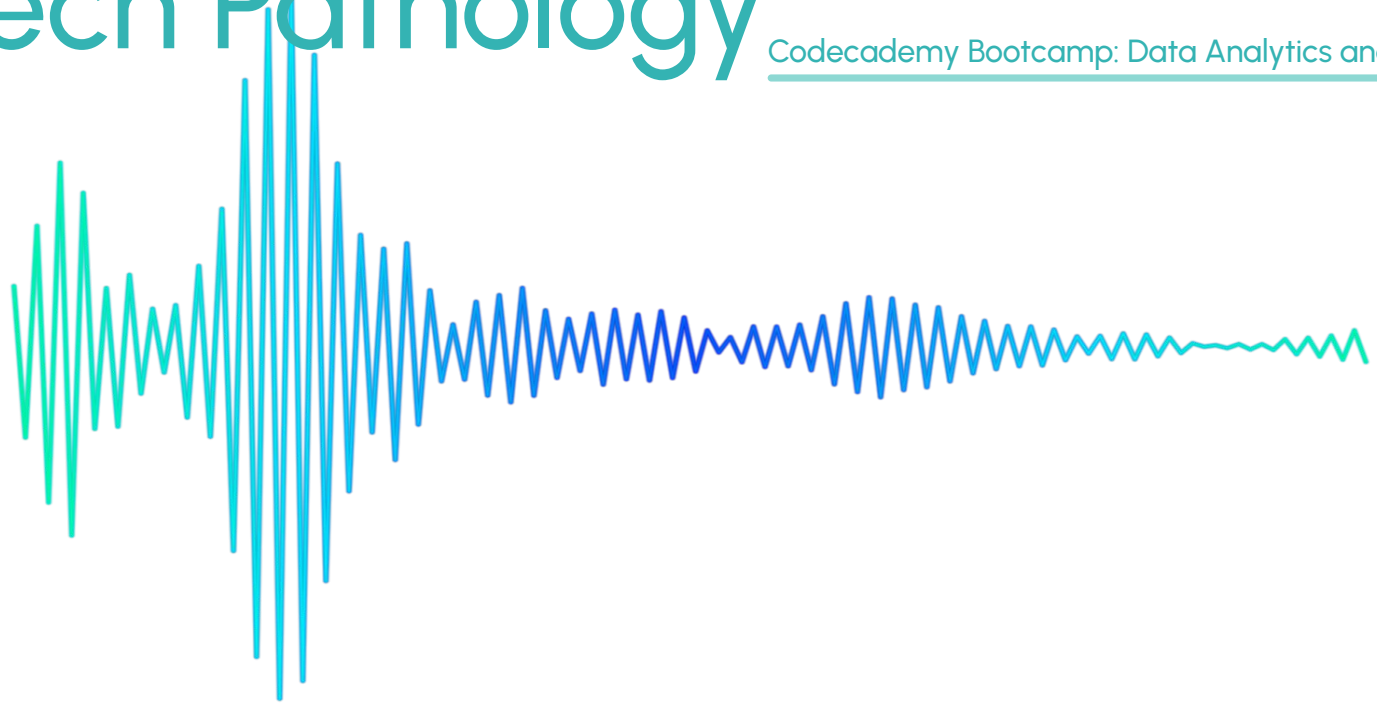
Capstone Project: Speech Pathology

Codecademy Bootcamp: Data Analytics and AI



Marian Pattugalan

Contents:

Project Overview:

In this capstone project, you will analyze a dataset containing audio-derived features (e.g., Mel filters, Fourier transformation) extracted from voice recordings of patients labeled as "Healthy" or "Unhealthy." The goal is to perform exploratory data analysis (EDA), clean the data, visualize insights in a Power BI dashboard, and build a predictive model to classify patients based on their health status.

Data Overview:

	filename	chroma_st	rmse	spectral_c	spectral_ba	rolloff	zero_cross	mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	mfcc8	mfcc9
0	1-a_h.wav	0.127341	0.355538	1587.507	2002.199	2737.972	0.066798	-233.566	48.50442	-34.5181	-36.2723	-42.9944	-34.279	23.74783	-11.7224	-32.8378
1	10-a_h.wav	0.215317	0.250342	1272.901	1850.303	1692.92	0.043325	-179.42	119.7057	0.882231	-25.1676	-11.8211	-6.27188	-14.576	31.81473	-12.769
2	100-a_h.wa	0.212512	0.251419	919.2038	1014.718	1275.967	0.047079	-209.186	167.1357	-37.3982	-12.4357	-13.8215	-1.58407	-2.15113	9.81761	0.882222
3	1000-a_h.v	0.198861	0.157396	1119.517	1529.491	1501.658	0.045487	-210.244	139.8195	-7.77401	13.63091	-52.2359	-10.8682	-6.9796	-7.40172	13.60316
4	1002-a_h.v	0.370113	0.143538	1060.678	1346.3	1524.942	0.050109	-175.04	181.9052	-33.1816	-0.09176	-12.4857	-26.2217	12.85998	9.708848	-12.7372
5	1003-a_h.v	0.232899	0.215562	873.0393	1041.309	1112.077	0.046657	-221.869	182.165	-18.4184	-24.692	-36.0539	-13.127	11.16934	6.71707	-2.89413
6	1004-a_h.v	0.331815	0.231105	1077.814	1303.157	1360.275	0.045667	-145.841	174.8312	-33.6699	-10.9282	-24.0189	-21.1933	18.40681	16.20833	-26.3515
7	1005-a_h.v	0.243202	0.21482	873.3272	1282.35	1068.378	0.040371	-189.152	186.3535	12.51912	6.089208	-36.3086	-10.7136	-0.00256	0.480074	-3.63637
8	1006-a_h.v	0.193382	0.18433	1066.099	1705.306	1183.004	0.043123	-256.187	127.9539	13.82222	-31.0976	-28.4454	12.86928	20.51917	13.28549	-13.7197
9	1007-a_h.v	0.226716	0.203022	1085.429	1284.242	1526.538	0.043089	-206.332	131.8645	-37.6106	-0.29648	-38.6851	-4.28765	13.09653	-15.7204	-6.38196
10	1008-a_h.v	0.191215	0.17789	620.3114	1130.468	707.3374	0.023662	-335.177	150.2773	26.24766	5.386874	-16.8598	-21.1406	14.13922	2.935519	-14.9148
11	1009-a_h.v	0.184429	0.217275	1180.663	1647.716	1362.872	0.053874	-179.649	135.8172	0.320124	-12.2397	-44.1717	-9.91046	-1.58335	11.41033	-28.1011
12	1010-a_h.v	0.210474	0.247952	845.9512	1019.606	1016.547	0.055105	-232.805	162.8062	-0.26342	1.913021	-56.6432	-14.0908	-16.8168	1.200328	13.09997
13	1011-a_h.v	0.273833	0.12768	993.7004	1137.781	1330.034	0.049121	-225.825	178.1395	-43.3519	-15.373	-24.0439	-18.2312	5.658623	2.796984	-5.1028
14	1012-a_h.v	0.261689	0.17097	1063.828	1252.11	1324.609	0.055161	-236.711	133.0197	-34.0601	25.86528	-49.1418	-14.034	13.45834	-12.6108	-23.145
15	1013-a_h.v	0.202576	0.220619	1058.036	1293.677	1341.96	0.052108	-208.979	137.4677	-26.343	3.381225	-52.538	-15.941	-20.1652	-9.92125	6.759032
16	1014-a_h.v	0.169917	0.250966	940.8586	1293.724	1037.47	0.049189	-197.065	166.4753	9.887601	-10.7365	-48.1591	-7.32388	-15.5345	-8.52953	19.48509
17	1015-a_h.v	0.190823	0.189345	1259.454	1781.977	1516.218	0.046419	-254.541	100.8923	4.8498	-24.0274	-34.6001	17.8183	-32.5624	-3.93736	-9.27433
18	1016-a_h.v	0.229435	0.143123	879.8541	1126.972	1101.053	0.052475	-257.978	164.5768	-16.1533	-12.1705	-18.558	-16.5197	-12.5038	-0.45713	10.10017
19	1017-a_h.v	0.254711	0.242359	878.0344	1394.277	976.1719	0.038974	-229.321	148.4775	25.692	2.367888	-48.2993	-0.27344	-12.3562	-12.788	-2.33135
20	1018-a_h.v	0.274134	0.223589	929.3645	1106.205	1028.305	0.05485	-194.77	185.265	-15.299	-0.62522	-41.7512	-47.2965	2.230133	13.45826	-15.4471
21	1019-a_h.v	0.200396	0.185512	882.8269	1294.231	1072.319	0.049309	-250.989	145.9424	8.310002	5.693005	-52.673	-27.5874	10.57889	-1.82043	-27.9694
22	102-a_h.wa	0.142181	0.272746	1257.555	1398.59	1321.941	0.076514	-238.868	90.16582	-52.0826	-41.2494	-43.3351	-18.067	9.842206	0.045936	-29.2259
23	1020-a_h.v	0.201941	0.215714	856.8432	1439.306	1122.5	0.025494	-212.237	172.557	5.367576	-23.0866	-12.1341	-14.8659	14.10978	-1.04833	-2.75207
24	1021-a_h.v	0.235878	0.236685	870.9978	1173.726	1061.433	0.03939	-185.126	178.0196	7.024727	7.230892	-53.4281	-15.5793	-1.14908	-3.3101	-4.23589
25	1022-a_h.v	0.171154	0.192371	1218.8	1896.835	1349.616	0.044894	-245.765	102.385	8.687534	-17.8823	-33.3439	9.419088	29.33423	3.165472	-5.14417
26	1023-a_h.v	0.248022	0.224855	1013.444	1423.527	1177.443	0.055152	-199.762	146.3226	-7.37952	1.869419	-33.1431	-21.1373	16.27794	5.727684	3.923112
27	1024-a_h.v	0.147948	0.224758	802.112	1155.726	1133.526	0.033293	-274.906	151.8642	-26.3951	-25.8336	-15.4312	5.831759	22.02791	14.64664	-21.1819
28	1025-a_h.v	0.29972	0.242206	1152.941	1320.318	2210.645	0.043167	-154.79	167.7584	-35.7273	22.17203	-32.9656	-54.556	5.368939	21.97403	-12.578
29	1026-a_h.v	0.338367	0.212788	1009.158	1228.026	1243.742	0.046097	-174.71	175.6335	-18.936	-20.742	-38.0512	-19.5037	15.90117	1.536258	-35.3872
30	1027-a_h.v	0.196924	0.232859	771.2162	1009.928	1033.369	0.044281	-231.947	186.9327	1.146849	-9.25401	-37.1973	-14.8941	-5.59404	8.185299	-2.57357
31	1028-a_h.v	0.226918	0.160969	1240.618	1437.617	1654.788	0.067732	-234.359	119.8309	-47.177	-0.91594	-37.1955	-29.4599	34.69217	-17.8443	1.2951
32	1029-a_h.v	0.195103	0.165637	1086.025	1535.501	1478.295	0.059719	-260.582	136.4148	-32.0684	-22.5926	-19.444	-21.0262	26.85797	-14.1687	-9.36114
33	103-a_h.wa	0.200965	0.304692	904.8353	1284.15	1203.975	0.042444	-179.986	165.7659	-6.17245	-20.8009	-11.5885	-26.4292	12.03552	11.53272	-33.5028
34	1030-a_h.v	0.378944	0.130136	1108.358	1641.703	1161.404	0.044717	-220.662	156.216	22.24621	-27.9255	-38.3758	-5.73903	-15.085	32.51913	7.575027
35	1031-a_h.v	0.287052	0.17267	869.4453	903.0472	1104.311	0.057773	-260.74	196.325	-23.152	-31.6364	-41.0581	-8.85357	-12.1091	4.110661	13.62833
36	1032-a_h.v	0.278662	0.233777	1130.739	1406.62	1686.377	0.063837	-169.844	149.863	-17.6253	-3.48178	-54.1305	5.366111	6.329211	-21.5671	9.195712
37	1033-a_h.v	0.350731	0.218185	1010.868	1391.996	1299.633	0.043221	-164.806	173.4043	-16.5673	-10.9732	-20.6208	-22.956	12.07322	3.257236	-10.3268
38	1034-a_h.v	0.381212	0.179185	705.5969	960.1029	983.8623	0.031308	-245.26	206.991	15.04247	-12.4355	-23.6822	-6.54356	-17.3566	17.40375	3.444344
39	1035-a_h.v	0.280319	0.125128	877.7336	1076.173	1239.661	0.046682	-266.643	192.5939	-43.952	8.743077	-21.0547	-33.936	20.8469	-14.9914	11.99274
40	1036-a_h.v	0.258747	0.260604	1180.558	1305.704	1856.825	0.052409	-138.966	168.6428	-46.8459	-6.52984	-37.0625	-20.7907	21.58622	-29.7246	5.56711
41	104-a_h.wa	0.237699	0.24498	1383.077	1686.215	2317.168	0.056401	-165.927	105.4509	-45.7755	-10.221	-33.642	-1.87026	12.51853	-6.14228	-14.5234
42	1059-a_h.v	0.175975	0.248959	1500.8	1812.353	1893.384	0.073493	-191.654	110.2036	-43.6058	-58.0022	-19.0147	-5.0368	-13.4689	25.99297	-15.9601
43	1060-a_h.v	0.250661	0.238288	1435.406	1908.537	2578.129	0.059665	-175.268	101.6755	-34.3488	-14.4912	-21.4772	-10.6241	7.593056	-1.70584	-23.9124
44	1061-a_h.v	0.2135	0.212023	886.8166	972.1254	1136.502	0.053052	-260.204	169.4857	-21.6067	-33.5397	-30.913	-18.7585	3.259392	5.509319	-7.22529
45	1062-a_h.v	0.306739	0.195488	998.9218	1435.645	1177.866	0.04137	-164.53	169.391	11.16918	-2.02643	-48.3616	3.873486	2.670718	8.687846	-6.93478
46	1063-a_h.v	0.304434	0.209477	1050.743	1259.658	1297.183	0.056684	-176.799	172.6538	-23.6477	-10.0126	-45.4617	-20.2728	-7.24243	8.144175	7.962858
47	1064-a_h.v	0.26337	0.14299	976.0512	1341.302	1220.491	0.042706	-228.309	175.8787	-18.1641	-30.9669	-11.6037	-20.2084	7.260695	9.020177	-4.99499
48	1065-a_h.v	0.168974	0.242934	1318.278	1554.382	1922.915	0.051351	-169.19	116.2374	-53.1499	-16.2693	-12.1796	4.151937	3.242763	-10.7902	-29.9112
49	1066-a_h.v	0.237819	0.122383	1689.199	2202.969	2749.99	0.067667	-184.047	110.5867	-27.2285	-42.2353	-18.6892	-15.4092	21.28083	-2.28208	-24.6895

	mfcc10	mfcc11	mfcc12	mfcc13	mfcc14	mfcc15	mfcc16	mfcc17	mfcc18	mfcc19	mfcc20	label	ID	Gender	Age
0	20.41746	-6.4932	4.146173	-12.5348	-13.1078	-28.1906	1.143172	4.134034	52.79483	39.0924	49.03706	Healthy	1	w	20
1	-5.63608	6.120925	7.275009	6.59109	-23.9271	7.79023	-7.03787	-9.1936	11.18356	-2.17361	-9.42094	Healthy	10	w	22
2	-10.4021	4.140114	24.70295	-5.14845	-0.60718	-20.0407	-4.15346	-2.72515	-10.6697	-2.33688	7.741264	Healthy	100	m	66
3	-40.3458	-7.15746	9.966608	-2.41478	12.74221	-0.16355	7.797343	-26.315	-6.62205	0.53512	0.749695	Healthy	1000	m	31
4	-9.50608	5.519818	1.367481	11.55102	-2.58331	-20.9398	13.30665	-26.5975	-8.32453	10.78601	-6.67751	Healthy	1002	m	25
5	-9.21666	-18.7284	11.29865	7.064672	-20.801	-15.1804	8.489902	-29.6265	-4.34001	-0.34296	-0.00415	Healthy	1003	m	35
6	5.541646	15.07505	-5.38708	11.02586	-9.16115	-24.2736	8.25876	-31.9161	17.1853	-4.90731	-9.37051	Healthy	1004	m	33
7	-1.41271	-31.1273	23.17631	15.69175	-5.01031	13.58097	-1.43191	-18.0889	-6.0785	-9.18721	7.610434	Healthy	1005	m	43
8	6.744126	-4.6204	-10.6621	-0.2249	-11.9769	-18.854	-0.96298	-9.12628	-11.1967	-3.24692	-7.80025	Healthy	1006	w	32
9	19.41811	-16.1083	23.30435	-1.67457	-18.9593	-9.16312	0.85383	-4.51475	-4.65488	2.379562	-11.3416	Healthy	1007	m	42
10	-1.6941	-18.7949	6.399647	9.350238	-5.11073	0.092198	-12.9934	-11.7596	-12.4347	-1.61468	2.077242	Healthy	1008	m	50
11	-7.19053	9.74223	14.78432	0.292139	-3.95393	-13.284	-9.33448	-28.5737	17.84802	-14.166	-8.21463	Healthy	1009	m	40
12	-24.9426	0.456464	13.9295	1.284007	3.095069	-0.5261	-2.78932	-14.5265	-22.2524	3.275426	2.73308	Healthy	1010	m	31
13	7.729187	-3.25297	17.6589	8.462935	-19.7108	-8.03039	2.284661	-13.2598	-3.37933	-6.63588	2.34766	Healthy	1011	m	28
14	-19.5293	-5.77098	15.25098	13.59159	-3.58402	13.18577	-23.0774	-15.6369	7.67544	-14.406	-5.7613	Healthy	1012	m	22
15	-12.9288	-19.6181	14.72008	8.839559	5.822886	-12.6371	3.165051	-21.6976	-14.3431	4.9837	-1.29662	Healthy	1013	m	44
16	-37.191	-8.00126	17.07967	-1.06497	6.281257	3.924834	2.404845	-18.3032	-13.8327	-7.72009	-1.84695	Healthy	1014	m	40
17	-19.0388	11.00425	18.33415	7.017072	-13.493	-8.29759	-9.37422	6.973831	-21.8699	4.040941	6.528244	Healthy	1015	m	37
18	-11.6412	-18.1618	6.995252	9.242693	12.14851	-22.3267	4.858014	-10.2933	-23.5521	2.343314	5.066183	Healthy	1016	m	38
19	-13.3278	-10.4264	8.923584	18.85686	-12.1371	-17.2474	6.235973	-20.6057	-24.7925	21.27742	-15.417	Healthy	1017	m	40
20	-2.72441	-3.45169	-6.4257	8.900938	7.262973	-6.61984	22.08159	-27.9281	-11.6575	2.577745	-21.7337	Healthy	1018	m	39
21	-10.5383	11.85348	-5.95109	-10.7766	18.43419	-12.6654	-6.09332	-28.7678	-1.03378	0.377638	-16.5948	Healthy	1019	m	48
22	-12.8706	-0.48118	-9.9711	-12.9075	-6.10186	9.20323	27.05626	40.47175	43.00858	9.747732	-16.3987	Healthy	102	w	39
23	1.742459	-24.8648	33.14514	-11.6761	-7.60025	-0.0053	-3.29222	-18.8172	-6.17507	12.73126	-13.87	Healthy	1020	m	25
24	-24.166	-15.2145	14.50845	0.262472	4.801905	11.60072	1.124244	-17.4257	-8.78706	0.999997	-8.72109	Healthy	1021	m	41
25	4.781082	-19.6813	-3.69641	-19.5549	-16.8053	-12.4454	-19.9065	-5.96716	-7.85587	-0.66018	3.276174	Healthy	1022	w	45
26	-2.5485	-7.67138	26.62555	-13.8686	0.682814	-11.2853	-2.51002	-21.7131	2.233833	6.165217	-4.38755	Healthy	1023	m	41
27	10.78058	-4.53367	-22.9127	-3.59706	-15.2564	-12.2868	-15.5208	-2.67643	-13.1634	-18.8123	-5.45984	Healthy	1024	w	37
28	13.04552	-11.536	1.671859	5.793391	3.916778	-12.7065	11.01613	-12.649	-20.726	-10.5844	6.759502	Healthy	1025	m	42
29	14.18629	8.851342	-2.07837	8.861491	-0.15425	-20.1169	-1.66257	-4.39172	-5.4821	6.100407	-5.4604	Healthy	1026	m	42
30	-16.2838	2.259925	3.424211	20.08404	-0.58738	-5.66813	-0.73017	-14.1097	-11.7302	-3.9947	-2.90851	Healthy	1027	m	48
31	5.86834	-3.13713	29.46584	-7.12955	-10.013	-21.871	-12.5716	-9.3286	0.822365	0.329878	-13.2649	Healthy	1028	m	41
32	10.26789	6.84524	-19.998	-14.0744	-16.1845	8.280423	-13.4395	-10.5534	-8.65157	-17.3771	-3.33857	Healthy	1029	w	41
33	16.45276	-1.07431	-7.31864	8.566545	-6.65559	-22.8361	-15.5522	-4.72134	3.146958	-13.4033	2.398429	Healthy	103	m	21
34	-23.7836	10.52436	3.594328	-2.09251	11.84021	-9.88639	5.044552	-22.9773	-2.32203	-0.37442	-2.77244	Healthy	1030	m	49
35	-14.1878	3.279861	-3.70491	20.21864	-3.76981	-27.3843	15.94819	-20.5367	-15.2754	9.327485	-5.57601	Healthy	1031	m	44
36	-2.32088	-15.6925	8.196841	23.52066	0.313465	-31.6956	7.393375	-10.22	-30.6184	14.10469	3.186529	Healthy	1032	m	37
37	-1.84262	-0.2714	9.624892	2.218447	-4.37326	-14.2185	4.983401	-27.8731	6.806544	-1.27296	-2.42726	Healthy	1033	m	45
38	-16.5388	1.591526	-4.61813	13.35868	-0.15118	-11.1717	9.259066	-6.7134	-27.913	10.12099	0.862846	Healthy	1034	m	29
39	8.721701	-14.0157	17.76728	10.92105	-17.4264	-13.4239	1.557464	-10.397	-6.2283	-2.35298	12.44739	Healthy	1035	m	24
40	2.775871	-20.8551	16.92366	19.73707	-3.22765	-33.3629	2.926944	-6.89521	-24.1453	5.067235	7.401256	Healthy	1036	m	39
41	2.22793	1.27281	2.559795	-27.4649	2.821136	-9.29032	-2.98816	-22.495	6.406731	-27.6169	3.936271	Healthy	104	w	48
42	-0.77002	14.03471	-15.2956	-8.50133	-15.5447	10.8967	-12.4034	-2.7117	-5.35229	4.550573	-8.88875	Healthy	1059	w	21
43	22.1754	-14.5593	7.486881	-33.4441	0.008508	-7.3808	-7.16548	1.997717	-2.96948	-10.7423	-1.08398	Healthy	1060	w	21
44	-0.563	6.481543	-0.7911	5.544901	-18.5771	-17.7951	-2.82717	-12.0709	1.507679	-3.74226	-0.29058	Healthy	1061	m	24
45	-7.85554	-6.14654	18.24474	-10.1383	6.705363	-1.1985	-12.5219	-15.6507	4.167094	-10.2716	-9.96438	Healthy	1062	m	31
46	-13.3043	-1.02623	5.053407	12.62896	2.05715	-27.5488	17.05694	-27.513	-13.8794	2.069075	3.631731	Healthy	1063	m	34
47	0.841605	-18.9375	20.8487	1.233056	-19.7688	3.714624	-6.96985	1.398311	-16.6323	1.250941	9.744197	Healthy	1064	m	22
48	19.26	-8.97381	-11.0038	-3.6429	0.008482	-9.72882	-8.86362	-9.68885	-19.9855	-2.21715	-2.97393	Healthy	1065	w	19
49	8.316171	1.922603	-5.09745	-11.1064	-6.66459	-0.02909	-5.9512	-6.90372	7.813918	-11.4109	-2.36718	Healthy	1066	w	21

This mixed-data type dataset contains 2,035 rows of participants whose audio files have had their spectral, temporal, and mel-frequency cepstral coefficients extracted from them. These extractions have their own respective column, and these are used to help determine the ‘status’ (healthy/unhealthy) of the participants. Characteristics (e.g. age and gender) have also been recorded, which will be useful for comparative analysis.

Approach

Phase 1: Understanding the Project Overview

- Define the objectives outlined in the project statement to ensure clarity with the end-goal of this assignment.
- Critically assess the provided dataset and deconstruct it into digestible components to improve comprehension.

Phase 2: Importing the data for Exploratory Data Analysis (EDA)

- Import the provided .csv file into Google Colab, assign it to a dataframe, and evaluate the schema of the dataset.
- Ensure each column is of the correct data type, impute any null values, check for any duplicated data, drop irrelevant columns, standardise column names to lowercase, and perform feature encoding where appropriate (e.g. with the gender column).
- Create visualisations for the dataset to uncover any trends, patterns, and relationships between the features.

Phase 3: Machine Learning

- Assign audio extractions as features (independent variables) and 'Status' as the target (dependent variable). Segregate the data into train-test split, scale the features where needed, and initialise several machine learning models to see which one yields the highest %.

Phase 4: Power BI

- Export the cleaned up dataset from Google Colab, so it can be imported into Power BI to transform the data and create interactive visualisations.

Phase 5: Power Presentation

- Compile the entire process into a power presentation to outline any notable findings.

Project Workflow:

Understand Objectives & Data →

Define the objectives and analyse the provided dataset to ensure that the goal of the project is clear.

Perform Exploratory Data Analysis →

Perform EDA on the .csv file, after assigning it to a dataframe, to extract insights and create visualisations.

Build a predictive model (machine learning) →

Create and compare several predictive models to analyse the data and predict the status outcome of participants.

Create an interactive results dashboard →

Export the .ipynb file and upload it into Power BI to create an interactive report that outlines the significant findings.

Design a presentation slide

Build a presentation deck to illustrate the process of the project and the key insights which have been extracted from the data set.

Data Cleaning

The dataset was cleaned by:

1. checking duplicate or irrelevant entries
2. executing a comprehensive age validation to ensure that there are no suspicious age values (e.g. -1 years old, or 125 years old)
3. standardising column names (converting column names into lower case) for consistency
4. dropping irrelevant columns
5. renaming columns for better visibility (e.g. g → gender)
6. feature engineering to transform the continuous, numerical values of the age column into a categorical column comprised of age bins (e.g. 0-12, 12-24, 24-36, etc.)
7. converting categorical variables like "gender" into numerical format (e.g., "gender_encoded") for analysis.
8. standardising “age_group” and “status” columns to ensure uniform labeling.

This process resulted in a structured, analysis-ready dataset suitable for further modeling or statistical evaluation.

Machine Learning:

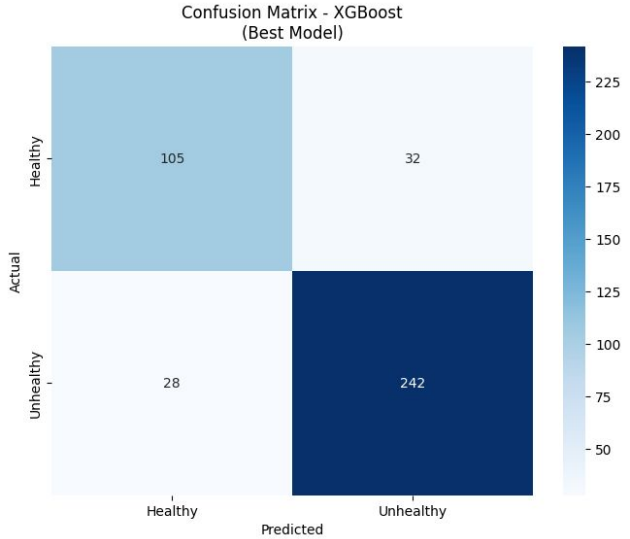
- Several predictive models were made to classify audio samples into two health status categories (Healthy and Unhealthy) and compared against each other to identify the most effective model.

Data Preparation

- This model utilises a comprehensive set of features extracted from the audio files. These include:
  - Spectral features (chroma stft, spectral centroid, spectral bandwidth, rolloff)
  - Temporal features: rmse and zero crossing rate
  - Cepstral features: 20 Mel-Frequency Cepstral Coefficients (MFCC1-MFCC2)
- The target variable was the 'Status' (Healthy/Unhealthy) column.
- Data splitting: the dataset of 2035 inputs was split into training (80%) and testing (20%) sets, using stratification, to ensure the class distribution was preserved in both splits. A fixed random_state = 42 was used for reproducibility.
- Feature Scaling: a StandardScaler was applied to normalise the features, which is important for models that are sensitive to the scale of data (e.g. K-Nearest Neighbors). Tree-based models (XGBoost, Random Forest Tree, Decision Tree) were trained on unscaled data.
- Label encoding: the string labels ('Healthy', 'Unhealthy') in the target variable were encoded to numerical values (0,1), using LabelEncoder to be compatible with scikit-learn models.
- 5 candidate models were selected to evaluate different algorithmic approaches: XGBoost, Random Forest, Logistic Regression, K-Nearest Neighbors, Decision Tree.
-

Machine Learning:

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	85.3%	85.2%	85.3%	85.2%
Random Forest	84.3%	84.0%	84.3%	84.1%
Logistic Regression	82.3%	82.7%	82.3%	82.5%
Decision Tree	76.9%	76.8%	76.9%	76.9%
K-Nearest Neighbors	74.4%	74.4%	74.4%	74.4%



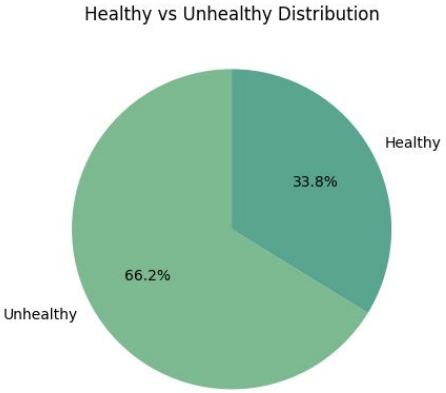
Performance Metrics and Evaluation Strategy

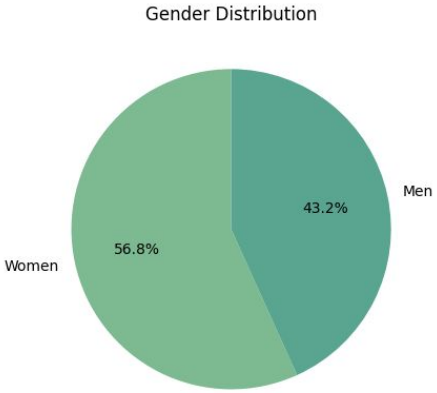
The primary metric for evaluating model performance was accuracy, defined as the proportion of correct predictions among the total predictions.

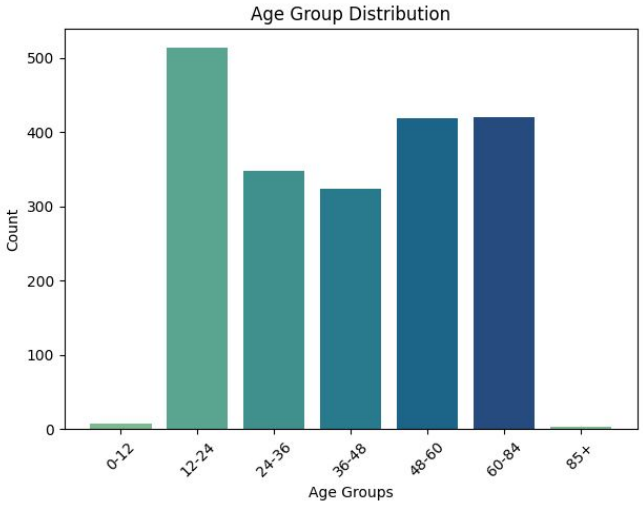
The results show two superior models: XGBoost and Random Forest, having the highest scores across all evaluation metrics. XGBoost's precision (85.2%) and recall (85.3%) indicates a well-balanced model that performs consistently across different classes. XGBoost is is slightly more accurate but can be sensitive to hyperparameters. Random Forest is generally more robust and less prone to overfitting on noisy data.

- On the confusion matrix, we see that XGBoost had 105 true negatives and 242 true positives.
- XGboost is better at identifying unhealthy cases (91% recall) than healthy cases (79% recall), indicating that this model is better at identifying unhealthy voices.

Data Analysis:



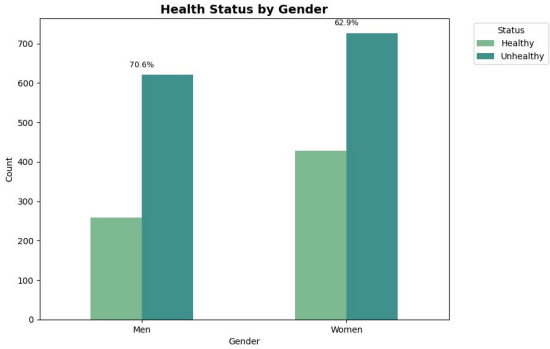


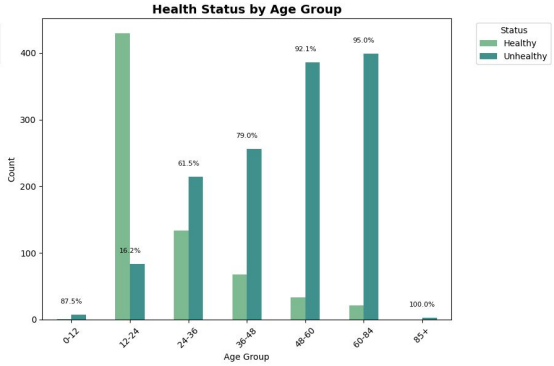


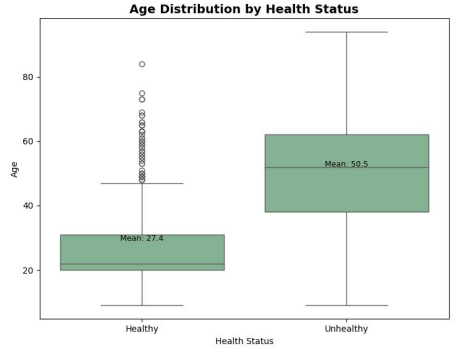
The sample was an inclusive sample, with participants ranging from children, teens, adults, and elderly. The demographic was also consistent of women (1150) and men (880). Each individual's health status also varied between 'Healthy' (687) and 'Unhealthy' (1348), with the latter being the more prominent class.

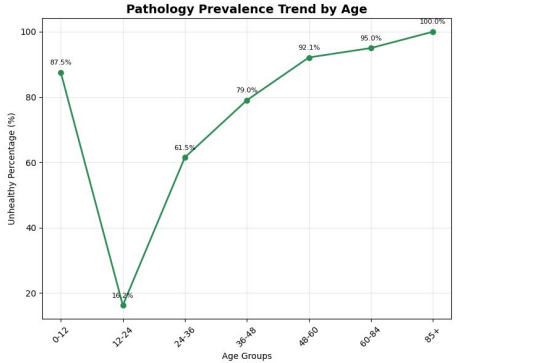
Data Analysis:

- There is a higher percentage of Unhealthy individuals in men (70.6%) compared to women (62.9%).
- There is a low pathology percentage in the younger demographic, particularly in the 12-24 year range, compared to their older counterparts.
- Mean age distribution shows that, on average, Unhealthy participants are older compared to Healthy participants.
- There is a disproportionately high pathology % on either spectrum (0-12 and 85+ ages), but these groups were underrepresented.
    - We need a bigger sample for these age groups to make any conclusive interpretations.



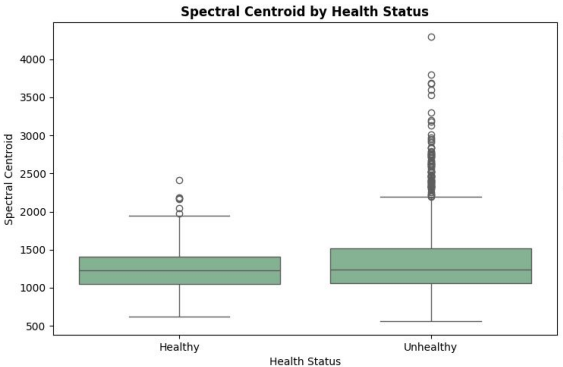


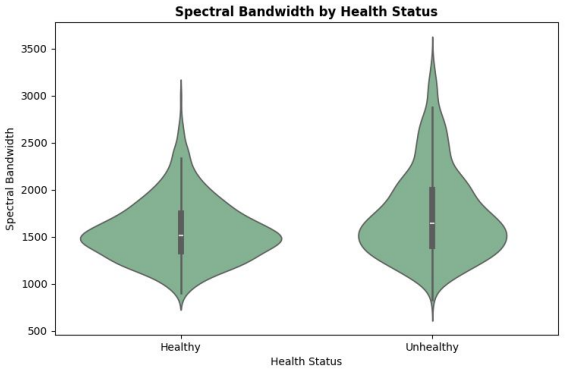


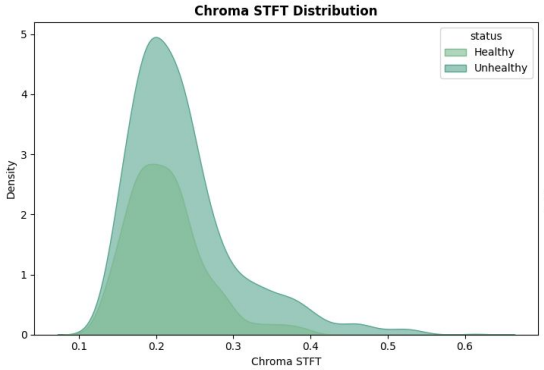


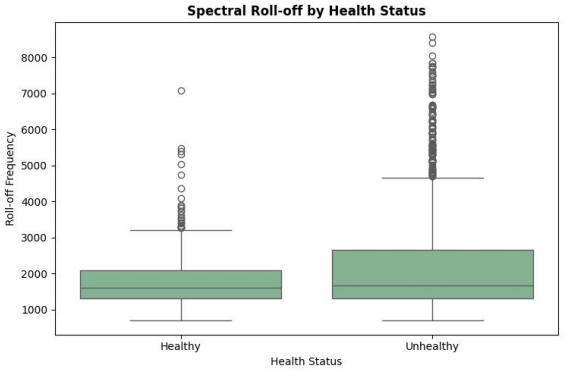
Data Analysis:

- Spectral Centroid is higher in Unhealthy participants compared to Healthy participants, with the Unhealthy group having outliers up to 4000 Hz.
    - Consequently, the Spectral Roll-Off is also higher in the Unhealthy group as they have more of a high frequency energy.
- The spectral bandwidth is wider, bimodal, in the Unhealthy group compared to the Healthy group.
- The Chroma STFT for the Unhealthy group has a higher density across many pitch classes. Meanwhile the Healthy group has a more concentrated and stable distribution.

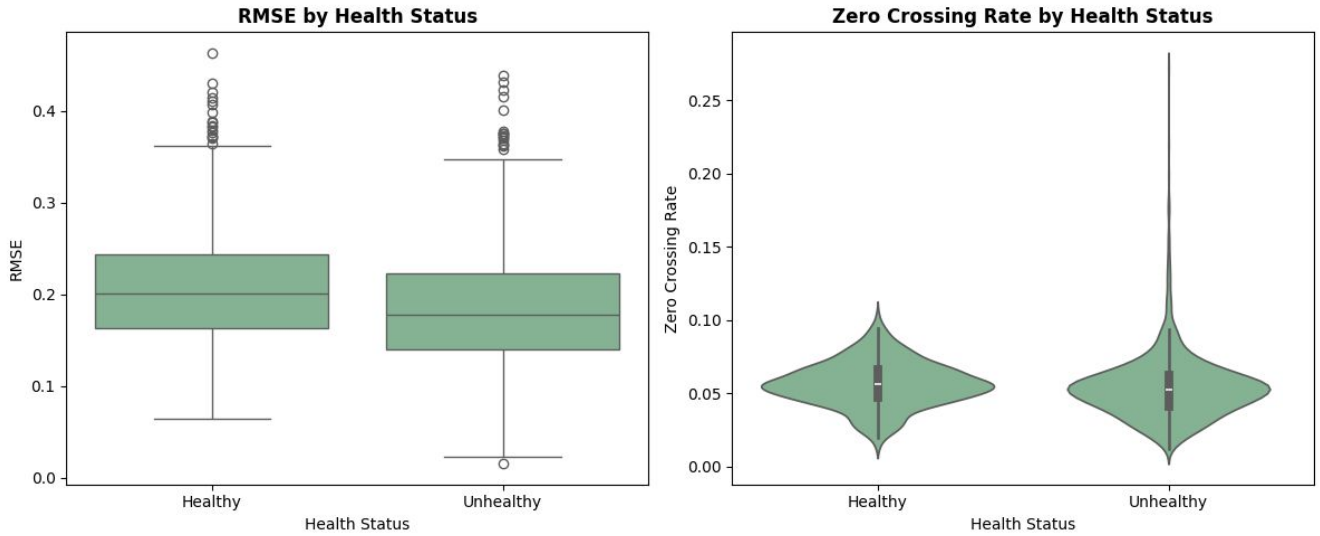






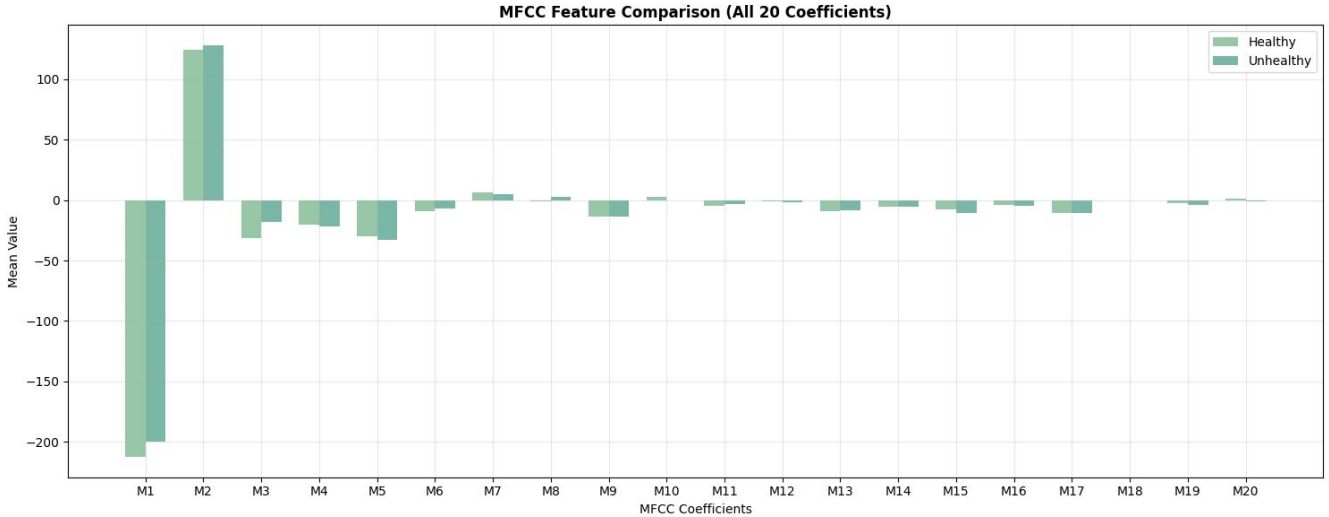


Data Analysis:



- RMSE (Root Mean Square Energy) is higher for the Healthy group compared to the Unhealthy group, suggesting that the Healthy group produces a stronger, more robust acoustic signal.
- In contrast, the Zero Crossing Rate (ZCR) is lower in the Healthy group, having a distribution tightly clustered at lower ZCR values. This elevated ZCR indicates high-frequency noise (signifying turbulent airflow, breathy phonation, hoarseness).
- Both graphs show that while the Unhealthy group is quieter overall (lower RMSE), their audio files have a noisier and sharper signal, with high-frequency components.

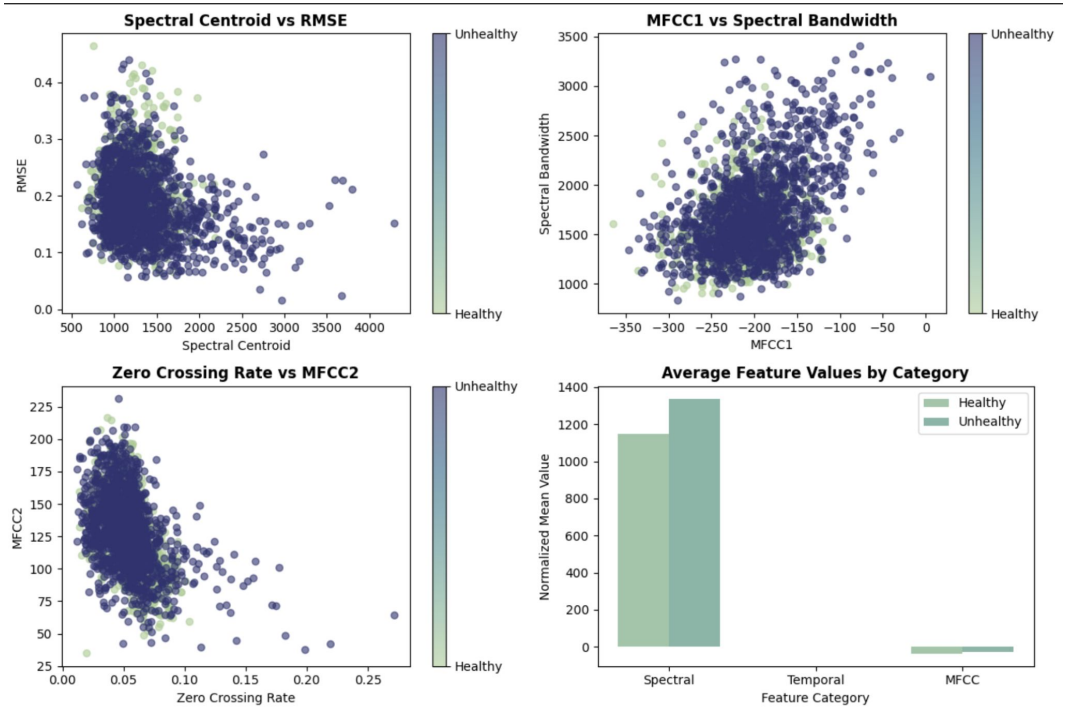
Data Analysis:



- Shows that for every coefficient, there is a significant difference in mean values for each group.
- MFCC1 shows that there is a difference in spectral tilt between the two groups: the Unhealthy participants have a higher (less negative) MFCC1, thus their voices product more low frequency energies.
- MFCC2-5 also consistently show that the Unhealthy group have higher mean values compared to the Healthy group, suggesting that the contour of the sound's spectrum varies in between states.
- As expected, the difference in MFCC values show that each group has a distinct acoustic structure. Vowels are defined by the shape of the vocal tract; a healthy individual will have one resonant signature, and an unhealthy individual with a different vocal tract shape (due to asthma, vocal fold nodules, etc)) will have a different resonant signature.

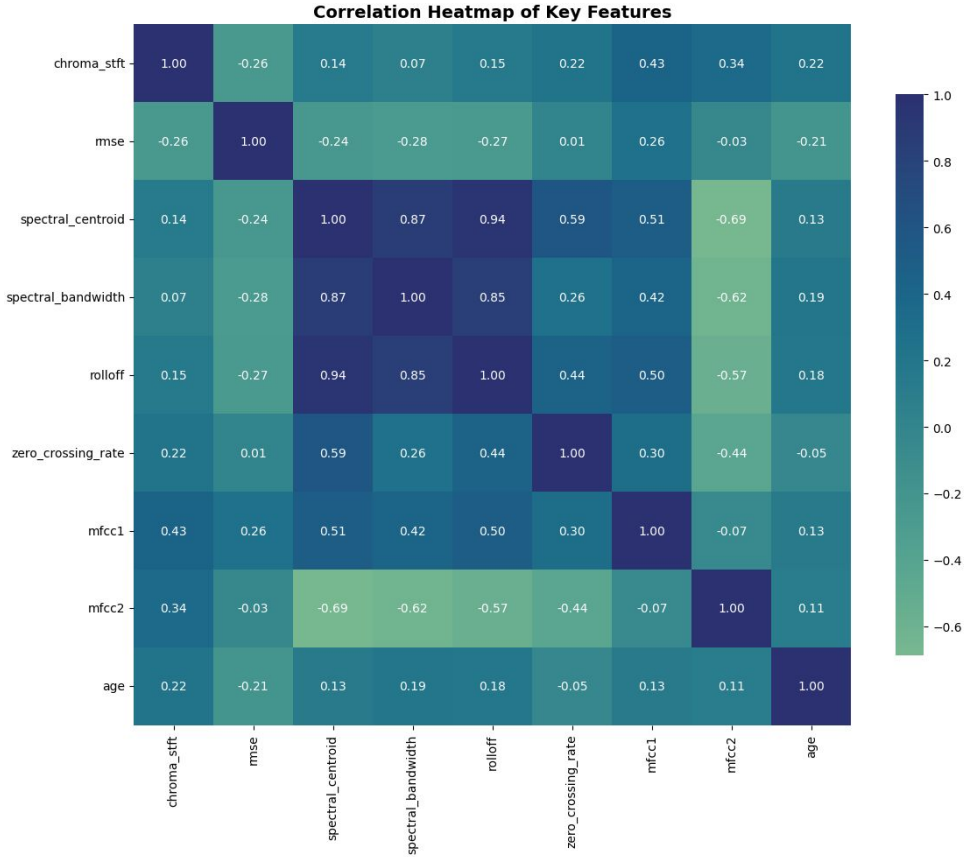
Data Analysis:

- **Spectral Centroid vs RMSE** shows two distinct clusters with a negative correlation between the features.
  - The Healthy cluster has a high RMSE and a low Spectral Centroid (dull, low frequency sound)
  - The Unhealthy cluster has a low RMSE and a high spectral centroid (bright, high-frequency sound)
- **MFCC1 vs Spectral Bandwidth:**
  - The Healthy cluster predominantly occupies a specific region of MFCC1 (consistent spectral tilt) and has a lower Spectral Bandwidth, indicating that the sound energy is more focused.
  - The Unhealthy cluster is more scattered around MFCC1, and it fills a wider range of the bandwidth (frequencies), consistent with noise and turbulent signals.
- **Zero Crossing Rate vs MFCC2**
  - The Healthy cluster has a low ZRC (tonal, smooth sound) and has a stable spectral shape for the MFCC2.
  - The Unhealthy cluster has a higher ZCR (noisy, chaotic sound) and an altered spectral shape, showing that their audio files has a noisier texture due to changes in timbral quality (MFCC2).
- **Average Feature Values by Category** show you that for every feature recorded, the Healthy group's average differs from the average of the Unhealthy group.
  - The temporal bars are not visible as they are too small to fit on the same scale as the spectral values.



Data Analysis:

- Strong positive correlations:
  - Spectral Centroid and rolloff (0.94): indicates that “brighter” (higher centroid) sounds have their energies spread out to higher frequencies.
  - Spectral Bandwidth and rolloff (0.85): indicates that there is a strong relation between high frequency cutoff and overall frequency.
- Moderate positive correlations:
  - Zero Crossing Rate and Spectral Centroid (0.59): as the sound becomes more high pitched and ‘brighter’, it also becomes more noisy and less tonal (high ZCR)
  - MFCC1 has moderate correlations with plenty of spectral features (centroid, bandwidth, rolloff), which shows that its spectral energy shape is influenced by broader properties.
- Strong negative correlations:
  - Spectral Centroid and MFCC2 (-0.69): indicates that as the sound becomes brighter, the value of MFCC2 decreases. Inverse relationship shows that the timbral texture (MFCC2) systematically changes as the pitch of the voice increases.



Conclusions & Recommendations

Conclusions:

- Men had a higher percentage of Unhealthy participants compared to women. This may be due to behavioural factors, such as men are more likely to engage in risk factors, like smoking. Additionally, there is an over-representation of men in construction and industrial work, which may exacerbate their risk of developing vocal pathologies.
- As expected, vocal pathologies became more prevalent in the older demographic . Presbyphonia (age-related voice age) leads to vocal folds thinning, loss of elasticity and mass, which can lead to tremor and hoarseness. Additionally, neurological decline is often observed in individuals as they age (e.g. stroke, dementia, Parkinson's), resulting in poor control of the larynx.
- The 'audio fingerprint' of Healthy participants is consistently and measurably different from Unhealthy participants, allowing them to produce acoustically distinct and mechanically classifiable sounds.
    - The Unhealthy group had an acoustic signature that was energetically weaker, spectrally busier, noisier, less harmonic, and had a different timbral texture than its Healthy counterpart.
- This study provides a strong foundation for developing an accurate, non-invasive, and automated diagnostic tool based on audio analysis.

Conclusions & Recommendations

Recommendations:

- **Have better class balance:** there was a significant imbalance between the number of Healthy and Unhealthy participants. This could lead to the model being biased towards the majority class, learning to predict "Unhealthy" most of the time and still maintaining high accuracy.
- **Have a more diverse dataset:** this dataset had more women than men, overlooking any anatomical and biological differences between the two groups. For example, women have smaller larynxes and shorter vocal tracts; physical aspects which will affect the measurements of their features.
    - A model trained on a female-skewed dataset will perform poorly when deployed on a general population as it will be systematically biased against men.
- **Have better representation of key demographic groups:** this dataset only had 8 young children (ages 0-12) and 4 elderly aged 85 and above, demonstrating paediatric and geriatric neglect. This means that the "Unhealthy" profile may be confounded by age and gender.
- **Add more confounding variables:** other factors, such as body habitus, can affect results. For example, an obese individual may have an impacted speech production, and therefore, will have differing MFCCs compared to a leaner individual.
    - Excess fat deposition in the neck, pharynx, and uvula changes the geometry of the upper airways, which could impact their MFCC values.