DSCI 403

Identifying Subglacial Lakes

Anastasia Horne, Spring 2024

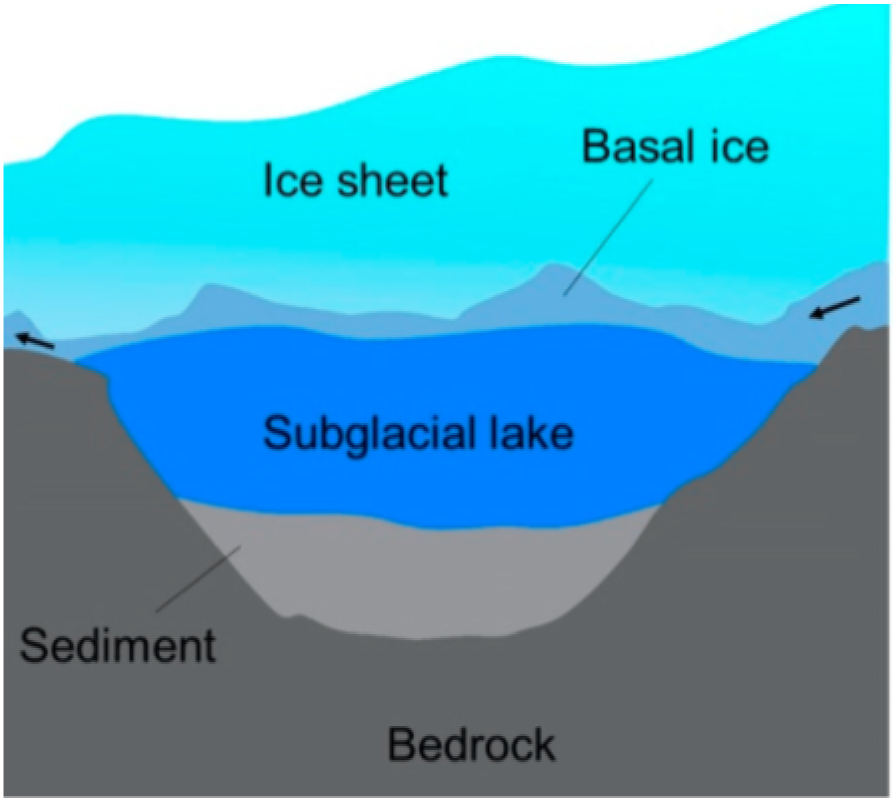


Figure 1. Diagram of Antarctic subglacial aquatic sediments (Gong et al., 2019)

Abstract

Subglacial lakes, hidden beneath the thick Antarctic ice sheet, play a crucial role in understanding the dynamics of the Antarctic ice sheet and potential impacts its melting has on global sea-level rise. Traditional methods for identifying subglacial lakes involve geophysical surveys which can be time-consuming, expensive, and in some cases impossible. This project presents an alternative approach to efficiently and accurately identify subglacial lakes using a machine learning algorithm.

Overview

I will be using BedMachine version 3 (Antarctica), a free, open source dataset from NASA's MEaSUREs program, which looks at properties of glaciers. The BedMachine dataset can be accessed using a free EarthAccess account. The data set provides a bed topography map of Antarctica along with firn air content, glacier/ice sheet thickness, ice surface elevation and other features. I am currently doing research in the Mines Glaciology LaborGatory, so I was hoping to work with a dataset that relates to my research that I haven't worked with before. This is an ambitious project as I am not sure what 'results' I can get from this dataset, however there is one particular question I am interested in answering. What hydrological features help predict the existence of a subglacial lake? The idea is to take a list of confirmed subglacial lakes (body of water under a glacier), and a list of non-lakes in Antarctica, and use a machine learning model to predict whether an area or point in Antarctica is a lake or not, given our BedMachine features.

Related Work

There has been limited research involving subglacial lakes. In a 2022 research paper that discusses subglacial lakes and their role in climate change, the researchers use one of the largest inventories of confirmed subglacial lakes and examines the crucial role the lakes play in providing climate history, supporting life, and influencing ice flow dynamics (Livingstone et al., 2022). This research looks at subglacial lakes, but does not actively look into how we can better identify them, which is what I hope to do.

Most of the research using BedMachine tries to create a better model for bed topography under specific glaciers. In a 2017 paper researchers used BedMachine data of Greenland to create a high-resolution bed topography map using ice thickness and bathymetry. From their bed topography, they were able to determine the sea level potential of the Greenland ice sheet, which provides future insights about the effect of global warming on sea level (Morlighem et al.). This research uses the BedMachine data of Greenland to create better bed topography maps, while my research will look at Antarctica and how the bed topography and accompanying parameters will influence the presence of subglacial lakes.

The research that relates most to the one I suggest is currently being conducted by PhD student Wilson Sauthoff at Colorado School of Mines, who is my research advisor in the Mines Glaciology Laboratory. The current phase of his research focuses on developing a machine learning model (and eventually a Bayesian inference network), that provides a comprehensive and highly accurate program to determine if an area contains a subglacial lake (Sauthoff). The research has produced a machine learning model that uses various Antarctica datasets, but does not use the entirety of BedMachine, which allows me to perform unique research. Additionaly, this project is an introduction into using BedMachine features in a machine learning algorithm and does not provide an exhaustive or an in-depth model to determine subglacial lakes.

Data Acquisition

BedMachine contains a bed topography map of Antarctica along with other features of interest: firn air content, surface, thickness, and x, y coordinates (see Figure A1). The bed topography map is values between 0 and 5000 meters which represent the elevation of the sediment or bedrock underneath the ice/glacier. Firn is the stage between snow and glacial ice, where our past-snow fall has been partially compacted, so the firn air content feature is the vertical height of our firn layer. Surface is the elevation of the ice surface in meters, which is the highest point of ice above the bedrock or sediment (see Figure 2).

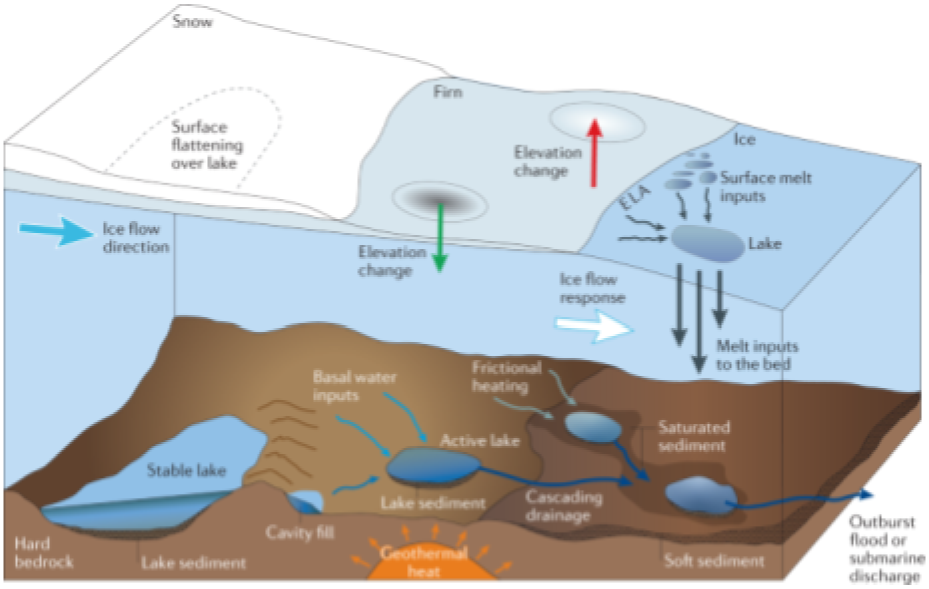


Figure 2. Diagram of Antarctica layers (Livingstone et al., 2022)

Thickness references the 'thickness' of our ice in meters. Ice thickness data was originally gathered from 47 flights using ice penetrating radar, however this data was sparse as the flights could not cover every inch of Antarctica. The dataset employs a mass conservation method that uses ice surface motion to extrapolate ice thickness data along each flight line and then interpolate and combine the maps to create one large-complete ice thickness map. Finally our x, y values are the Antarctic Polar Stereographic coordinates of our features. We will use these BedMachine features as the features in our model.

We still need our target data, which is a list of confirmed subglacial lakes and a list of confirmed non-subglacial lakes. The list of confirmed subglacial lakes is a csv file obtained from code written by my research advisor, Wilson Sauthoff, who contributed to a github repository that pulls lake outlines from the paper "Illuminating Active Subglacial Lake Processes With ICESat-2 Laser Altimetry" (Siegfried & Fricker 2021). The paper contains an extensive list of confirmed subglacial lakes discovered in different projects and papers. I imported directly from a csv instead of importing the code from the github because it is cleaner and is better suited for the scope of this project. The dataset contains the areas, represented by polygons in Antarctic Polar

Stereographic coordinates, of confirmed subglacial lakes along with the their names and citations that inform us who discovered them, and where they were published (see Figure A2).

Similarly, for the list of non-subglacial lakes I obtained a csv file from code written by Wilson Sauthoff. The csv contains Antarctic Polar Stereographic POINT(x, y) coordinates of places we know there is not a subglacial lake, the type of lake (ie. ‘nonlake’), and the sampling name of the nonlake (see Figure A3). These coordinates were created by taking a simple random sample of Antarctica with areas of confirmed and suspected lakes removed.

Finally, there is one more dataset we need. There are two primary types of ice in Antarctica, land ice and sea ice, both of which behave differently from the other. Subglacial lakes are only found under land ice, however BedMachine contains values from both land and sea ice, so we need to limit or clip our BedMachine values to only contain land ice. We do this by clipping to the grounding line. The grounding line is where land ice stops and sea ice begins, we can think of it as the coastal line of Antarctica. The data can be accessed from the Scripps Institution of Oceanography (Depoorter et al.).

Data Preprocessing

We have all of the necessary datasets in either an xarray or geopandas dataframe. What we don’t have is a singular dataset that we can split and use within the model. So we need to edit each dataset individually in order to combine them. Starting with the lake geopandas dataframe, we need to remove all features except our ‘geometry’ column that holds the location of our lakes, and we need to add a type column representing that the polygons are lakes (ie. ‘lake’). We can thus clip our lake dataframe to our bedMachine xarray and view the areas of our confirmed subglacial lakes in relation to ice surface elevation.

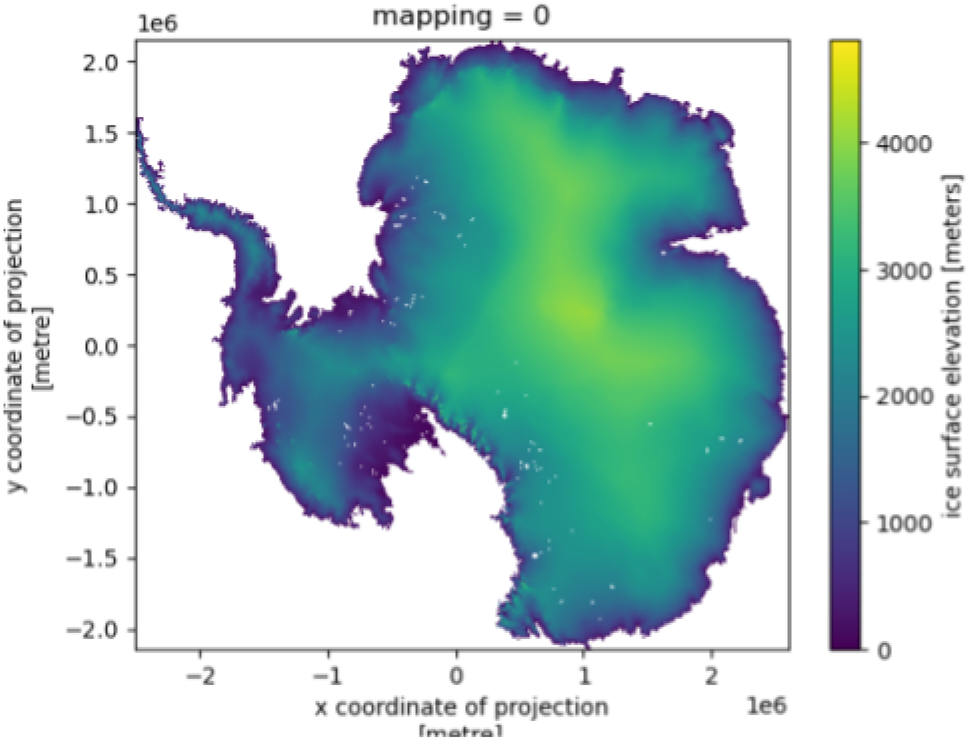

mapping = 0
1e6
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
y coordinate of projection [metre]
−2
−1
0
1
2
x coordinate of projection [metre]
1e6
4000
3000
2000
1000
0
ice surface elevation [meters]

Figure 3. Ice Surface Elevation values, excluding subglacial lakes.

We see that the areas within the Antarctica continent in white represent our confirmed subglacial lakes, and these areas tend to be closer to the coast, where our ice elevation is less (see Figure 3).

We can perform the same process for our nonlakes dataframe. The only feature we need to remove is the name feature, and then we can view our nonlakes in relation to ice surface via the red dots.

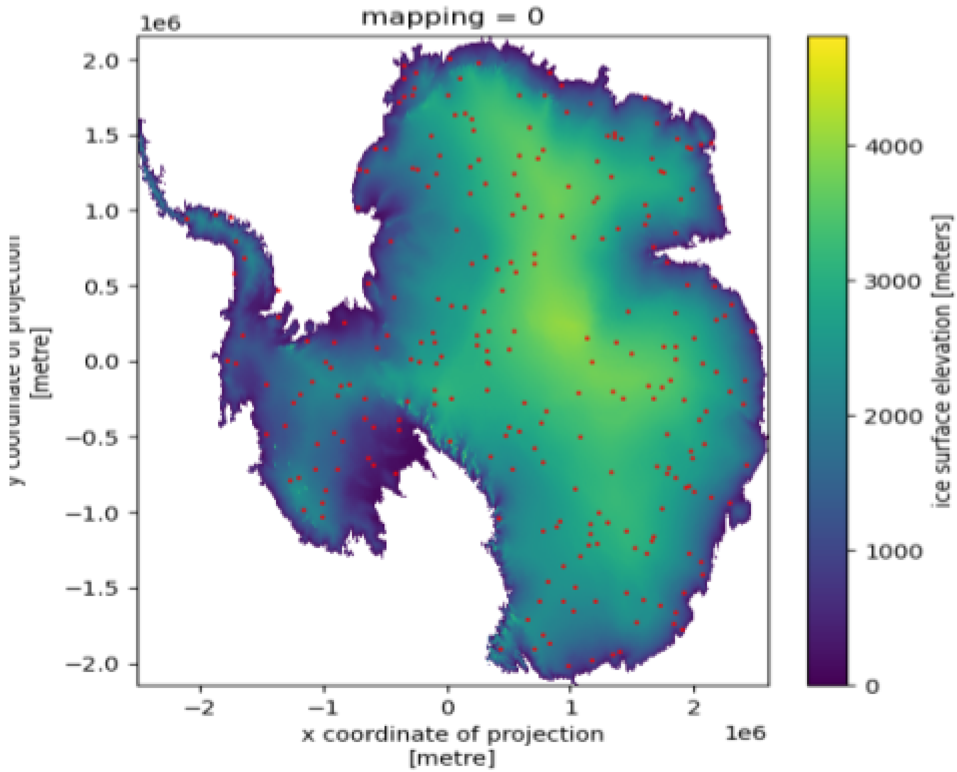


Figure 4. Ice Surface Elevation values, excluding non-subglacial lakes.

We see that these points are randomly dispersed, as we hoped, and the points represent both high and low values of ice surface elevation. We have some dots that are very close to subglacial lakes and some dots that are very far away (see Figure 4). It is important to capture the behavior of all nonlakes throughout the entire continent so we can understand the range of values for nonlakes and generalize our model in the future to classify areas we believe subglacial lakes exist.

Now we can combine our datasets. First we concatenate our lake and nonlake geopandas dataframe, creating a singular list of target objects. To combine the BedMachine objects (rows in the dataset) with their corresponding lake/nonlake type it is necessary to loop through our combined lakes dataframe and find the closest bedMachine row based on the geometry column. Then we can select the features of interest and append that to a final dataframe. We now have a singular dataframe with our feature and target values (see Figure A4).

Looking at the feature values, each individual feature has a different range of values (see Table 1).

Feature	Minimum Value	Maximum Value
Firn	0	38.0924
Ice Surface Elevation	68.6411	3967.4246
Ice Thickness	0	4126.224
Bed Elevation	-2394.0723	2527.3254

Table 1. Range for features of interest

We need to scale the data to balance the impact of all variables. This is vital if the model we use is a distance-based algorithm like K-Neighbors. After applying a standard scaler transformation to our datapoints, we see the ranges of each feature is more uniform (see Table 2).

Feature	Minimum Value	Maximum Value
Firn	-3.38	2.3346
Ice Surface Elevation	-1.8365	1.9857
Ice Thickness	-2.3344	2.0721
Bed Elevation	-2.8013	3.6137

Table 2. Scaled range for features of interest

Now we can compare our different features to each other and to our target value. Bed elevation and ice surface elevation seem like they might be correlated based on their definitions. Furthermore, if bed elevation rises, then the peak height of the ice would seemingly be higher. Plotting these two features we see a positive linear relationship between bed elevation and ice surface elevation (see Figure 5).

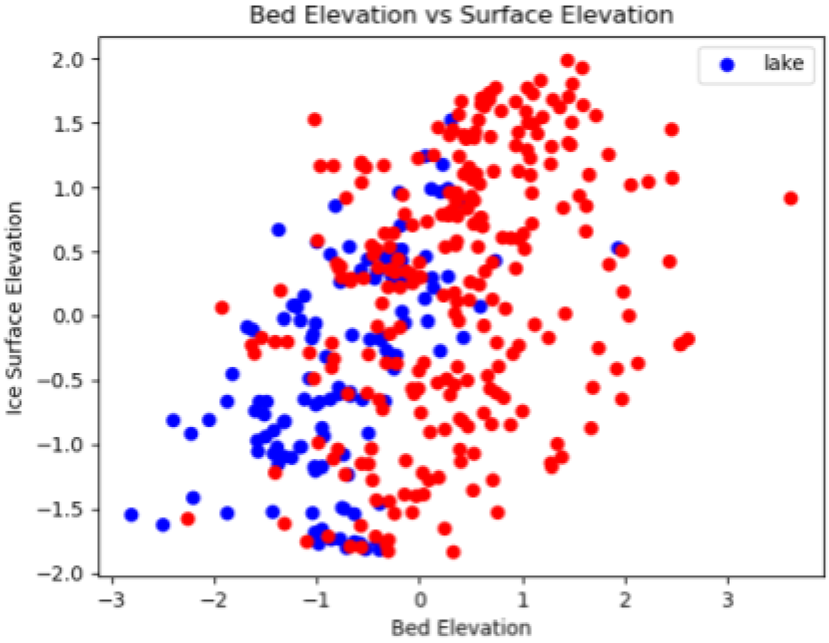


Figure 5.

From the graph, we see that the blue dots (lakes) and red dots (nonlakes) have alot of overlap, but we also see that the lake points tend towards a steeper linear relationship and have more negative bed elevation values. Meanwhile nonlake points have more points at a positive bed elevation and a more moderate linear relationship. It is possible that bed elevation and ice surface elevation play a role in the existence and thus classification of subglacial lakes, but it is also clear there is enough overlap that using only these two features would result in an inaccurate model.

Next we compare our firn values to our target values by plotting two horizontal lines, each representing a lake (0 value) or nonlake (1 value).

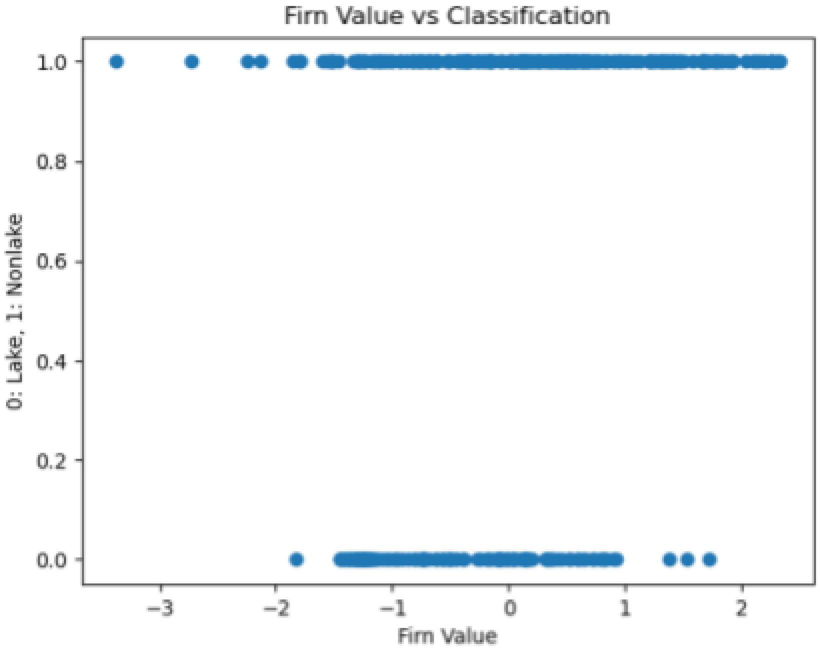


Figure 6.

We see that there is still overlap between lakes and nonlakes, however the range of values for nonlakes is larger than the range for lakes (see Figure 6). Thus firn value might be an indicator of a subglacial lake, but it is clear that firn value alone cannot accurately classify a lake.

Now we will compare our three features that seem to be heavily correlated, ice surface elevation, ice thickness and bed elevation (see Figure 7).

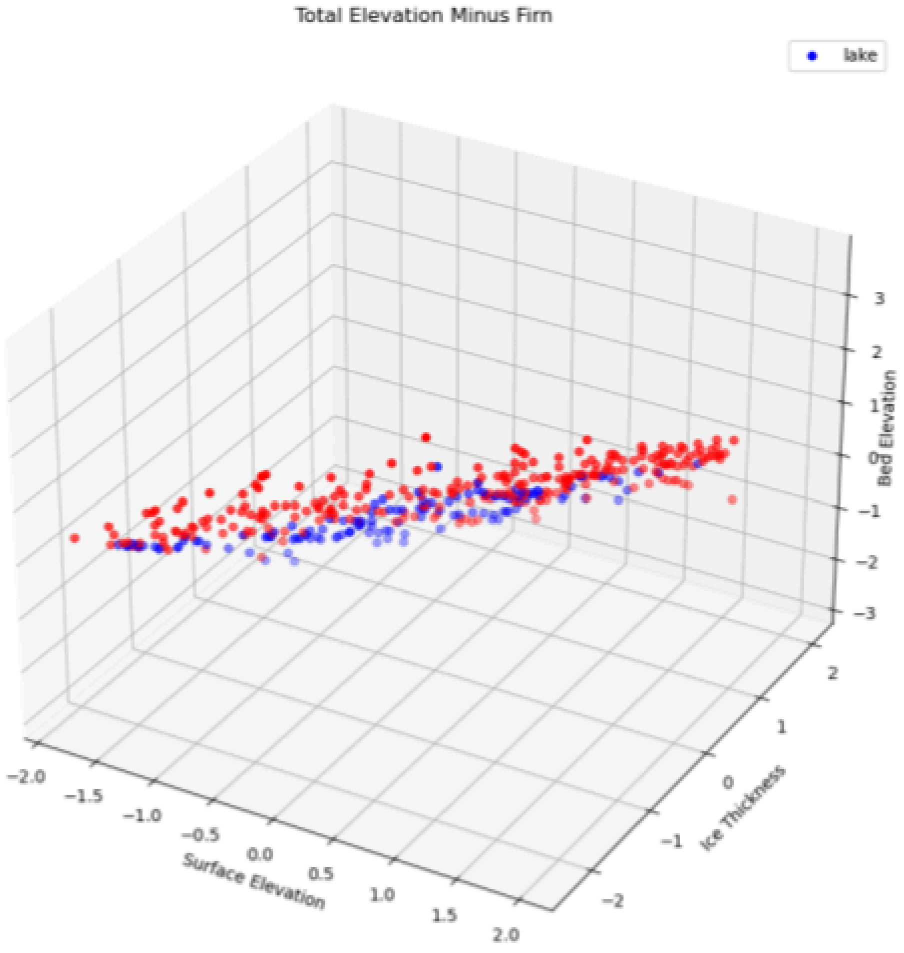


Figure 7.

There is some overlap, but similar to Figure 5, we see a separation between our lake and nonlake points. Furthermore the points form a seemingly straightline suggesting a linear relationship between our features.

All the features seem to slightly separate our lake and nonlake points, with some overlap. Thus it is necessary to use all of them when determining which machine learning model is best at classifying subglacial lakes.

Model Selection

The goal of any of our models is to accurately predict whether a point is a subglacial lake or not. We are looking at classifying our objects, thus we will use a supervised classification machine learning algorithm. We will look at three different classification models, K-Neighbors Classifier, Support-Vector Machines (SVM), and Logistic Regression (LR). Additionally, our data was initially separated so 30% is removed for testing, and 70% is reserved for training the model.

The KNeighborsClassifer was trained with k=4 because it provided the highest accuracy while not overfitting the data. The SVM model was trained with kernel=’rbf’ due to its ability to capture complex decision boundaries. The results of each model are below.

Model	Precision	Recall	f1-score	Accuracy	Support
KNeighbors	0.69	0.80	0.74	0.78814	44
SVM	0.68	0.43	0.53	0.71186	44
LR	0.75	0.55	0.63	0.76271	44

Table 3. Classification Reports for target=’lake’

Model	Precision	Recall	f1-score	Accuracy	Support
KNeighbors	0.87	0.78	0.82	0.78814	74
SVM	0.72	0.88	0.79	0.71186	74
LR	0.77	0.89	0.82	0.76271	74

Table 4. Classification Reports for target=’nonlake’

Since the overarching goal of this project is to expand the current inventory of confirmed subglacial lakes, this model should have a larger recall score for the 'lake' target. We see that the K-Neighbors classifier has the highest recall for 'lakes' while still having a reasonable recall score for 'nonlakes'. It is important to note that while SVM and LR have decent accuracy scores, their low 'lake' recall scores mean there is a high proportion of false negatives, thus we will use K-Neighbors Classifier as our machine learning algorithm.

Results and Evaluation

The K-Neighbors Classifier performs best for our model compared to the other model types, but how well does it truly perform? Referencing Tables 3 and 4 and our confusion matrix (see Table 5), we see the following results for precision, recall and F1-Score.

Table 5. Confusion Matrix for KNeighbors Model

		Predicted Label	
		Lake	Nonlake
True Label	Lake	35	9
	Nonlake	16	58

Precision measures our models ability to not have false positive predictions. For example, the closer the precision value is to 1, the less false positives we have. For our model, the 'lake' precision score is fairly low at 0.69 because we had 16 false positives, meanwhile the 'nonlake' precision score is very high at 0.87, meaning we have less false positives for nonlakes.

Recall measures our models ability to find all the true positive objects in our dataset. The closer the value to 1, the less false negatives we have. For our model the 'lake' recall score is high at 0.80, meaning we did not have many false negatives (9 to be exact). The 'nonlake' recall score was lower and there were 16 false negatives for nonlakes.

F1-Score combines our precision and recall values. The higher our value, the more balance between recall and precision. A high F1-score tells us that we have a model that weights false positives and false negatives similarily. For our model the F1-score of nonlakes is better than the score of lakes. This shows that for our model the balance between false negatives and false positives was better for nonlakes. However, the lake F1-score is still at an acceptable value.

Overall, there are apparent shortcomings in the KNeighbors model, and while the accuracy score of 0.78814 is acceptable for the scope of this project and better than the other models, it is not a good score if the next step is to apply the model to a list of points that are suspected subglacial lakes.

Ethics

My project does not have any clear ethical or societal impacts because it is located in an area of the world where very few humans reside, and those that do live there are not permanent residents, instead they are mainly researchers and military personnel. However, there are two main impacts we can consider The impact subglacial lakes have on climate change, and what we can do with knowing where these subglacial lakes are.

If the glaciers in Antarctica and Greenland were to all melt, sea level would rise by approximately 210 feet. This would cause a large increase in coastal erosion and an increase in storm surges, as warming air and water temperatures contribute to more frequent storms, such as hurricanes. So understanding any and all causes of glacial melt is very important. The faster glaciers flow/move the more frictional heating we have and this leads to more ice melt. The flow of glaciers is extremely dependent on their subglacial hydrology, including subglacial lakes. So if we know exactly where subglacial lakes are, and what geophysical characteristics and features represents them, we can better track them and potentially visit them and collect other valuable data. While this information would not help slow or fix climate change, what it can do is allow us to better track how quickly ice is melting, and what climate change is doing to the ice sheets.

Additionally, because places like Antarcitica are so cold, they are able to preserve information about the past very well. Subglacial lakes contain information on ancient climate, and ancient microorganisms who have been able to survive for over thousands of years. Similar to how ice cores provide information on how greenhouse gasses have changed throughout the past, we can look at the chemical composition of the water in these lakes to gain vital information on atmospheric conditions.

While neither of these considerations directly impact our daily lives, they can provide scientists with vital information on how climate change is impacting the world, and what that might mean for our future.

Future Work

The ice-hydrology of Antarctica is very complex, with many features that go far beyond what we have used in this model. A valuable next step would be to find these other features and incorporate them into a future model. Another possible step would be to explore more complex classification algorithms such as random forests or neural networks. KNN classifies an unclassified object by determining its closest classified neighbor who shares similar feature values, whereas neural networks utilize interconnected and deep layers to identify patterns between features and then classify our objects. Thus a neural network might be able to better analyze the complex connections between our features.

There are two types of subglacial lakes, active and inactive. Active subglacial lakes drain and fill continuously, contributing to short and long-term glacier flow dynamics. Inactive lakes remain stagnant, not draining or filling and thus behave very differently from active lakes. Thus a next step could be to take our classifed subglacial lakes and further break them down into active and inactive subglacial lakes. We would probably need more data to determine this, specifically one showing ice height changes.

References:

Depoorter, Mathieu A; Bamber, Jonathan L; Griggs, Jennifer; Lenaerts, Jan T M; Ligtenberg, Stefan R M; van den Broeke, Michiel R; Moholdt, Geir (2013): Antarctic masks (ice-shelves, ice-sheet, and islands), link to shape file. PANGAEA, https://doi.org/10.1594/PANGAEA.819147,

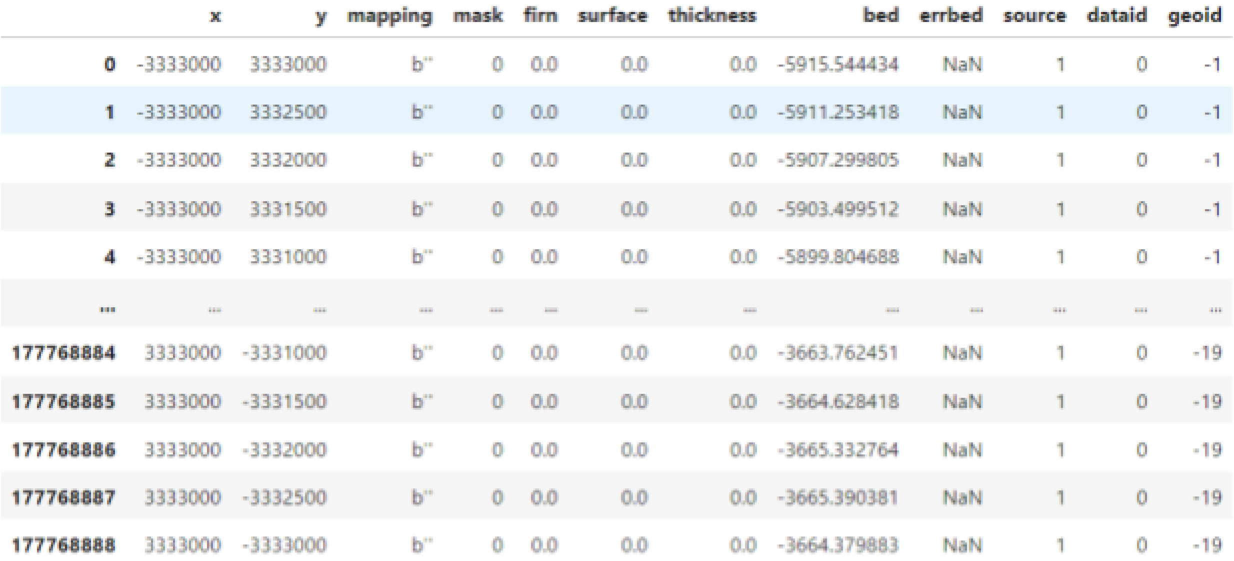
Livingstone, S.J., Li, Y., Rutishauser, A. et al. Subglacial lakes and their changing role in a warming climate. Nat Rev Earth Environ 3, 106–124 (2022). https://doi.org/10.1038/s43017-021-00246-9

Morlighem, M., Williams, C. N., Rignot, E., An, L., Arndt, J. E., Bamber, J. L., ... Zinglersen, K. B. (2017). BedMachine v3: Complete bed topography and ocean bathymetry mapping of Greenland from multibeam echo sounding combined with mass conservation. Geophysical Research Letters, 44, 11,051–11,061. https://doi.org/10.1002/2017GL074954

Sauthoff, Wilson. “Multi-mission altimetry of variable Antarctic active subglacial lake geometries and causal inference networks to inform geostatistics-based subglacial water modeling”. 2022.

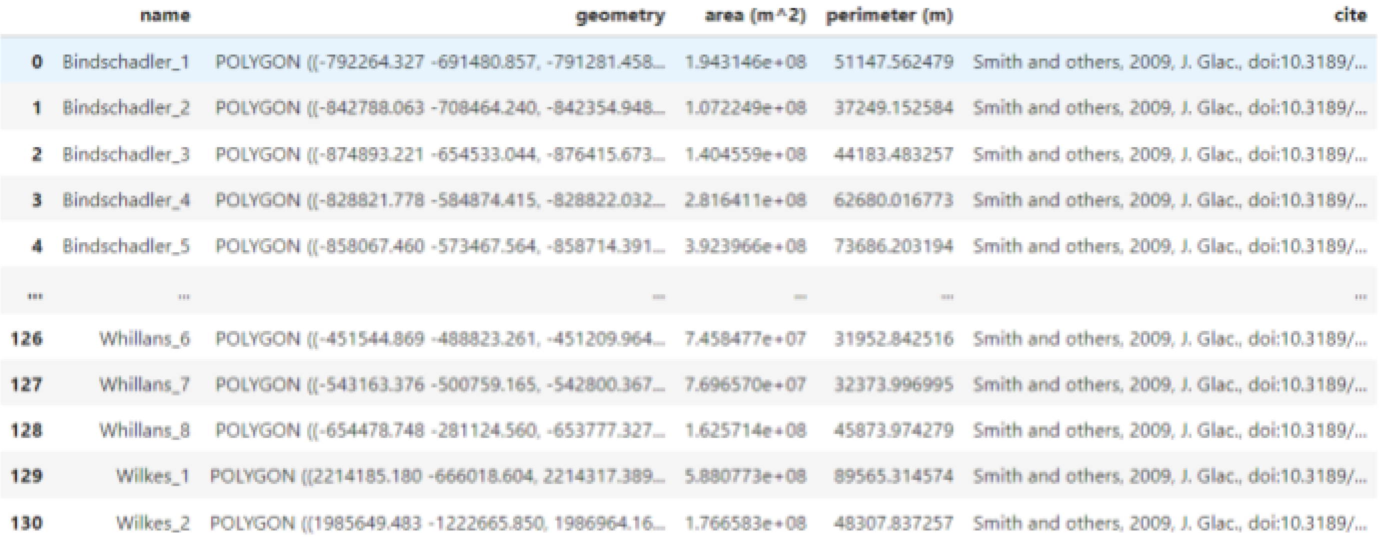
Gong D, Fan X, Li Y, Li B, Zhang N, Gromig R, Smith EC, Dummann W, Berger S, Eisen O, et al. Coring of Antarctic Subglacial Sediments. *Journal of Marine Science and Engineering*. 2019; 7(6):194. https://doi.org/10.3390/jmse7060194

Appendix A



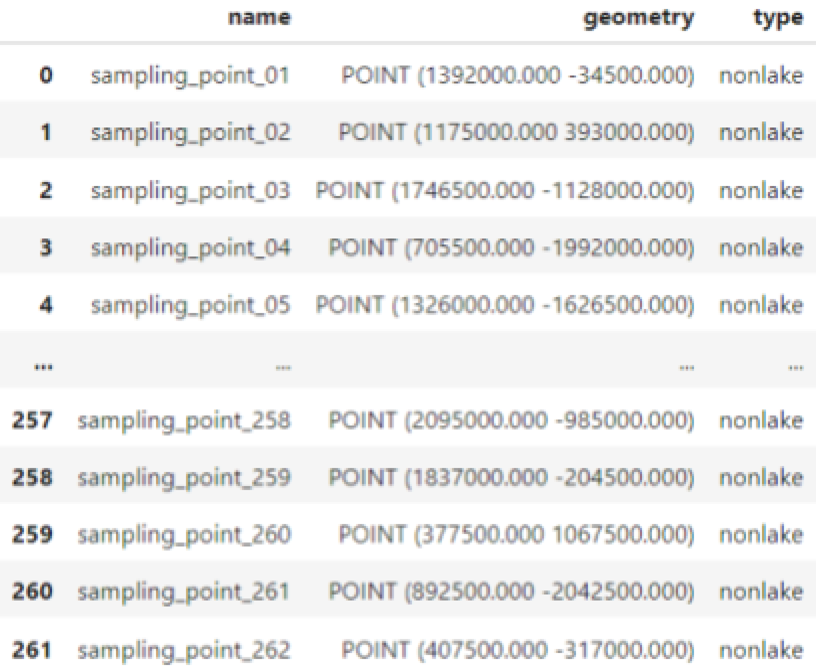
	x	y	mapping	mask	firn	surface	thickness	bed	errbed	source	dataid	geoid
0	-3333000	3333000	b''	0	0.0	0.0	0.0	-5915.544434	NaN	1	0	-1
1	-3333000	3332500	b''	0	0.0	0.0	0.0	-5911.253418	NaN	1	0	-1
2	-3333000	3332000	b''	0	0.0	0.0	0.0	-5907.299805	NaN	1	0	-1
3	-3333000	3331500	b''	0	0.0	0.0	0.0	-5903.499512	NaN	1	0	-1
4	-3333000	3331000	b''	0	0.0	0.0	0.0	-5899.804688	NaN	1	0	-1
...	...	...	...	...	...	...	...	...	...	...	...	...
177768884	3333000	-3331000	b''	0	0.0	0.0	0.0	-3663.762451	NaN	1	0	-19
177768885	3333000	-3331500	b''	0	0.0	0.0	0.0	-3664.628418	NaN	1	0	-19
177768886	3333000	-3332000	b''	0	0.0	0.0	0.0	-3665.332764	NaN	1	0	-19
177768887	3333000	-3332500	b''	0	0.0	0.0	0.0	-3665.390381	NaN	1	0	-19
177768888	3333000	-3333000	b''	0	0.0	0.0	0.0	-3664.379883	NaN	1	0	-19

Figure A1. Output of reading in BedMachine into a Xarray Dataset.



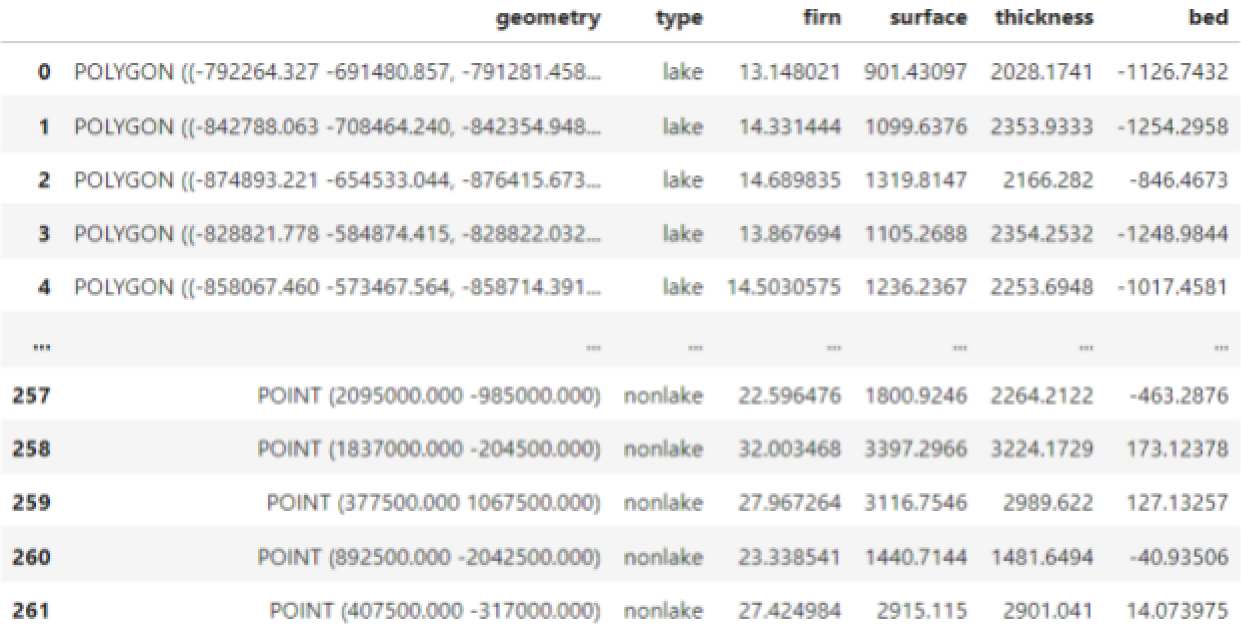
	name	geometry	area (m^2)	perimeter (m)	cite
0	Bindschadler_1	POLYGON ((-792264.327 -691480.857, -791281.458...	1.943146e+08	51147.562479	Smith and others, 2009, J. Glac., doi:10.3189/...
1	Bindschadler_2	POLYGON ((-842788.063 -708464.240, -842354.948...	1.072249e+08	37249.152584	Smith and others, 2009, J. Glac., doi:10.3189/...
2	Bindschadler_3	POLYGON ((-874893.221 -654533.044, -876415.673...	1.404559e+08	44183.483257	Smith and others, 2009, J. Glac., doi:10.3189/...
3	Bindschadler_4	POLYGON ((-828821.778 -584874.415, -828822.032...	2.816411e+08	62680.016773	Smith and others, 2009, J. Glac., doi:10.3189/...
4	Bindschadler_5	POLYGON ((-858067.460 -573467.564, -858714.391...	3.923966e+08	73686.203194	Smith and others, 2009, J. Glac., doi:10.3189/...
...	...	...	...	...	...
126	Whillans_6	POLYGON ((-451544.869 -488823.261, -451209.964...	7.458477e+07	31952.842516	Smith and others, 2009, J. Glac., doi:10.3189/...
127	Whillans_7	POLYGON ((-543163.376 -500759.165, -542800.367...	7.696570e+07	32373.996995	Smith and others, 2009, J. Glac., doi:10.3189/...
128	Whillans_8	POLYGON ((-654478.748 -281124.560, -653777.327...	1.625714e+08	45873.974279	Smith and others, 2009, J. Glac., doi:10.3189/...
129	Wilkes_1	POLYGON ((2214185.180 -666018.604, 2214317.389...	5.880773e+08	89565.314574	Smith and others, 2009, J. Glac., doi:10.3189/...
130	Wilkes_2	POLYGON ((1985649.483 -1222665.850, 1986964.16...	1.766583e+08	48307.837257	Smith and others, 2009, J. Glac., doi:10.3189/...

Figure A2. Output of reading outlines.csv into a geopandas dataframe.



	name	geometry	type
0	sampling_point_01	POINT (1392000.000 -34500.000)	nonlake
1	sampling_point_02	POINT (1175000.000 393000.000)	nonlake
2	sampling_point_03	POINT (1746500.000 -1128000.000)	nonlake
3	sampling_point_04	POINT (705500.000 -1992000.000)	nonlake
4	sampling_point_05	POINT (1326000.000 -1626500.000)	nonlake
...	...	...	...
257	sampling_point_258	POINT (2095000.000 -985000.000)	nonlake
258	sampling_point_259	POINT (1837000.000 -204500.000)	nonlake
259	sampling_point_260	POINT (377500.000 1067500.000)	nonlake
260	sampling_point_261	POINT (892500.000 -2042500.000)	nonlake
261	sampling_point_262	POINT (407500.000 -317000.000)	nonlake

Figure A3. Output of reading in nonlakePTS.csv into a geopandas dataframe



	geometry	type	firn	surface	thickness	bed
0	POLYGON ((-792264.327 -691480.857, -791281.458...	lake	13.148021	901.43097	2028.1741	-1126.7432
1	POLYGON ((-842788.063 -708464.240, -842354.948...	lake	14.331444	1099.6376	2353.9333	-1254.2958
2	POLYGON ((-874893.221 -654533.044, -876415.673...	lake	14.689835	1319.8147	2166.282	-846.4673
3	POLYGON ((-828821.778 -584874.415, -828822.032...	lake	13.867694	1105.2688	2354.2532	-1248.9844
4	POLYGON ((-858067.460 -573467.564, -858714.391...	lake	14.5030575	1236.2367	2253.6948	-1017.4581
...	...	...	...	...	...	...
257	POINT (2095000.000 -985000.000)	nonlake	22.596476	1800.9246	2264.2122	-463.2876
258	POINT (1837000.000 -204500.000)	nonlake	32.003468	3397.2966	3224.1729	173.12378
259	POINT (377500.000 1067500.000)	nonlake	27.967264	3116.7546	2989.622	127.13257
260	POINT (892500.000 -2042500.000)	nonlake	23.338541	1440.7144	1481.6494	-40.93506
261	POINT (407500.000 -317000.000)	nonlake	27.424984	2915.115	2901.041	14.073975

Figure A4. Output of our combined dataset, with feature and target data.