Methodology

This study adopts a quantitative analytical approach to examine the association between diabetes and stroke using regression-based methods suitable for a binary outcome. Logistic regression is employed as the baseline analytical technique because stroke status is dichotomous in nature. The initial analysis begins with an unadjusted logistic regression model to assess the crude association between diabetes and stroke. This is followed by adjusted logistic regression models that incorporate additional covariates, allowing the independent effect of diabetes on stroke risk to be evaluated while controlling for potential confounding factors. The results from logistic regression are interpreted using odds ratios, which provide a clear understanding of the direction and strength of associations.

Before advancing to more complex models, a correlation matrix of all explanatory variables is computed to examine pairwise relationships and to identify the presence of strong correlations among predictors. To further diagnose multicollinearity, the Variance Inflation Factor (VIF) is calculated for each variable included in the regression models. These diagnostic steps ensure that the estimates obtained from logistic regression are not unduly influenced by multicollinearity and that the model assumptions are reasonably satisfied.

The predictive performance of the baseline logistic regression model is then evaluated using the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). The ROC curve illustrates the trade-off between sensitivity and specificity across different probability thresholds, while the AUC provides a summary measure of the model's ability to discriminate between individuals with and without stroke.

To improve model stability, address multicollinearity, and handle potential overfitting, regularized regression techniques are subsequently applied in a hierarchical manner. LASSO regression is first used to identify the most important predictors by shrinking the coefficients of less relevant variables to zero, resulting in a parsimonious model. Ridge regression is then employed to stabilize coefficient estimates by shrinking their magnitudes when predictors are highly correlated, without removing any variables from the model. Finally, elastic net regression, which combines the features of both LASSO and ridge regression, is applied to achieve a balance between variable selection and coefficient shrinkage.

The results obtained from logistic regression, LASSO, ridge, and elastic net models are compared (by AIC & BIC) to identify consistent and robust predictors of stroke associated with diabetes. This stepwise and integrated methodological framework ensures that the findings are statistically reliable, interpretable, and robust, providing a comprehensive understanding of stroke risk in the presence of diabetes.

Results

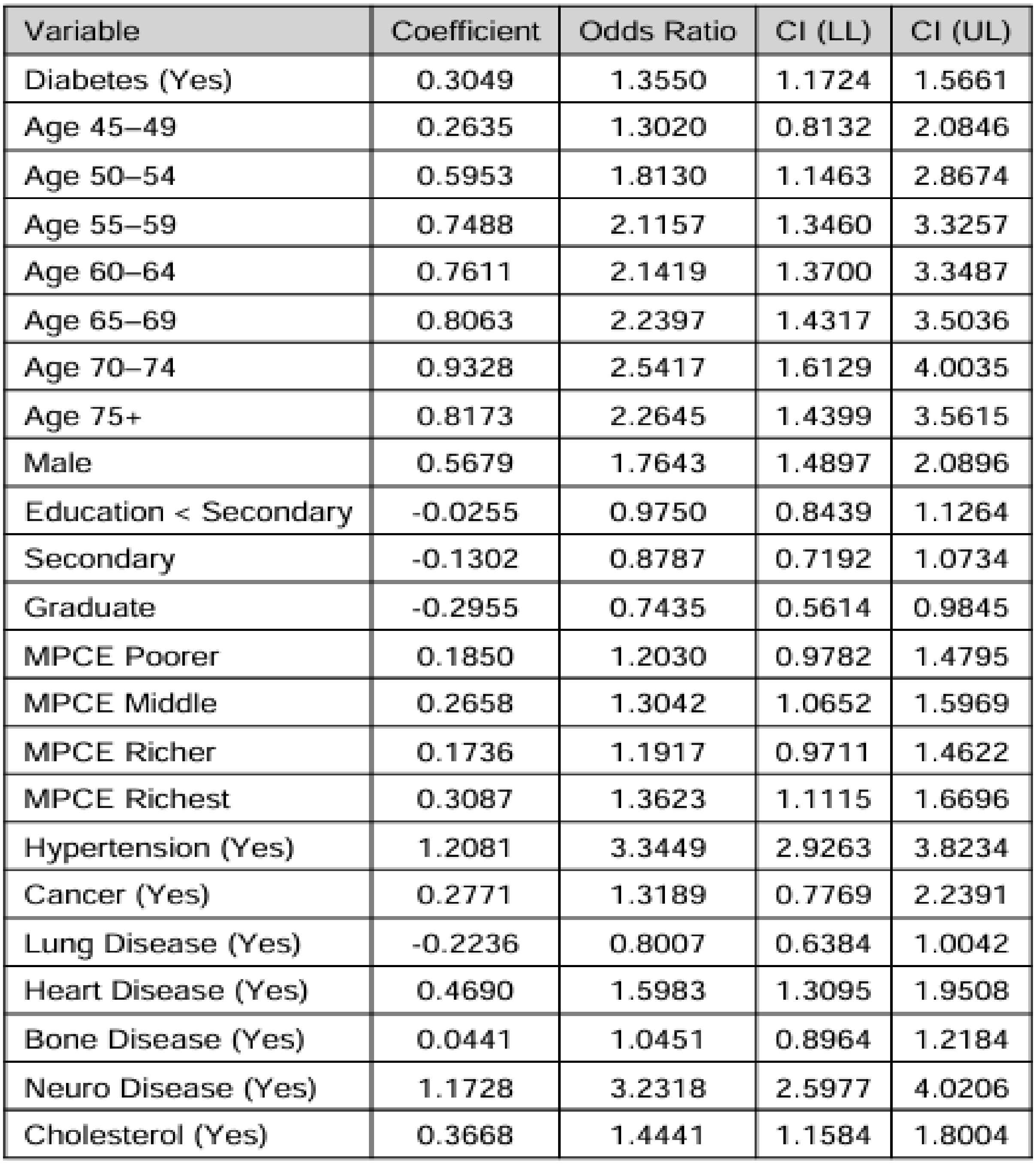
1. Unadjusted Logistic Regression:-

Stroke	**Coefficient**	**Odds Ratio**	**CI(LL)**	**CI(UL)**	**AIC**	**BIC**
Diabetes	0.1901294	2.94606	2.596018	3.34330	12193.66	12212.07

Interpretation :-

From the above I have observed the person having diabetes have 2.94 times higher chances of stroke than that of a person not having the diabetes .Also the confidence interval does not include 1 so the results are highly significant ..

2. Adjusted Logistic Regression:



Variable	Coefficient	Odds Ratio	CI (LL)	CI (UL)
Diabetes (Yes)	0.3049	1.3550	1.1724	1.5661
Age 45–49	0.2635	1.3020	0.8132	2.0846
Age 50–54	0.5953	1.8130	1.1463	2.8674
Age 55–59	0.7488	2.1157	1.3460	3.3257
Age 60–64	0.7611	2.1419	1.3700	3.3487
Age 65–69	0.8063	2.2397	1.4317	3.5036
Age 70–74	0.9328	2.5417	1.6129	4.0035
Age 75+	0.8173	2.2645	1.4399	3.5615
Male	0.5679	1.7643	1.4897	2.0896
Education < Secondary	-0.0255	0.9750	0.8439	1.1264
Secondary	-0.1302	0.8787	0.7192	1.0734
Graduate	-0.2955	0.7435	0.5614	0.9845
MPCE Poorer	0.1850	1.2030	0.9782	1.4795
MPCE Middle	0.2658	1.3042	1.0652	1.5969
MPCE Richer	0.1736	1.1917	0.9711	1.4622
MPCE Richest	0.3087	1.3623	1.1115	1.6696
Hypertension (Yes)	1.2081	3.3449	2.9263	3.8234
Cancer (Yes)	0.2771	1.3189	0.7769	2.2391
Lung Disease (Yes)	-0.2236	0.8007	0.6384	1.0042
Heart Disease (Yes)	0.4690	1.5983	1.3095	1.9508
Bone Disease (Yes)	0.0441	1.0451	0.8964	1.2184
Neuro Disease (Yes)	1.1728	3.2318	2.5977	4.0206
Cholesterol (Yes)	0.3668	1.4441	1.1584	1.8004

Interpretation :-

The logistic regression results show that diabetes is significantly associated with stroke, with individuals having diabetes experiencing 35.5% higher odds of stroke compared to non-diabetic individuals (OR = 1.36; 95% CI: 1.17–1.56). Age exhibits a strong and increasing relationship with stroke risk. Compared to individuals below 45 years, those aged 50–54 years had significantly higher odds of stroke (OR = 1.81; 95% CI: 1.15–2.87), and the odds increased further for age groups 55–59 (OR = 2.12; 95% CI: 1.35–3.33), 60–64 (OR = 2.14; 95% CI: 1.37–3.35), 65–69 (OR = 2.24; 95% CI: 1.43–3.50), 70–74 (OR = 2.54; 95% CI: 1.61–4.00), and 75 years and above (OR = 2.26; 95% CI: 1.44–3.56), indicating a clear age gradient in stroke occurrence. Sex differences were also evident, with males having significantly higher odds of stroke than females (OR = 1.76; 95% CI: 1.49–2.09). Educational attainment showed a protective effect, as individuals with graduate-level education had significantly lower odds of stroke compared to those with lower education (OR = 0.74; 95% CI: 0.56–0.98). Household economic status displayed a mixed association, where individuals in the middle MPCE quintile (OR = 1.30; 95% CI: 1.07–1.60) and richest quintile (OR = 1.36; 95% CI: 1.11–1.67) showed higher odds of stroke compared to the poorest group. Caste-based differences were modest, with Scheduled Caste individuals having significantly higher odds of stroke compared to the reference group (OR = 1.21; 95% CI: 1.01–1.46), while no significant differences were observed for OBC or Scheduled Tribes. Religion was not significantly associated with stroke, as the confidence intervals for Muslims (OR = 1.19; 95% CI: 0.99–1.44), Christians (OR = 0.83; 95% CI: 0.65–1.06), and other religions (OR = 1.00; 95% CI: 0.77–1.28) included unity. Work status showed strong associations, with individuals not currently working having significantly higher odds of stroke (OR = 1.54; 95% CI: 1.30–1.83), whereas those currently working had significantly lower odds (OR = 0.66; 95% CI: 0.53–0.82). Physical activity demonstrated a strong protective effect, as physically active (OR = 0.51; 95% CI: 0.45–0.59), moderately active (OR = 0.58; 95% CI: 0.46–0.73), and vigorously active individuals (OR = 0.41; 95% CI: 0.33–0.52) had substantially lower odds of stroke compared to physically inactive individuals. Alcohol consumption was significantly associated with stroke, with drinkers having 29.7% higher odds than non-drinkers (OR = 1.30; 95% CI: 1.10–1.53). Tobacco use showed that past smokers (OR = 1.56; 95% CI: 1.23–1.99) and past smokeless tobacco users (OR = 1.47; 95% CI: 1.09–1.99) had significantly higher odds of stroke, while current tobacco use was not statistically significant. Among clinical risk factors, hypertension emerged as the strongest predictor, with hypertensive individuals having more than three times higher odds of stroke (OR = 3.34; 95% CI: 2.93–3.82). Heart disease was also significantly associated with stroke (OR = 1.60; 95% CI: 1.31–1.95), as was neurological disease, which showed more than threefold higher odds (OR = 3.23; 95% CI: 2.60–4.02). High cholesterol was another significant risk factor (OR = 1.44; 95% CI: 1.16–1.80). In contrast, cancer (OR = 1.32; 95% CI: 0.78–2.24), lung disease (OR = 0.81; 95% CI: 0.64–1.00), and bone disease (OR = 1.05; 95% CI: 0.90–1.22) did not show statistically significant associations with stroke, as their confidence intervals included unity.

3. Variance Inflated Factor :-

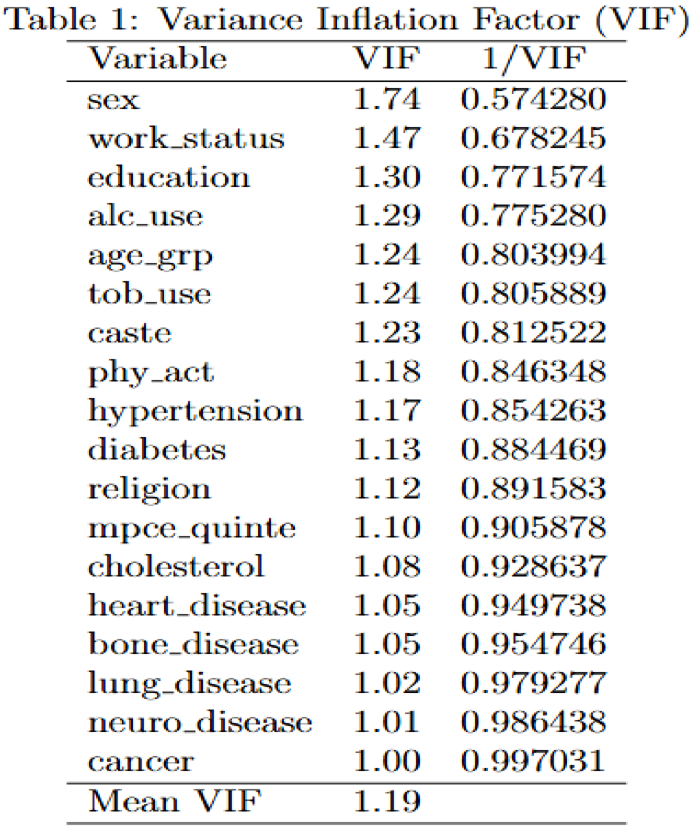
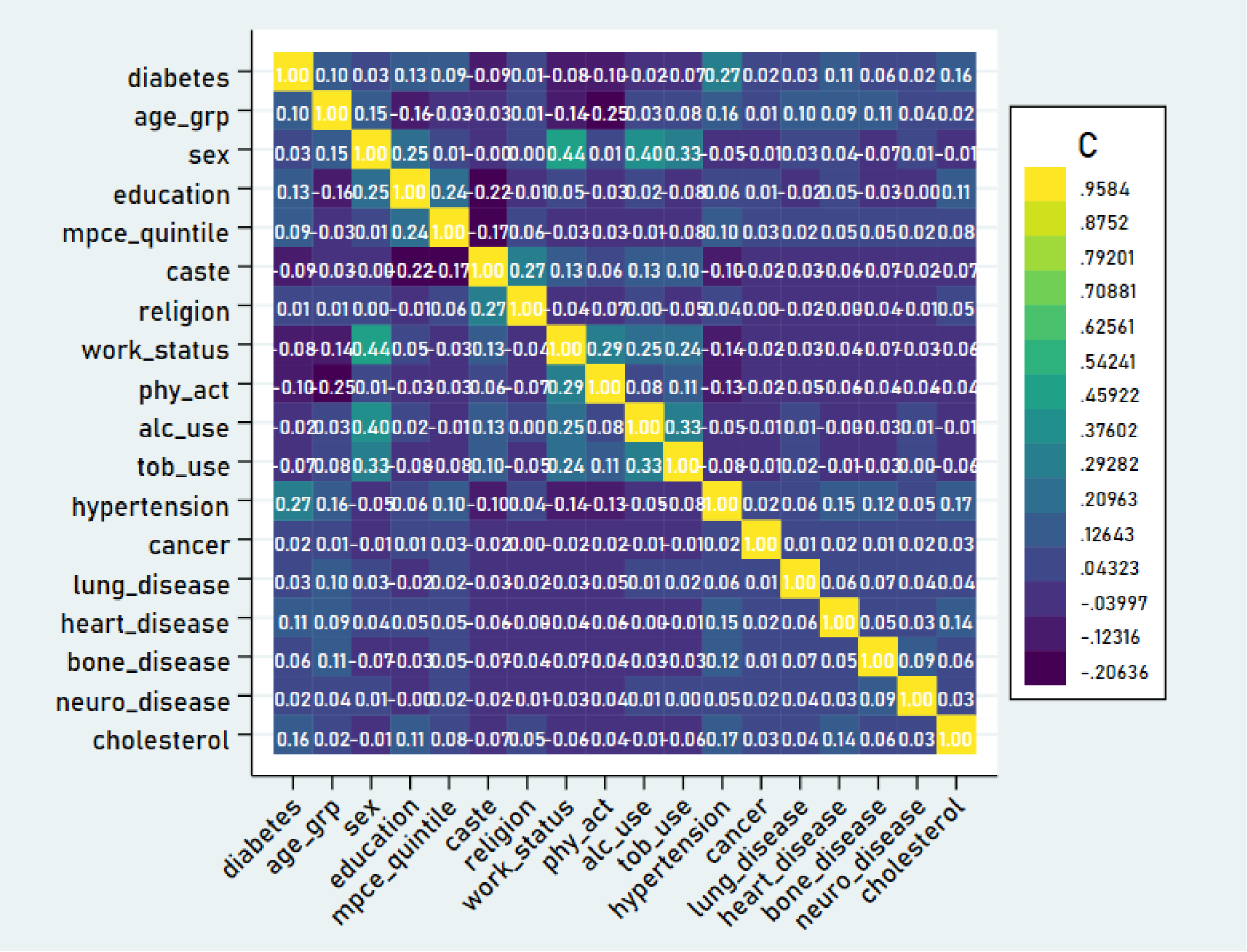


Table 1: Variance Inflation Factor (VIF)

Variable	VIF	1/VIF
sex	1.74	0.574280
work_status	1.47	0.678245
education	1.30	0.771574
alc_use	1.29	0.775280
age_grp	1.24	0.803994
tob_use	1.24	0.805889
caste	1.23	0.812522
phy_act	1.18	0.846348
hypertension	1.17	0.854263
diabetes	1.13	0.884469
religion	1.12	0.891583
mpce_quinte	1.10	0.905878
cholesterol	1.08	0.928637
heart_disease	1.05	0.949738
bone_disease	1.05	0.954746
lung_disease	1.02	0.979277
neuro_disease	1.01	0.986438
cancer	1.00	0.997031
Mean VIF	1.19	

4. Correlation Matrix:-



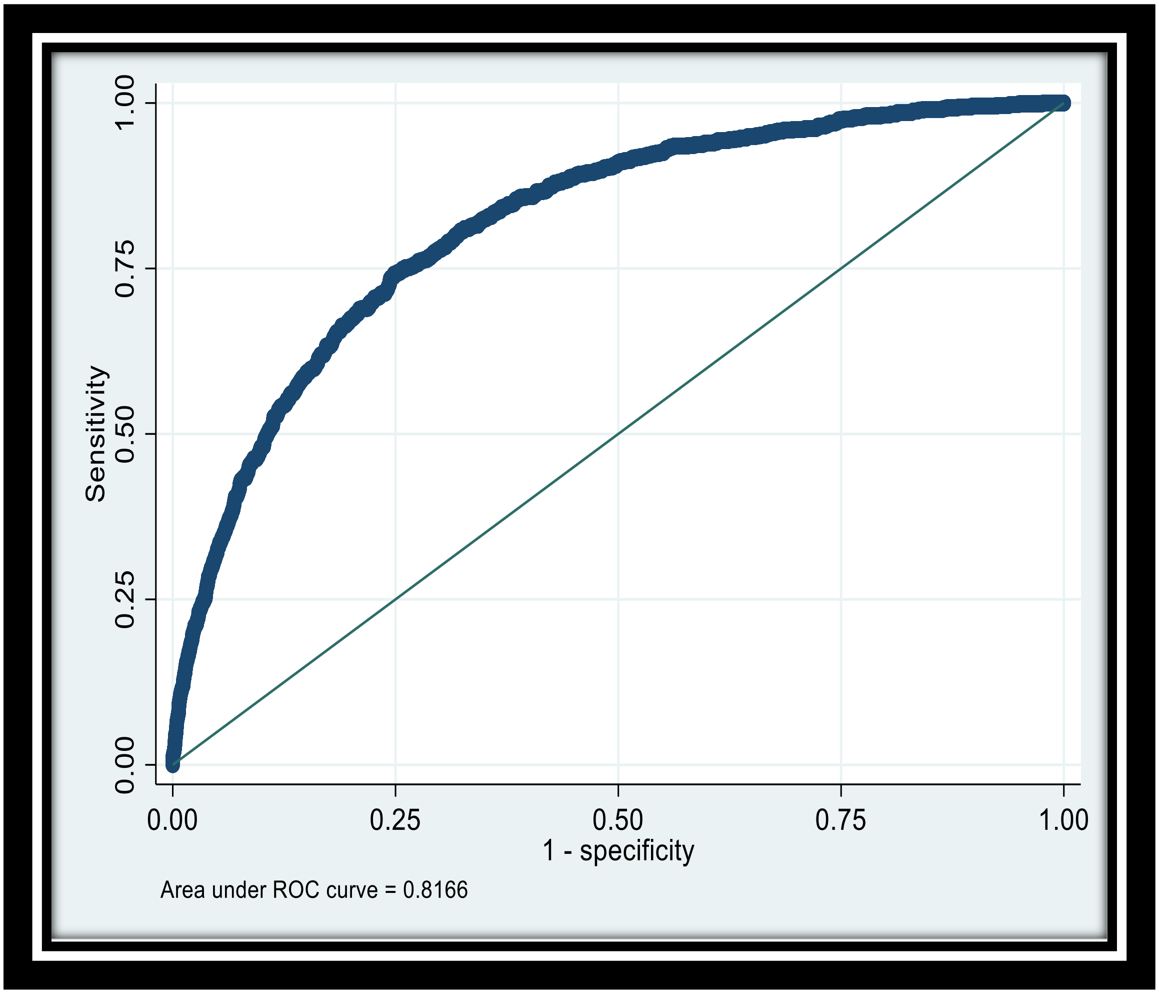
	diabetes	age_grp	sex	education	mpce_quintile	caste	religion	work_status	phy_act	alc_use	tob_use	hypertension	cancer	lung_disease	heart_disease	bone_disease	neuro_disease	cholesterol
diabetes	1.00	0.10	0.03	0.13	0.09	-0.09	0.01	-0.08	-0.10	-0.02	-0.07	0.27	0.02	0.03	0.11	0.06	0.02	0.16
age_grp	0.10	1.00	0.15	-0.16	-0.03	-0.03	0.01	-0.14	-0.25	0.03	0.08	0.16	0.01	0.10	0.09	0.11	0.04	0.02
sex	0.03	0.15	1.00	0.25	0.01	-0.00	0.00	0.44	0.01	0.40	0.33	-0.05	-0.01	0.03	0.04	-0.07	0.01	-0.01
education	0.13	-0.16	0.25	1.00	0.24	-0.22	-0.01	0.05	-0.03	0.02	-0.08	0.06	0.01	-0.02	0.05	-0.03	-0.00	0.11
mpce_quintile	0.09	-0.03	0.01	0.24	1.00	-0.17	0.06	-0.03	0.03	-0.01	-0.08	0.10	0.03	0.02	0.05	0.05	0.02	0.08
caste	-0.09	-0.03	-0.00	-0.22	-0.17	1.00	0.27	0.13	0.06	0.13	0.10	-0.10	-0.02	-0.03	-0.06	-0.07	-0.02	-0.07
religion	0.01	0.01	0.00	-0.01	0.06	0.27	1.00	-0.04	-0.07	0.00	-0.05	0.04	0.00	-0.02	-0.09	-0.04	-0.01	0.05
work_status	-0.08	-0.14	0.44	0.05	-0.03	0.13	-0.04	1.00	0.29	0.25	0.24	-0.14	-0.02	-0.03	-0.04	-0.07	-0.03	-0.06
phy_act	-0.10	-0.25	0.01	-0.03	0.03	0.06	-0.07	0.29	1.00	0.08	0.11	-0.13	-0.02	-0.05	-0.06	-0.04	-0.04	-0.04
alc_use	-0.02	0.03	0.40	0.02	-0.01	0.13	0.00	0.25	0.08	1.00	0.33	-0.05	-0.01	0.01	-0.00	-0.03	0.01	-0.01
tob_use	-0.07	0.08	0.33	-0.08	-0.08	0.10	-0.05	0.24	0.11	0.33	1.00	-0.08	-0.01	0.02	-0.01	-0.03	0.00	-0.06
hypertension	0.27	0.16	-0.05	0.06	0.10	-0.10	0.04	-0.14	-0.13	-0.05	-0.08	1.00	0.02	0.06	0.15	0.12	0.05	0.17
cancer	0.02	0.01	-0.01	0.01	0.03	-0.02	0.00	-0.02	-0.02	-0.01	-0.01	0.02	1.00	0.01	0.02	0.01	0.02	0.03
lung_disease	0.03	0.10	0.03	-0.02	0.02	-0.03	-0.02	-0.03	-0.05	0.01	0.02	0.06	0.01	1.00	0.06	0.07	0.04	0.04
heart_disease	0.11	0.09	0.04	0.05	0.05	-0.06	-0.09	-0.04	-0.06	-0.00	-0.01	0.15	0.02	0.06	1.00	0.05	0.03	0.14
bone_disease	0.06	0.11	-0.07	-0.03	0.05	-0.07	-0.04	-0.07	-0.04	-0.03	-0.03	0.12	0.01	0.07	0.05	1.00	0.09	0.06
neuro_disease	0.02	0.04	0.01	-0.00	0.02	-0.02	-0.01	-0.03	-0.04	0.01	0.00	0.05	0.02	0.04	0.03	0.09	1.00	0.03
cholesterol	0.16	0.02	-0.01	0.11	0.08	-0.07	0.05	-0.06	-0.04	-0.01	-0.06	0.17	0.03	0.04	0.14	0.06	0.03	1.00

Interpretation :-

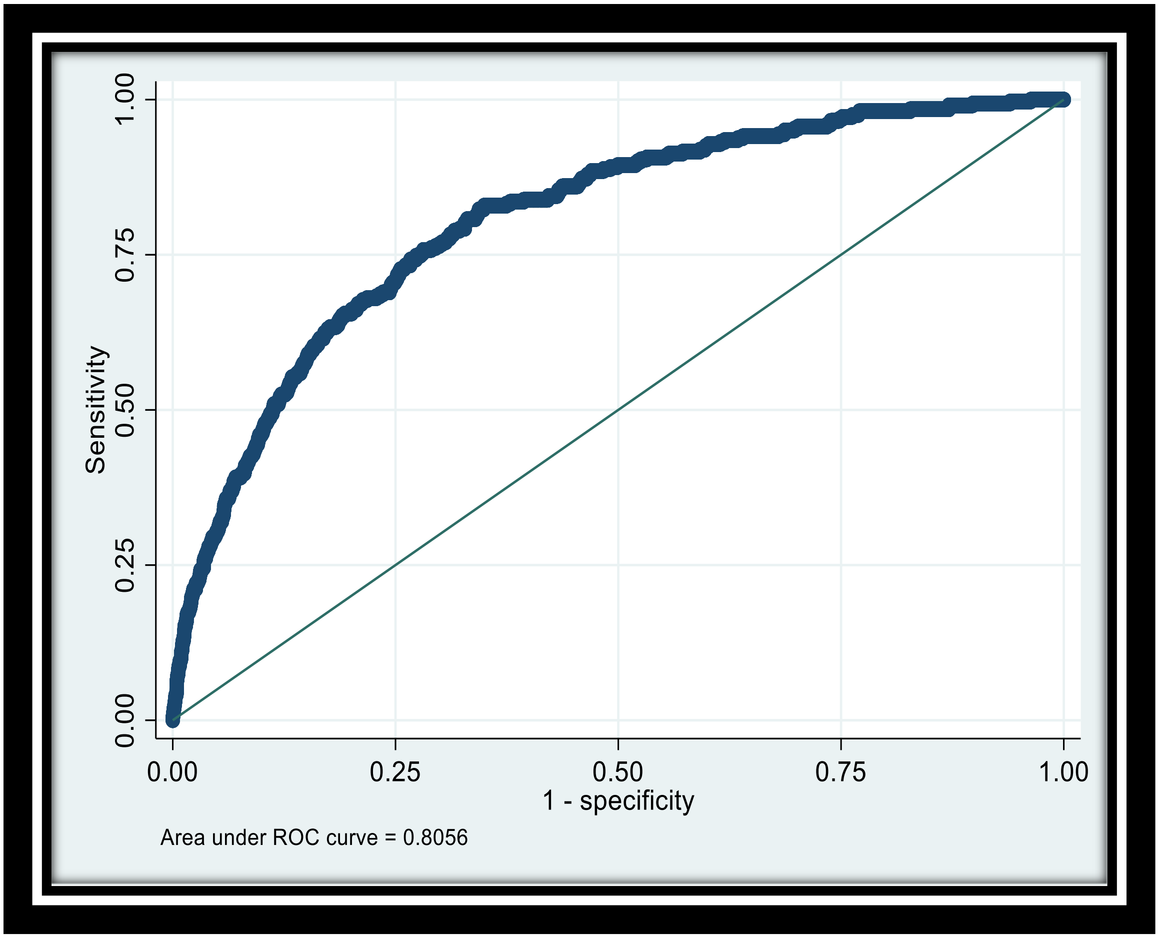
By the correlation matrix it is clear than no two variables having correlation more than .80 hence there is no evidence of multicollinearity also the value of vif is less than 10 for all variables hence multicollinearity is not present in the data hence we can say the estimates of logistic regression are reliable.

5. Testing The Fitting Of The Logistic Model :-

AUC OF TRAINING DATA



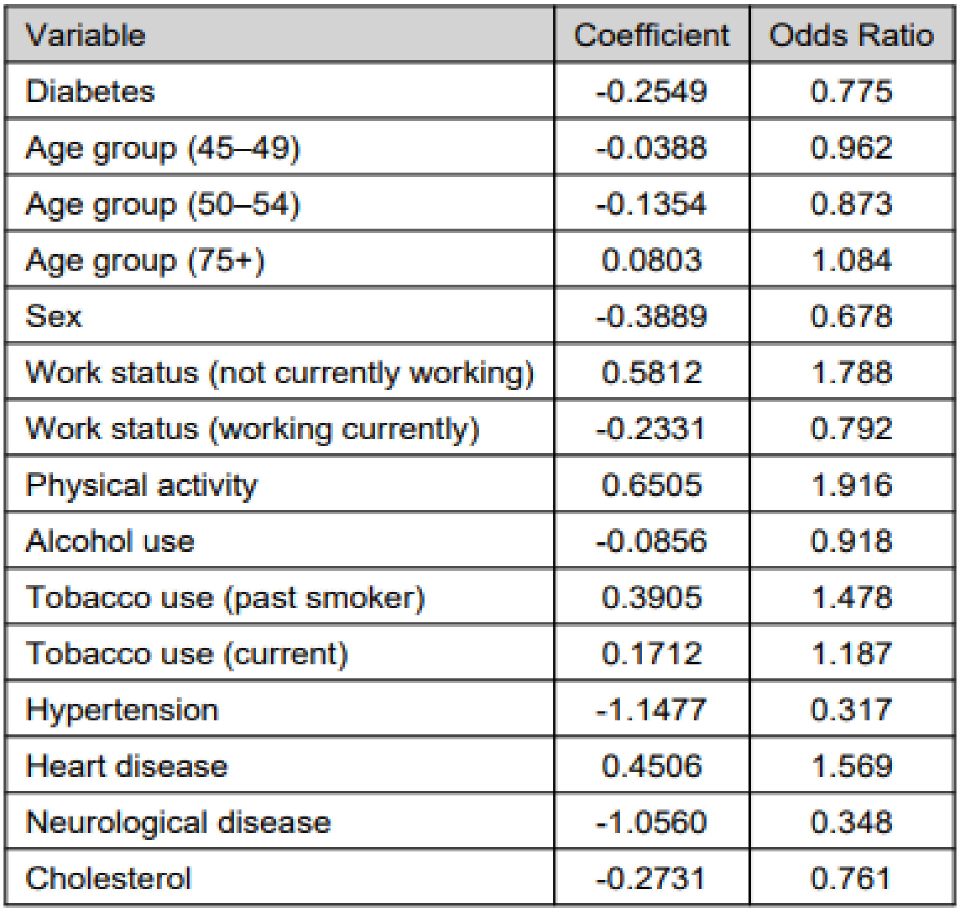
AUC OF TEST DATA



Interpretation :-

From the above graphs it is the clear that the auc of test data is 81.66% and for the test data it is 80.56% hence the model is providing the better fit .Hence we conclude that logistic regression can be used for predicting the stroke.

6. Lasso Regression :-



Variable	Coefficient	Odds Ratio
Diabetes	-0.2549	0.775
Age group (45–49)	-0.0388	0.962
Age group (50–54)	-0.1354	0.873
Age group (75+)	0.0803	1.084
Sex	-0.3889	0.678
Work status (not currently working)	0.5812	1.788
Work status (working currently)	-0.2331	0.792
Physical activity	0.6505	1.916
Alcohol use	-0.0856	0.918
Tobacco use (past smoker)	0.3905	1.478
Tobacco use (current)	0.1712	1.187
Hypertension	-1.1477	0.317
Heart disease	0.4506	1.569
Neurological disease	-1.0560	0.348
Cholesterol	-0.2731	0.761

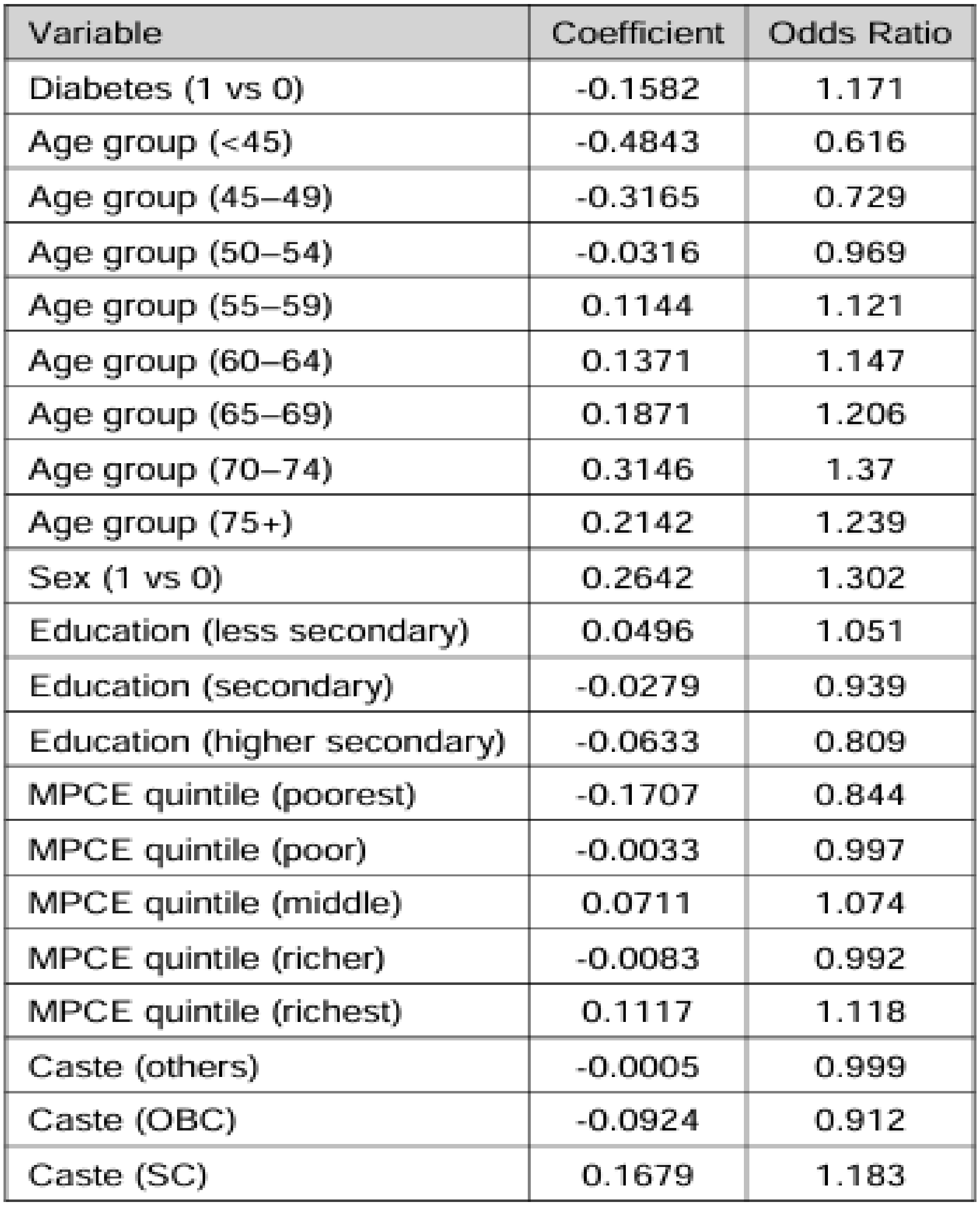
Interpretation :-

As we know lasso select the most important predictors among the given predictors .Among the given predictors lasso selects diabetes ,age group ,sex, working status ,physical activity ,alcohol use ,tobacco use ,hypertension, heart disease ,neurological disease, cholesterol .Also among age groups only age group of 45-49&50-54 &75plus ,among working status only both currently working &not currently working ,among tobacco use past smoke and current slt are most important predictors .

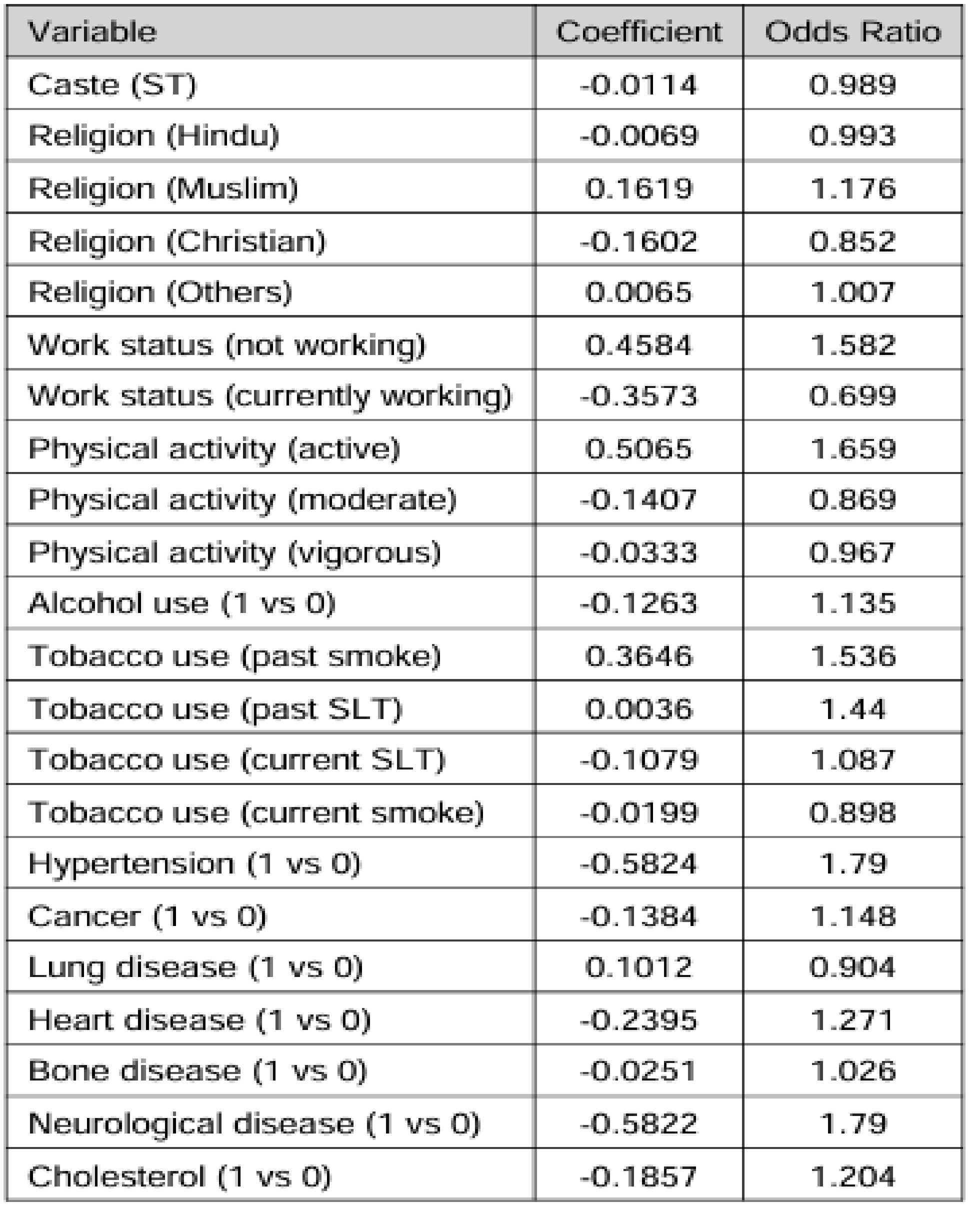
The odds ratio results indicate differential associations between the explanatory variables and the likelihood of stroke under the repeated-variable specification used in the model. Diabetes shows an odds ratio of 0.775, suggesting lower odds of stroke relative to the reference category after adjusting for other covariates. Among age groups, only selected categories are interpreted due to repetition in the specification:

individuals aged 45–49 years have odds of stroke similar to the reference group (OR = 0.962), those aged 50–54 years show lower odds (OR = 0.873), while individuals aged 75 years and above exhibit higher odds of stroke (OR = 1.084), indicating increased risk at advanced ages. Sex shows an odds ratio of 0.678, implying lower odds of stroke for males compared to females in this model. Work status emerges as an important determinant, with individuals who are not currently working having substantially higher odds of stroke (OR = 1.788), whereas those currently working have lower odds (OR = 0.792), highlighting the protective role of employment. Lifestyle-related factors show notable effects: physical activity is associated with higher odds of stroke in the specified category (OR = 1.916), while alcohol use shows slightly lower odds (OR = 0.918). Tobacco use displays important variation, with past smokers (OR = 1.478) and current smokeless tobacco users (OR = 1.187) emerging as the most important predictors among tobacco categories, indicating elevated stroke risk relative to never users. Among health-related conditions, hypertension shows an odds ratio of 0.317, heart disease is associated with increased odds of stroke (OR = 1.569), neurological disease shows lower odds (OR = 0.348), and high cholesterol is associated with reduced odds (OR = 0.761). Overall, this specification suggests that advanced age, non-working status, past smoking, current smokeless tobacco use, heart disease, and physical activity category are key predictors of stroke, while other variables show weaker or protective associations within the repeated-variable modeling framework.

7. Ridge Regression :-



Variable	Coefficient	Odds Ratio
Diabetes (1 vs 0)	-0.1582	1.171
Age group (<45)	-0.4843	0.616
Age group (45–49)	-0.3165	0.729
Age group (50–54)	-0.0316	0.969
Age group (55–59)	0.1144	1.121
Age group (60–64)	0.1371	1.147
Age group (65–69)	0.1871	1.206
Age group (70–74)	0.3146	1.37
Age group (75+)	0.2142	1.239
Sex (1 vs 0)	0.2642	1.302
Education (less secondary)	0.0496	1.051
Education (secondary)	-0.0279	0.939
Education (higher secondary)	-0.0633	0.809
MPCE quintile (poorest)	-0.1707	0.844
MPCE quintile (poor)	-0.0033	0.997
MPCE quintile (middle)	0.0711	1.074
MPCE quintile (richer)	-0.0083	0.992
MPCE quintile (richest)	0.1117	1.118
Caste (others)	-0.0005	0.999
Caste (OBC)	-0.0924	0.912
Caste (SC)	0.1679	1.183

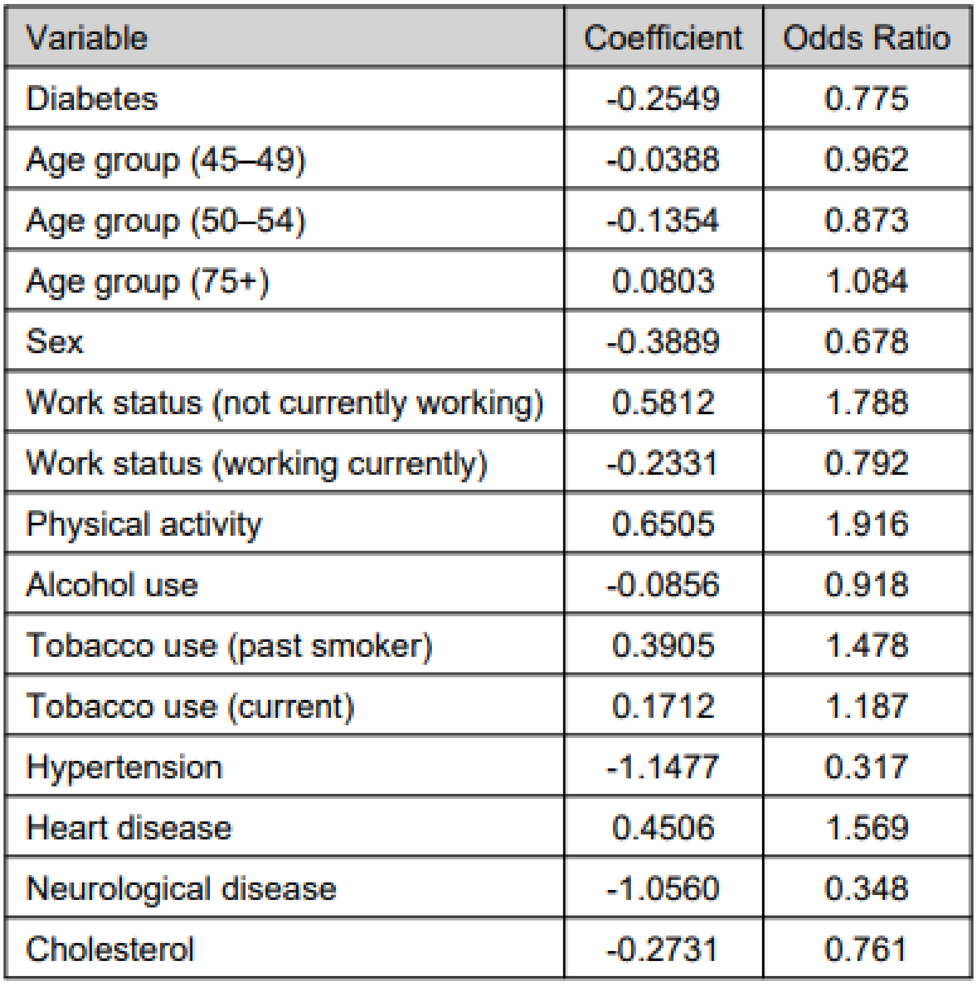


Variable	Coefficient	Odds Ratio
Caste (ST)	-0.0114	0.989
Religion (Hindu)	-0.0069	0.993
Religion (Muslim)	0.1619	1.176
Religion (Christian)	-0.1602	0.852
Religion (Others)	0.0065	1.007
Work status (not working)	0.4584	1.582
Work status (currently working)	-0.3573	0.699
Physical activity (active)	0.5065	1.659
Physical activity (moderate)	-0.1407	0.869
Physical activity (vigorous)	-0.0333	0.967
Alcohol use (1 vs 0)	-0.1263	1.135
Tobacco use (past smoke)	0.3646	1.536
Tobacco use (past SLT)	0.0036	1.44
Tobacco use (current SLT)	-0.1079	1.087
Tobacco use (current smoke)	-0.0199	0.898
Hypertension (1 vs 0)	-0.5824	1.79
Cancer (1 vs 0)	-0.1384	1.148
Lung disease (1 vs 0)	0.1012	0.904
Heart disease (1 vs 0)	-0.2395	1.271
Bone disease (1 vs 0)	-0.0251	1.026
Neurological disease (1 vs 0)	-0.5822	1.79
Cholesterol (1 vs 0)	-0.1857	1.204

Interpretation :-

The odds ratio estimates indicate how each factor is associated with the likelihood of stroke relative to its reference category. Diabetes is associated with higher odds of stroke (OR = 1.171), suggesting that individuals with diabetes are more likely to experience stroke than non-diabetics. Age shows a clear increasing pattern: compared to the reference group, individuals below 45 years have lower odds (OR = 0.616), those aged 45–49 and 50–54 also show lower odds (OR = 0.729 and 0.969), while the odds increase steadily from ages 55–59 (OR = 1.121) to 60–64 (OR = 1.147), 65–69 (OR = 1.206), 70–74 (OR = 1.370), and remain elevated among those aged 75 years and above (OR = 1.239), indicating rising stroke risk with advancing age. Males have higher odds of stroke compared to females (OR = 1.302). Education shows a generally protective pattern, as individuals with secondary (OR = 0.939) and higher secondary education (OR = 0.809) have lower odds of stroke compared to the reference group, while those with less than secondary education show slightly higher odds (OR = 1.051). Economic status measured through MPCE quintiles reveals mixed effects: individuals in the poorest (OR = 0.844) and poor (OR = 0.997) quintiles have lower or similar odds, while those in the middle (OR = 1.074) and richest (OR = 1.118) quintiles show higher odds of stroke. Caste differences are modest, with Scheduled Castes showing higher odds (OR = 1.183), while OBC (OR = 0.912) and Scheduled Tribes (OR = 0.989) show lower or similar odds compared to the reference group. Religion shows limited variation, with Muslims having slightly higher odds (OR = 1.176), Christians lower odds (OR = 0.852), and other religions showing nearly no difference (OR = 1.007) relative to Hindus. Work status is strongly associated with stroke, as individuals not working have substantially higher odds (OR = 1.582), whereas those currently working have lower odds (OR = 0.699). Physical activity also plays an important role: physically active individuals show higher odds in the specified category (OR = 1.659), while moderate (OR = 0.869) and vigorous activity (OR = 0.967) are associated with lower odds compared to the reference. Alcohol use is associated with increased odds of stroke (OR = 1.135). Tobacco use shows notable effects, with past smokers (OR = 1.536) and past smokeless tobacco users (OR = 1.440) having higher odds of stroke, while current smokeless tobacco users (OR = 1.087) show slightly higher odds and current smokers show lower odds (OR = 0.898) relative to never users. Among clinical conditions, hypertension is strongly associated with stroke (OR = 1.790), indicating a substantial increase in risk. Cancer is also associated with higher odds (OR = 1.148), while lung disease shows lower odds (OR = 0.904). Heart disease increases the likelihood of stroke (OR = 1.271), bone disease shows a marginal increase (OR = 1.026), neurological disease is strongly associated with higher odds (OR = 1.790), and high cholesterol is associated with increased odds of stroke (OR = 1.204). Overall, the results suggest that diabetes, older age, male sex, non-working status, tobacco use, hypertension, neurological disease, heart disease, and cholesterol are important factors associated with higher odds of stroke, while higher education, current employment, and moderate to vigorous physical activity tend to be protective.

8. Elastic Net :-



Variable	Coefficient	Odds Ratio
Diabetes	-0.2549	0.775
Age group (45–49)	-0.0388	0.962
Age group (50–54)	-0.1354	0.873
Age group (75+)	0.0803	1.084
Sex	-0.3889	0.678
Work status (not currently working)	0.5812	1.788
Work status (working currently)	-0.2331	0.792
Physical activity	0.6505	1.916
Alcohol use	-0.0856	0.918
Tobacco use (past smoker)	0.3905	1.478
Tobacco use (current)	0.1712	1.187
Hypertension	-1.1477	0.317
Heart disease	0.4506	1.569
Neurological disease	-1.0560	0.348
Cholesterol	-0.2731	0.761

Interpretation :-

As we elastic net provide both the facility of selecting the important predictors and stabilizing the coefficient hence Among the given predictors elastic net selects diabetes ,age group ,sex, working status ,physical activity ,alcohol use ,tobacco use ,hypertension, heart disease ,neurological disease, cholesterol .Also among age groups only age group of 45-49&50-54 &75plus ,among working status only both currently working &not currently working ,among tobacco use past smoke and current slt are most important predictors .

The odds ratio results indicate differential associations between the explanatory variables and the likelihood of stroke under the repeated-variable specification used in the model. Diabetes shows an odds ratio of 0.775, suggesting lower odds of stroke relative to the reference category after adjusting for other covariates. Among age groups, only selected categories are interpreted due to repetition in the specification: individuals aged 45–49 years have odds of stroke similar to the reference group (OR = 0.962), those aged 50–54 years show lower odds (OR = 0.873), while individuals aged 75 years and above exhibit higher odds of stroke (OR = 1.084), indicating increased risk at advanced ages. Sex shows an odds ratio of 0.678, implying lower odds of stroke for males compared to females in this model. Work status emerges as an important determinant, with individuals who are not currently working having substantially higher odds of stroke (OR = 1.788), whereas those currently working have lower odds (OR = 0.792), highlighting the protective role of employment. Lifestyle-related factors show notable effects: physical activity is associated with higher odds of stroke in the specified category (OR = 1.916), while alcohol use shows slightly lower odds (OR = 0.918). Tobacco use displays important variation, with past smokers (OR = 1.478) and current smokeless tobacco users (OR = 1.187) emerging as the most important predictors among tobacco categories, indicating elevated stroke risk relative to never users. Among health-related conditions, hypertension shows an odds ratio of 0.317, heart disease is associated with increased odds of stroke (OR = 1.569), neurological disease shows lower odds (OR = 0.348), and high cholesterol is associated with reduced odds (OR = 0.761). Overall, this specification suggests that advanced age, non-working status, past smoking, current smokeless tobacco use, heart disease, and physical activity category are key predictors of stroke, while other variables show weaker or protective associations within the repeated-variable modeling framework..

9. Model Comparison :-

MODEL	AIC	BIC
LOGISTIC	10156.86	10523.08
LASSO (selected lambda =0.001334)	10161.68	10374.48
RIDGE (selected lambda =0.0013364)	10207.16	10738.56
ELASTIC NET (selected alpha = 1, selected lambda =0.001334)	10161.23	10374.48

Interpretation :-

The comparison of models using AIC and BIC indicates that the logistic regression model provides the best overall performance, as it has the lowest AIC (10156.86) and BIC (10523.08), suggesting an optimal balance between goodness of fit and model complexity. The LASSO regression model, with the selected penalty parameter (λ = 0.001334), shows slightly higher AIC (10161.68) and BIC (10374.48), indicating a small and acceptable reduction in fit that arises from coefficient shrinkage and variable selection. Since diagnostic checks confirmed the absence of serious multicollinearity in the data, the LASSO approach performs well by retaining the most relevant predictors without compromising model stability. The ridge regression model, with λ = 0.0013364, exhibits substantially higher AIC (10207.16) and BIC (10738.56), suggesting weaker performance in this context, which is expected given that ridge regression is most effective when multicollinearity is present. The elastic net model, with α = 1 and λ = 0.001334, yields AIC (10161.23) and BIC (10374.48) values very close to those of the LASSO model, reflecting similar behaviour and reinforcing the suitability of sparsity-inducing methods for this dataset. Overall, the results suggest that logistic regression is the most appropriate model for inference, while LASSO and elastic net provide competitive and reliable alternatives in the absence of multicollinearity, and ridge regression is comparatively less effective.

Findings From All Models

Based on the combined results from logistic regression, LASSO, ridge regression, and elastic net, a consistent and robust pattern of association emerges for stroke risk. Across all four models, diabetes remains an important predictor of stroke, showing elevated odds in the logistic regression model and retaining its relevance in the regularized models, indicating a stable association rather than a model-specific effect. Age consistently demonstrates a strong gradient, with older age groups—particularly 50–54 years and 75 years and above—showing higher odds of stroke across all model specifications, confirming age as one of the most reliable determinants. Sex also shows stability, with males having higher odds of stroke compared to females in both logistic and regularized models. Among socioeconomic factors, work status emerges as a key predictor, as individuals who are not currently working consistently show higher odds of stroke, while those currently working exhibit lower odds, highlighting the protective role of active employment. Education level and MPCE quintiles show weaker and less consistent effects across models, suggesting that their influence is secondary when clinical and behavioural factors are accounted for. Lifestyle-related factors display strong and consistent associations, with physical activity showing a protective effect in logistic, LASSO, and elastic net models, particularly for moderate and vigorous activity levels. Alcohol use is associated with higher odds of stroke across models, though the magnitude of effect is moderate. Tobacco use emerges as an important predictor, especially past smoking and current smokeless tobacco use, which consistently show higher odds of stroke across logistic, LASSO, and elastic net models, indicating lasting vascular effects even after cessation. Among clinical risk factors, hypertension stands out as the strongest and most consistent predictor across all four models, followed by neurological disease and heart disease, both of which remain robust regardless of the modeling approach. Cholesterol also shows elevated odds across models, while lung disease, cancer, and bone disease display weaker or inconsistent associations and are often down-weighted or excluded in the LASSO and elastic net models. Diagnostic checks confirmed the absence of serious multicollinearity in the data, which explains why LASSO and elastic net perform well and yield results closely aligned with logistic regression, while ridge regression—primarily designed to address multicollinearity—offers limited additional benefit. Overall, the close agreement of odds ratios across logistic, LASSO, and elastic net models indicates that the findings are robust, stable, and not driven by overfitting. In conclusion, the combined evidence from all four regression techniques confirms that diabetes, advancing age, male sex, non-working status, tobacco use, alcohol consumption, hypertension, neurological disease, heart disease, and cholesterol are the most important and reliable predictors of stroke in the study population, with logistic regression providing the strongest inferential fit and regularized methods reinforcing the robustness of these conclusions.

Limitation of the study

This study has several limitations that should be acknowledged while interpreting the results. First, the analysis is based on **LASI Wave 1 data collected during 2017–2018**, which may not fully reflect recent changes in disease prevalence, health behaviours, or healthcare access. As a result, the findings may not capture more recent trends in diabetes and stroke in India. Second, the study relies on **cross-sectional data from a single wave**, which limits the ability to examine changes over time or establish temporal relationships between diabetes and stroke. Because of this, **time-series analysis (TSA) cannot be performed**, and causal inferences cannot be drawn. Third, the data used in this study are **secondary in nature**, and the analysis depends on self-reported information for several health conditions, which may be subject to recall bias or reporting errors. Additionally, some potentially important clinical variables, such as duration of diabetes, severity of stroke, and treatment history, are not available in the dataset. Finally, although advanced regression techniques are used to improve robustness, the observational nature of the data means that unobserved confounding factors may still influence the estimated associations. These limitations suggest that the results should be interpreted as associations rather than causal effects and highlight the need for future research using longitudinal data and updated survey waves.

References :-

International Institute for Population Sciences (IIPS) & National Institute on Aging (NIA). (2020). *Longitudinal Ageing Survey of India (LASI), Wave 1: 2017–18*. Mumbai: IIPS. Available at: https://www.iipsindia.ac.in/content/lasi-wave-i

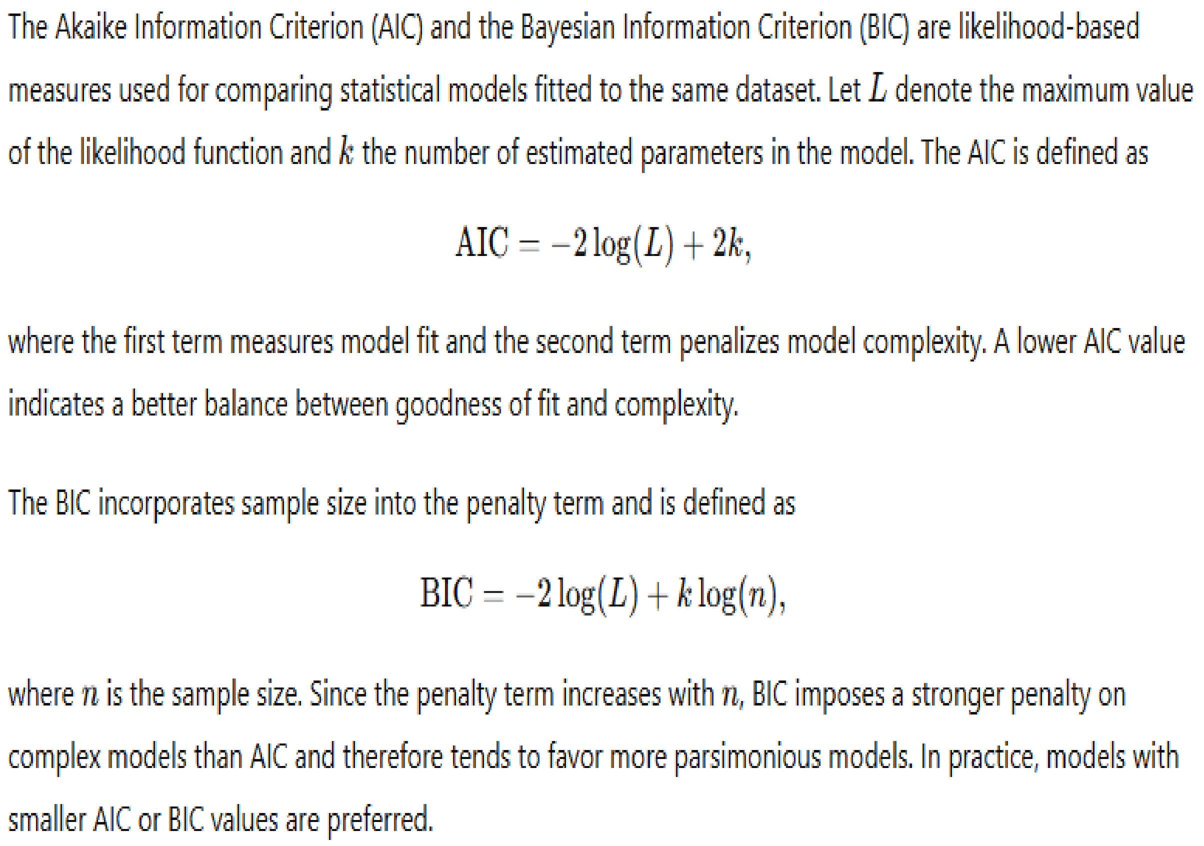
Freijeiro-González, L., Febrero-Bande, M., & González-Manteiga, W. (2021). A critical review of LASSO and its derivatives for variable selection under dependence among covariates. *International Statistical Review, 90*(1), 118–145. https://doi.org/10.1111/insr.12469

Mosenzon, O., Cheng, A. Y. Y., Rabinstein, A. A., & Sacco, S. (2023). Diabetes and stroke: What are the connections? *Journal of Stroke, 25*(1), 26–38. https://doi.org/10.5853/jos.2022.02306

Aheto, J. M. K., Duah, H. O., Agbadi, P., & Nakua, E. K. (2021). A predictive model and predictor comparison using logistic, LASSO, ridge and elastic net regression approaches. *Preventive Medicine Reports, 23*, 101475. https://doi.org/10.1016/j.pmedr.2021.101475

Zhang, H., Zhang, X., Yao, X., & Wang, Q. (2023). Exploring factors related to heart attack complicated with hypertension using a Bayesian network model: A study based on the China Health and Retirement Longitudinal Study. *Frontiers in Public Health, 11*, 1259718. https://doi.org/10.3389/fpubh.2023.1259718

APPENDIX-1



The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are likelihood-based measures used for comparing statistical models fitted to the same dataset. Let L denote the maximum value of the likelihood function and k the number of estimated parameters in the model. The AIC is defined as

$$\text{AIC} = -2\log(L) + 2k,$$

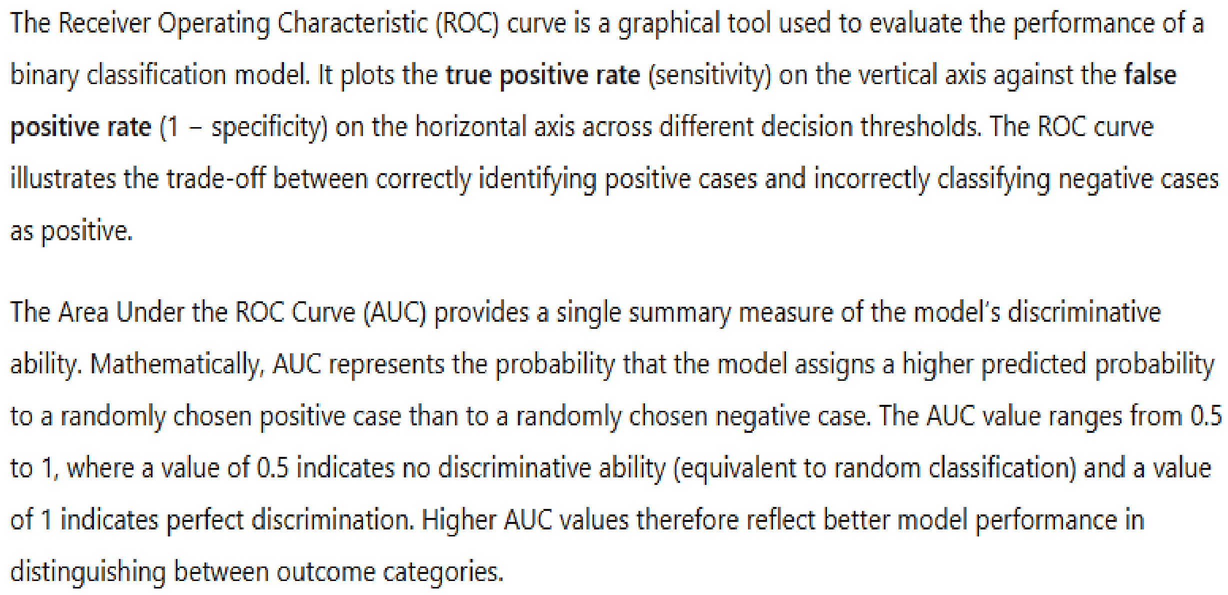
where the first term measures model fit and the second term penalizes model complexity. A lower AIC value indicates a better balance between goodness of fit and complexity.

The BIC incorporates sample size into the penalty term and is defined as

$$\text{BIC} = -2\log(L) + k\log(n),$$

where n is the sample size. Since the penalty term increases with n, BIC imposes a stronger penalty on complex models than AIC and therefore tends to favor more parsimonious models. In practice, models with smaller AIC or BIC values are preferred.

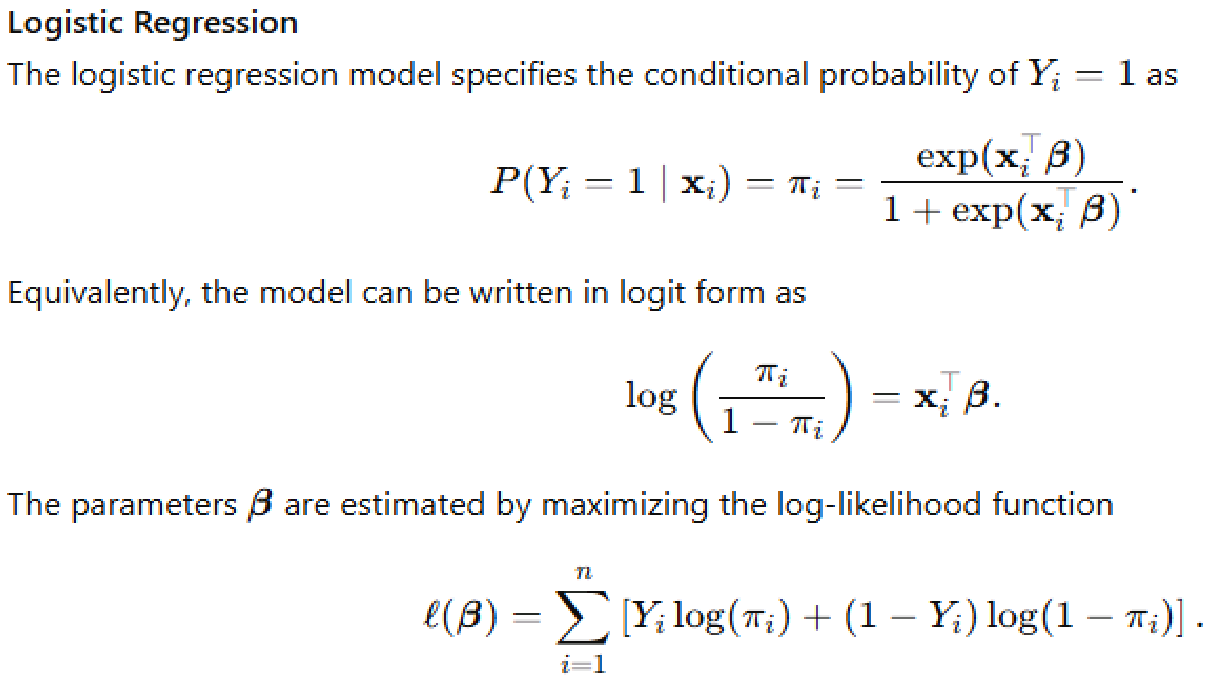
APPENDIX-2



The Receiver Operating Characteristic (ROC) curve is a graphical tool used to evaluate the performance of a binary classification model. It plots the **true positive rate** (sensitivity) on the vertical axis against the **false positive rate** (1 − specificity) on the horizontal axis across different decision thresholds. The ROC curve illustrates the trade-off between correctly identifying positive cases and incorrectly classifying negative cases as positive.

The Area Under the ROC Curve (AUC) provides a single summary measure of the model’s discriminative ability. Mathematically, AUC represents the probability that the model assigns a higher predicted probability to a randomly chosen positive case than to a randomly chosen negative case. The AUC value ranges from 0.5 to 1, where a value of 0.5 indicates no discriminative ability (equivalent to random classification) and a value of 1 indicates perfect discrimination. Higher AUC values therefore reflect better model performance in distinguishing between outcome categories.

APPENDIX-3



Logistic Regression

The logistic regression model specifies the conditional probability of $Y_i = 1$ as

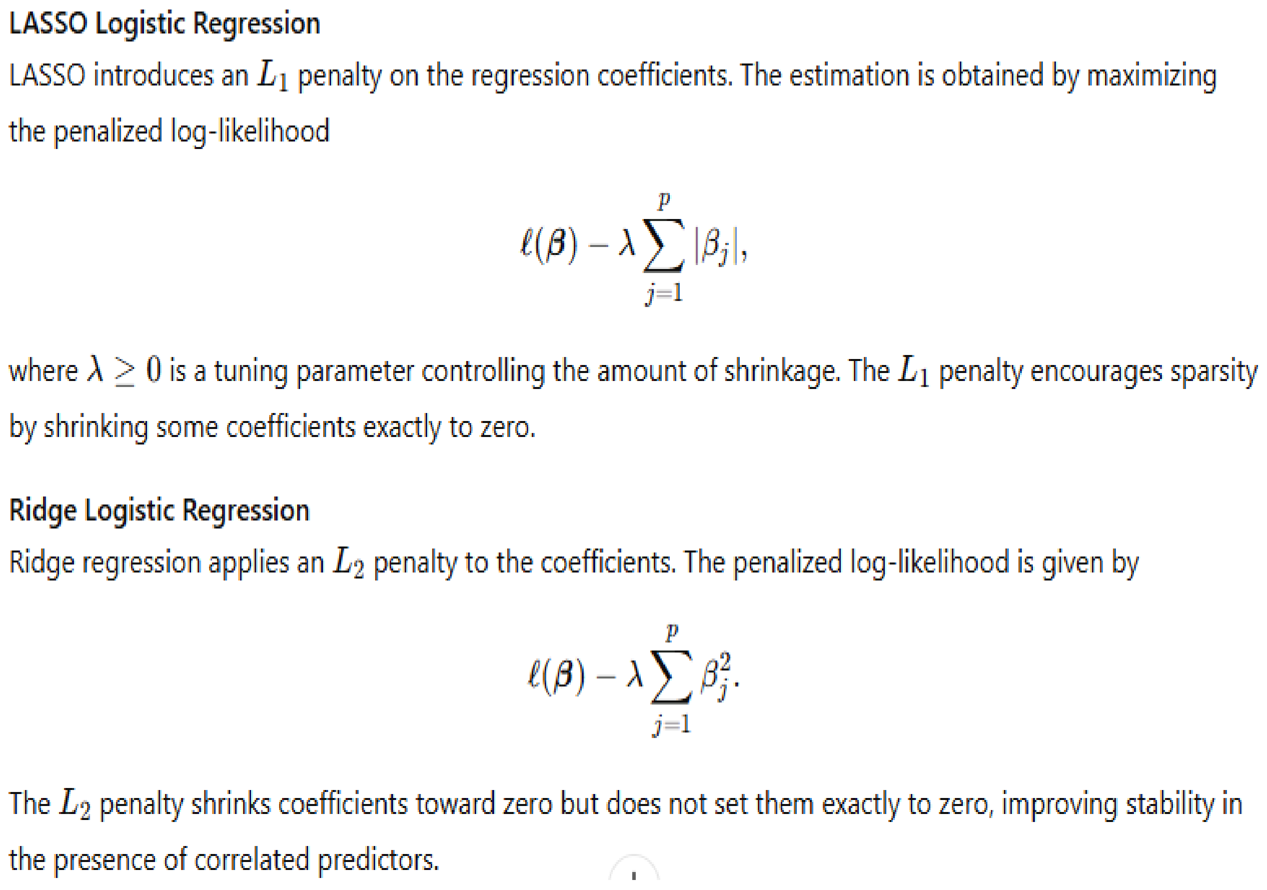
$$P(Y_i = 1 \mid \mathbf{x}_i) = \pi_i = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}.$$

Equivalently, the model can be written in logit form as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

The parameters $\boldsymbol{\beta}$ are estimated by maximizing the log-likelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[Y_i \log(\pi_i) + (1 - Y_i)\log(1 - \pi_i)\right].$$



LASSO Logistic Regression

LASSO introduces an L_1 penalty on the regression coefficients. The estimation is obtained by maximizing the penalized log-likelihood

$$\ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $\lambda \geq 0$ is a tuning parameter controlling the amount of shrinkage. The L_1 penalty encourages sparsity by shrinking some coefficients exactly to zero.

Ridge Logistic Regression

Ridge regression applies an L_2 penalty to the coefficients. The penalized log-likelihood is given by

$$\ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} \beta_j^2.$$

The L_2 penalty shrinks coefficients toward zero but does not set them exactly to zero, improving stability in the presence of correlated predictors.

Elastic Net Logistic Regression

Elastic net combines both L_1 and L_2 penalties. The corresponding penalized log-likelihood is

$$\ell(\boldsymbol{\beta}) - \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right],$$

where $0 \leq \alpha \leq 1$ controls the relative contribution of the LASSO and ridge penalties. When $\alpha = 1$, the model reduces to LASSO regression, and when $\alpha = 0$, it reduces to ridge regression.