بسمه تعالی

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

موضوع :

گزارش پروژه درس یادگیری ژرف

دانشجو : متین قاسمی

شماره دانشجویی : 402112098

استاد : دکتر مصطفی شمسی – دکتر زاهد رحمتی

بهار ۱۴۰۳

فهرست

مقدمه 2
پیش بردازش 3
معرفی دیتاست 3
اصلاح داده های خالی 3
تمیز کردن دیتاست 4
تحلیل داده ها 6
مدل های پایه یادگیری ماشین 10
پیاده سازی 10
نتایج 10
مدل یادگیری ماشین مبتنی بر گراف 10
تشکیل گراف 11
مدل GCN 11
مقایسه با دیگر مدل ها 13
نتایج 14

مقدمه

هدف از تعریف این پروژه پیاده سازی انواع مدل های یادگیری ماشین پایه ای روی حیوانات برای دسته بندی رژیم غذایی آن ها و سپس استفاده از مدل های مبتنی بر گراف در یادگیری ماشین بر روی داده ها و مقایسه نتایج آن ها است. این گزارش در چهار بخش تدوین شده است که شامل پیش پردازش داده ها برای اصلاح دیتاست اولیه و آماده سازی داده ها جهت پیاده سازی الگوریتم های یادیگری ماشین، آنالیز و تحلیل داده ها، پیاده سازی مدل های پایه ای یادگیری ماشین و در نهایت پیاده سازی مدل مبتنی بر گراف و مقایسه نتایج است.

به همراه فایل گزایش یک مجوعه فایل پیوست ارایه شده است که شامل ژوپیتر نوت بوک هایی مربط به هر بخش می باشد. در این نوت بوک ها به تفکیک اعمال انجام شده توضیح داده شده است و این فایل، گزارش تکمیلی است فلذا در این گزارش از پرداختن به جز به جز اعمال انجام پرهیز میکنیم و تنها به کلیات موضوع اشاره می شود.

پیش بردازش

در این بخش ابتدا به صورت کلی با دیتاست انتخاب شده آشنا میشویم سپس و مشکلات موجود در داده ها را شناسایی میکنیم و در نهایت این مشکلات را حل میکنیم. همچنین در طی همه مراحل سعی میکنیم حداقل داده ها را حدف کنیم و از جداکثر پتانسیل دیتاست انتخاب شده استفاده کنیم.

معرفی دیتاست

دیتاست حیوانات شامل ۱۶ ستون و ۲۰۵ ردیف است. در دیتاست اولیه حتی رغم تعریف عددی بعضی از ستون ها مانند قد، همه ی آن ها به شکل کاراکتر تعریف شده اند. هم چنین در بعضی ستون ها عمارت های تکراری مانند 'Ups to' دیده میشود که نیازمند اصلاح است. در ستون هایی که بازه عددی نوشته شده است از میانگین آن ها استفاده میکنیم. هم چنین در دیتاست اولیه داده های تکراری از حیوانات مشاهده میشود. برای رفع این مشکل ستون تکراری که داده های NAN بیشتری دارد از مجموعه داده حذف شده است. با کمی دقت میتوان متوجه شده در بعضی ستون ها به جای مقدار خالی عبارت هایی مانند 'Not Applicable' دیده میشود. در مجموعه میتوان نشان داده در دیتاست ۱۵۳ خانه خالی وجود دارد. مورد بعدی تنوع ۱۶ تایی در ستون 'Diet' است که به عنواع برچسب دسته بندی باید مورد استفاده قرار بگیرد بنابراین نیاز است تعداد آن به ۳ تا کاهش پیدا کند. همچنین تعدادی از خانه ها دارای اطاعات غلط بودند که نیازمند اصلاح بود.

اصلاح داده های خالی

در جدول زیر میتوان تعداد خانه های خالی هریک از سطر های دیتاست را در ابتدای پیش پردازش مشاهده کرد.

Column	NAN
Animal	0
Height (cm)	3
Weight (kg)	3
Color	5
Lifespan (years)	6
Diet	0
Habitat	1
Predators	14
Average Speed (km/h)	26
Countries Found	0
Conservation Status	12
Family	0
Gestation Period (days)	19
Top Speed (km/h)	38
Social Structure	1
Offspring per Birth	25

در ابتدا سطر هایی که بیشتر از ۳۰ درصد از داده های آن ها خالی بود از دیتاست حذف شدند. همچنین برای کاهش حدف سطر های بقیه خانه های خالی با استفاده اطلاعات ChatGPT تکمیل شد.

تمیز کردن دیتاست

برای اصلاح دیتاست ابتدا ستون هایی که باید به شکل عددی تعریف شوند اصلاح شدند. همچنین عبارت های ثابت موجود در ستون ها جذف شدند. ستون 'Diet' که به عنوان برچست دسته بندی مورد استفاده قرار میگیرد اصلاح شده به سه دسته ی گیاه خوار،گوشت خوار و همه چیز خوار تقسیم شدند که توزیع داده ها را میتوان در جدول زیر مشاهده کرد. تعداد حیوانات گوشت خوار حدود نیمی از دیتاست است که این موضوع میتواند مشکل ساز باشد اما بنا به تعریف پروژه که خواسته شده است در این باره کاری انجام نشود داده ها به همان شکل باقی مانده اند.

Diet	Count
Carnivore	94
Herbivore	51
Omnivore	42

با توجه به همبستگی بسیار زیاد ستون سرعت میانگین و حداکثر سرعت تصمیم گرفته شده یکی از آن ها حدف شود که به دلیل کمتر بودن مقدار خانه های خالی ستون ستون میانگین این ستون نگه داشته شد.

در نهایت پس از انجام اصلاحات مورد نیاز و تکمیل خانه های خالی و اصلاح خانه های غلط دیتاست شامل ۱۵ ستون و ۱۸۷ سطر است که هیچ خانه ی خالی ای ندارد.

```
#   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
0   Animal                  187 non-null    object
1   Height (cm)             187 non-null    float64
2   Weight (kg)             187 non-null    float64
3   Color                   187 non-null    object
4   Lifespan (years)        187 non-null    float64
5   Diet                    187 non-null    object
6   Habitat                 187 non-null    object
7   Predators               187 non-null    object
8   Average Speed (km/h)    187 non-null    float64
9   Countries Found         187 non-null    object
10  Conservation Status     187 non-null    object
11  Family                  187 non-null    object
12  Gestation Period (days) 187 non-null    float64
13  Social Structure        187 non-null    object
14  Offspring per Birth     187 non-null    float64
```

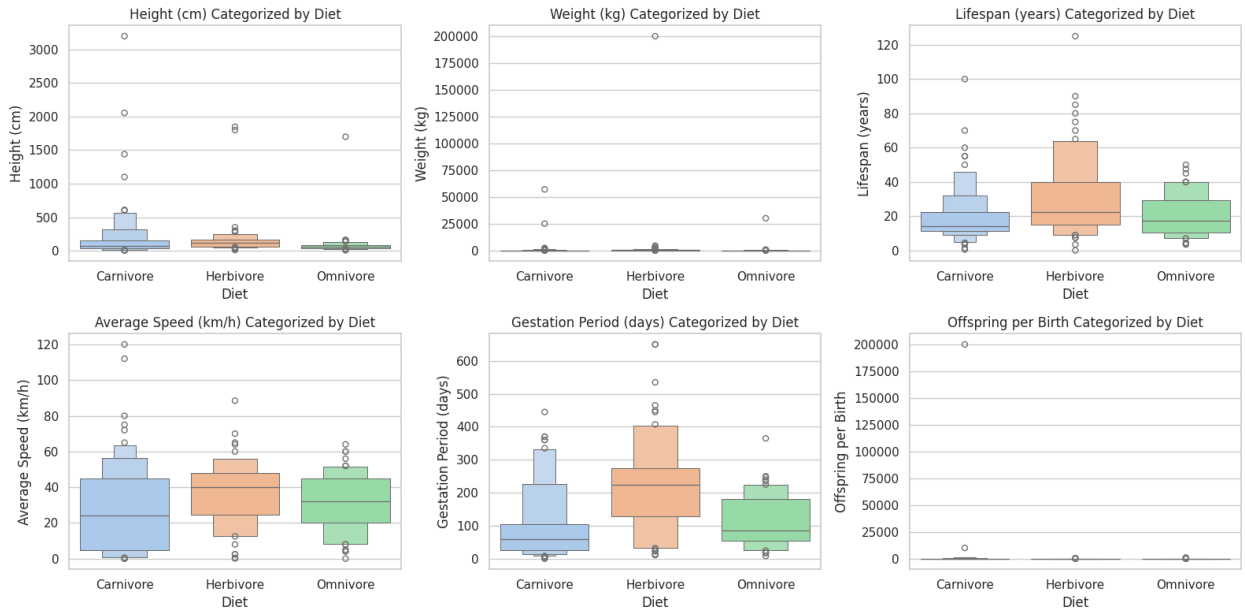
همچنین در جدول زیر میتوان اطلاعات آماری ستون های عددی را مشاهده کرد. فاصله داشتن میان از مقادیر بیشینه نشانه ی وجود داده های پرت است که خودشان را در نمودار ها نیز نشان خواهند داد.

Col / Stat	Height (cm)	Weight (kg)	Lifespan (years)	Average Speed (km/h)	Gestation Period (days)	Offspring per Birth
count	187.000000	187.000000	187.000000	187.000000	187.000000	187.000000
mean	175.828877	1889.816746	22.933983	31.377995	134.829259	1182.791444
std	373.586782	15409.975050	18.961394	22.355671	129.028395	14636.520588
min	4.500000	0.006000	0.071429	0.025000	0.071429	1.000000
25%	40.000000	3.325000	11.500000	14.000000	37.000000	1.000000
50%	78.500000	31.500000	17.500000	30.000000	90.000000	2.500000
75%	150.000000	180.000000	27.500000	45.750000	207.500000	15.750000
max	3200.000000	200000.000000	125.000000	120.000000	650.000000	200000.000000

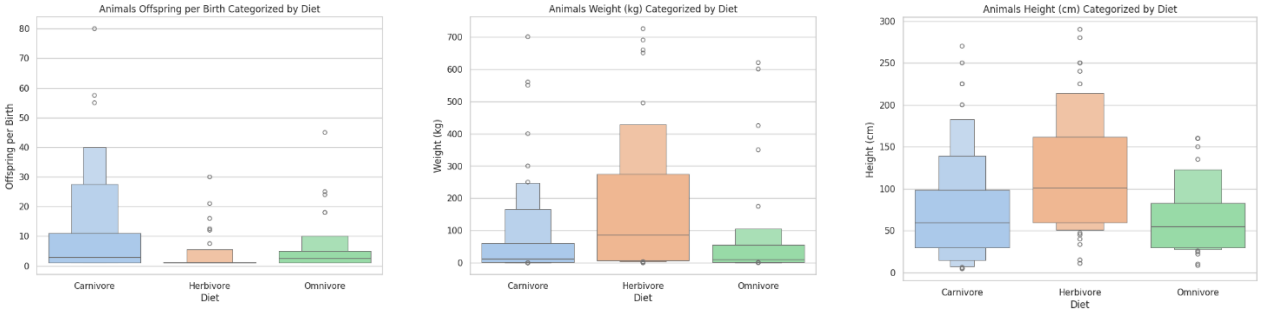
تحلیل داده ها

در این بخش ابتدا ستون های مختلف را با نمودارهای مختلف تصویر میکنیم تا به ابعاد مختلف دیتاست دست پیدا کنیم سپس به استفاده از ابزار های آماری دیتاست را بررسی میکنیم و به دنبال الگوهایی میگردیم که در آینده بتواند به ما کمک کند.

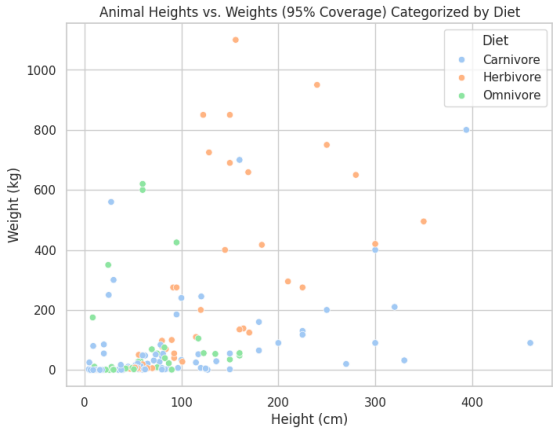
نمودار جعبه ای متواند به خوبی ویژگی هایی از توزیع داده ها نشان دهد. همانطور که مشاهده میکنید در تعداد از ستون ها داده های پرت نسبت به بقیه داده ها هستند به عنوان مثال وال آبی با وزنی در حدود ۲۰۰ تن بزرگ ترین موجود روی کره زمین است که وزن آن به مینگین بقیه حیوانات بسیار فاصله دارد. همچنین موجوداتی وجود دارند که توانایی تخم گذاری در تعداد چند ده هزار هستند که نسبت به پستنداران که انگشت شمار فرزند به دنیا می اورند فاصله بسیار دارند.



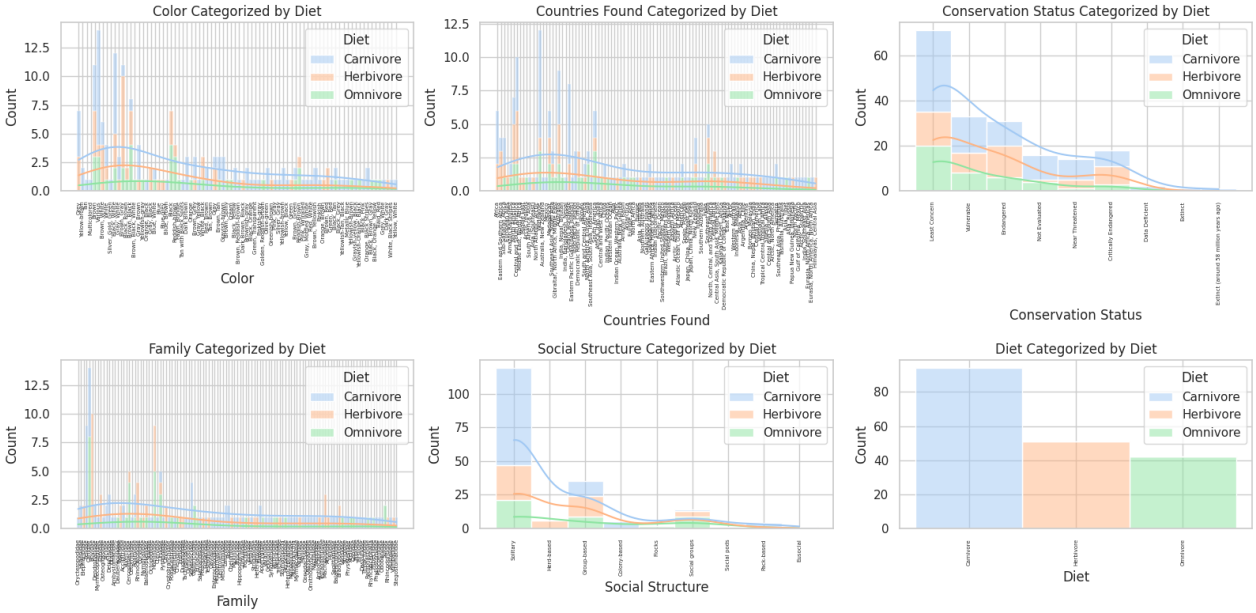
برای نشان داده بهتر توزیع داده ها در سه ستونی که در بالا واضح نیستند از نمودار های پایین استفاده میکنیم که شامل ۹۵ درصد داده های موجود در دیتاست میشوند.



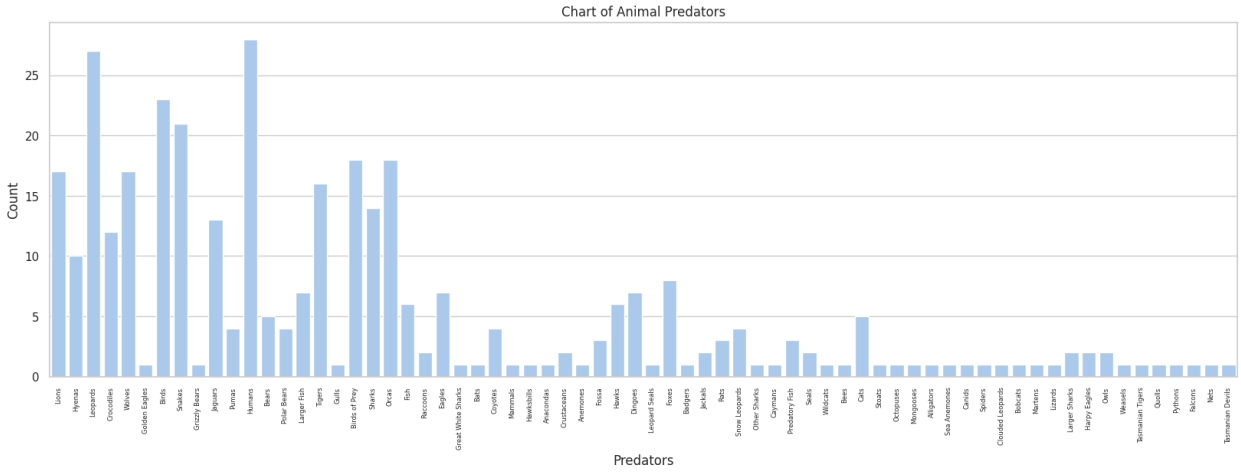
همچنین در تصویر زیر میتوان تاثیرجثه موجودات بر روی نوع رژیم غذایی آن ها را مشاهده کرد.



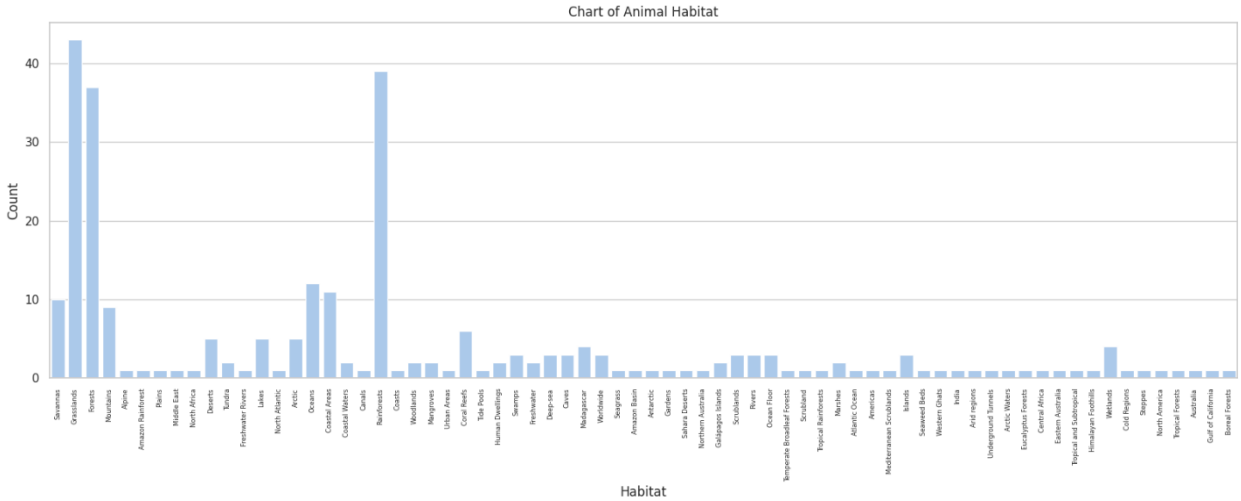
توزیع داده ها بر روی ستون های کتگوریکال نیز به شکل زیر است.



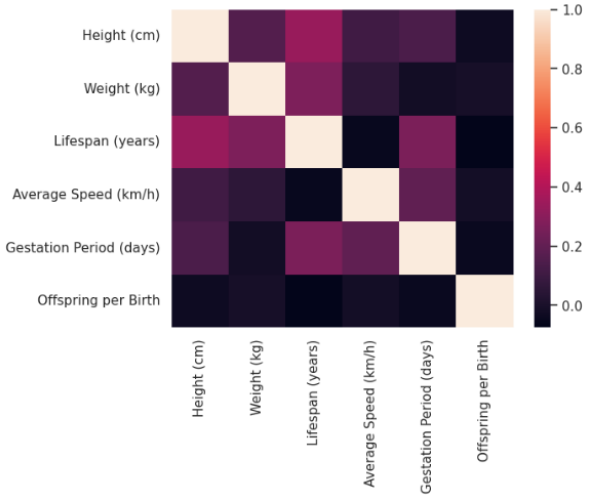
همچنین در نمودار زیر میتوانید انواع موجودات شکارچی را مشاده میکنیم. نکته قابل آن است که انسان خطرناک ترین موجود است و بعد از کفتار در رتبه دوم قرار دارد.



همچنین حیوانات موجود در دیتاست را براساس زیستگاه میتون به شکل زیر تصویر کرد. بیشتر موجودات داخل دیتاست متعلق به چمن زار ها و جنگل های بارانی هستند.



همبستگی میان ستون های عددی نیز به شرح زیر است. در این نمودار همبستگی بسیار بالایی دیده نمیشود.



در نهایت امتیاز آزمون خی ۲ برای ستون های مختلف نسبت به ستون برچست برحسب زیر است. عدد بالاتر نشان دهنده همبستگی بیشتر بین این دو ستون است.

Column	Cramér's V
Height (cm)	0.782
Weight (kg)	0.915
Color	0.702
Lifespan (years)	0.638
Habitat	0.715
Predators	0.808
Average Speed (km/h)	0.645
Countries Found	0.695
Conservation Status	0.228
Family	0.909
Gestation Period (days)	0.843
Social Structure	0.394
Offspring per Birth	0.546

مدل های پایه یادگیری ماشین

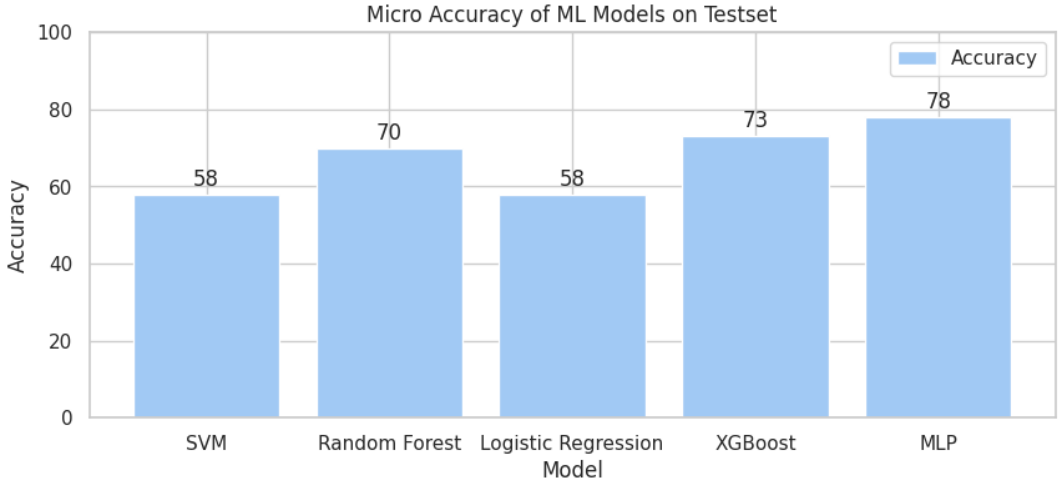
در این بخش از پنج مدل یادگیری ماشین مختلف برای پیش بینی برچسب داده ها استفاده شده است. همه این مدل ها روی یک مجموعه آموزشی و مجموعه تست آموزش و ارزیابی شده اند.

پیاده سازی

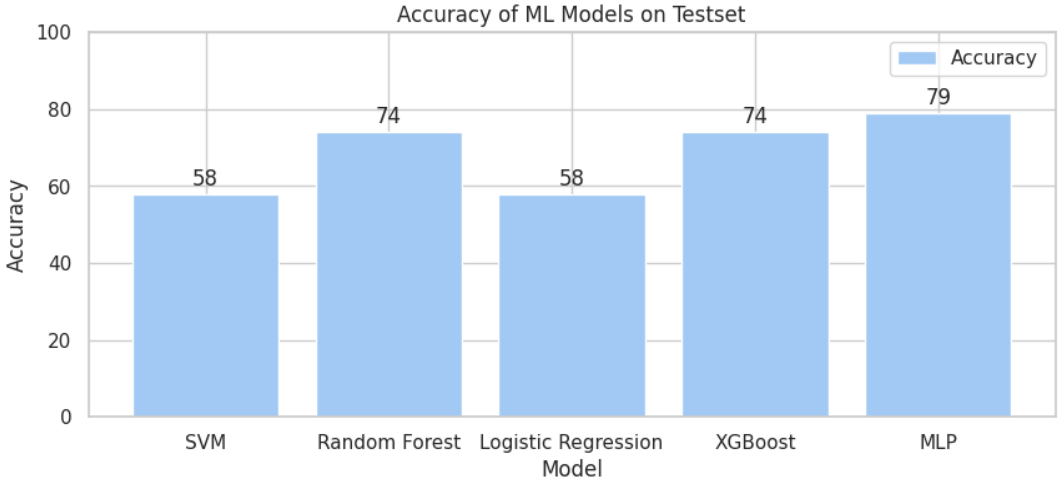
برای پیاده سازی این مدل های از کتابخانه `sklearn` استفاده شده است. همچنین ۱۰ درصد داده ها به عنوان تست در ابتدا کنار گذاشته شده است و سپس با استفاده از `k-fold` مجموعه به ۹ قسمت تقسیم شده است که همواره ۱۰ درصد کل مجموعه به عنوان داده `validation` استفاده شده است. دلیل استفاده از `k-fold` کوچیک بودن حجم دیتاست است که ارزیابی دقیق تری از مدل به ما میدهد. جزییات پیاده سازی هر مدل در فایل ژوپیتر موجود است.

نتایج

در نمودار زیر نتایج هر پنج مدل با یگدیگر مقایسه شده است. این ارزیابی با میار `micro accuracy` انجام شده است که بدلیل نامتوازن بودن دیتاست اولیه معبار معتبری است. همانطور که مشاهده میشود مدل چند لایه پرسپترون موفق شده بیشترین دقت را ثبت کنید.



همچنین در نمودار زیر نیز میتوانیم مدل ها را بر اساس معیار دقت مشاهده کنیم. همانطور که واضع است مدل `MLP` بهترین هملکرد را داشته است.

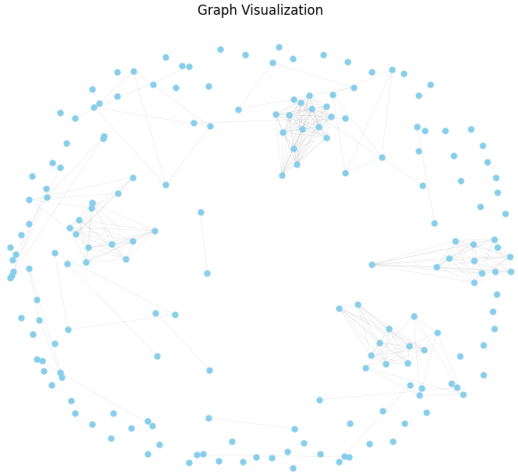


مدل یادگیری ماشین مبتنی بر گراف

در این بخش ابتدا دیتاست را به شکل گراف درمی اوریم و سپس با استفاده از یک مدل GCN داده ها را دسته بندی و نتایج را با دیتاست های گذشته مقایسه میکنیم.

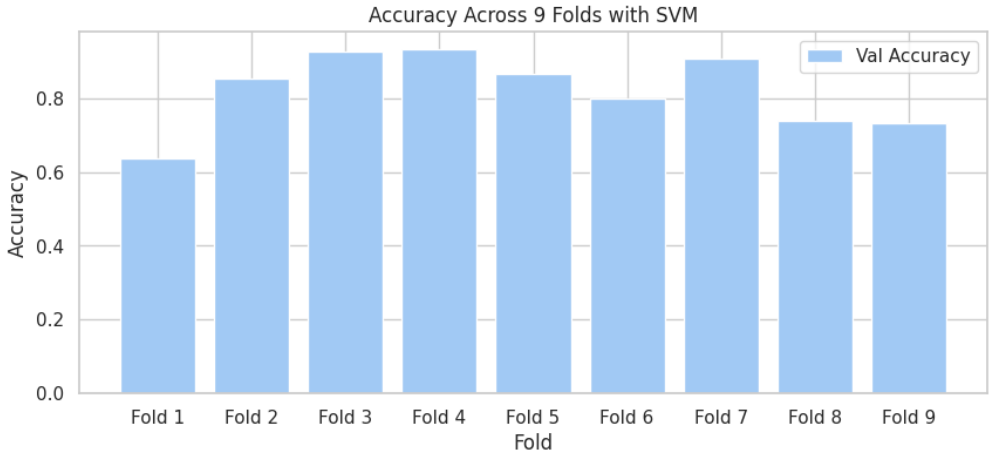
تشکیل گراف

برای تشکیل گراف نیازمند تعریف راس ها و یال ها هستیم. در این گراف راس ها حیوانات هستند و در صورت وجود رابطه فامیلی بین دو حیوان بین آن ها یا کشیده اینم. نتیجه آن گرافی با ۱۸۷ راس و ۲۶۶ یال شده است که در زیر مشاهده میشود. این گراف شامل تعدادی راس با درجه صفر است که دلیل آن کم بودن تعداد نمونه ها است. گفته میشود در طبیعت قریب به ۱.۴ میلیون گونه جانوری وجود دارد اما دیتاست ما صرفا شامل حدود ۲۰۰ گونه است به همین دلیل بسیاری از حیواناتی که با حیوانات درون دیتاست رابه فامیلی دارند در این مجموعه داده حضور ندارند. بنابراین یک از بهترین راه ها برای افزایش دقت میتوان همین موضوع یعنی اضافه کردم راس های جدید و در نتیجه ایجاد یال های بیشتر است.



مدل GCN

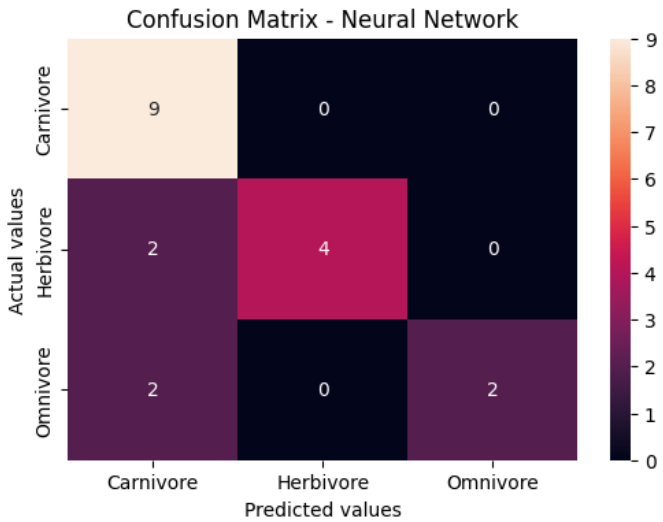
معماری مدل استفاده شده شامل سه لایه است. دو لایه ی GCNConv که در نهایت به یک لایه ی Linear می رسد. همچنین دراین شبکه از دو لایه Dropout استفاده شده است تا جلوی بیش برازشی را بگیرد. برای آموزش نیز به شکل دیگر مدل ها عمل کردیم و ۱۰ درصد داده را به عنوان تست و باقی داده ها به صورت 9-fold برای آموزش و ارزیابی مورد استفاده قرار گرفته اند. همچنین با کلاس توقف زود هنگام آموزش مدل را قبل از اینکه دچار بیش برازشی شود متوقف میکنیم. در نهایت دقت مدل بر روی فولد های مختلف به شرح زیر است.



همچنین دقت مدل بر روی داده های تست نیز به شکل زیر است.

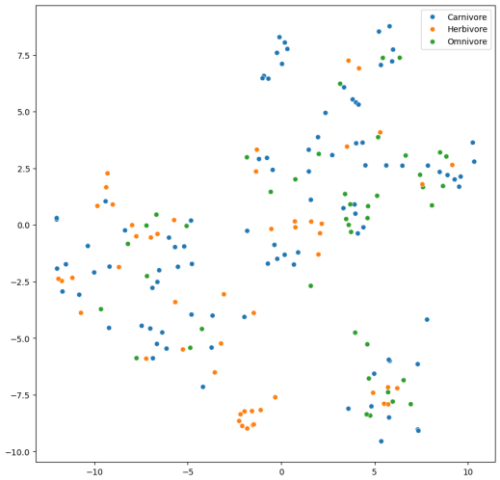
	precision	recall	f1-score	support
Carnivore	0.69	1.00	0.82	9
Herbivore	1.00	0.67	0.80	6
Omnivore	1.00	0.50	0.67	4
accuracy			0.79	19
macro avg	0.90	0.72	0.76	19
weighted avg	0.85	0.79	0.78	19

رسم نمودار Confusion Matrix نیز میتواند به ارزیابی بهتر مدل کمک کند.

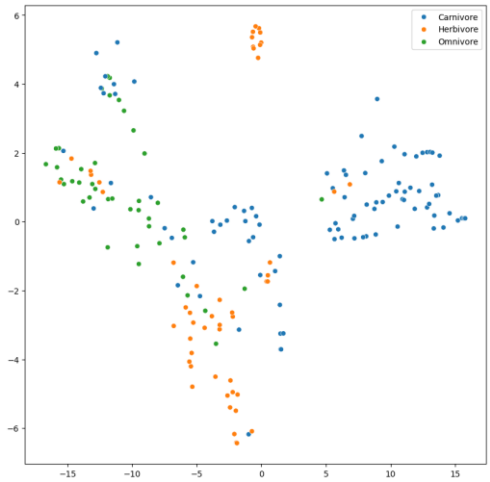


دو تصویر زیر با استفاده از کاهش ابعاد خروجی مدل به کمک t-SNE به دو بعد رسم شده است. یکی از این تصاویر قبل از آموزش مدل و دیگری بعد از آموزش مدل رسم شده است. همانطور که مشاهده میشود مدل توانسته است به خوبی سه گروه از یکدیگری جدا کند. نقطه های سبز نشان دهنده حیوانات همه چیز خوار است که نزدیک شدن بخشی از حیوات گوشت خوار و گیاهان خوار به آن ها جالب است.

تصویر مدل قبل از آموزش

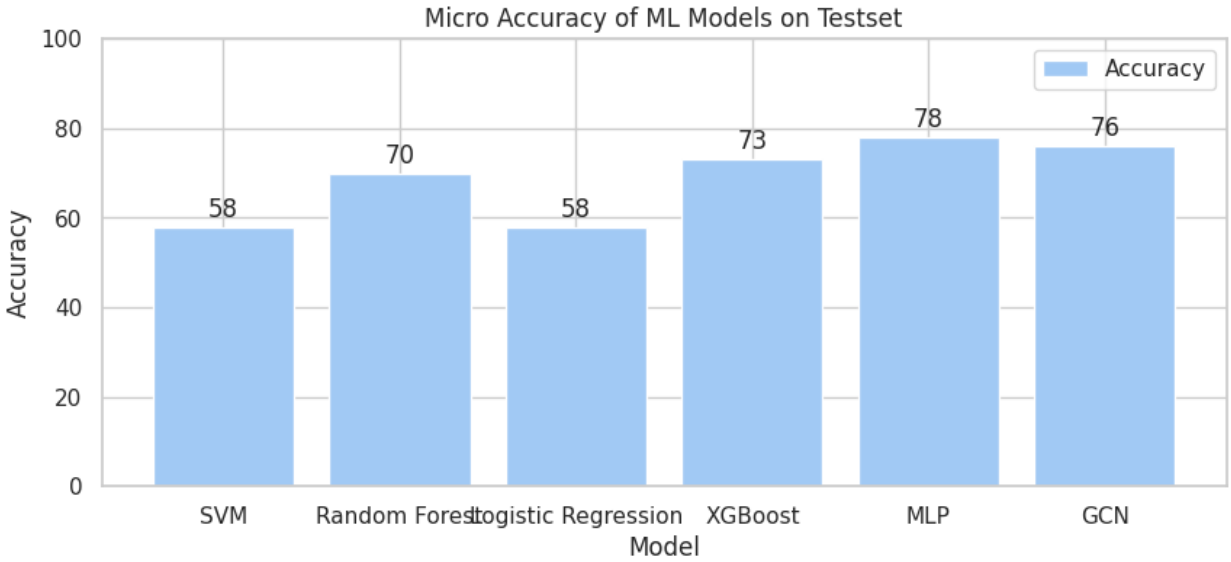


تصویر مدل بعد از آموزش



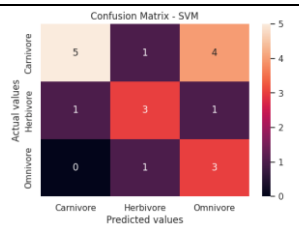
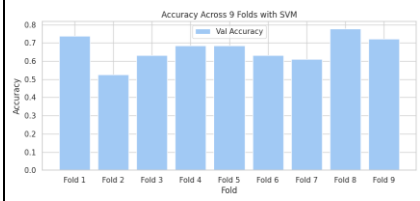
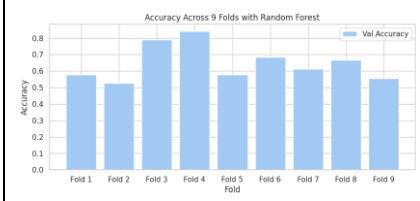
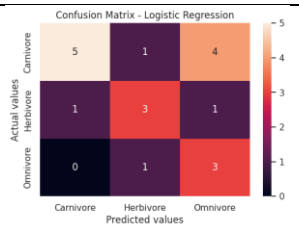
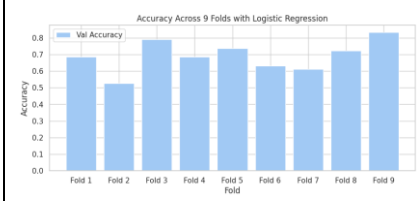
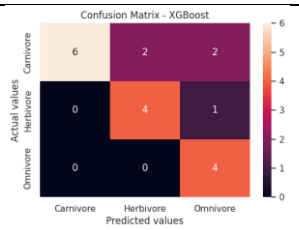
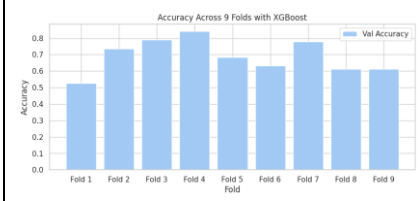
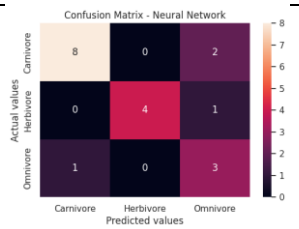
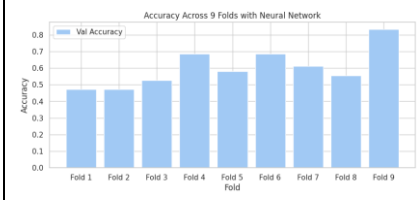
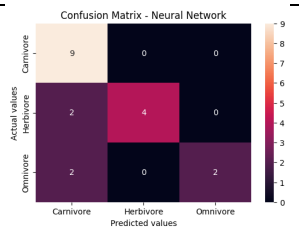
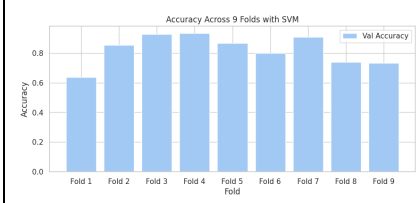
مقایسه با دیگر مدل ها

عملکرد مدل مبتنی بر گراف نسبت به دیگر مدل های خوب بوده است. در صورت افزایش تعداد راس ها و درنتیجه آن ایجاد روابط بهتر شاید بتوان بهتر حیوانات را از هم تفکیک کرد. نمودار زیر دقت مدل های مختلف را نسبت به هم نشان میدهد.



نتایج

در این پژوهش شش مدل یادگیری ماشین برای دسته بندی حیوات به سه گروه از نظر رژیم غذایی آن ها انجام شد. جدول زیر جزییات بسیار بیشتری از نتایج بدست آمده از هر مدل را در اختیار ما میگذارد. نتایج نشان دهنده عملکرد خوب مدل GCN بر روی داده ها است.

Model	Accuracy	Confusion Matrix	Accuracy Across 9 Folds Validation
SVM	precision recall f1-score support Carnivore 0.83 0.50 0.62 10 Herbivore 0.60 0.60 0.60 5 Omnivore 0.38 0.75 0.50 4 accuracy 0.58 19 macro avg 0.60 0.62 0.58 19 weighted avg 0.68 0.58 0.59 19		
Random Forest	precision recall f1-score support Carnivore 0.69 0.90 0.78 10 Herbivore 1.00 0.60 0.75 5 Omnivore 0.67 0.50 0.57 4 accuracy 0.74 19 macro avg 0.79 0.67 0.70 19 weighted avg 0.77 0.74 0.73 19		
Logistic Regression	precision recall f1-score support Carnivore 0.83 0.50 0.62 10 Herbivore 0.60 0.60 0.60 5 Omnivore 0.38 0.75 0.50 4 accuracy 0.58 19 macro avg 0.60 0.62 0.58 19 weighted avg 0.68 0.58 0.59 19		
XGBoost	precision recall f1-score support Carnivore 1.00 0.60 0.75 10 Herbivore 0.67 0.80 0.73 5 Omnivore 0.57 1.00 0.73 4 accuracy 0.74 19 macro avg 0.75 0.80 0.73 19 weighted avg 0.82 0.74 0.74 19		
MLP	precision recall f1-score support Carnivore 0.89 0.80 0.84 10 Herbivore 1.00 0.80 0.89 5 Omnivore 0.50 0.75 0.60 4 accuracy 0.79 19 macro avg 0.80 0.78 0.78 19 weighted avg 0.84 0.79 0.80 19		
GCN	precision recall f1-score support Carnivore 0.69 1.00 0.82 9 Herbivore 1.00 0.67 0.80 6 Omnivore 1.00 0.50 0.67 4 accuracy 0.79 19 macro avg 0.90 0.72 0.76 19 weighted avg 0.85 0.79 0.78 19