**Contents**

**1 Project Topic 3**

1.1 Task . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2 Goals . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

**2 Data 3**

2.1 Data Source . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

2.2 Data Description . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

2.3 Data Attributes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

2.4 Data Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.4.1 Missing Values and Loading the Dataset . . . . . . . . . . . . . . . . . . . . . 4

2.4.2 Removing ‘fnlwgt’ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.4.3 Data Summary Statistics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

**3 Data Cleaning 7**

3.1 Data Type Change . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

3.2 Whitespace Removal . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

3.3 Missing Values . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

3.4 Check of Imbalanced Dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

3.4.1 Cleaning target variable . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

3.5 Conclusion of Data Cleaning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

**4 Exploratory Data Analysis 11**

4.1 Univariate Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

4.1.1 Histogram of age . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

4.1.2 Histogram of hours worked per week . . . . . . . . . . . . . . . . . . . . . . . 13

4.1.3 Histogram of capital gain . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

4.1.4 Histogram of capital loss . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

4.1.5 Histogram of education . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

4.1.6 Histogram of workclass . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

4.1.7 Histogram of education number . . . . . . . . . . . . . . . . . . . . . . . . . . 17

4.1.8 Histogram of marital status . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

4.1.9 Histogram of occupation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

4.1.10 Histogram of relationship . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20

4.1.11 Histogram of race . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

4.1.12 Histogram of sex . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22

4.1.13 Histogram of native country . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

4.2 Bivariate Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

4.2.1 Relationship between age and annual income . . . . . . . . . . . . . . . . . . 24

4.2.2 Relationship between capital gain and annual income . . . . . . . . . . . . . 26

4.2.3 Relationship between capital loss and annual income . . . . . . . . . . . . . . 28

4.2.4 Feature Engineering . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

4.2.5 Relationship between delta capital and annual income . . . . . . . . . . . . . 32

4.2.6 Relationship between hours per week and annual income . . . . . . . . . . . . 35

4.2.7 Relationship between workclass and annual income . . . . . . . . . . . . . . . 37

4.2.8 Relationship between education and annual income . . . . . . . . . . . . . . . 39

4.2.9 Relationship between education num and annual income . . . . . . . . . . . . 42

4.2.10 Relationship between marital status and annual income . . . . . . . . . . . . 45

4.2.11 Relationship between occupation and annual income . . . . . . . . . . . . . . 47

4.2.12 Relationship between relationship and annual income . . . . . . . . . . . . . 50

4.2.13 Relationship between race and annual income . . . . . . . . . . . . . . . . . . 52

4.2.14 Relationship between sex and annual income . . . . . . . . . . . . . . . . . . 55

4.2.15 Relationship between native country and annual income . . . . . . . . . . . . 57

4.3 Multivariate Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59

4.3.1 Correlation Matrix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59

4.3.2 Multivariate Categorical Analysis . . . . . . . . . . . . . . . . . . . . . . . . . 60

4.4 Conclusion of Exploratory Data Analysis . . . . . . . . . . . . . . . . . . . . . . . . 61

**5 Models 61**

5.1 Marital status and Relationship: Multicollinearity? . . . . . . . . . . . . . . . . . . . 62

5.2 Encoding categorical variables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63

5.3 Standardizing Feature Set . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67

5.4 Multicollinearity Test . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 68

5.5 SMOTE: Synthetic Minority Over-sampling Technique . . . . . . . . . . . . . . . . . 71

5.6 Train-Test split . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73

5.7 Evaluation Metric . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74

5.7.1 Logistic Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74

5.7.2 Logistic Regression using SMOTE . . . . . . . . . . . . . . . . . . . . . . . . 76

5.7.3 Random Forest . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 78

5.7.4 Random Forest using SMOTE . . . . . . . . . . . . . . . . . . . . . . . . . . 79

5.7.5 Gaussian Naive Bayes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 81

5.7.6 Gaussian Naive Bayes using SMOTE . . . . . . . . . . . . . . . . . . . . . . . 83

**6 Results and Analysis 85**

6.1 Logistic Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

6.1.1 Without SMOTE . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

6.1.2 With SMOTE . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

6.1.3 AUC . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

6.2 Random Forest . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

6.2.1 Without SMOTE . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

6.2.2 With SMOTE . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

6.2.3 AUC . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

6.3 Gaussian Naive Bayes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87

6.3.1 Without SMOTE . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87

6.3.2 With SMOTE . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87

6.3.3 AUC . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87

**7 Discussion and Conclusion 87**

7.1 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87

7.2 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87