First Recipe
1st Prompt
Accuracy 9/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No explicit check (e.g., value counts or plots of class distribution)
Sampling type	Yes	`train_test_split` with default behavior = random sampling
Outliers removal	Yes	Outliers in `'hsc_p'` removed using IQR filtering
Check for duplicates	No	No `.duplicated()` or `.drop_duplicates()` present
Imputation of missing values	Yes	`'salary'` filled with `0` → replace with text/numeric placeholder
Drop columns	Yes	`['sl_no', 'ssc_b', 'hsc_b']` dropped as irrelevant
Encoding	Yes	Mixture of encoding → `LabelEncoder` for binary; `get_dummies()` for multiclass
Create new columns	Yes	Dummies for `'hsc_s'`, `'degree_t'` created
Feature selection	Yes	Low-importance dummy features dropped before model fitting
Data scaling/standardisation	No	No scaler or standardisation applied before modeling
Hyperparameter tuning	No	No grid/random search used; only default parameters tested

2nd Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Used `sns.countplot(x="status")` and `value_counts()` on `"status"` column to check placement distribution

Sampling type	Random	Used `train_test_split()` without stratification or oversampling
Outliers removal	Yes	IQR method on `hsc_p` to filter outliers
Check for duplicates	No	No duplicate check (`.duplicated()` or `drop_duplicates()`) found
Imputation of missing values	Replace with 0	`salary` missing values filled with 0, justified by domain context
Drop columns	Yes	Dropped `sl_no`, `ssc_b`, and `hsc_b`
Encoding	Mixture	Used `LabelEncoder` for binary, `pd.get_dummies` for multi-class
Create new columns	Yes	Dummy columns added (not derived from transformation of values)
Feature selection	Yes	Pruned features based on low RandomForest importance scores
Data scaling or standardisation	No	No `StandardScaler`, `MinMaxScaler`, or related used
Hyperparameter tuning	No	Parameters like `max_depth`, `n_estimators` manually set without systematic tuning or search

3rd Prompt
Accuracy 10/11

#	Question	Answer	Snippet / Explanation
1	Is data checked for balance?	Yes	Countplot on `status`
2	Sampling type?	Random	`train_test_split` used
3	Have outliers been removed?	Yes	IQR method on `hsc_p`
4	Does the code check for duplicates?	No	No duplicate checks present
5	Does the code impute missing values?	Yes	`fillna(value=0)` on `salary`
6	Does the code drop columns?	Yes	Dropped `sl_no`, `ssc_b`, `hsc_b`
7	What encoding is used?	LE, OHE	Label Encoding + Dummies via `pd.get_dummies()`

8	Does the code create new columns?	Yes	Dummy vars from categorical features
9	Does the code perform feature selection?	Yes	Manual pruning of low-importance features
10	Does the code perform standardization?	No	No scaler used
11	Does the code perform hyperparameter tuning?	No	No cross-validation tuning or parameter grid search

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Countplot of 'status' in 5a and 5e visualizations
Sampling type	Random	`train_test_split` without `stratify`
Outliers removal	Yes	IQR-based filtering on `hsc_p` before creating `placement_filtered`
Check for duplicates	No	No use of `duplicated()` or `drop_duplicates()`
Imputation of missing values	replace with text	Missing salaries filled with `0` to denote unplaced students
Drop columns	Yes	Dropped `sl_no`, `ssc_b`, `hsc_b`
Encoding	mixture of encoding	LabelEncoder for binary cols; get_dummies for multi-class categories
Create new columns	No	Dummies and engineered columns derived from existing ones
Feature selection	Yes	Dropped dummy vars after Random Forest feature importance
Data scaling/standardisation	No	No scaler used; raw features passed to models
Hyperparameter tuning	No	K value manually selected; no GridSearchCV/RandomSearch/Optuna used

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	`sns.countplot(x="status")` was used to check class distribution.
Sampling type	Random	Used `train_test_split` without `stratify`.
Outliers removal	Yes	Outliers in `hsc_p` removed using IQR filtering.
Check for duplicates	No	No check for `.duplicated()` or `.drop_duplicates()` detected.
Imputation of missing values	replace with text	`salary` column NaNs replaced with `0` based on domain reasoning.
Drop columns	Yes	`'sl_no','ssc_b','hsc_b' columns are dropped.`
Encoding	mixture of encoding	`LabelEncoder` for binary vars + `pd.get_dummies()` for multi-class.
Create new columns	No	All derived columns (dummy vars) are based on existing features.
Feature selection	Yes	Features dropped based on model-based importance scores from Random Forest.
Data scaling/standardisation	No	No scaler (e.g., StandardScaler, MinMaxScaler) applied anywhere.
Hyperparameter tuning	No	Fixed parameters used; no GridSearchCV or similar method.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	`sns.countplot(x="status")` to view class distribution.

Sampling type	Random	`train_test_split(X, y, train_size=0.8, random_state=1)` without `stratify`.
Outliers removal	Yes	IQR filter on `hsc_p` to create `placement_filtered`.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()`.
Imputation of missing values	Replace with text/numeric placeholder	`placement_copy['salary'].fillna(0)` (0 as “no income”).
Drop columns	Yes	Dropped `['sl_no','ssc_b','hsc_b']` during cleaning (not reused).
Encoding	Mixture	`LabelEncoder` for binary (`gender`,`workex`,`specialisation`,`status`) and `pd.get_dummies()` for `hsc_s`,`degree_t`.
Create new columns	No	One-hot dummies are derived from existing columns, so not counted as “new”.
Feature selection	Yes	Model-based: dropped low-importance `dummy_*` features after RF importances; also dropped `salary` post-visualisation (counts as feature selection).
Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) used.
Hyperparameter tuning	No	No systematic search (no `GridSearchCV`/`RandomizedSearchCV`); parameters set manually (e.g., `max_depth=3`, K sweep without CV search).

Second Recipe

1st Prompt

Accuracy 11/11

Data Wrangling Step	Technique Used	Details

Check for balanced data	Yes	`placement.status.value_counts(normalize=True)` checks target class balance
Sampling type	No	No train-test split or sampling performed in this code
Outliers removal	No	No filtering or removal using IQR or other criteria
Check for duplicates	No	No `.duplicated()` check in code
Imputation of missing values	No	`placement.isnull().sum()` called, but no action taken
Drop columns	No	No column dropped anywhere
Encoding	No	No encoding applied; all categorical variables remain as-is
Create new columns	No	No new features or derived variables added
Feature selection	No	No indication of dropping low-importance variables
Data scaling/standardisation	No	No scaler, standardizer, or normalization used
Hyperparameter tuning	No	No model training, hence no tuning involved

2nd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Used `value_counts(normalize=True)` on target `status`
Sampling type	No	No train-test split performed
Outliers removal	No	Visualizations done but no removal
Check for duplicates	No	Not checked
Imputation of missing values	No	`.isnull()` checked but not handled
Drop columns	No	None dropped
Encoding	No	No encoding performed

Create new columns	No	None created
Feature selection	No	No selection done
Data scaling or standardisation	No	No scaling performed
Hyperparameter tuning	No	No modeling applied

3rd Prompt
Accuracy 11/11

#	Question	Answer	Snippet / Explanation
1	Is data checked for balance?	Yes	`placement.status.value_counts(normalize=True)`
2	Sampling type?	No	No train/test split present
3	Have outliers been removed?	No	No IQR/z-score/capping applied
4	Does the code check for duplicates?	No	No `.duplicated()` or similar used
5	Does the code impute missing values?	No	Nulls identified but not filled
6	Does the code drop columns?	No	All columns retained
7	What encoding is used?	None	No `LabelEncoder`, `get_dummies`, etc.
8	Does the code create new columns?	No	No feature engineering
9	Does the code perform feature selection?	No	No column pruning
10	Does the code perform standardization?	No	No use of `StandardScaler` or similar
11	Does the code perform hyperparameter tuning?	No	No `GridSearchCV` or param variation

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Countplots on `status`, `workex`, `specialisation` with `hue='status'`
Sampling type	None	No `train_test_split` or sampling code present
Outliers removal	No	No filtering, capping, or IQR logic used
Check for duplicates	No	No `duplicated()` or `drop_duplicates()` used
Imputation of missing values	none	No handling or filling of missing `salary` values
Drop columns	No	No column removed from dataset
Encoding	none	No encoding applied to categorical variables
Create new columns	No	No derived or additional columns created
Feature selection	No	No correlation/importance-based or manual column pruning
Data scaling/standardisation	No	No scaling functions or transformers used
Hyperparameter tuning	No	No models used; hence no tuning

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Used `placement.status.value_counts(normalize=True)` to inspect class balance.
Sampling type	None	No train-test split or model training present in this script.
Outliers removal	No	No IQR, boxplot filtering, or removal code used; only visual analysis.

Check for duplicates	No	No `.duplicated()` or similar check in the notebook.
Imputation of missing values	none	Although `placement.isnull().sum()` is shown, no imputation or handling was applied.
Drop columns	No	All columns retained.
Encoding	none	No LabelEncoder or OneHotEncoder or dummy variables applied.
Create new columns	No	No new columns created, all analysis done on original variables.
Feature selection	No	No columns dropped based on correlation or feature importance.
Data scaling/standardisation	No	No scaling or normalization applied.
Hyperparameter tuning	No	No models were trained, hence no tuning attempted.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Used `placement.status.value_counts(normalize=True)` to inspect class distribution.
Sampling type	No	No train–test split (no `train_test_split` or similar).
Outliers removal	No	Only visual checks (pairplot/boxplots); no filtering applied.
Check for duplicates	No	No `.duplicated()` or `.drop_duplicates()` used.
Imputation of missing values	No	Missingness only inspected via `placement.isnull().sum()`; no imputation.
Drop columns	No	No columns removed from analysis.
Encoding	No	No label/one-hot/CatBoost encoding performed.

Create new columns	No	No new features created.
Feature selection	No	No correlation/model-based drops; only visual EDA.
Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) used.
Hyperparameter tuning	No	No systematic search (no Grid/Random/Optuna).

Third recipe
1st Prompt
Accuracy 9/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Class ratio checked using `value_counts()` for gender and status
Sampling type	Yes	`train_test_split(..., shuffle=True)` → random sampling
Outliers removal	No	No IQR or z-score based outlier filtering
Check for duplicates	No	No `.duplicated()` or `.drop_duplicates()` used
Imputation of missing values	Yes	`salary.fillna(0)` → replaced with numeric placeholder (zero)
Drop columns	Yes	Dropped `sl_no`, `ssc_b`, `hsc_b`, `salary`, and original percent columns after binning
Encoding	Mixture of encoding	Manual mapping + one-hot (`get_dummies`) on categorical variables
Create new columns	Yes	New binned categorical columns: `ssc_p_c`, `hsc_p_c`, etc.
Feature selection	Yes	Columns with low predictive value (like board names) were dropped
Data scaling/standardisation	No	No use of `StandardScaler`, `MinMaxScaler`, or normalization

Hyperparameter tuning	Yes	`GridSearchCV` used for tuning Random Forest hyperparameters

2nd Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Distribution of `status` and gender analyzed
Sampling type	Random	Used `train_test_split()` with shuffle
Outliers removal	No	No filtering or IQR/Z-score logic used
Check for duplicates	No	No deduplication method used
Imputation of missing values	Replace with 0	Salary missing values replaced with `0`
Drop columns	Yes	Dropped `sl_no`, `ssc_b`, `hsc_b`, `salary`
Encoding	Mixture	Binary encoding with `.map()`, one-hot with `pd.get_dummies()`
Create new columns	No	New columns derived from existing ones (percentage binning)
Feature selection	Yes	Removed features via heatmap + domain knowledge
Data scaling or standardisation	No	No scaling techniques used
Hyperparameter tuning	Yes	Used `GridSearchCV` on `RandomForestClassifier`

3rd Prompt
Accuracy 9/11

#	Question	Answer	Snippet / Explanation
1	Is data checked for balance?	Yes	`df['status'].value_counts()` and catplots
2	Sampling type?	Random	`train_test_split(..., shuffle=True)`
3	Have outliers been removed?	No	No outlier removal logic present

4	Does the code check for duplicates?	No	No `.duplicated()` used
5	Does the code impute missing values?	Yes	`df['salary'].fillna(0, inplace=True)`
6	Does the code drop columns?	Yes	Dropped many including `sl_no`, `salary`, `ssc_b`, etc.
7	What encoding is used?	LE, OHE	Used `map()` and `get_dummies()`
8	Does the code create new columns?	Yes	Created `*_p_c` binned columns
9	Does the code perform feature selection?	Yes	Dropped based on heatmap & logic
10	Does the code perform standardization?	No	No scaling/scaler used
11	Does the code perform hyperparameter tuning?	Yes	Used `GridSearchCV` for RandomForestClassifier

4th Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Countplots of `status` with hue; placement ratios computed by gender
Sampling type	Random	`train_test_split(..., shuffle=True)` without stratify
Outliers removal	No	No IQR filtering or outlier logic applied
Check for duplicates	No	No `duplicated()` or `drop_duplicates()` used
Imputation of missing values	replace with text	Filled `salary` NaNs with 0
Drop columns	Yes	Dropped `sl_no`, `ssc_b`, `hsc_b`, `salary` without replacement

Encoding	mixture of encoding	Label encoding + pd.get_dummies for `hsc_s`, `degree_t`
Create new columns	No	All derived columns from existing values (e.g. `_c` binning)
Feature selection	Yes	Dropped columns based on correlation heatmap (`ssc_b`, `hsc_b`)
Data scaling/standardisation	No	No scaler or normalization applied
Hyperparameter tuning	Yes	GridSearchCV used for Random Forest `n_estimators` and `max_depth`

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Value counts used on `status` and gender-specific placement ratios calculated.
Sampling type	Random	Used `train_test_split` without `stratify`.
Outliers removal	No	No filtering or removal of extreme values based on IQR or z-score.
Check for duplicates	No	No `.duplicated()` or `.drop_duplicates()` call found.
Imputation of missing values	replace with text	Replaced missing `salary` values with `0`.
Drop columns	No	Columns like `sl_no`, `ssc_b`, `hsc_b`, and `salary` dropped post-correlation or EDA → counted under feature selection.
Encoding	mixture of encoding	Used `.map()` for binary features and `pd.get_dummies()` for multi-class (e.g., `degree_t`, `hsc_s`).
Create new columns	No	Created score bands (e.g., `ssc_p_c`) derived directly from existing columns.
Feature selection	Yes	Columns removed after correlation heatmap (e.g., `sl_no`, `ssc_b`, etc.).

Data scaling/standardisation	No	No scaling methods (StandardScaler, MinMaxScaler) applied.
Hyperparameter tuning	Yes	`GridSearchCV` used for Random Forest (`n_estimators`, `max_depth`).

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	`df['status'].value_counts()` used to inspect target distribution; multiple count plots with hue=`'status'` in EDA.
Sampling type	Random	`train_test_split(X, Y, test_size=0.3, random_state=0, shuffle=True)` with no `stratify`.
Outliers removal	No	No IQR/quantile/rule-based filtering applied.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()`.
Imputation of missing values	Replace with text/numeric placeholder	`df['salary'].fillna(0, inplace=True)` (0 to indicate no salary for not placed).
Drop columns	No	Drops of `sl_no`, `ssc_b`, `hsc_b`, `salary` are counted under Feature selection (post-EDA).
Encoding	Mixture	Binary mapping via `.map()` for `gender`, `ssc_b`, `hsc_b`, `workex`, `specialisation`, `status`; one-hot with `pd.get_dummies()` for `hsc_s` and `degree_t`.
Create new columns	No	Binned score features (`ssc_p_c`, `hsc_p_c`, `degree_p_c`, `mba_p_c`, `etest_p_c`) are derived from existing columns; this does not count as creating new columns.
Feature selection	Yes	After correlation heatmap/EDA, dropped `sl_no`, `ssc_b`, `hsc_b`, `salary` before modelling

		(drop-after-visualisation counts as feature selection).
Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) applied before KNN/SVC.
Hyperparameter tuning	Yes	Systematic search with `GridSearchCV` for `RandomForestClassifier` over `n_estimators` and `max_depth`.

Fourth Recipe
1st Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No visual or statistical check for class imbalance performed
Sampling type	Yes	Used `train_test_split()` → random sampling
Outliers removal	No	No filtering or IQR/Z-score outlier detection
Check for duplicates	No	No use of `.duplicated()` or `.drop_duplicates()`
Imputation of missing values	No	Missing values printed, but not handled or filled
Drop columns	Yes	Dropped `sl_no` via `index_col` in `read_csv()`
Encoding	Mixture of encoding	`OneHotEncoder` on features, `LabelEncoder` on target
Create new columns	No	No new feature engineering or column creation
Feature selection	No	All available features used without pruning
Data scaling/standardisation	No	No scaling or normalization before model training
Hyperparameter tuning	No	Models trained with fixed parameter values only

2nd Prompt
Accuracy 9/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Status distribution was examined
Sampling type	Random	`train_test_split()` used
Outliers removal	No	No filtering or outlier detection used
Check for duplicates	No	No deduplication checks performed
Imputation of missing values	No	Null values identified but not handled
Drop columns	Yes	`sl_no` dropped via index on import
Encoding	Mixture	Label + OneHot Encoding applied
Create new columns	No	No derived columns
Feature selection	No	No dimensionality reduction or feature removal
Data scaling or standardisation	No	No normalization used
Hyperparameter tuning	No	Models have manually set hyperparameters, no tuning strategy

3rd Prompt
Accuracy 10/11

#	Question	Answer	Snippet / Explanation
1	Is data checked for balance?	No	No check for class distribution (`status`)
2	Sampling type?	Random	`train_test_split(..., random_state=20)`
3	Have outliers been removed?	No	No removal or visualization
4	Does the code check for duplicates?	No	No `.duplicated()` or equivalent used
5	Does the code impute missing values?	No	Salary missing values are left as-is
6	Does the code drop columns?	Yes	Dropped `sl_no` (via `index_col`)

7	What encoding is used?	LE, OHE	LabelEncoder on y; OneHotEncoder on x
8	Does the code create new columns?	No	No engineered or binned features
9	Does the code perform feature selection?	No	No pruning based on correlation or importance
10	Does the code perform standardization?	No	No scaling applied
11	Does the code perform hyperparameter tuning?	No	Static hyperparameters (e.g., C=0.6 for SVM)

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No countplots or distribution checks for class imbalance
Sampling type	Random	`train_test_split(..., random_state=20)` without stratify
Outliers removal	No	No IQR filtering or outlier logic present
Check for duplicates	No	No use of `duplicated()` or `drop_duplicates()`
Imputation of missing values	none	Missing values in `salary` noted but not handled
Drop columns	No	`sl_no` used as index, but not dropped within feature matrix explicitly
Encoding	mixture of encoding	`OneHotEncoder` on features, `LabelEncoder` on target `status`
Create new columns	No	No new features created beyond encoded columns
Feature selection	No	No features dropped based on correlation, importance, or logic
Data scaling/standardisation	No	No scalers (StandardScaler, MinMax, etc.) applied

Hyperparameter tuning	No	Fixed parameters passed manually (e.g. `n_neighbors=29`, `C=0.6`)

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No value_counts or class distribution check for `status` seen.
Sampling type	Random	`train_test_split` used without `stratify`.
Outliers removal	No	No handling or filtering of outliers present.
Check for duplicates	No	No use of `.duplicated()` or `.drop_duplicates()`.
Imputation of missing values	none	`isnull().sum()` checked but no imputation applied.
Drop columns	No	Columns are retained (even `salary`, which is unused but undropped).
Encoding	mixture of encoding	Used `OneHotEncoder` for features and `LabelEncoder` for target.
Create new columns	No	No new columns created.
Feature selection	No	All original columns retained.
Data scaling/standardisation	No	No scaling/standardisation applied.
Hyperparameter tuning	No	All models used fixed parameters; no GridSearch or tuning method used.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No value counts or class-distribution plots for `status`.

Sampling type	Random	`train_test_split(x, y, test_size=0.2, random_state=20)` with no `stratify`.
Outliers removal	No	No IQR/quantile/rule-based filtering present.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()`.
Imputation of missing values	No	Missingness only inspected via `data.isnull().sum()`; no imputation performed.
Drop columns	No	`sl_no` used by loading as index (`index_col='sl_no'`).
Encoding	Mixture	`OneHotEncoder()` applied to features `x`; `LabelEncoder()` applied to target `y`.
Create new columns	No	One-hot outputs are derived from existing columns, so not counted as “new”.
Feature selection	No	No correlation/model-based selection; no post-EDA drops.
Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) used.
Hyperparameter tuning	No	Only manual parameter settings (e.g., `n_neighbors=29`, `C=0.6`); no Grid/Random/Optuna search.

Fifth recipe
1st Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Gender-based placement comparisons; visual and Highcharts-based analysis
Sampling type	Yes	`train_test_split(..., shuffle=True)` → random sampling
Outliers removal	No	No IQR/Z-score or manual removal

Check for duplicates	No	No `.duplicated()` or `.drop_duplicates()` call
Imputation of missing values	No	Acknowledged missing salary values but not imputed
Drop columns	Yes	Dropped: `sl_no`, `Others`, `Arts`, `ssc_b`, `hsc_b`, `salary`, etc.
Encoding	Mixture of encoding	LabelEncoding for binary categories, OneHotEncoding for multiclass
Create new columns	No	No feature engineering or binning
Feature selection	Yes	Dropped low-correlation variables based on heatmap
Data scaling/standardisation	Yes	Used `StandardScaler()` before logistic regression
Hyperparameter tuning	No	Used default parameters for all classifiers

2nd Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Done via value counts
Sampling type	Random	`train_test_split()` used
Outliers removal	No	Not applied
Check for duplicates	No	Not applied
Imputation of missing values	Yes (drop)	Salary dropped to handle NaNs
Drop columns	Yes	Low-correlation and ID columns dropped
Encoding	Mixed	Label + One-Hot Encoding used
Create new columns	Yes	Dummy vars added
Feature selection	Yes	Based on correlation
Data scaling or standardisation	Yes	StandardScaler used

Hyperparameter tuning	No	Defaults used in LogisticRegression

3rd Prompt
Accuracy 9/11

#	Question	Answer	Snippet / Explanation
1	Is data checked for balance?	Yes	`value_counts()` and multiple bar plots
2	Sampling type?	Random	`train_test_split(..., shuffle=True)`
3	Have outliers been removed?	No	No IQR/z-score or filters used
4	Does the code check for duplicates?	No	No `.duplicated()` or `.drop_duplicates()`
5	Does the code impute missing values?	No	Salary missing values acknowledged but ignored
6	Does the code drop columns?	Yes	Dropped via `data.drop([...], axis=1)`
7	What encoding is used?	LE, OHE	Used both label and one-hot encoding
8	Does the code create new columns?	No	No binning or derived variables
9	Does the code perform feature selection?	Yes	Based on correlation matrix analysis
10	Does the code perform standardization?	Yes	Used `StandardScaler()`
11	Does the code perform hyperparameter tuning?	No	All models use default settings

4th Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Multiple barplots showing placement counts by gender, board, etc.

Sampling type	Random	`train_test_split(..., shuffle=True)` without stratify
Outliers removal	No	No IQR-based filtering or explicit removal
Check for duplicates	No	No `duplicated()` or `drop_duplicates()` used
Imputation of missing values	none	Missing `salary` values acknowledged but not filled or removed
Drop columns	Yes	Dropped `sl_no`, `salary`, `ssc_b`, `hsc_b`
Encoding	mixture of encoding	Label encoding + `get_dummies` for `hsc_s`, `degree_t`
Create new columns	No	Only encoding done; no derived or engineered features
Feature selection	Yes	Features dropped based on correlation matrix (negative correlation)
Data scaling/standardisation	Yes	Applied `StandardScaler` to feature matrix before train-test split
Hyperparameter tuning	No	No systematic tuning; models used with fixed parameters

5th Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Multiple `value_counts()` checks on `status`, `gender`, etc., used to inspect class distribution.
Sampling type	Random	Used `train_test_split` with `shuffle=True` and no `stratify`.
Outliers removal	No	No code to filter or remove outliers via IQR, z-score, etc.
Check for duplicates	No	No `.duplicated()` or `.drop_duplicates()` check.

Imputation of missing values	none	Acknowledged `salary` has missing values but no imputation applied.
Drop columns	No	Columns like `salary`, `ssc_b`, `hsc_b`, `Others`, etc., dropped post-correlation analysis → counts as feature selection.
Encoding	mixture of encoding	Used `LabelEncoder` for binary variables and `pd.get_dummies()` for multiclass (`hsc_s`, `degree_t`).
Create new columns	No	No additional columns created beyond encoding.
Feature selection	Yes	Removed columns based on correlation heatmap and domain knowledge (`salary`, `board`, etc.).
Data scaling/standardisation	Yes	Used `StandardScaler` on input features before modeling.
Hyperparameter tuning	No	All models used default or manually specified parameters without `GridSearchCV` or tuning framework.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	`value_counts()` and countplot of `status`.
Sampling type	Random	`train_test_split(X, y, test_size=0.2, random_state=42, shuffle=True)` with no `stratify`.
Outliers removal	No	No IQR/quantile/rule-based filtering performed.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()`.
Imputation of missing values	Ignore	Salary missing values acknowledged but ignored

Drop columns	Yes	`sl_no` dropped at load. Drops of `degree_t`/`hsc_s` excluded here since they were replaced by dummies.
Encoding	Mixture	`LabelEncoder` for `gender`, `workex`, `specialisation`, `status`; `pd.get_dummies()` for `hsc_s`, `degree_t`.
Create new columns	No	One-hot dummies are derived from existing columns; not counted as “new”.
Feature selection	Yes	Correlation-guided post-visualisation drops: `['Others','Arts','degree_t','hsc_s','ssc_b','hsc_b','salary']`. Dropping after heatmap counts as feature selection.
Data scaling/standardisation	Yes	`StandardScaler()` used (`X = sc.fit_transform(data)`).
Hyperparameter tuning	No	No systematic search (no Grid/Random/Optuna); `LogisticRegression()` used with defaults.

Sixth recipe
1st Prompt
Accuracy 9/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No checks for target distribution
Sampling type	Yes	`train_test_split(..., shuffle=True)` → random sampling
Outliers removal	No	No IQR/Z-score/manual removal
Check for duplicates	No	Not addressed in the notebook
Imputation of missing values	No	Salary column acknowledged as missing but not imputed
Drop columns	Yes	Categorical and identifier columns dropped
Encoding	One-hot encoding only	Done sequentially with `pd.get_dummies()`
Create new columns	No	No feature engineering or transformations

Feature selection	Yes	Dropped `Mkt&HR` and `Degree other` based on correlation
Data scaling/standardisation	No	Not applied
Hyperparameter tuning	No	Default parameters used for logistic regression

2nd Prompt
Accuracy 9/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Implied from use of `y.value_counts()`
Sampling type	Random	Used `train_test_split()`
Outliers removal	No	No filtering done
Check for duplicates	No	No use of `.duplicated()` or `.drop_duplicates()`
Imputation of missing values	Drop rows	Used `dropna()`
Drop columns	Yes	Dropped categorical and non-predictor columns
Encoding	One-hot encoding	`pd.get_dummies()`
Create new columns	Yes	Via dummy variables
Feature selection	Yes	Used correlation heatmap
Data scaling or standardisation	No	No scaler used
Hyperparameter tuning	No	Used LogisticRegression with default parameters

3rd Prompt
Accuracy 7/11

#	Question	Answer	Snippet / Explanation

1	Is data checked for balance?	No	No countplot or `value_counts` for class distribution
2	Sampling type?	Random	Used `train_test_split(..., shuffle=True)`
3	Have outliers been removed?	No	No IQR, z-score, or condition applied
4	Does the code check for duplicates?	No	No `.duplicated()` check
5	Does the code impute missing values?	No	Used `dropna()` to remove null salary rows
6	Does the code drop columns?	Yes	Dropped: categorical fields, `sl_no`, `salary`, etc.
7	What encoding is used?	OHE	`pd.get_dummies(..., drop_first=True)`
8	Does the code create new columns?	Yes	OHE created new binary columns
9	Does the code perform feature selection?	Yes	Dropped columns after correlation matrix analysis
10	Does the code perform standardization?	No	No use of `StandardScaler()` or normalization
11	Does the code perform hyperparameter tuning?	No	LogisticRegression used with default parameters

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No class distribution check for `Placed` status
Sampling type	Random	`train_test_split(..., shuffle=True)` without stratify
Outliers removal	No	No IQR filtering or specific outlier handling
Check for duplicates	No	No use of `duplicated()` or `drop_duplicates()`

Imputation of missing values	drop the missing value rows	data1.dropna() used
Drop columns	Yes	Dropped categorical cols post encoding + `sl_no` and `salary`
Encoding	One hot encoding	`get_dummies` used for all categorical variables
Create new columns	No	All new columns are derived dummy encodings
Feature selection	Yes	Removed `Mkt&HR` and `Degree other` based on correlation heatmap
Data scaling/standardisation	No	No use of StandardScaler or other scaling methods
Hyperparameter tuning	No	Logistic regression used with default parameters

5th Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No class distribution check (e.g., `value_counts` on `status`) performed.
Sampling type	Random	Used `train_test_split` without stratification.
Outliers removal	No	No explicit outlier filtering via IQR, z-scores, etc.
Check for duplicates	No	No duplicate check or removal observed.
Imputation of missing values	Drop the missing value rows	`data1.dropna(inplace=True)`
Drop columns	No	Columns removed post-correlation analysis, so counts as feature selection.
Encoding	dummy	`pd.get_dummies(..., drop_first=True)` used for all categorical columns.
Create new columns	No	Only dummy variables created from existing columns.

Feature selection	Yes	Dropped low-correlation features (`Mkt&HR`, `Degree other`) based on correlation heatmap.
Data scaling/standardisation	No	No scaler applied (e.g., StandardScaler not used).
Hyperparameter tuning	No	LogisticRegression used without tuning or cross-validation.

Ground Truth

Data Wrangling Step	**Technique Used**	**Details**
Check for balanced data	No	No explicit class-distribution check for `status` (no `value_counts`/countplot).
Sampling type	Random	`train_test_split(x, y, test_size=0.3, random_state=101, shuffle=True)` with no `stratify`.
Outliers removal	No	No IQR/quantile/rule-based filtering present.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()`.
Imputation of missing values	Drop the missing value rows	`data1.dropna(inplace=True)` applied (twice) prior to modelling.
Drop columns	Yes	Dropped `sl_no` and `salary` without replacement; categorical sources later dropped were replaced by dummies (not counted here).
Encoding	One hot encoding	Multiple `pd.get_dummies(..., drop_first=True)` on categorical features (incl. `status` → `Placed`).
Create new columns	No	One-hot dummies are derived from existing columns; not “new”.
Feature selection	Yes	After correlation heatmap, dropped `['Mkt&HR','Degree other']` (drop-after-visualisation ≡ feature selection).

Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) used.
Hyperparameter tuning	No	`LogisticRegression()` used with defaults; no Grid/Random/Optuna search.

Seventh Recipe
1st Prompt
Accuracy 8/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No target distribution or imbalance check
Sampling type	Yes	`train_test_split` with `random_state=1`
Outliers removal	No	No filtering, clipping, or outlier tests
Check for duplicates	No	No use of `.duplicated()`
Imputation of missing values	No	`salary` dropped due to nulls; no imputation elsewhere
Drop columns	Yes	Dropped `sl_no`, `salary`, and some uncorrelated features
Encoding	Mixed encoding	`get_dummies` for features, `LabelEncoder` for target
Create new columns	No	No feature engineering
Feature selection	Yes	Dropped `hsc_b`, `ssc_b`, `hsc_s` based on heatmap
Data scaling/standardisation	Yes	Used `StandardScaler` on X_train and X_test
Hyperparameter tuning	No	Default logistic regression settings

2nd Prompt
Accuracy 8/11

Data Wrangling Step	Technique Used	Details

Check for balanced data	No	No explicit balance check using `y.value_counts()`
Sampling type	Random	`train_test_split()` with `random_state`
Outliers removal	No	No filtering or outlier logic
Check for duplicates	No	Not addressed
Imputation of missing values	Drop column	Dropped column `salary` with missing values
Drop columns	Yes	Dropped `salary`, `sl_no`, and other low-impact vars
Encoding	Mixture of encoding	Used both one-hot encoding and label encoding
Create new columns	Yes	Dummy variables added for categorical fields
Feature selection	Yes	Dropped low-value features based on correlation
Data scaling or standardisation	Yes	Used `StandardScaler()`
Hyperparameter tuning	No	No grid/random search or tuning used

3rd Prompt
Accuracy 8/11

#	Question	Answer	Code / Justification
1	Is data checked for balance?	Yes	`sns.countplot(..., hue='status')` for gender, degree, etc.
2	Sampling type?	Random	`train_test_split(..., random_state=1)`
3	Have outliers been removed?	No	No use of IQR, z-score, or filtering
4	Does the code check for duplicates?	No	No `.duplicated()` or related check
5	Does the code impute missing values?	No	`salary` dropped due to nulls

6	Does the code drop columns?	Yes	Dropped `salary`, `sl_no`, `hsc_b`, `ssc_b`, etc.
7	What encoding is used?	OHE, LE	`pd.get_dummies(...)`, `LabelEncoder()` for `status`
8	Does the code create new columns?	Yes	Created dummy columns d1 to d4
9	Does the code perform feature selection?	Yes	Dropped columns after `sns.heatmap(df.corr())`
10	Does the code perform standardization?	Yes	`StandardScaler()` used
11	Does the code perform hyperparameter tuning?	No	Logistic Regression used without tuning

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Countplots of `status` vs gender, workex, degree type, etc.
Sampling type	Random	`train_test_split(..., random_state=1)` without stratify
Outliers removal	No	No IQR filtering or outlier-specific handling
Check for duplicates	No	No `duplicated()` or `drop_duplicates()` used
Imputation of missing values	Drop the missing value columns	`salary` dropped instead of imputed
Drop columns	Yes	`salary, sl_no` are dropped without replacement
Encoding	mixture of encoding	`get_dummies` + LabelEncoder for `status`
Create new columns	No	Only derived dummies from existing columns

Feature selection	Yes	Dropped three columns before modeling based on correlation/EDA
Data scaling/standardisation	Yes	Applied `StandardScaler` to feature matrix
Hyperparameter tuning	No	Logistic regression used with default settings

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	`sns.countplot` on `'status'` indicates checking target distribution.
Sampling type	Random	Used `train_test_split` without stratification.
Outliers removal	No	No IQR, z-score, or other filtering for outliers.
Check for duplicates	No	No duplicate checks applied.
Imputation of missing values	Drop the missing value column	`salary` column was dropped entirely due to NaNs.
Drop columns	Yes	`salary` and `sl_no` are dropped at the beginning.
Encoding	mixture of encoding	Used `pd.get_dummies(..., drop_first=True)` for features and `LabelEncoder` for target (`status`).
Create new columns	No	Only derived dummy variables used.
Feature selection	Yes	Dropped low-impact features (`hsc_b`, `ssc_b`, `hsc_s`) based on domain understanding.
Data scaling/standardisation	Yes	Used `StandardScaler` before logistic regression.
Hyperparameter tuning	No	Logistic regression used with default parameters; no tuning attempted.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Multiple plots with `hue='status'` (pairplot; countplots for `workex`, `gender`, `specialisation`, `degree_t`) allow visual inspection of class balance (no `value_counts()` on `status`).
Sampling type	Random	`train_test_split(X, y, test_size=0.2, random_state=1)` with no `stratify`.
Outliers removal	No	No IQR/quantile/rule-based filtering in code.
Check for duplicates	No	No `.duplicated()` / `.drop_duplicates()` used.
Imputation of missing values	Drop the missing value columns	Dropped `salary` (many nulls) instead of imputing.
Drop columns	Yes	Dropped `sl_no` before modelling; post-EDA drops are captured under Feature selection.
Encoding	Mixture	One-hot via `pd.get_dummies(..., drop_first=True)` for `gender`, `degree_t`, `specialisation`, `workex`; `LabelEncoder` for target `status`.
Create new columns	No	One-hot dummies are derived from existing columns, so not counted as “new”.
Feature selection	Yes	After correlation heatmap, dropped `hsc_b`, `ssc_b`, `hsc_s` (“don’t affect classification”); drop-after-visualisation counts as feature selection.
Data scaling/standardisation	Yes	`StandardScaler()` fit on train, applied to test.
Hyperparameter tuning	No	`LogisticRegression()` used without Grid/Random search.

Eighth Recipe
1st Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No check on target distribution
Sampling type	Yes	Used `sample.split()` from `caTools`
Outliers removal	No	No outlier detection or filtering
Check for duplicates	No	No mention of `duplicated()` or equivalent
Imputation of missing values	Yes (Mean)	Imputed `salary` with mean using `ifelse()` + `ave()`
Drop columns	No	All columns retained
Encoding	Label encoding	Used `factor(..., labels=...)` for all categorical vars
Create new columns	No	No engineered features
Feature selection	Yes (manual)	Only `ssc_p` and `hsc_p` used in regression
Data scaling/standardisation	No	Predictors not scaled
Hyperparameter tuning	No	Not applicable in linear regression used here

2nd Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	Not explicitly done
Sampling type	Random sampling	`sample.split()` from `caTools`
Outliers removal	No	Not addressed
Check for duplicates	No	Not addressed
Imputation of missing values	Mean imputation	Only on `salary`, using `ave()`
Drop columns	No	None dropped

Encoding	Label encoding	`factor()` with labels used
Create new columns	No	No dummy or engineered columns
Feature selection	Yes	Only `ssc_p` and `hsc_p` used to predict `mba_p`
Data scaling or standardisation	No	Not used
Hyperparameter tuning	No	Not used

3rd Prompt
Accuracy 10/11

#	Question	Answer	Code / Justification
1	Is data checked for balance?	No	No `table(status)` or similar used
2	Sampling type?	Random	`sample.split(..., SplitRatio=0.80)`
3	Have outliers been removed?	No	No IQR or filtering logic
4	Does the code check for duplicates?	No	No `duplicated()` check
5	Does the code impute missing values?	Yes	`ifelse(is.na(salary), ave(...))`
6	Does the code drop columns?	No	None dropped explicitly
7	What encoding is used?	Label	Used `factor(..., labels=...)`
8	Does the code create new columns?	No	No new columns introduced
9	Does the code perform feature selection?	Yes	Only `ssc_p` and `hsc_p` used in regression
10	Does the code perform standardization?	No	No `scale()` or normalization used
11	Does the code perform hyperparameter tuning?	No	Basic linear regression only

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No bar plots or proportion checks for `status`
Sampling type	Random	`sample.split(..., SplitRatio=0.80)` used on continuous variable
Outliers removal	No	No filtering or outlier logic present
Check for duplicates	No	No `duplicated()` or `distinct()` used
Imputation of missing values	use summary statistics	`salary` imputed with column-wise `mean()`
Drop columns	No	No columns were dropped
Encoding	Label Encoder	All categorical variables mapped to numeric factor levels
Create new columns	No	No new or derived features created
Feature selection	No	No feature selection performed
Data scaling/standardisation	No	No `scale()`, normalization, or standardisation applied
Hyperparameter tuning	No	No tuning applied to linear model

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	No value count or proportion check on `status`.
Sampling type	Random	Used `sample.split()` from `caTools` without stratification.
Outliers removal	No	No filtering or outlier handling performed.

Check for duplicates	No	No check for duplicated records using `duplicated()` or similar.
Imputation of missing values	use summary statistics	Missing values in `salary` column imputed using mean imputation.
Drop columns	No	No columns were removed.
Encoding	Label Encoder	All categorical features (e.g., gender, board, stream, etc.) converted using `factor(..., labels=...)`.
Create new columns	No	No derived or engineered columns created.
Feature selection	No	All features retained; linear model only used subset for regression, but no columns were dropped.
Data scaling/standardisation	No	No use of normalization or scaling methods.
Hyperparameter tuning	No	No model tuning or cross-validation applied.

Ground Truth

Data Wrangling Step	**Technique Used**	**Details**
Check for balanced data	No	No value counts or class-distribution plots for `status`.
Sampling type	Random	Random split via `caTools::sample.split(..., SplitRatio = 0.80)`; train/test subsets created (both with `split = TRUE` in code).
Outliers removal	No	No IQR/quantile/rule-based filtering present.
Check for duplicates	No	No use of `duplicated()` / `dplyr::distinct()`.
Imputation of missing values	Use summary statistics (mean)	`salary` NAs replaced with overall mean via `ifelse(is.na(salary), ave(salary, FUN = mean, na.rm = TRUE), salary)`.
Drop columns	No	No columns removed from analysis.

Encoding	Label Encoder	Multiple variables recoded to numeric factor levels (`gender`, `ssc_b`, `hsc_s`, `hsc_b`, `degree_t`, `workex`, `specialisation`, `status`).
Create new columns	No	Only recoding/labeling of existing columns; no truly new features created.
Feature selection	No	No correlation/model-based drops or post-EDA removals.
Data scaling/standardisation	No	No scaling step applied.
Hyperparameter tuning	No	Simple linear model `lm(mba_p ~ ssc_p + hsc_p)`; no systematic search.

Ninth Recipe
1st Prompt
Accuracy 9/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes (Manual Review)	Crosstab of `status` with categorical features
Sampling type	Yes	Used `train_test_split()`
Outliers removal	No	No IQR, z-score, or filtering
Check for duplicates	No	No check like `.duplicated()` or `.drop_duplicates()`
Imputation of missing values	No (Avoided)	`salary` was simply dropped
Drop columns	Yes	Dropped `sl_no`, `ssc_b`, `hsc_b`, `salary`
Encoding	Yes (Manual dummies)	Manually used `get_dummies()` for each categorical feature
Create new columns	No	No derived features created
Feature selection	Yes (SelectKBest + manual)	Used `SelectKBest(chi2)` and manual selection of top 6 features

Data scaling/standardisation	Yes	Used `StandardScaler()` on numerical features
Hyperparameter tuning	Yes (basic sweep)	Varied k in `SelectKBest`, assessed metrics

2nd Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	No	Not explicitly done
Sampling type	Random	`train_test_split()`
Outliers removal	No	No method used
Check for duplicates	No	Not checked
Imputation of missing values	No	Dropped salary instead of imputing
Drop columns	Yes	Dropped `sl_no`, `ssc_b`, `hsc_b`, `salary`
Encoding	Dummy	`pd.get_dummies()`
Create new columns	No	All columns derived from existing ones
Feature selection	Yes	Used `SelectKBest(chi2)` with parameter sweep
Data scaling or standardisation	Yes	`StandardScaler()`
Hyperparameter tuning	No	No tuning methods used

3rd Prompt
Accuracy 8/11

#	Question	Answer	Code / Justification
1	Is data checked for balance?	Yes	`pd.crosstab(..., dataset['status'])`
2	What sampling type is used?	Random	`train_test_split(..., random_state=0)`
3	Are outliers removed?	No	No filtering or IQR checks

4	Does the code check for duplicates?	No	No `duplicated()` or similar used
5	Does the code impute missing values?	No	Dropped `salary` column
6	Does the code drop columns?	Yes	`drop('sl_no'...)`, `drop('hsc_b'...)`
7	What encoding is used?	Dummy	`pd.get_dummies(...)` across categorical vars
8	Does the code create new columns?	Yes	Dummy variables like `UG_Comm`, `HS_Sci`
9	Does the code perform feature selection?	Yes	`SelectKBest(chi2, k=...)`
10	Does the code perform standardization?	Yes	`StandardScaler()` used
11	Does the code perform hyperparameter tuning?	Yes	Parameter sweep over k in feature selection

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Countplots by `status`, gender, workex, degree type, etc.
Sampling type	Random	`train_test_split(..., random_state=0)` without stratify
Outliers removal	No	No outlier-specific logic or filtering seen
Check for duplicates	No	No use of `duplicated()` or `drop_duplicates()`
Imputation of missing values	none	Missing values in `salary` acknowledged but dropped
Drop columns	Yes	Dropped `sl_no`, `salary`
Encoding	One hot encoding	Used `get_dummies` for all categorical vars including `status`

Create new columns	No	All new columns were dummy encodings derived from original columns
Feature selection	Yes	Used `SelectKBest(chi2)` to select optimal 6 features via sweeps
Data scaling/standardisation	Yes	`StandardScaler` used before modeling
Hyperparameter tuning	No	All models used fixed/default settings, no grid/random search

5th Prompt
Accuracy 10/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Performed multiple `crosstab` checks and countplots on `status` against various features.
Sampling type	Random	Used `train_test_split` without `stratify`.
Outliers removal	No	No explicit outlier filtering applied.
Check for duplicates	No	No `.duplicated()` or `drop_duplicates()` used.
Imputation of missing values	none	`salary` column was dropped due to missing values; no imputation done.
Drop columns	No	Columns like `ssc_b`, `hsc_b`, and `sl_no` dropped after EDA or correlation analysis → counts as feature selection.
Encoding	Dummy	Used `pd.get_dummies()` for all categorical features, including custom binary coding (e.g., `HS_Comm`, `UG_Sci`).
Create new columns	No	Only dummy/binary encodings used, no standalone new features.
Feature selection	Yes	Used `SelectKBest(chi2)` for univariate selection; identified optimal k=6.
Data scaling/standardisation	Yes	Used `StandardScaler` before training.

Hyperparameter tuning	No	Default parameters used in all models; no grid/random search. Cross-validation done only for final logistic regression.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Crosstabs by `status` and multiple plots with `x='status'`/`hue='status'` (violin/countplots) to view class distribution.
Sampling type	Random	`train_test_split(X, y, test_size=0.3, random_state=0)` without `stratify`.
Outliers removal	No	No IQR/quantile/rule-based filtering applied.
Check for duplicates	No	No use of `.duplicated()` / `.drop_duplicates()`.
Imputation of missing values	Ignore	Missing `salary` noted; excluded from modelling (no imputation performed).
Drop columns	Yes	`sl_no` dropped without replacement. (Drops of `ssc_b`/`hsc_b` counted under Feature selection due to post-EDA reasoning.)
Encoding	One hot encoding	Manual `pd.get_dummies` for `gender`, `hsc_s`→(`HS_Sci`,`HS_Comm`), `degree_t`→(`UG_Sci`,`UG_Comm`), `workex`, `specialisation`; `status`→binary dummy.
Create new columns	No	One-hot dummies are derived from existing columns, so not “new”.
Feature selection	Yes	Post-EDA manual drops (`ssc_b`, `hsc_b`); univariate chi^2 `SelectKBest` with k-sweep; final model uses k=6 features.
Data scaling/standardisation	Yes	`StandardScaler()` applied to train and test (note: test scaled with a separate `fit_transform`, but scaling is present).

Hyperparameter tuning	No	No systematic search (`GridSearchCV`/`RandomizedSearchCV`); model hyperparameters not tuned.

Tenth Recipe
1st Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Used `data.status.value_counts()`
Sampling type	Yes (Manual Train-Test Split)	Used `data[:175]` for train and `data[175:]` for test
Outliers removal	No	No IQR, z-score, or filtering
Check for duplicates	No	No `.duplicated()` check
Imputation of missing values	Yes (Domain-based)	Filled `salary` with 0 after verifying all had status "Not Placed"
Drop columns	Yes	Dropped `sl_no`
Encoding	Yes (Manual Label Encoding)	Manually converted all categorical variables to numeric values
Create new columns	No	No feature engineering
Feature selection	Yes (Manual subset)	Selected 7 features based on barplots and correlation matrix
Data scaling/standardisation	No	Did not scale numerical values
Hyperparameter tuning	No	Logistic Regression with fixed `max_iter=150`, no grid/random search

2nd Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Used `value_counts()` on `status`
Sampling type	Random (manual)	`train_data = data[:175]`

Outliers removal	No	No filtering applied
Check for duplicates	No	Not checked
Imputation of missing values	Use summary statistics	Filled missing salary with 0 based on domain logic
Drop columns	Yes	Dropped `sl_no`
Encoding	Mixture of encoding	Used label encoding via `loc[...]` + `pd.get_dummies()`
Create new columns	No	All derived from existing columns
Feature selection	Yes	Visual-based manual exclusion of weak features
Data scaling or standardisation	No	No use of scaler like `StandardScaler()`
Hyperparameter tuning	No	Logistic Regression used with fixed parameters

3rd Prompt
Accuracy 11/11

#	Question	Answer	Code Snippet / Reasoning
1	Is data checked for balance?	Yes	`data.status.value_counts()` + markdown comment
2	What sampling type is used?	Manual	`train_data = data[:175]`
3	Are outliers removed?	No	No IQR/z-score/filtering seen
4	Does the code check for duplicates?	No	No `.duplicated()` or equivalent check
5	Does the code impute missing values?	Yes	`data['salary'].fillna(value=0)`
6	Does the code drop columns?	Yes	`drop(['sl_no'])`
7	What encoding is used?	Label, Dummy	Manual binary encoding + `pd.get_dummies`
8	Does the code create new columns?	No	Only replaced or transformed columns

9	Does the code perform feature selection?	Yes	Manual: `"features = [...]"` based on correlation/plots
10	Does the code perform standardization?	No	No scaler used
11	Does the code perform hyperparameter tuning?	No	LogisticRegression with fixed `penalty` and `max_iter`, no tuning

4th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	`data.status.value_counts()` used; imbalance noted in comments
Sampling type	None	No `train_test_split`; manual slicing (`[:175]`, `[175:]`) used
Outliers removal	No	No IQR filtering or explicit outlier logic
Check for duplicates	No	No use of `duplicated()` or `drop_duplicates()`
Imputation of missing values	replace with text	Filled all missing `salary` with 0 after confirming `status == 0`
Drop columns	Yes	Dropped `sl_no`
Encoding	Label Encoder	Manual label mapping for all categorical columns
Create new columns	No	All columns directly converted or selected; no new derived features
Feature selection	Yes	Removed columns after correlation and barplot analysis
Data scaling/standardisation	No	No use of `StandardScaler` or similar scaler
Hyperparameter tuning	No	Logistic Regression used with fixed parameters (`penalty='none'`)

5th Prompt
Accuracy 11/11

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	Class imbalance in `status` noted explicitly; label counts printed and implications discussed.
Sampling type	None	Manual slicing (`train_data = data[:175]`) used instead of `train_test_split`; no random or stratified sampling.
Outliers removal	No	No outlier detection or removal (IQR/z-score) applied.
Check for duplicates	No	No duplicate check or removal seen.
Imputation of missing values	replace with text	Filled missing `salary` values with `0`, reasoning that status = "Not Placed".
Drop columns	Yes	Dropped columns like `sl_no`, but others like `gender`, `ssc_b`, `hsc_b`, etc. dropped *only after correlation/EDA* → counts as feature selection.
Encoding	Dummies	Used `.loc[...] = 0/1/2` assignments to manually encode all categorical columns.
Create new columns	No	No columns created except for encodings.
Feature selection	Yes	Selected features manually (`["workex", "ssc_p", "degree_p", "hsc_p", "specialisation", "etest_p", "mba_p"]`) based on correlation/EDA.
Data scaling/standardisation	No	No `StandardScaler` or other scaling used.
Hyperparameter tuning	No	`LogisticRegression(penalty='none')` used without systematic tuning.

Ground Truth

Data Wrangling Step	Technique Used	Details
Check for balanced data	Yes	`data.status.value_counts()` printed; note explicitly flags class imbalance.
Sampling type	Manual	Manual index-based split: `train_data = data[:175]`, `test_data = data[175:]` (not random/stratified/oversampling).
Outliers removal	No	No IQR/quantile/rule-based filtering applied.
Check for duplicates	No	No `.duplicated()` or `.drop_duplicates()` used.
Imputation of missing values	Replace with text/numeric placeholder	`data['salary'].fillna(0, inplace=True)` after verifying missing salaries correspond to `status = Not Placed`.
Drop columns	Yes	`sl_no` dropped and not reused.
Encoding	Dummies	Manual label-style recoding (e.g., `gender`, `workex`, `status`, etc.) plus `pd.get_dummies()` on selected `features` for modelling.
Create new columns	No	One-hot dummies are derived from existing variables;not counted as “new”.
Feature selection	Yes	Post-EDA exclusion: after heatmap/plots, `['gender','ssc_b','hsc_b','hsc_s','degree_t']` not selected in final `features` list. (Drop-after-visualisation ⇒ feature selection.)
Data scaling/standardisation	No	No scaler (`StandardScaler`, `MinMaxScaler`, etc.) applied.
Hyperparameter tuning	No	Logistic regression fit with fixed params (`penalty='none'`, `max_iter=150`); no systematic search.