No | No GridSearch or parameter sweep; default models used |

**2nd Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Imbalance in `income` detected and visualised using bar plot |
| Sampling type | Random | `train_test_split` used with `random_state=0` |
| Outliers removal | No | Visualised (e.g., heatmaps, histograms), but not explicitly removed |
| Check for duplicates | Yes | Used `.drop_duplicates()` |
| Imputation of missing values | mixture of imputation techniques | `"?"` replaced with `NaN`; imputed by **group-wise mode** per class |
| Drop columns | No | No explicit drop of unused columns during preprocessing |
| Encoding | mixture of encoding | Binary manual encoding + `pd.get_dummies()` with `drop_first=True` |
| Create new columns | No | No new derived columns created |
| Feature selection | Yes | Based on correlation with target > 0.05 threshold |
| Data scaling/standardisation | Yes | `StandardScaler` used on all numeric columns |
| Hyperparameter tuning | No | Default models used; no tuning applied for KNN, SVC, or LogisticRegression |

**3rd Prompt**
**Accuracy 10/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Explicitly checked and visualised class imbalance in `income`. |
| Sampling type | Oversampling | Oversampled the minority class (`income == 1`) to rebalance the dataset. |
| Outlier removal | No | No outlier filtering was performed. |
| Check for duplicates | Yes | Used `drop_duplicates()` to remove duplicate records. |
| Imputation of missing values | multivariate | Imputed by mode **within each income class** using `.groupby("income").transform(lambda: mode)`. |
| Drop columns | No | All columns retained for initial modeling; feature reduction done later, not hard drops. |
| Encoding | mixture of encoding | Binary encoding for 2-class columns + one-hot encoding for other categorical variables. |
| Create new columns | No | No new feature columns were engineered. |
| Feature selection | Yes | Used correlation with `income` target to retain high-impact variables only. |
| Standardization | Yes | Applied `StandardScaler()` to numeric columns using z-score scaling. |
| Hyperparameter tuning | No | Trained multiple models (KNN, SVM, LR), but no tuning (e.g., GridSearchCV) applied. |

**4th Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | `value_counts()` and bar plots used before and after oversampling |

| | | |
|---|---|---|
| Sampling type | Random | `train_test_split(..., test_size=0.33, random_state=0)` used without `stratify` |
| Outliers removal | No | No explicit outlier filtering (e.g., IQR or z-score based removal) |
| Check for duplicates | Yes | `drop_duplicates()` used on full dataframe |
| Imputation of missing values | mixture of imputation techniques | Replaced `"?"` with `np.nan`, then used class-wise mode filling via `groupby("income")` |
| Drop columns | No | No drop of ID or irrelevant columns (removed only highly correlated ones during reduction) |
| Encoding | mixture of encoding | Binary encoding for 2-level categorical + one-hot encoding for others |
| Create new columns | No | Feature expansion only via encoding; no new features constructed |
| Feature selection | Yes | Applied correlation-based selection (target correlation > 0.05 threshold) |
| Data scaling/standardisation | Yes | Used `StandardScaler` on all numerical columns |
| Hyperparameter tuning | No | Fixed K for KNN; SVM and Logistic Regression used with default parameters |

**5th Prompt**
**Accuracy 10/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Class distribution (`income`) visualised before and after oversampling using `value_counts().plot(kind='bar')`. |
| Sampling type | Oversampling | Minority class (`income = 1`) was oversampled to balance the classes. |

|  |  |  |
|---|---|---|
| Outliers removal | No | Z-score normalisation applied, but no explicit outlier filtering (e.g., IQR or capping) done. |
| Check for duplicates | Yes | Explicit call to `df.drop_duplicates()` to remove duplicate records. |
| Imputation of missing values | multivariate | NA imputation via grouped mode: `groupby("income")[col].transform(lambda x: x.fillna(x.mode()[0]))`. |
| Drop columns | No | No columns were dropped manually beyond correlation- and importance-based pruning. |
| Encoding | mixture of encoding | Binary columns encoded manually; rest transformed using `pd.get_dummies()` with `drop_first=True`. |
| Create new columns | No | No additional features created beyond preprocessing; feature space refined by selection and reduction. |
| Feature selection | Yes | Two-stage: (1) correlation-based feature reduction; (2) target correlation filtering using `corr_feature_selection(...)`. |
| Data scaling/standardisation | Yes | `StandardScaler` (Z-score normalisation) applied to all numeric columns. |
| Hyperparameter tuning | No | KNN k auto-derived as √n, but no GridSearchCV or RandomSearchCV used; models trained with default hyperparameters. |

**Ground Truth**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Target distribution checked via `df['income'].value_counts()` and bar plot of `income`. |
| Sampling type | Random | `train_test_split(X, y, random_state=0, test_size=0.33)` used without `stratify`. |

| | | |
|---|---|---|
| Outliers removal | No | No IQR/quantile/z-score masking or row deletion for outliers present. |
| Check for duplicates | Yes | Duplicates dropped with `df = df.drop_duplicates()`. |
| Imputation of missing values | use summary statistics | Replaced `'?'`→`NaN`, then **class-wise mode imputation** for `workclass`, `occupation`, `native-country` via `groupby('income').transform(lambda x: x.fillna(x.mode()[0]))`. |
| Drop columns | No | No pre-EDA unreplaced column drops; any correlation-based removals are handled under Feature selection. |
| Encoding | mixture of encoding | Manual binary mapping for 2-level categoricals, plus `pd.get_dummies(..., drop_first=True)` for remaining categorical columns. |
| Create new columns | No | One-hot dummies are derived from existing columns; no truly new features introduced. |
| Feature selection | Yes | (i) Correlation-threshold routine (`corr_reduction`) to drop highly correlated features (none removed at τ=0.8 here); (ii) Target-correlation filter keeps features with |
| Data scaling/standardisation | Yes | Z-score standardisation of **all numeric columns** using `StandardScaler` before modeling. |
| Hyperparameter tuning | No | Models (KNN with heuristic k, default SVC/LogisticRegression) trained without grid/random/Bayesian search. |

**NINTH RECIPE**
**1st Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|

| | | |
|---|---|---|
| Check for balanced data | Yes | Used `value_counts()` and visualised imbalance in `income`, followed by stratified split |
| Sampling type | Stratified | Used `train_test_split(..., stratify = y)` to maintain class distribution |
| Outliers removal | No | Visual exploration via histograms and skew checks, but no outlier filtering applied |
| Check for duplicates | No | Not performed (`duplicated()` or similar not used) |
| Imputation of missing values | drop the missing value rows | Replaced `'?'` with `np.nan`, then removed missing rows using `dropna()` |
| Drop columns | Yes | Dropped `'education rank'`, `'relationship'`, and `'country'` |
| Encoding | mixture of encoding | Applied `.map()` to manually encode all categorical features numerically |
| Create new columns | No | No engineered features added |
| Feature selection | Yes | Dropped multicollinear and low-importance columns; `relationship` removed for collinearity |
| Data scaling/standardisation | No | No use of `StandardScaler()` or other scaling |
| Hyperparameter tuning | Yes | Used `cross_val_score` with KFold and `n_neighbors` tuning in KNN from 1 to 30 |

**2nd Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Target variable `income` distribution explicitly checked and visualised |
| Sampling type | Stratified | `train_test_split` used with `stratify=y` |

| | | |
|---|---|---|
| Outliers removal | No | Outliers visualised (e.g., `age`, `Final Weight`) but not removed |
| Check for duplicates | No | No `.drop_duplicates()` call found |
| Imputation of missing values | drop the missing value rows | Replaced `"?"` with `np.nan`, then dropped missing rows with `dropna()` |
| Drop columns | Yes | Dropped `education rank`, `country`, and `relationship` without reuse |
| Encoding | mixture of encoding | Used `.map()` for manual encoding of categorical variables |
| Create new columns | No | No new features engineered beyond renaming or transformation |
| Feature selection | Yes | Dropped features based on correlation and multicollinearity |
| Data scaling/standardisation | No | No use of `StandardScaler` or normalization |
| Hyperparameter tuning | Yes | Used `cross_val_score` for Logistic Regression, Random Forest, and KNN; KNN also tuned for optimal k |

**3rd Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Class imbalance of `income` explicitly observed and visualised. |
| Sampling type | Stratified | Used `train_test_split(..., stratify=y)` to preserve class distribution. |
| Outlier removal | No | Skewness observed but no capping, trimming, or filtering applied. |
| Check for duplicates | No | No duplicate handling (e.g., `drop_duplicates()`) was used. |

| | | |
|---|---|---|
| Imputation of missing values | drop the missing value rows | Replaced `"?"` with `np.nan` and used `dropna()` to remove missing entries. |
| Drop columns | Yes | Dropped `education rank`, `relationship`, and `country` features. |
| Encoding | Map to ordinal values | All categorical features encoded using `map()` to ordinal integers. |
| Create new columns | No | No new columns created; all transformations were encodings. |
| Feature selection | Yes | Removed multicollinear and weak features using correlation matrix (e.g., dropped `relationship`). |
| Standardization | No | Did not apply scaling (e.g., `StandardScaler`) prior to training models. |
| Hyperparameter tuning | Yes | K-fold CV used with Logistic Regression, Random Forest, and KNN to select optimal parameters. |

**4th Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | `value_counts()` and visual plots used for income distribution |
| Sampling type | Stratified | `train_test_split(..., stratify=y)` used to preserve class balance |
| Outliers removal | No | Right-skewed distributions identified, but no capping/removal applied |
| Check for duplicates | No | No `drop_duplicates()` used |
| Imputation of missing values | drop the missing value rows | Replaced `"?"` with `np.nan`, then removed using `dropna()` |
| Drop columns | No | Dropped `education rank`, `relationship`, and `country` before feature coding |

| | | |
|---|---|---|
| Encoding | mixture of encoding | Categorical variables mapped to integers using hard-coded dictionaries |
| Create new columns | No | No new features added |
| Feature selection | Yes | Selected 11 features manually; used `feature_importances_` from RF to guide interpretation |
| Data scaling/standardisation | No | No scaler used (raw features passed to LR, DT, RF, KNN) |
| Hyperparameter tuning | No | Cross validation performed but not for hyperparameter tuning. |

**5th Prompt**
**Accuracy 11/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Target (`income`) balance explored via `value_counts`, visualisations, and justification for stratification in train-test split. |
| Sampling type | Stratified | `train_test_split(..., stratify=y)` used to preserve class distribution. |
| Outliers removal | No | Although skew and distributions were discussed, no IQR or Z-score based removal or adjustment was applied. |
| Check for duplicates | No | No `drop_duplicates()` or `duplicated()` call in pipeline. |
| Imputation of missing values | drop the missing value rows | Replaced ? with `NaN`, then dropped all rows with any missing value using `dropna(how='any')`. |
| Drop columns | No | Columns dropped: `education rank`, `relationship`, and `country` after EDA |
| Encoding | mixture of encoding | All categorical columns manually mapped using `.map()` dictionaries with hardcoded ordinal and nominal schemes. |

| | | |
|---|---|---|
| Create new columns | No | No additional features were derived or engineered beyond existing columns. |
| Feature selection | Yes | `relationship` dropped due to multicollinearity (with `marital status`), and low-impact variables like `country` removed based on correlation. |
| Data scaling/standardisation | No | No use of `StandardScaler` or other scaling methods observed across any model pipeline. |
| Hyperparameter tuning | No | Cross-validation (`cross_val_score`) used for model evaluation;not for hyperparameter tuning |

**Ground Truth**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Target distribution inspected via `Salary['income'].value_counts(normalize=True)` and `sns.countplot(Salary['income'])`. |
| Sampling type | Stratified | `train_test_split(..., stratify=y, test_size=0.3, random_state=30)`. |
| Outliers removal | No | No IQR/quantile/z-score filtering or row deletion for outliers. |
| Check for duplicates | No | No use of `.duplicated()` / `.drop_duplicates()` found. |
| Imputation of missing values | drop the missing value rows | Replaced `'?'` with `NaN` in `workclass/occupation/native-country`, then `Salary.dropna(how='any', inplace=True)`. |
| Drop columns | No | Columns removed for modelling were dropped **after EDA** (`['education rank','country','relationship']`), so counted under Feature selection (not Drop columns). |

| | | |
|---|---|---|
| Encoding | Label Encoder | Manual integer mapping for categoricals (`education`, `workclass`, `marital status`, `occupation`, `gender`, `race`, and `income`). |
| Create new columns | No | No truly new features; recodes/mappings are transformations of existing columns. |
| Feature selection | Yes | Post-EDA pruning: dropped `['education rank','country','relationship']`; justification via correlation/EDA commentary. |
| Data scaling/standardisation | No | No scaler (`StandardScaler`, `MinMaxScaler`, etc.) applied. |
| Hyperparameter tuning | No | CV used for evaluation; manual K sweep for KNN but no `GridSearchCV`/`RandomizedSearchCV`/Optuna. |

**TENTH RECIPE**
**1st Prompt**
**Accuracy 8/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Used `value_counts()` and `StratifiedShuffleSplit` due to class imbalance |
| Sampling type | Stratified | Used `StratifiedShuffleSplit` to maintain income class distribution |
| Outliers removal | No | No IQR or Z-score filtering used; distributions explored visually |
| Check for duplicates | No | No use of `duplicated()` or `.drop_duplicates()` |
| Imputation of missing values | none | Null percentages checked; no `dropna()` or imputation used |
| Drop columns | Yes | Dropped original categorical columns after one-hot encoding |

|  |  |  |
|---|---|---|
| Encoding | mixture of encoding | Used `.map()` for ordered categories and `OneHotEncoder()` for others |
| Create new columns | Yes | Converted education, occupation, and country into ordered categories |
| Feature selection | Yes | Selected subset of features for training (`x_train`, `x_test`) |
| Data scaling/standardisation | Yes | Applied `MinMaxScaler` to continuous features |
| Hyperparameter tuning | No | Compared models (Logistic, Random Forest, KNN, etc.) but no tuning done |

**2nd Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Class imbalance detected and stratified sampling used |
| Sampling type | Stratified | `StratifiedShuffleSplit` used to maintain class distribution |
| Outliers removal | No | Outliers visualised (via `describe`), not removed |
| Check for duplicates | No | No `.drop_duplicates()` or equivalent used |
| Imputation of missing values | None | Null values were checked but not imputed or dropped |
| Drop columns | Yes | Dropped all object columns after one-hot encoding |
| Encoding | mixture of encoding | Used manual ordinal encoding + `OneHotEncoder` |
| Create new columns | No | No new columns derived or added beyond encoding |
| Feature selection | Yes | Manual selection of numeric columns + binary encoding and removal of unused categorical fields |

| | | |
|---|---|---|
| Data scaling/standardisation | Yes | MinMaxScaler used on non-categorical features |
| Hyperparameter tuning | No | Models trained with fixed parameters (max_depth, etc.) |

**3rd Prompt**
**Accuracy 9/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Value counts and stratified splitting indicate imbalance in target (income). |
| Sampling type | Stratified | Used StratifiedShuffleSplit() to preserve class distribution. |
| Outlier removal | No | Extensive analysis performed, but no removal or capping of outliers. |
| Check for duplicates | No | No use of drop_duplicates() detected. |
| Imputation of missing values | replace with text | Filled missing categorical values (object) by assigning meaningful codes and collapsing classes. |
| Drop columns | Yes | Dropped many columns including remaining object features after encoding. |
| Encoding | mixture of encoding | Manual ordinal encoding + one-hot encoding (OneHotEncoder) for object columns. |
| Create new columns | No | No additional engineered features; only transformed existing ones. |
| Feature selection | Yes | Removed high-cardinality or low-impact variables like country, education-num, fnlwgt, etc. |
| Standardization | Yes | Scaled continuous features using MinMaxScaler. |
| Hyperparameter tuning | No | Trained multiple models (Logistic, KNN, RF, DT) but did not tune hyperparameters (e.g., GridSearch not used). |

**4th Prompt**

**Accuracy 10/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | `value_counts()` and bar plots used; class imbalance explicitly addressed |
| Sampling type | Stratified | `StratifiedShuffleSplit(..., stratify=target)` used for balanced train-test split |
| Outliers removal | No | Distributions and bar charts visualised, but no outlier removal (IQR/z-score) performed |
| Check for duplicates | No | No use of `drop_duplicates()` or similar |
| Imputation of missing values | none | No imputation or NA/“?” handling observed |
| Drop columns | No | Columns dropped and replaced with processed data |
| Encoding | mixture of encoding | Ordinal encoding for categorical ranking + OneHotEncoder for other categorical columns |
| Create new columns | No | Encoding expands column space but no engineered features added |
| Feature selection | Yes | Selected `x_train` and `x_test` features manually after categorical encoding |
| Data scaling/standardisation | Yes | `MinMaxScaler` applied to numeric columns before modeling |
| Hyperparameter tuning | No | Models used with fixed parameters; no `GridSearchCV` or tuning strategy |

**5th Prompt**
**Accuracy 10/11**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Class distribution of `income` checked using `value_counts()` and justified stratification based on imbalance. |

| | | |
|---|---|---|
| Sampling type | Stratified | `StratifiedShuffleSplit` used to split the dataset while preserving class distribution. |
| Outliers removal | No | No outlier detection or filtering applied, though `describe()` and categorical effects were analysed extensively. |
| Check for duplicates | No | No `drop_duplicates()` or similar function used. |
| Imputation of missing values | none | No explicit imputation for missing values; presumed clean after load, with no ? to `NaN` conversion. |
| Drop columns | No | Columns were transformed but not permanently dropped except during `drop(object_columns)` step pre-encoding. |
| Encoding | mixture of encoding | Applied ordered `pd.Categorical(...).codes`, manual replacements (e.g. country to binary), and one-hot encoding for remaining object columns. |
| Create new columns | No | No engineered features created; only transformed encodings and reshaped input matrix. |
| Feature selection | Yes | Kept only influential variables (e.g., dropped some high-cardinality or low-informative ones such as `country`, then selected top 10 manually). |
| Data scaling/standardisation | Yes | `MinMaxScaler()` used to scale continuous numeric features before modeling. |
| Hyperparameter tuning | No | Fixed hyperparameters passed (e.g., `max_depth=7`), but no tuning loop or search strategy implemented. |

**Ground Truth**

| Data Wrangling Step | Technique Used | Details |
|---|---|---|
| Check for balanced data | Yes | Used value_counts() and bar plots |

| | | |
|---|---|---|
| Sampling type | Stratified | `StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)` used to preserve class ratios in train/test. |
| Outliers removal | No | No IQR/quantile/z-score filtering or row deletion performed. |
| Check for duplicates | No | No use of `.duplicated()` / `.drop_duplicates()`. |
| Imputation of missing values | none | Nulls are inspected, but no `fillna`/imputer or NA row/column drops applied. |
| Drop columns | No | Object columns are dropped only after being replaced by their One-Hot encoded matrix (`adult.drop(a, ...)` after `encoder.fit_transform(adult[a])`), so not counted as unreplaced drops. |
| Encoding | mixture of encoding | Manual ordinal encodes for workclass/education/occupation; binary recode for country (US vs rest); OneHotEncoder applied to remaining object columns. |
| Create new columns | No | One-hot features appended via `pd.concat` are derived from existing categoricals (not “new” under the rubric). |
| Feature selection | No | No correlation/variance/model-importance pruning or post-EDA column removal; features assembled as all except the label. |
| Data scaling/standardisation | Yes | `MinMaxScaler()` fit to a subset of non-categorical columns (`adult[not_cat] = scaler.fit_transform(...)`). |
| Hyperparameter tuning | No | Models trained with fixed/default parameters (LogisticRegression/RandomForest/DecisionTree/KNN); no `GridSearchCV`/`RandomizedSearchCV`/Optuna. |